A Study on the Use of Summaries and Summary-based Query Expansion for a Question-answering Task

Ian Ruthven, Anastasios Tombros and Joemon M. Jose

Department of Computing Science, University of Glasgow,

Glasgow, G12 8QQ, Scotland

igr, tombrosa, jj@dcs.gla.ac.uk

Abstract

In this paper we report an initial study on the effectiveness of query-biased summaries for a question-answering task. Our summarisation system presents searchers with short summaries of documents. The summaries are composed of a set of sentences that highlight the main points of the document as they relate to the query. These summaries are also used as evidence for a query expansion algorithm to test the use of summaries as evidence for interactive and automatic query expansion. We present the results of a set of experiments to test these two approaches and discuss the relative success of these techniques.

1 Introduction

Query formulation is commonly regarded as one of the demanding activities in information seeking. The complexity of verbalising an information need can increase when the need is vague [17], when the document collection is unfamiliar, [14], or when the searcher is inexperienced with information retrieval (IR) systems, [3].

Relevance feedback methods are designed to overcome the difficulties in selecting query terms by detecting which terms in a collection are good at retrieving relevant documents. The main source of evidence for relevance feedback techniques are the documents that the user has assessed as containing relevant material. To achieve the benefits of relevance feedback a user must first, however, assess a number of documents retrieved in response to their initial query.

However, in real life situations a user may not be willing to read and assess the full-text of entire documents: the documents may be lengthy, the user may have time restrictions or the initial query may have retrieved a poor set of documents. The alternative strategy, which we explore in this paper, is to present the user with short summaries of retrieved documents. Summaries allow users to assess a set of documents and thus enter relevance feedback more quickly.

In this paper we report on a set of experiments to investigate the use of summaries for interactive searching. Our experiments used a form of the query-biasing summarisation technique proposed by Tombros and Sanderson [19], to create short document summaries that are tailored to the user's query. The summaries are based on highly matching sentences, allowing users to view the context in which query terms are used within the document.

We hypothesised that this form of summarisation would allow users to filter out non-relevant documents more effectively and target potentially relevant documents more quickly than either title alone or the full text of the documents.

The summaries themselves form an important source of evidence for relevance feedback algorithms, one that is potentially better than the full-text of the relevant documents. This is because the summaries display the *context* of the query terms, so feedback is based only on the sections of the document that pertain to the query. In addition to investigating the effectiveness of summaries in interactive searching, we investigated the use of summaries for relevance feedback, using the content of the summaries, rather than the full-text of the documents, to generate query expansion terms.

In the remainder of this paper we shall discuss the system and experiments we designed in order to test the effectiveness of summaries in interactive searching and relevance feedback. The experiments we present in this paper were carried as part of the TREC-9 interactive track [9]. In this paper, we expand our original analysis of our experiments to provide a deeper insight into our results. We also indicate potential strategies for improving the efficiency of our summarisation methods and discuss extensions of our approach.

In section 2 we present details of the overall system, including the relevance feedback and summarisation components. We outline the methodology used in our experiments in section 3 and the results of our experiments in section 4. We conclude and discuss our findings in section 5.

2. System

In section 2.1 we outline the overall architecture of our system, in section 2.2 we describe the interface component and in sections 2.3 and 2.4 we describe in more detail the summarisation and relevance feedback techniques we used.

2.1 System architecture

Our experimental system was composed of three units:

1. *retrieval system*. The retrieval system (SMART) performs an initial query run using the query terms passed from the interface. The list of retrieved document identifiers and document titles are passed to the interface for display.
2. *summariser*. For each retrieved document, the summariser (described in section 2.3) generates a query-biased summary, which is passed to the interface on demand.
3. *interface*. The interface displays the retrieved document identifiers, document titles and summary. The overall look and feel of the interface is described in section 2.2. The interface is also responsible for logging user interaction and generating query expansion terms (described in section 2.4).

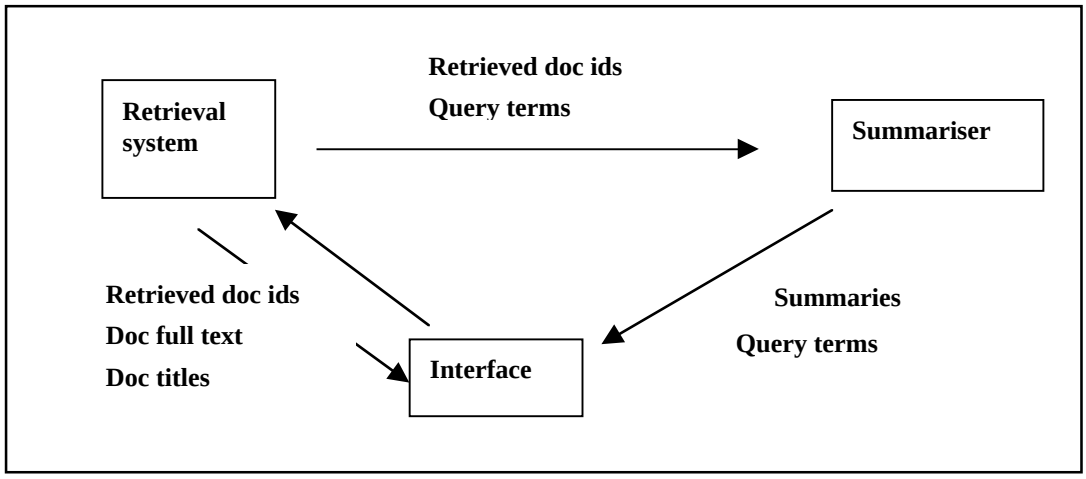

Figure 1: System architecture

2.2 Interface

A screen-shot of the web-based interface used in these experiments is to be found in Appendix A, Figure A.1.

The interface consists of four main components: a list of 20 retrieved document titles (labelled 1 in Figure A.1) and associated check boxes for marking documents relevant (labelled 2), summary display (labelled 3), query box (labelled 4) and suggested expansion terms (labelled 5).

Summaries were generated on demand. To request a document summary, the user passed the mouse pointer over a document title. The complete document could be viewed, in a separate window, by clicking on the document title.

Query expansion was both automatic and interactive. The user marked a document/summary relevant by clicking a check box next to the document title. The user could request suggestions for new query terms by clicking the 'Get More Terms' button, the suggested terms appearing in the 'More Terms' box on the bottom right of the screen. Each query, and suggested, term was displayed in this box; check-boxes were used to add/delete terms from the query. Users could also add their own query terms directly into the query box.

2.3 Summariser

A document summary conventionally refers to a condensed version of a document that succinctly presents the main points of the original document. *Query-biased* summarisation methods generate summaries in the context of an information need expressed as a query by a user. Such methods aim to identify and present to the user individual parts of the text that are more focused towards this particular information need than a generic, non-

query-sensitive summary. In this way summaries can serve an *indicative* function, providing a preview format to support relevance assessments on the full text of documents [15].

Query-biased text summarisation is an emerging area of summarisation research that had not been addressed until recently. Tombros and Sanderson looked into the application of such methods in information retrieval, evaluating the indicative function of the summaries [19]. Their study showed that users were better able to identify relevant documents when using the summaries than when using the first few sentences of a document.

The summaries generated by our system were indicative and query-biased, aiming to provide users working on an interactive IR system with information on the relevance of documents retrieved in response to their query. The system is based on a number of sentence extraction methods that utilise information both from the documents of the collection and from the queries submitted, and is a simplified version of the system described in [19].

Each document that was retrieved in response to a specific query was passed through the summarisation system, and as a result a score for each sentence of each document was computed. This score represents the sentence's importance for inclusion in the document's summary. Scores are assigned to sentences by examining the structural organisation of each document, and by utilising the inverse document frequency (IDF) weights assigned to each term. Information from the structural organisation of the documents was utilised in two ways. Terms occurring in the title section of a document were assigned a positive weight (title score) in order to reflect the fact that headlines of news articles tend to reveal the major subject of the article. In addition, a positive ordinal weight was assigned to the first two sentences of each article, capturing the informativeness of the leading text of news articles. The IDF weights of the terms in the collection were used as a source of evidence of attributing an overall measure of importance for each sentence of the source documents. In order to establish such a measure, the sum of the IDF weights of all the terms comprising a sentence was divided by the total number of terms in that sentence. In that way an importance score was attributed to each sentence in the collection.

In addition to the scores assigned to sentences, information from the query submitted by the user was also employed in order to compute the final score for each sentence. A query score was thus computed, intended to represent the distribution of query words in a sentence. The rationale for this choice was that, by allowing users to see the context in which the query terms occurred, they could better judge the relevance of a document to the query. The computation of that score was based on the distribution of query terms in each sentence. This was based on the belief that the larger the number of query terms in a sentence, the more likely that sentence conveyed a significant amount of the information need expressed in the query. The actual measure of significance of a sentence in relation to a specific query, was derived by dividing the square of the number of query terms included in that sentence by the total number of the terms comprising the query.

The final score for each sentence is calculated by summing the partial scores discussed above. The summary for each document is then generated by selecting the top-scoring sentences, and outputting them in the order in which they appear in the original document. Summary length was defined to be 20% of the document's length, up to a maximum of 6 sentences. Such a value seems to be in general agreement with suggestions made by [5, 2].

Figure 3 shows the summary produced from the document in Figure 2, retrieved in response to the query ‘*America national parks redwood trees’*. In Figure 2 bold type marks those sentences that were extracted to form the summary.

<DOC>

<DOCNO> SJMN91-06312178 </DOCNO>

<ACCESS> 06312178 </ACCESS>

<DESCRIPT> CALIFORNIA; TREE; PARK; US </DESCRIPT>

<LEADPARA> **California's majestic redwood parks may be ceded to the federal government under a cost-cutting proposal under study by state Parks and Recreation Department officials, the officials said Wednesday**.; The three parks -- Jedediah Smith Redwoods State Park and Del Norte Coast Redwoods State Park in Del Norte County, and Prairie Creek Redwoods State Park in Humboldt County -- are the crown jewels of the state park system and home to 2,000-year-old redwoods, among the oldest living things on Earth. </LEADPARA>

<SECTION> California News </SECTION>

<HEADLINE> STATE MAY CEDE ITS 3 REDWOOD PARKS TO U.S. </HEADLINE>

<TEXT> The proposal emerged from a review ordered by state Parks and Recreation Director Henry R. Agonia after the Wilson administration sent a directive to, state agencies asking them to identify budget cuts. The state is facing a staggering $2 billion deficit in this year's $55.7 billion budget.; The prospect of transferring California's redwood parks to the National Park Service drew praise and criticism from environmentalists and park rangers Wednesday as word spread.; "I'm strongly against it," state parks Superintendent Bill Beap

said in a telephone interview from Eureka. "These are the prime jewels of the state park system."; But the proposal was welcomed by Sierra Club officials, who called it "a splendid idea." **Edgar Wayburn, the club's vice president for conservation, noted that the neighboring National Redwood Park's boundaries touch all three state parks.** </TEXT>

<BYLINE> Los Angeles Times </BYLINE>

Figure 2: Document SJMN91-06312178

<SUMMARY>

<TITLE> STATE MAY CEDE ITS 3 REDWOOD PARKS TO U.S.

</TITLE>

<DOCID> SJMN91-06312178 </DOCID>

<LI>California's majestic redwood parks may be ceded to the federal government under a cost-cutting proposal under study by state Parks and Recreation Department officials, the officials said Wednesday. </LI>

<LI> " Edgar Wayburn, the club's vice president for conservation, noted that the neighboring National Redwood Park's boundaries touch all three state parks. </LI>

</SUMMARY>

Figure 3: Summary produced from document SJMN91-06312178

The sentences that appear in a summary were used as the basis for the relevance feedback component of our system which we shall explain next.

2.4 Relevance feedback

Relevance feedback techniques generally have two components: *relevance weighting* of query terms and *query expansion* [7].

The relevance weighting component assigns a weight to all terms in the relevant documents. This weight reflects how well the term discriminates relevant from non-relevant documents: terms with higher weights appear in a higher proportion of the relevant documents than the non-relevant documents. Relevance weighting, then, is used to prioritise retrieval of documents that contain 'good' query terms – those that help retrieval of relevant documents.

Terms that have high relevance weights, but do not appear in the user's query, may also be useful in retrieving new relevant documents. This is the basis behind the query expansion or query reformulation component: terms that have high relevance weights are likely to improve retrieval effectiveness if they are added to the user's query.

Both these components have been shown to be powerful techniques for improving retrieval effectiveness, [7], with query expansion often being the more successful of the two, e.g. [8, 16].

Query expansion may be *automatic*, in which case the system selects which terms to add to the user's query and how many terms to add. An alternative is *interactive* query expansion, where the user selects the new expansion terms, based on a set chosen by the system. Current experimental evidence indicates that better overall results are achieved with automatic rather than interactive query expansion, [1, 11] but this success is dependent on how the query expansion is presented and the search task that the user is given, e.g. [10].

In our experiments we concentrated on query expansion only. Query expansion was both automatic (the top 6 expansion terms were automatically added to the query when the user requested more documents), and interactive (the user selected terms from the list of suggested terms at the bottom right of the interface).

Expansion terms were selected from the summaries of marked relevant documents using Porter's term weighting function [13]. Equation 1 shows the calculation of the Porter weight for term t. r_t equals the number of relevant summaries containing term t, R is the number of documents marked relevant by the user, n_t is the number of documents in the collection containing t, and N is the number of documents in the collection.

$$Porter_t = r_t / R - n_t / N$$

Equation 1: Porter term weighting function

3.Experimental details

Our experiments were carried as part of the TREC-9 Interactive Track. In section 3.1 we describe the basic experimental methodology that was used and in section 3.2 we give details on the subjects who took part in our experiments.

3.1 Interactive TREC

TREC [20] is a cross-site initiative which provides a forum for the development and evaluation of new IR techniques. The interactive session (or track) of TREC has a special interest in studying user interaction with IR systems. The interactive track organisers provide a common experimental protocol, baseline system, data and searching tasks that each participating group use to run a set of experiments.

For the TREC-9 experiments the data to be searched consisted of a set of newspaper and newswire collections, approximately 2.5 GB in total, composed of roughly 900, 000 individual articles. All participating groups were supplied with the ZPRISE system [4], which they were to use as a control system in their experiments. ZPRISE offered automatic and interactive query expansion facilities.

This year the interactive track investigated a question answering task: subjects were asked to find documents that answered a set of predetermined questions. The questions were of two types:

i. *Comparison questions*. This type of question, e.g.*' Is Denmark larger or smaller in population than Norway* ?', asked the user to compare two items (*Denmark* and *Norway*) according to a given criterion (size of *population*).

ii. *Multiple part questions*. This type of question, e.g. '*Name four films in which Orson Welles appeared*', asked the user to compile a list of *n* answers, where the set of possible answers is usually greater than *n.*

Each user was asked to answer eight questions, all users being given the same eight questions. The user was asked to answer four questions using the summariser system and four questions using ZPRISE. The order in which the questions were given to the users, and the allocation of which IR system the user was given for each question was randomised according to an experimental matrix supplied by the track organisers. Each group is given a different matrix. Figure 4 shows a sample of one of the experimental matrices.

Subject		
1	**System 2** - Questions 4-7-5-8	**System 1** - Questions 1-3-2-6
2	**System 1** - Questions 3-5-7-1	**System 2** - Questions 8-4-6-2
3	**System 1** - Questions 1-3-4-6	**System 2** - Questions 2-8-7-5

Figure 4: Sample of experimental matrix

Each participating group was asked to follow the same experimental methodology employing the same questionnaires and giving the subject similar instructions.

Subjects were only given five minutes to attempt to find an answer for a question. If they found an answer they were asked to write down the answer and the number of the document that supplied the answer.

3.2 Subjects

Ten experimental subjects took part in our experiments, all subjects were educated to graduate level in a non-computing, non-LIS discipline, and, with two exceptions, all our subjects were postgraduate students recruited from the Information Technology course at Glasgow University. None of the subjects had any formal training in information searching or retrieval, beyond basic training on the university library search facilities.

The average age of the subjects was 23 years, and the average previous search experience was 3 years. All subjects reported some experience with library systems but the majority of reported experience was gained using web search engines using a point-and-click interface. These subjects were relatively regular searchers, performing searches either daily or weekly.

None of the subjects had used the control system (ZPRISE), the experimental system or an IR system with summarisation facilities. Prior to searching on a system, all subjects were given a short tutorial on the system.

4. Analysis

In this section we present two analyses of our results: a quantitative analysis and a qualitative analysis. This looks at the factors that may have caused a difference in search results between the two systems. In section 4.2 we examine the effectiveness of searching on the two systems.

The qualitative analysis, section 4.3, examines the responses to open-ended interview questions. This section also examines possible reasons for the results outlined in section 4.2.

4.1 Quantitative analysis – search factors and statistics

In this section we present results regarding the searchers' certainty on the topics (section 4.1.1) and their views on the two systems used (section 4.1.2). These results are based primarily on the answers to the pre- and post-search questionnaires supplied by TREC. The questionnaires used a 5 point semantic scale in which 1 means Not At All and 5 means Extremely. In section 4.1.3 we present some statistics on the searches.

4.1.1 Topics

The subjects were generally unfamiliar with the topics before searching, the reported certainty before searching being on average between 1.24 – 1.9. The certainty of the users increased after searching for all topics, the final certainty ranging from 2.97 for topic 6, to 4.4 for topic 1. Although the pre-search certainty for both types of topics were roughly the same (1.60 for multiple part topics, 1.59 for comparison topics), searchers reported a greater degree of post-search certainty for multiple part topics (3.51 multiple part against 3.15 for comparison topics).

Users found slightly more relevant documents for the multiple part topics (1.05 per search vs. 1.02 for comparison topics), however the greater reported certainty seems to come from the interaction rather than search success. For the multiple part topics, subjects reported that these topics were easier to start a search on, and easier to search for[1]. However they reported that they were less satisfied with searches on the multiple part topics (3.00 vs. 3.27 comparison) and would have liked more time to search on these topics (3.41 multiple part vs. 3.45 comparison). There was a greater pre-search familiarity with the multiple part topics (1.96 vs. 1.78) not reflected in pre-search certainty.

4.1.2 Systems

Users marked slightly more relevant documents on average with the control system (1.04 vs. 1.01 per query), and found it easier to start a search with the control system (3.58 vs. 3.36) but were overall less satisfied with the control system (2.62 vs. 2.51).

From the exit questionnaires, the subjects claimed a relatively high level of understanding of the task (4.4), and a fair similarity with other searching tasks (3.5). The subjects did not feel there was a great difference between systems (3.4).

Of the ten subjects tested 6 claimed the experimental system was easier to learn to use (1 for ZPRISE, 3 undecided), 7 found the experimental system easier to use (3 ZPRISE, none undecided), and 7 preferred the simplicity of the experimental system (3 opting for ZPRISE).

4.1.3 Search statistics

In this section we analyse the search statistics regarding the number of documents retrieved, query terms entered and documents assessed relevant for both systems. Table 1 summarises the basic search statistics.

[1] Ease to start (multiple part 4.46 vs. 3.97 comparison), ease to search (3.77 multiple part vs. 3.54 comparison).

	Control System	Experimental System
Initial query terms	3.54	4.69
Query terms added	1.24	1.55
Iterations (including initial ranking)	1.5	2.16
Documents assessed relevant by subject	1.04	1.01
Full texts viewed	2.94	1.19
Summaries generated	-	5.48

Table 1: Average search statistics per query

The searchers tended to use more query terms on the experimental interface than the control system and more terms were added through query expansion. Very few terms were added through the interactive query expansion facility. As noted before the searchers assessed slightly more relevant documents with the control system than the experimental system.

Summaries were often used by the subjects; out of a total of 672 unique documents retrieved by the experimental system over all the searches, 167 unique summaries were viewed and only 47 unique full-texts were viewed.

The searchers appeared to do more iterations of feedback with the experimental system although this is slightly deceptive as in fact they tended to do a new search more often than modify their existing query.

4.2 Quantitative analysis - search results

In this section we examine the search effectiveness of the two systems. Before we do this we outline how our searchers' answers are verified by the interactive track organisers.

Each of the users, for each question, returns two items of data: an answer to the question and a list of documents that supports the answer, i.e. the document(s) they used to obtain the answer. Figure 5[2] shows an example of this, in which user *2*, searching for an answer to question *3* using the control system (*SysC*), found the answers *Citizen Kane* and *Terminator 2* using documents *AP890211-0126* and *AP890211-0259* to answer the question.

SysC;2;3;Citizen Kane, Terminator 2;AP890211-0126, AP890211-0259

Figure 5: TREC answer format

All answers and supporting documents are returned to the track organisers and are (manually) assessed in two ways. Firstly the answers are checked to see how many answers are correct. The correctness or completeness of an answer is given by one of three possible assessments, each of which is associated with a numerical score: all of the answers are correct (**2**), some of the answers are correct (**1**), or none of the answers are correct (**0**). The answer given in Figure 5 is only partially correct (**1**) as Orson Welles did appear in *Citizen Kane* but not in *Terminator 2*.

The correct answers are then checked against the associated documents to check whether the documents did, in fact, support the answers, i.e. the document clearly provided an answer to the question. The degree of support is also given by one of three categories: all the correct answers are supported (**2**), some of the correct answers are supported (**1**), or none of the correct answers are supported (**0**) by the documents cited.

The combination of these two methods of analysis give seven possible assessments for a document[3], i.e. a 2:2 assessment means all answers were correct and supported, a 1:0 assessment means that some answers were correct but none were supported, etc.

4.2.1 Control vs experimental systems

In Table 2 we outline the overall search results for our experiment, comparing our subjects on the experimental versus the control system. The users returned a higher proportion of non-supporting documents on the control

[2] The answer was given in response to the question, '*Name four films in which Orson Welles appeared*'.

[3] An answer set in which all answers were incorrect can only receive a 0 for the support.

system (41.67%[4]) than on the experimental system (36.84%), and a lower proportion of supporting documents with the control system (58.33% control vs 63.16% experimental[5]).

The subjects using the experimental system returned more correct answers (79% of answers were fully or partially correct) compared with the control system (63%). Our subjects, then, found a higher percentage of correct answers with the experimental system and a higher percentage of these answers were supported by the returned documents.

	2:2	**2:1**	**2:0**	**1:2**	**1:1**	**1:0**	**0:0**
Control average	33.33%	-	-	25.00%	-	4.17%	37.50%
Experimental average	31.58%	-	5.26%	31.58%	-	10.53%	21.05%

Table 2: Summarised results – control system versus experimental system

4.2.2 Our results vs TREC average

In Table 3 we compare our results from both systems with the average from all the interactive track participants. We perform this analysis for the two types of question:

i. *multiple part questions* (1-4). If we take our subjects' answers as a complete set, i.e. the answers found using either system, then our subjects are performing as well as the subjects used in the other interactive track experiments. That is, our subjects found approximately as many fully or partially correct answers (64%) as the TREC average, and had a similar percentage of fully or partially supported answers (54% TREC, 52% us). This indicates that our searchers themselves were not a significant factor in the success of this type of search.

Examining the difference between our subjects on the control and experimental systems for this kind on topic, there is a clear difference. We found that our subjects found more fully or partially correct answers using the experimental system (60% experimental vs 47% ZPRISE) and a higher percentage of these answers were supported (58% experimental vs 46% control).

ii. *Comparison questions* (topics 5 – 8). Comparing our subjects' answers as a complete set against the interactive track average, our subjects found a higher proportion of fully or partially correct answers (78%) then the TREC average (61%) and had a higher percentage of fully or partially supported answers (72% us, 46% TREC). This suggests that our searchers found this kind of question easier to answer, regardless of which system they were using. As noted in section 4.1.1, our searchers claimed a greater degree of post-search satisfaction with this type of question.

This is reinforced by the analysis between experimental and control system which shows that our subjects found the same percentage of fully or partially supporting documents using both systems (72%) and a similar percentage of fully or partially correct answers (72%).

		2:2	**2:1**	**2:0**	**1:2**	**1:1**	**1:0**	**0:0**
Topics 1 -4	**TREC average**	13.91%	2.16%	0.96%	28.06%	9.35%	9.83%	35.73%
	Our average	4.00%	-	-	48.00%	-	12.00%	36.00%
	Our control average	-	-	-	46.15%	-	0.72%	4.32%
	Our experimental average	8.33%	-	-	50.00%	-	1.40%	2.10%
Topics 5 - 8	**TREC average**	46.60%	-	14.56%	-	-	-	38.84%
	Our average	72.22%	-	5.56%	-	-	-	22.22%
	Our control average	72.73%	-	-	-	-	-	6.47%
	Our experimental average	71.43%	-	0.70%	-	-	-	0.70%

Table 3: Summarised results – Glasgow results versus average results

[4] Sum of columns 4, 7 and 8 in Table 2

[5] Sum of columns 2 and 5 in Table 2

In Table 4, we present an analysis of the overlap of the supporting documents found with the two systems. These results attempt to capture the performance of subjects using either system to discover new documents that support the answer to each query. Any document returned by a subject for a specific query that was marked by TREC assessors as 'supporting the right answer for this query' was used in the calculations, irrespective of the score assigned to that response. That means that even if a subject provided a wrong answer for a query based on a document marked as 'supporting' by the assessors, that subject would still be credited with finding a document that supports the correct answer for that query, and would be included in the calculations.

Topic	Control unique	Control total	Experimental unique	Experimental total	Total unique for query	Total TREC	Averages
1	1	5	1	3	1	13	7.62%
2	1	1	-	-	1	7	14.29%
3	2	6	2	5	3	17	17.65%
4	2	3	9	11	11	39	28.21%
5	3	7	3	4	5	7	71.43%
6	2	2	2	5	2	3	66.67%
7	5	9	1	1	6	23	26.09%
8	1	1	-	-	1	15	6.67%
Total	17	34	17	29	30		
Average	**50.00%**		**58.62%**				

Table 4: Document analysis results

The **Control unique** and **Experimental unique** columns indicate the number of unique supporting documents found by subjects for each of the two systems (Control or Experimental) for a specific query. The **Control total** and **Experimental total** columns indicate the total number of documents marked by subjects for a specific query for each of the two systems. The **Total unique for query** column displays the total number of unique supporting documents for each query returned by both systems. Finally, the **Total TREC** column indicates the number of documents for each query that were marked as 'supporting the right answer' from the TREC assessors.

The results indicate that, on average, subjects using the experimental system performed better at discovering documents that could potentially support the right answer for a query. Subjects under both systems discovered the same number of unique supporting documents (17), however subjects using the experimental system discovered these documents in fewer attempts (29 vs. 34).

4.2 Qualitative analysis

In this section we summarise our findings on the two major components of our system: relevance feedback (section 4.2.1) and summarisation (section 4.2.2).

4.2.1 Relevance feedback

Relevance feedback was not popular amongst our searchers. In particular the subjects were not convinced about the benefits of RF with an average response of 2.7 for the question "*Was relevance feedback useful*?", 2.1 for the question "*Did the system add good terms to your query?"* and 2.5 for "*How well did you understand the relevance feedback option?"*.

Most users tried the automatic expansion technique once, but often only used it once. No user seriously used the interactive option. Only two users actually used the interactive query expansion facility as a source of new terms, each only used it for one search. One user used the suggested terms to add variants of existing terms, the other to add place names to a search on sites of Roman ruins[6].

As noted before searchers on the experimental system tended to enter new search terms rather than work with the modified query. However, searchers on the control system were just as unlikely to use relevance feedback. From our interviews with the searchers we extracted four main reasons for this:

[6] The terms added were place names that were unfamiliar to the user, who thought these places might be relevant to the search.

i. *Understanding of relevance feedback*. Although the subjects were given short tutorials on both systems and relevance feedback was explained to them, the searchers still did not feel they understood enough about how relevance feedback operated. In particular they felt they lacked enough information about how to make relevance decisions: which documents to mark relevant, how many to mark relevant, the effect of marking a document relevant, etc.

ii. *Poor choice of expansion terms*. Several users criticised the choice of expansion terms in the interactive query expansion option.

Our method of calculating relevance weights has previously been shown to give good results [12, 6] for automatic and interactive query expansion. The expansion terms produced in our case, however, were not always useful. This was for two reasons: low numbers of relevant documents, and the processing of summaries for relevance weighting.

The searchers tended to find relatively low numbers of relevant documents. If a searcher found only one relevant document in the first display of documents (as many did) then the relevance weighting prioritised those terms that only appeared in the relevant document. In this case, not only were these terms not useful for retrieving further relevant documents but occasionally the terms turned out to be spelling errors in the original documents, e.g. '*armioffici*', '*withth*', and '*sovietunion*'.

A flaw in our preparation of the summaries for relevance weighting was not to remove stop words from the summaries before weighting. Consequently, a proportion of the suggested expansion terms consisted of labels such as, '*docno*' and '*bylin*'. This was shown to affect a small number of queries.

For this experiment we used only the summary to determine possible expansion terms, even if the user had viewed the full-text of the document. A natural enhancement of this may be to vary the source of the expansion terms according to the representation(s) of the document viewed by the user. For example if the user has only viewed a summary before making an assessment, then the summary should be used for feedback, if the user has viewed the full text of the document, then the full-text may be a better source of evidence for feedback.

iii. *cognitive load*. Although relevance feedback has been demonstrated to provide a useful technique to improve retrieval effectiveness, it does require additional actions on behalf of the user, in particular marking documents relevant. Our searchers were not keen to provide this additional effort, especially without seeing any actual benefit.

iv. *Search experience*. We believe that the high familiarity with web search engines in our subjects may have also contributed to the lack of uptake of relevance feedback. The web engines that our subjects were familiar with typically do not provide relevance feedback, users being forced instead to generate new query terms themselves. Our subjects may have found it easier to submit new queries than modify the existing one.

4.2.2 Summarisation

The subjects were, on average, in favour of the use of summaries with an average score of 3.7 for the question "*Were the document summaries useful in answering the questions?"* and 3.5 for the question "*Were the document summaries a good representation of the full document text?"*. However, it is doubtful whether this second result was valid, as few users actually compared the full-text with the summary.

9 of 10 subjects judged the length of the summary as "*About right*", the remaining subject said the summaries were too short.

All subjects liked the simplicity of the summariser interface, however they felt the summaries took too long to produce. Summaries are produced on request, section 2.1, consequently speed is an important issue for this system. The average time taken to produce a summary was 5 seconds, the range being 0.5-20 seconds. However in practice summaries took longer than average to generate, somewhere of the order of 10 seconds on average. The main reason for this is that the summarisation time is dependent on the length of the original document.

In our experiments, the retrieved documents tended to be longer than the average document length, consequently the summaries took longer than average to produce. This was criticised by several users.

One way to speed-up the on-line generation of the summaries is to compute the static part of the sentence scores (section 2.3, i.e. IDF scores, title scores, ordinal weights) at indexing time, and leave only the query-biasing scores to be added at retrieval-time.

Although analysis is not conclusive, we believe that summaries may have resulted in some false positive answers. In this case the searcher marks the document as supporting based solely on the summary, whereas in fact the document does not support the answer, something the searcher might have identified had he read the full-text.

One contributing factor to this misleading behaviour of summaries is the lack of coherence. This is a well-known problem of sentence-extraction approaches [12]. Our query-biased approach is based on the belief that coherency problems can be tackled by the customisation of the summary to the query. For the purposes of a time-limited interactive search task, users seek relevance clues in order to achieve their goals, especially the context in which query terms are used in the documents [18]. It is in our intentions to further analyse our results, and identify the cases where summaries lead users to false positive answers, our aim being to improve user interaction.

A good example of this is the assessment of document AP890215-0071 for the topic "*Name four films in which Orson Welles appeared*". One summary produced for this document contained the sentence "*Turner Entertainment Co. said it will not colorize Orson Welles' black-and-white film classic "Citizen Kane" because the late director's estate may have the right to prohibit it*." which correctly identified the film Citizen Kane as being one of the films of Orson Welles.

However the summary also contained the last sentence of the document, "*Movie purists have previously lamented the colorizing of such classics as "It's a Wonderful Life", "Casablanca" and "A Christmas Carol*."" which led the searcher to credit Mr Welles as appearing in these films, even though this was not supported by the text of the full document. Although this may have affected some of our searchers, it may not be a real concern in an operational environment in which users are not so restricted by time limitations.

5. Conclusion

Our research aim was to investigate the use of summarisation techniques for a question-answering task, and the use of summaries for relevance feedback.

Our main hypothesis: that summaries could provide a useful source of evidence for relevance feedback was neither proved nor disproved due to the lack of uptake of relevance feedback on our experimental system. This lack of use of relevance feedback on both systems is disappointing and further demonstrates that encouraging users to interact with the system still remains a fundamental problem.

We achieved rather more success with the use of summaries. Although our subjects returned a low number of documents, the analysis showed that subjects returned a higher proportion of supporting documents and a lower proportion of non-supporting documents with our summarisation system. The subjects also viewed fewer full documents per relevant document found with the experimental system than the control system. Although our experiments were only partially completed, these two findings indicate that our experimental hypothesis that summaries can help users target relevant documents and eliminate non-relevant documents more effectively is worth investigating further. In addition, the positive response from our subjects towards the use of summaries indicates that summaries are not only effective, but can also be a popular aid to searching.

Our research on this area is continuing with three main aims: improving the efficiency of the summarisation-creation techniques, improving the interface, and increasing the uptake of RF. We have outlined a possible solution to the first problem in section 4.2.2, namely pre-processing the documents offline. The second and third issues we are investigating by a large-scale evaluation of users' interaction with summarisation systems, and techniques for eliciting relevance information from users. This study is considering a wider range of searching tasks and document types, to gain a broader insight into the role of summaries and RF in interactive searching.

Acknowledgements

Ian Ruthven is currently supported by the Library and Information Commission funded project '*Retrieval through explanation'*, Anastasios Tombros is supported by a University of Glasgow Postgraduate Scholarship. We gratefully acknowledge the work of Neill Alexander and Craig Brown for their contribution to this research. We would also like to acknowledge the Computing Science Research Committee for funding this research and all the experimental subjects who took part in these experiments. Thanks are also due to the anonymous reviewers and the IR Group at the University of Glasgow for their helpful comments and input.

References

1. M. Beaulieu. Experiments with interfaces to support query expansion. Journal of Documentation. **53**. 1. pp 8-19. 1997.

2. R. Brandow, K. Mitze, L. and Rau. Automatic condensation of electronic publications by sentence selection. Information Processing and Management. **31**. 5. pp 675 - 685. 1995.

3. C. Cool, S. Park, N. J. Belkin, J. Koenemann, K. B. Ng. Information seeking behavior in new searching environment. CoLIS 2. Copenhagen. pp 403-416. 1996

4. L. Downey and D. Tice. A Usability Case Study Using TREC and ZPRISE. Information Processing and Management. **35.** 5. pp 589 – 603. 1999.

5. H. Edmundson. Problems in automatic abstracting. Communications of the ACM. **7**. 4. pp 259 - 263. 1964.

6. E. N. Efthimiadis. User-choices: a new yardstick for the evaluation of ranking algorithms for interactive query expansion. Information Processing and Management. **31**. 4. pp 605 - 620. 1995.

7. D. Harman. Relevance feedback and other query modification techniques. In: Information retrieval: data structures & algorithms . W. B. Frakes and R. Baeza-Yates, ed. pp 241-263. 1992

8. D. Harman. *Relevance feedback revisited.* Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 1-10. Copenhagen. 1992.

9. W. Hersh and P. Over. TREC-9 Interactive Track Report. NIST Special Publication: The Ninth Text REtrieval Conference (TREC 9). 2001. (in press).

10. J. Koenemann and N. J. Belkin, N. A case for interaction: a study of interactive information retrieval behavior and effectiveness. CHI ‘96. 1996.

11. M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 324-331. Philadelphia. 1997

12. C. D. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. Information Processing and Management. **26**. 1. pp 171-186. 1990.

13. M. Porter and V. Galpin. Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute. Program. **22**. 1. pp 1 - 20. 1988.

14. J. J. Rocchio. Relevance feedback in information retrieval. In: The SMART retrieval system: experiments in automatic document processing. G. Salton, ed. pp 313-323. Prentice-Hall.

15. J. Rush, J, R. Salvador and A. Zamora Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. Journal of the American Society for Information Science. **22**. 4. pp 260 - 274. 1971.

16. G. Salton and C. Buckley. *Improving retrieval performance by relevance feedback.* Journal of the American Society for Information Science. **41**. 4. pp 288-297. 1990.

17. A. Spink, and T. D. Wilson. Toward a theoretical framework for information retrieval (IR) evaluation in an information seeking context. In Mira '99. electronic Workshops in Computing. S. Draper, M. Dunlop, I. Ruthven and C. J. van Rijsbergen (ed).

18. A. Tombros. Reflecting User Information Needs Through Query Biased Summaries. MSc Thesis. Technical Report (TR-1997-35) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK. 1997.

19. A. Tombros and M. Sanderson. The advantages of query-biased summaries in Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2 - 10. 1998.

20. E. Voorhees and D. Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8). pp 1. 2000.

Appendix A

Figure A.1: Interface