'Show me more': incremental length summarisation using novelty detection

Simon Sweeney
Dept. Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
simon@cis.strath.ac.uk

Fabio Crestani
Dept. Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland, UK
fabioc@cis.strath.ac.uk

David E. Losada
Depto. de Electronica y Computacion
Universidad de Santiago de Compostela, Spain
dlosada@usc.es

29th April 2007

Abstract

The paper presents a study investigating the effects of incorporating novelty detection in automatic text summarisation. Condensing a textual document, automatic text summarisation can reduce the need to refer to the source document. It also offers a means to deliver device-friendly content when accessing information in non-traditional environments. An effective method of summarisation could be to produce a summary that includes only novel information. However, a consequence of focusing exclusively on novel parts may result in a loss of context, which may have an impact on the correct interpretation of the summary, with respect to the source document. In this study we compare two strategies to produce summaries that incorporate novelty in different ways: a constant length summary, which contains only novel sentences, and an incremental summary, containing additional sentences that provide context. The aim is to establish whether a summary that contains only novel sentences provides sufficient basis to determine relevance of a document, or if indeed we need to include additional sentences to provide context. Findings from the study seem to suggest that there is only a minimal difference in performance for the tasks we set our users and that the presence of contextual information is not so important. However, for the case of mobile information access, a summary that contains only novel information does offer benefits, given bandwidth constraints.

1 Introduction

As device technologies improve and advance, so too does the services that they provide. Combined with the wealth of electronic information currently available, additional digital services add to the problem often referred to as information overload. Frequently associated with the information consumer, information overload describes the effects of having more information available than can be readily assimilated. It is not only the quantity of information items that contribute to this problem, but also the way in which information is presented.

The design and presentation of content is of particular importance when accessing information in a non-traditional setting, a case being, mobile information access. Here, to meet the demands of anytime, anywhere information access, information has to be delivered in a form that can be readily and easily digested whilst on the move. Further, additional considerations are needed to account for the inherent interaction constraints of mobile devices.

Automatic summarisation can be employed to condense a textual document, presenting only the important parts of a full text thereby reducing the need to refer to the source document. Therefore, at a document level, summarisation may be considered as a means of reducing overheads in digesting information. Traditionally, summaries can be classified as, those that are indicative of the content of the source document, and summaries that are informative, providing information contained in the document (Brandow, Mitze, & Rau, 1995). A summary can also be described according to its orientation, being document-based, containing generic information from the author's perspective, or query-based, containing content tailored to a particular user's interests. Other dimensions of summarisation have been highlighted in (Mani, 1999), but are not relevant to the work reported here.

The intended use, and consequent type of summarisation employed is an important characteristic, another is the length of the summary. Summary length is particularly important for mobile information access, given restrictions in screen displays and the associated navigational costs of scrolling vertically, or 'paging' to view content. Vertical scrolling describes the action of viewing content in a progressive manner, serially. By contrast, paging permits access to the next full screen worth of content without any further action by the user. In

terms of an optimal summary size, according to findings of our previous work (Sweeney & Crestani, 2006), it would appear that short summaries (7% of the document length) perform well for a range of display screen sizes.

Aside from summary length, another factor that could improve the effectiveness of summaries, particularly in the task of identifying relevant items with respect to an information need, is the novelty of information. In this paper we consider summarisation with novelty detection, where information is not only condensed but also an attempt is made to remove redundant information. Whilst the combination of summarisation paired with novelty detection is not a new concept (Carbonell & Goldstein, 1998), we concern ourselves with the mechanism of delivering the information. Our focus is the notion of 'show me more', where given interest in a topic, or theme of a document a reader wishes to satisfy further interests in that document. To satisfy the request to show me more we generate and deliver novel summaries, with the aim to provide additional novel information. However, there may be a negative effect by concentrating on only novel information, as there is a greater potential to misrepresent the content of document if essential contextual information is removed from summaries. With this in mind, we investigate whether summaries that contain novel sentences alone provide sufficient basis to determine relevance of a document, or if we need to include additional sentences in the summaries to provide context.

The framework we adopt to investigate novelty detection in summarisation is a user study. Our, objective therefore, is to evaluate whether or not successful novelty detection methods at sentence level are useful for the purpose of fulfilling a practical information seeking task, as reflecting in the tasks we set our users.

The remainder of the paper is structured as follows. Section 2 describes our use of 'show me more' as a framework for displaying summaries and briefly outlines the motivations of our work. Section 3 expands the motivations, and presents the research questions that the paper sets out to investigate. Section 4 outlines existing work in novelty detection and how it can be combined with summarisation. Section 5 describes the methods used to generate our query-biased novel summaries. In Section 6, we present the details of the user study we carried out to evaluate query-biased novel summaries. And in Section 7, we present the

results and analysis of the experiments. Finally, Section 8 concludes the paper with a short discussion of the implications of our findings and indicates directions for future work.

2 'Show me more'

The strategy we adopt to deliver information can be characterised as fulfilling the request to 'show me more'. In particular, we focus on the situation where, given interest in a topic, or theme of a document a reader wishes to satisfy further interests in that document. An example is a news brief that provides details of a capturing news headline, or the body text of an article that expands the details of a news brief. The 'show me more' paradigm then assumes a willingness to invest effort in viewing additional iterations of content to get more information. In our case 'show me more' describes the delivery of additional summaries. In the example of the news article, while the body text provides more details of the news contained in the brief, there will undoubtedly be overlapping, and possibly repetition of information contained in the brief. By adopting a strategy to detect novelty in the generation of summaries, we aim to reduce the amount of redundant information contained in subsequent summaries. In previous studies, we have investigated the use of hierarchical query-biased summarisation using summaries of increasing length on a mobile phone (Sweeney, Crestani, & Tombros, 2002) and a PDA (Sweeney & Crestani, 2003). In those studies our assumption of 'show me more' was simply as providing a summary of increasing length. By contrast, in the work we report here 'more' is taken to be not just a function of summary length, the size of the summary, but also the information content. This then can be considered a more intuitive approach, where 'more' (the next summary to be shown) will not only be query-biased (presenting those sentences that are relevant to the query) but also contain only novel information with respect to previously seen content.

If we consider the full text of a document consists of 3 types of sentences: (i) relevant sentences, (ii) novel sentences and the (iii) remaining sentences. A summary based on relevance will have sentences that contain content relevant to an information need. Within a summary based on relevance there may be redundant information since sentences appearing

later in the summary may repeat earlier concepts. In contrast, a summary based on novelty will contain only sentences that are both relevant and novel. However, a possible shortcoming of a summarisation strategy that focuses on presenting only novel information is the potential for a loss of context. In this sense, we refer to context as the background, or more specifically the information digested from previously seen content, which may have bearing on the correct interpretation given the source document. This constitutes the basis of the research questions that the work presented in this paper sets out to investigate. These research questions will be explained in more detail in the next section.

3 Relevance, novelty and context: research questions

To assess the notion of 'show me more' and any potential for a loss of context in novel summaries, we adopt two strategies to produce summaries that incorporate novelty in different ways; an incremental length summary, and a constant length summary. Constant length summaries contain only novel sentences, whereas the incremental length summaries contain additional sentences that provide context. To evaluate the performance of both strategies we carried out a set of user experiments. In the experiments we measure users' perception of relevance of displayed documents, in the form of the automatically generated summaries, in response to a simulated submitted query. We measure performance as users' ability to correctly identify relevant documents. The aim is to study experimentally how users' perception of relevance varies depending on the type of summary used. This should permit us to determine if summaries that contain novel sentences alone provide sufficient basis to determine relevance, or if we need to include additional sentences in the summaries to provide context.

Our aim can be characterised by the following research questions. Given the task we set our users:

1. Do query-biased summaries that take account of novelty perform better than those without novelty?

2. Do query-biased summaries that have a constant length, containing only novel sentences, perform better than those with an increasing length, where the additional sentences

provide context?

3. Finally, which of the summary configurations achieve the highest level of performance?

Answering the above questions will allow us to fulfil our underlying overall objective, which is to determine if there is an optimal strategy for showing summaries to users in response to the request to 'show me more'.

4 Background and Related Work

4.1 Novelty detection and summarisation

A large proportion of work in novelty detection has been carried out in Topic Detection and Tracking (TDT) (Allan, 2002). In the domain of news, TDT refers to the detection of breaking news in the form of new event, or first story detection and tracking the reappearance and evolution of these stories from a news stream (Allan, Carbonell, Doddington, Yamron, & Yang, 1998). Since the application of TDT to news is concerned with event-based novelty detection, the emphasis then is on detecting overlaps in event coverage in news stories, and to identify whether two news stories cover the same event. It is often the case that many of the techniques applied in TDT to detect events make use of temporal clues and other features that are particular to the structure of stories in news reporting.

Another area where novelty detection research has been actively pursued is at the Novelty tracks of the Text REtrieval Conferences '02-04 (TREC)[1]. In contrast to TDT, the novelty track is concerned with topic-based novelty detection. Here, the focus is novelty detection at a sentence level where the importance is not only on finding whether two sentences discuss the same topic, but also identifying where there is new information on the topic. Track participants are required to build a ranked list of novel relevant sentences, which consists of a two part process: (i) identify relevant sentences from a set of retrieved documents for a topic; and (ii) using the list of relevant sentences, identify those that contain new information. It is

[1]For a more details of the TREC Novelty track, and listing of other techniques submitted to the (more recent) novelty tracks refer to http://trec.nist.gov/pubs.html.

implicitly assumed that the process of topic learning happens within the task, and effects of prior knowledge are ignored.

Techniques that have been demonstrated at the Novelty track include those that are word-based and those that make use of other textual features. Using TREC'02 data, UMass experimented with a range of techniques from a simple count of new words to more complex approaches that use language models and Kullback-Leibler (KL) divergence with different smoothing strategies (Allan, Wade, & Bolivar, 2003). In their study, Allan et al. (2003) found that simple word counting methods (e.g. *NewWords*) performed no worse than other tested techniques to detect novelty at a sentence level; indeed performed best in the case where non-relevant sentences were present. More recent approaches have investigated features in sentences, such as various types of patterns of word combinations ranging from named entities and phrases, to other natural language structures.

A successful recent technique used at the track detected focus discourse in combination with new words counts (Schiffman & McKeown, 2005). This additional evidence aims to improve performance particularly in regards to achieving high precision with high compression rates, a key goal in summarisation. Another approach makes use of query-related patterns to detect expected answer types (Li & Croft, 2005). For this approach, the task of recognising novelty is interpreted as new answers to potential questions posed in a query that expresses a users' information need. Similar to techniques applied in the open-domain automatic Question Answering[2] (QA) an initial stage is required to transform the query into a question to establish the expected answer type(s). The aim then is to dramatically reduce the set of relevant items by removing those that do not match the expected answer type. Novelty is further boosted by accepting only those sentences containing answers that have not already been seen.

Other research in novelty detection at a topic-level is in adaptive information filtering (Zhang, Callan, & Minka, 2002) where document streams are monitored to find documents that match changing information needs specified by user profiles.

In terms of summarisation paired with novelty detection, early work combining query-relevance and information-novelty was in (Carbonell & Goldstein, 1998). Here, Maximal

[2]An active forum for work in open-domain question answering is at the TREC QA track, http://trec.nist.gov/data/qa.html

Marginal Relevance (MMR) was used to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passages for text summarisation. However, for the work reported in this paper, we approach novelty detection in a slightly different way. Rather than treat each sentence independently and assess novelty at a sentence level, we instead apply novelty detection at a summary level, with respect to previously seen summaries. In this way we provide the most relevant important parts of the document in response to the query first, and for any subsequent requests for more content, we present only novel information with respect to what has been already seen.

In relation to other work in novelty detection, we make use use of a similar approach to *NewWords*, used by Allan et al. (2003), as our first account of detecting novelty in generating summaries. We justify this decision based on its simplicity to implement, and on the basis that this approach performed no worse then more complex methods in situations that more accurately reflect use in a real environment, which is applicable in our work. The state of the art in novelty detection at the sentence level is actually quite simplistic. Whilst one might expect that an elaborated model should be able to capture the different subtopics of the text and, then, produce the novel sentences accordingly, the simpler"bag-of-sentences model of text" works as good as any other more evolved method (Allan et al., 2003). In the future we might expect that more evolved methods (e.g. based on subtopic structure in a text) can reach reasonable performance, and hence, outperform simple word-based methods. Nevertheless, as demonstrated by Allan et al. (2003) this was not the case in the reported experiments.

4.2 Information access in non-traditional environments

The experience of accessing information in non-traditional computing environments is very different from that in a conventional setting, using a desktop PC. Comparing information access in a mobile environment to the conventional setting there are substantial differences (Loudon, Sacher, & Kew, 2002). Aside from inherent device constraints there are additional factors, such as, user multi-tasking, carrying out a number of tasks concurrently; increased potential for distractions from outside factors, such as, noise and interruptions; and the need to fulfil users' requests in a timely manner given increased temporal and/or locational dependencies of

a transient environment. There exists a large volume of research into provisions and support for accessing information on mobile phones, PDAs, mobile communicators (telephone/PDA) and Pocket PCs. For the purposes of this paper, we restrict the scope of our review to consider devices that have display screens and can present textual information. However, we recognise that there are many other effective modes to communicate information on a mobile devices (e.g. interaction via aural interfaces).

Research investigating the effects of using small screen devices on search task performance have found that increased scrolling due to limited display size can lead to an increase in cognitive load, and impeded search task performance (Jones, Marsden, Mohd-Nasir, & Boone, 1999b). In particular, the negative effects associated with horizontal scrolling to view content outside the screen display area. When required to scroll horizontally, users reported experiencing disorientation within the information space. In another study investigating the effects of interacting with content on handheld devices, Albers and Kim (2000) also found that traversing pages led to an increase in cognitive load. Other studies have found that increased within page navigation also has the effect of increasing task completion times (Jones et al., 1999b; Kim & Albers, 2001; Jones, Buchanan, & Thimbleby, 2003). However, this may not translate to an increase in task error rates Kim and Albers (2001). These studies illustrate the issues encountered when accessing information on small screen devices. They suggest that many of problems can be associated with the within-page navigation in order to view content. The studies also provide evidence that content for small screen devices should undergo some form of processing to reduce effects due to inherent device constraints.

A variety of strategies have been investigated to make content more device-friendly when viewed on the small screen. Techniques range from the manual creation of device-specific content, to automated re-authoring approaches that apply transcoding, or transformation strategies. For a discussion of a range of adaption techniques for the small screen refer to (MacKay & Watters, 2003). Many of the approaches that exist concentrate on presenting web pages. However, other uses include access to digital library services (Buchanan, Jones, & Marsden, 2002) and support for email processing/viewing (Corston-Oliver, 2001).

Among the first to directly addresses the need of automatic tools for layout adaptation

is the Digestor project (Bickmore & Schilit, 1997). Classifying automatic adaptation techniques into two categories: syntactic techniques, based on the structure of a web page, and semantic techniques, accounting for the content of a web page, the authors describe a number of alternatives to automatically adapt the content of web pages. One of the methods described relies on a mechanism for text outlining, which supports linking to paragraphs of text within a document, and is aimed to permit quicker access to content in small devices. More recently WebTwig (Jones, Buchanan, & Mohd-Nasir, 1999a) and PowerBrowser (Buyukkokten, Garcia-Molina, Paepcke, & Winograd, 2000) have adopted a similar strategy. Both are designed to take account of limited display screens by allowing collapsing views of textual content. Here, the mechanism is used to provide an outline view to convey high-level information, while details are concealed/revelled to display further text regions of the original document. The result is a more direct and systematic approach to viewing content that requires much less scrolling. Interestingly both these schemes have more recently incorporated features that use forms of summarisation (Buyukkokten, Garcia-Molina, & Paepcke, 2001; Jones, Jones, & Deo, 2004).

Using summarisation to adapt content for small screen delivery, Buyukkokten et al. (2001) used an approach that, given an initial phase of content segmentation and extraction, can hide, partially display, make fully visible, or summarise text units. Described as an "accordian" structure the method combines summarisation with supporting the outlining action of being able to reveal/conceal content. They experimented with a variety of methods to summarise the text units of a web page, evaluating the relative performance of the summarisation methods in a user study involving information searching tasks.

An alternative, but similar style of presentation is hierarchical text summarisation. Sweeney et al. (2002; Sweeney & Crestani, 2003), investigate the use of automatically-generated hierarchical query-biased text summaries of newspaper articles presented to on WAP mobile phones, and PDAs. They describe hierarchical text summaries as having a root, or top level summary, which corresponds to the minimum level of information; each hierachy, or summary level, is then intended to provide more information. Proceeding down the hierarchy, more and more information is made available, up to a maximum, which corresponds to the

full-text of the document. For the studies, summaries were produced using a query-biased sentence extraction algorithm, where a score was assigned to a sentence to reflect its importance for inclusion in the document's summary. Scores were assigned based on examining the structural organisation of a document, utilising within-document term frequency information, and the distribution of contained query words. The final summary being generated as the desired number of top-scoring sentences, outputted in the order in which they appear in the original document. To evaluate the utility of the hierarchical query-biased summaries a user study carried out in a task-based setting. Summaries evaluated in the study ranged from title only, 7%, 15%, to 30% of the original document length. Results from the study suggest that hierarchical query-biased summaries are useful when dealing with small screens.

Radev, Kareem, and Otterbacher (2005) also use hierarchical text summarisation to summarise web documents for viewing on small, mobile devices. They describe the top level summary as presenting the most important sentences in an document, providing a gist of the content contained in the reminder of the document. Following this initial summary, users can then choose to "drill down" into the details by expanding nodes. To generate summaries, document sentences are first ranked in order of salience; a tree is then constructed using the ranking, where the root node is the highest scoring sentences. Summaries then consists of the highest scoring sentences up to a cut-off salience level. Each sentence in a summary may also act as a node that links to other sentences that have a lower salience. The salience of a sentence is computed as a linear combination of four features: centroid (similarity of the sentence to the overall document); position; length and SimWithFirst (similarity to the first sentence of the document, most cases the title/headline). In a later paper, (Otterbacher, Radev, & Kareem, 2006), evaluate the approach when used to summarise news articles sent to a web mail account (in plain text format) and accessed via a cellular phone. Comparing the hierarchical text summaries to that in which subjects were given the full text articles, there was no significant difference in task accuracy or the time taken to complete the task. Also, compared to three other summarisation methods, their users achieved significantly better accuracy on the tasks when using hierarchical summaries.

Other related work that employs summarisation as a mechanism for delivering textual

information for small screen viewing include, the integration of linguistic analysis in the summarisation process for the custom information delivery for hand-held devices (Boguraev, Bellamy, & Swart, 2001); and the application of fractal theory to summarisation for the delivery of financial news (Yang & Wang, 2003).

5 Query-biased summarisation using novel detection

We now report the methods used to generate our query-biased novel summaries. We start by describing query-biased summarisation, which forms an initial phase of the overall process, and then describe how we include novelty detection in the summary generation process.

5.1 Query-biased summarisation

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 a document's text that are more focused towards this particular information need rather than a generic, non-query-sensitive summary. Summaries of this type can then serve as an indicative function, providing a preview format to support relevance assessments on the full text of documents (Rush, Salvador, & Zamora, 1971).

The application of query-biased summarisation to aid information retrieval tasks was investigated by Tombros and Sanderson (1998). The summarisation system employed in the study we report in this paper is similar to one described in (Sweeney & Crestani, 2006) and is based on the one developed by Tombros and Sanderson. The system uses a number of sentence extraction methods (Paice, 1990) that utilise information both from the documents of the collection and from the queries used.

The underlying summarisation process relies on scoring sentences in a document to reflect their importance for inclusion in the document's summary. Scores are assigned based on evidence from the structural organisation of the document (title, leading text and heading scores), within document term-frequency information (significant term score) and the presence of query terms (query score). The final score for a sentence is computed as the sum of the

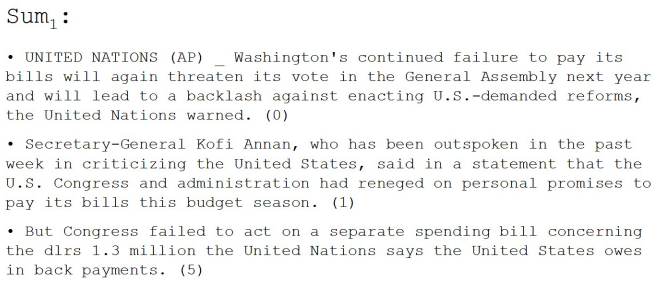

Figure 1: Query-biased summary for a typical document taken from the experiment. Query terms were: '*U.S. U.N. funding impact withhold*'.

partial scores. The inclusion of a query score, which is based on the distribution of query words in a sentence, is of particular importance and distinguishes a query-biased summary from a generic summary. Finally, the summary for a document is generated by selecting the desired number of top-scoring sentences, and outputting them in the order in which they appear in the original document. Summary length, that is, the number of sentences picked, can be controlled to restrict the level of information a user would be presented with in relation to the original document.

Figure 1 provides an example of a query-biased summary generated for a sample newswire document, which is taken from the Associated Press Wire of TREC (refer to later Section 6.3). This document was used in the experiment, which we describe later in the paper. Annotations have been added, for the purposes of reporting here, to denote the summary and to identify sentences according to their ordinal position in the source document.

5.2 Summarisation with novelty detection

The starting point for generating our novel summaries is an initial seed summary, Sum_1, which is a query-biased summary. The length of this summary, l_1, is determined as a percentage of the original document length. Given a ranked set of sentences, $s_{r_1}, s_{r_2}, ..., s_{r_n}$ (relevance-based ranking), Sum_1 is composed of the top l_1 sentences ordered as they appear in the original document.

Subsequent summaries are generated to include only novel information, and reflect previously seen summary content. In this way, a request to 'show me more' would produce a novel

summary, $SumN_2$, that contained sentences with minimal overlap with those constituting the first summary. To avoid the presentation of material that the user has already seen the focus is on the sentences which, in the original (relevance-based) rank, were ranked right after the ones selected for Sum_1. That is, $SumN_2$ will be composed of sentences selected from $s_{r_{l_1+1}}, s_{r_{l_1+2}}, ..., s_{r_n}$. Similarly, further requests to 'show me more' would continue the process of selecting sentences from lower rank positions.

To estimate how novel the candidate sentences are, a history log, composed of previously seen sentences is formed. Each candidate sentence has a relevance score greater than zero. Sentences with a zero relevance score are not included to remove those sentences considered 'not relevant' which, may be novel but off-topic with respect to the query. For the case where the number of candidate sentences is less than the number required to generate a summary then those candidate sentences available are used, supplemented with additional sentences from the history log to the number required. When there are no candidate sentences available, we assume the summary to be the same as the previous summary. However, this is a pathological case which could happen only for very short documents; steps were taken to avoid such occurrence in the evaluation.

Next, a WordsSeen list is generated from the history log. The novelty score is based on the proportion of new words with respect to the WordsSeen and compared to all words in the sentence. We compute this as the count of the number of new words divided by the sentence size, including only those words in the sentence that have been stopped and stemmed. To combine the novelty score with the relevance-based score we apply weighting to the novelty score to emphasize novelty scoring over the previous scoring matrix for a sentence. The final score for a candidate sentence is then, the sum of the novelty score with the existing relevance score. Candidate sentences are then ranked according to the combined score.

On the basis of the score ranking and on the required size, a summary is produced. The top scoring candidate sentences form the final summary. The final stage of the process involves reordering summary sentences according to their ordinal position as they occurred in the original document. Figure 2 provides example query-biased novel summaries generated for the same sample newswire document shown in the previous section. In the figure, the

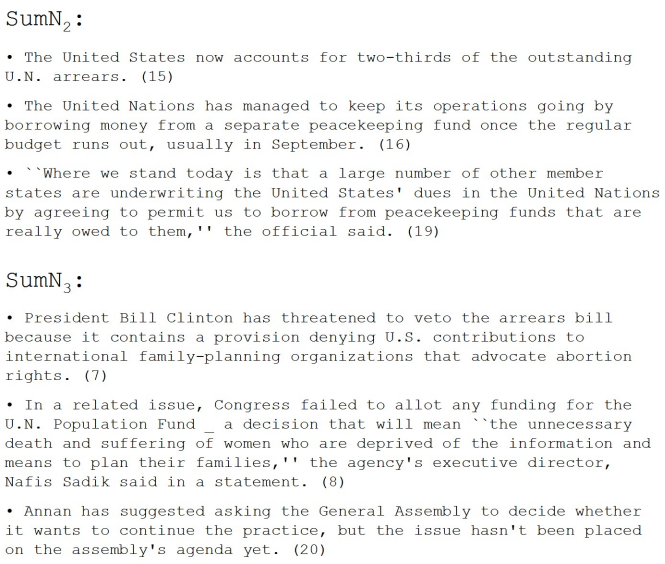

Figure 2: Query-biased novel summaries for a sample document taken from the experiment (same document as earlier figure).

summary $SumN_2$ serves the request to 'show me more' following the query-biased summary (Figure 1). The summary $SumN_3$ is then the result of a further request, and contains novel information with respect to both Sum_1 and $SumN_2$.

6 Experimental Settings

We now provide details of the experimental framework used to investigate the application of query-biased novel summarisation. To illustrate the process, a listing of summary sentences for the baseline and novel summaries of a typical document from the experimental collection is reported. This is followed by details of the document collection, and the measures used to compare the performance of the experimental summaries. The section concludes with an outline of the experimental procedure that was employed.

6.1 Experimental Arrangement

To evaluate our method of query-biased novel summarisation we carried out a set of user experiments with groups of users. To fulfil the requirements set out in our research questions, the experimental arrangement comprises two parts. Firstly, to assess the relative performance

of query-biased novel summarisation compared to query-biased summarisation. And secondly, to evaluate any effects on performance due to a loss of context in a summary that contains only novel information.

To provide a point of reference for the rest of this section it is helpful to first illustrate the complete range of summaries built for the user study. Figure 3 serves to describe the arrangement of the summaries that were generated for the experiment, whilst Table 1 serves to illustrate the summaries created for a typical document.

Figure 3 shows both the levels and types of summaries prepared. Reading in a vertical perspective the diagram can be divided along an imaginary central axis (beneath Sum_1) to show two approaches: one that incorporates novelty (left of centre), $SumN_i$ and $SumN_c$; and the baseline query-biased summaries, which do not (right of centre), $SumB_i$ and $SumB_c$. The horizontally dotted lines indicate additional levels of summary, which depending on their type may increase in length ($SumN_{i_2}$, $SumN_{i_3}$) or maintain a constant length ($SumN_{c_2}$, $SumN_{c_3}$). Example summaries, again for the same sample document, are given in Figure 4.

Key decisions made at the outset, and influence the production of summaries, relate to the number of summary levels and the length of summaries. We restrict the number of summary levels to 3, primarily to avoid overburdening users in the experimental tasks. Also, including the document title with summaries we aim to assist users in associating summary levels with the source text. In terms of summary length, for each document a number of sentences equal to 7% of its length (with a minimum of 2 sentences and maximum of 6 sentences) were used. This is supported by our previous experiments with summary length, where we found short summaries to be suited and performed well in similar tasks (Sweeney & Crestani, 2006).

Level	Novel		Baseline	
	$SumN_i$	$SumN_c$	$SumB_i$	$SumB_c$
1	0,1,5	0,1,5	0,1,5	0,1,5
2	0,1,5,15,16,19	15,16,19	0,1,5,9,10,15	9,10,15
3	0,1,5,7,8,15,16,19,20	7,8,20	0,1,3,4,5,6,9,10,15	3,4,6

Table 1: Summary sentences for document a typical document, e.g. APW19981020.1368.

A further feature shown in the diagram (Figure 3) is an indication of differences in how

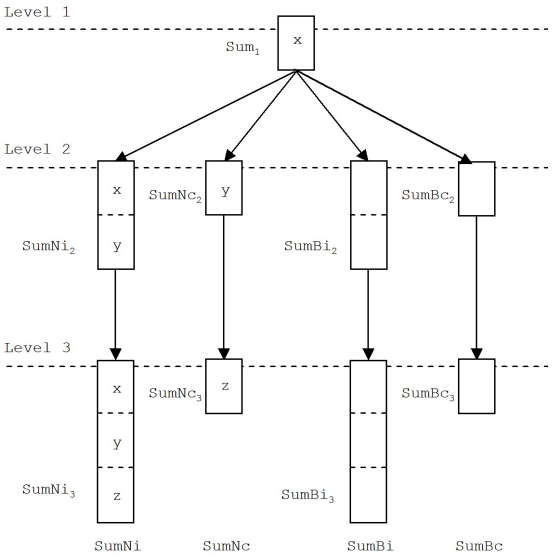

Figure 3: Illustrating the summary types built for the user study.

information content is presented. In the diagram, x represents information gained from the summary at level 1. The contrasting methods of delivery are apparent then at levels 2 and 3. For $SumN_i$, they consist of the union of what has been seen previously, and the additional new information, whereas for $SumN_c$, only the new information is shown. A similar situation happens for the baseline summaries. The overall pattern then is that the same information is conveyed in both cases, and only the method of delivery is varied.

We now describe the process of generating the summaries for the experiment. The idea is that Sum_1 is the first summary presented to the user and, then, they can ask to see more information. There are two different ways to produce the next summaries. The first method increases length (N_i) and increments the size of the next summary to be l_2, $= K * l_1$, where $K = 2$, for example, as is the case reported here. This method produces a new summary where all of the material which appeared in Sum_1 is also present in $SumN_{i_2}$. The second method maintains a constant length (N_c) and takes a very different approach producing a new summary, $SumN_{c_2}$, whose size l_2 is equal to l_1. The idea here is to avoid the presentation of material that the user has already seen, and instead focus on the sentences which, in the original (relevance-based) rank, were ranked right after the ones selected for Sum_1. That is, $SumN_{c_2}$ will be composed of sentences selected from $s_{r_{l_1+1}}, s_{r_{l_1+2}}, ..., s_{r_n}$. In contrast, the

increasing length method includes both the new sentences and the material already seen, which we consider as the context.

The generation process for both $SumN_i$ and $SumN_c$ is for the most part the same with the key difference at the final stage. The generation process then differs depending on the summary type, as follows:

- **Increasing length summaries:** A combination of the sentences taken from the history log, and the top N scoring candidate sentences form the final summary. Therefore, given $SumN_{i_1} = x$, then $SumN_{i_2} = x + y$ and $SumN_{i_3} = x + y + z$, where x, y and z represent the information content of summaries;

- **Constant length summaries:** The top N scoring candidate sentences form the final summary. Given $SumN_{c_1} = x$, then $SumN_{c_2} = y$ and $SumN_{c_3} = z$.

We used query-biased summarisation to generate the baseline summaries, and they form the basis of our comparisons.

6.2 Sample summaries for a typical document

To illustrate the described process for building novel and baseline summaries, Table 1 shows the output of the summarisation processes for a typical document. The table highlights the difference between the summaries generated using the different settings, and at each distinct level, the associated sentence identifiers. The differences between Sum_c and Sum_i are clearly shown, with the increasing length summary containing previously seen summary sentences. Also evident is the shared seed summary at level 1, which is a generic query-biased summary (recall Sum_1 in Figure 3). A final point of interest is the overlap in summary sentences between the novelty and baseline methods. The overlap, sentence 15 occurring at level 2, is most easily seen in the constant length summaries ($SumN_c$ and $SumB_c$). For some documents, the number of overlapping sentences is greater.

Figure 4 contains the summaries generated for the sample document. Annotations marking the type of summary have been added for the purposes of reporting here. For ease of cross-referencing with Table 1, sentence identifiers have also been included in the summary

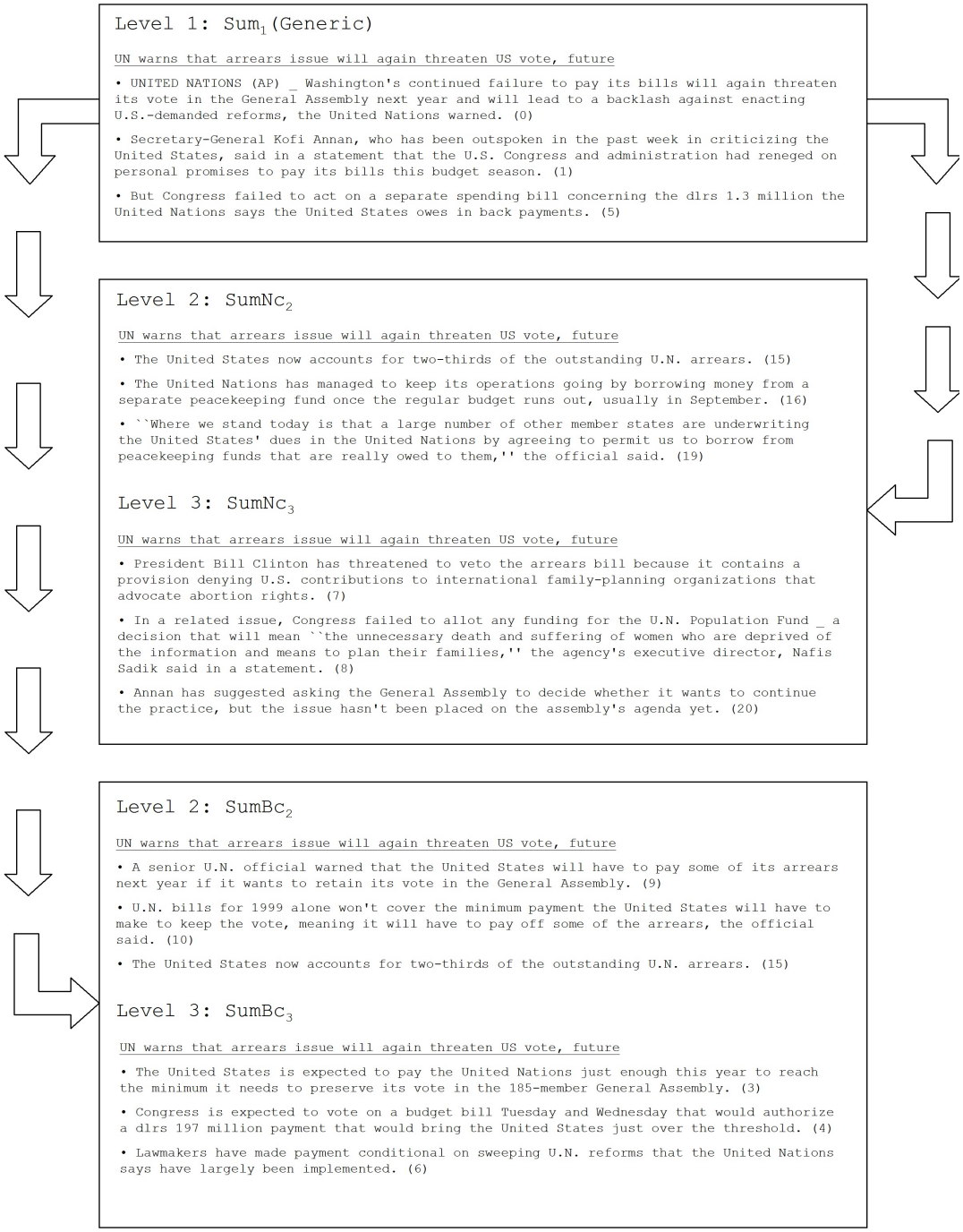

Figure 4: Summary text for typical document, e.g. APW19981020.1368 ($SumN_c$ and $SumB_c$ only).

text.

6.3 The Test Collection

Documents used in the experiment were taken from the AQUAINT collection, from the Novelty track, and consist of newswire stories from the New York Times (NYT) and Associated Press Wire (APW). Topics selected were used as a data source, providing users the necessary orientation and background with which to make their decisions. The TREC relevance assessments that are part of the collection, and made by TREC assessors, are used to enable precision and recall figures to be calculated.

It is worth noting that a feature of the novelty track is the assessment of relevance at the sentence level. In the novelty track assessors review each sentence in a document and mark it as either relevant or not relevant with respect to the topic. Therefore, we are able to make use of this sentence level information in our experiment.

A total of 5 randomly selected TREC queries and for each query, the 10 top-ranking documents were used as an input to the summarisation system. To ensure suitability of the documents for the experiment, a minimum of 5 relevant documents were present in each test set. The test collection then consisted of a total of 50 news articles.

6.4 Experimental Measures

The experimental measures used to assess the effectiveness of user relevance assessments were the *time to complete the task* and *accuracy*. We quantify accuracy as precision, recall and decision-correctness. In the experiment we focused on the variation of these measures in relation to the different experimental conditions ($SumN_i$, $SumN_c$, $SumB_i$, $SumB_c$). This is in contrast to the absolute values normally used in information retrieval (IR) research.

We define *precision* (P) as the number of documents marked correctly as relevant (in other words, found to be relevant in agreement with the TREC judges' assessments) out of the total number of documents marked. This definition corresponds to the standard definition of precision. The use of TREC relevance assessments as the "ground truth" towards which the users' decisions are measured is a procedure used by many researchers in IR. In fact, for

example, we used the same procedure in our papers in (Tombros & Crestani, 2000). *Recall* (R) is defined as the number of documents marked correctly as relevant out of the total number of relevant documents seen. A further measure we used to quantify the accuracy of a user's assessment was *decision-correctness* (DC), that is users' ability to identify correctly both the relevant document and the non-relevant (irrelevant) documents. We define decision-correctness as the sum of the number of documents marked correctly as relevant, plus the number of documents correctly marked as non-relevant out of the total number of documents marked for that query.

6.5 Experimental Design

For the experiment we recruited 20 users to form four experimental groups ($Group_1$ to $Group_4$). Participants were recruited from members of staff and postgraduate students of the Department of Computer and Information Sciences at the University of Strathclyde. The experiment was divided into two sessions, with two of the user groups completing the experimental tasks in each of the sessions. Care was taken to ensure consistency in the conditions experienced by all groups.

For the experiment, each user was given 5 queries, and for each query, the top 10 retrieved documents. There were on average 5.6 relevant documents among the documents for queries. The 10 documents were represented as 5 documents summarised using technique which included novelty, $SumN$, and 5 summarised using the baseline query-biased summarisation, $SumB$. For each document there are three summary levels as, Sum_1, Sum_2, and Sum_3 (Figure 3).

The experiment was conducted in such a way that each user group used the different system settings. The system configurations that were shown to users alternated so as to mix the different types. For example, the first document might be $SumB_i$, then the next document $SumN_c$, and then $SumB_i$, and so on. Table 2 depicts the experimental conditions used. The allocation of summary types were assigned in such a way as to avoid users' gaining preference for a type of summary over another. Both the user group and session assignments were selected randomly.

Order	Group			
	1	2	3	4
1	$SumB_i$	$SumN_i$	$SumB_c$	$SumN_c$
2	$SumN_c$	$SumB_c$	$SumN_i$	$SumB_i$
3	$SumB_i$	$SumN_i$	$SumB_c$	$SumN_c$
...	...	...	...	...

Table 2: Assignment of summaries to the experimental user groups ($Group_1$: users 1-5; $Group_2$: users 6-10, $Group_3$: users 11-15; and $Group_4$: users 16-20).

To summarise, each user was given a total of 50 documents to work through, each represented by 3 summaries. At the end of the experiment, a user had visited a total of 150 document summaries (75 with novelty $SumN$ and 75 without novelty $SumB$).

6.6 Experimental Procedure

Each user was presented with a retrieved document list in response to a simulated query (TREC topic), and tasked with identifying correctly relevant and non-relevant documents for that particular query. Further, so as not to biased quick decisions, the importance of making accurate responses was stressed to users. The information presented for each document was the automatically generated summaries.

Following an initially briefing about the experimental process and instructions by the experimenter, users were presented with a list of 5 queries. To start the experiment users were asked to select the first query from the list. The title and the description of each query (i.e., the "title" and "description" fields of the respective TREC topic[3]) provided the necessary background to their 'information need' to allow users to make relevance assessments. For each query, an initial period was allowed to read and digest the query details. Following this, the first of the 10 highest ranked documents were presented and timing for that specific document started.

Users were shown documents from the list where the content for a document consisted of the level 1, 2 and 3 summaries (e.g. $SumN_{c1}$, $SumN_{c2}$, and $SumN_{c3}$). This order, based on level, was the order that the content was presented to users. Having seen summary $SumN_{c3}$

[3]Examples of TREC topics are available at http://trec.nist.gov/data/testq_eng.html

users were required to make a decision as to whether to mark the document as relevant, or non-relevant. After indicating their decision users were presented with the first summary of the next document. On completing the final document for a query users were returned to the list of queries. The process was repeated until all queries have been evaluated.

Once all query tasks were completed a simple online questionnaire was given to the users. The key quantative data of interest, user decisions and the individual summary timing data, were recorded in logs file.

Some shortcomings to the methodology used in our experiment relate to the use of TREC topics to simulate information needs imposes an unnatural overhead on users to carry out relevance assessments. Added to this, is the use of TREC relevance assessments as the basis for comparing user decisions in order to obtain precision and recall values. However, despite this limitation the same experimental conditions applied to all of the test systems. A further factor imposed as part of the experimental design corresponds to permitting relevance decisions only after viewing all of the summaries, and not at individual summary levels. In removing the ability to make an early decision, it could be argued that we are not giving users a true representation of the case for 'show me more'. The motivation for the restriction was to ensure a consistent basis for comparing all systems. It was an assumption of the study that users would make better decisions if shown more of the original document contents. With this in mind, we evaluate the best strategy for showing the user more. Therefore, we do not expect to evaluate the system at intermediate steps (before presenting the 3rd summary) but our aim is instead to evaluate the different production strategies (incremental versus constant length) with and without novelty. To evaluate the effects of intermediatory decisions, before the 3rd summary, we carried out some additional experiments, which are reported in Section 7.5.

7 Results

We now report the results of the experiment described in the previous section. The results are reported from a number of view points. We start from an overall view of users' performance, and then consider the performance at both query and document levels. Isolating relevant

documents, we report how the make-up of these summaries may have influenced users' decision making. We then consider performance at the different summary levels. Finally, we end the section by discussing the findings from the results.

7.1 Overall performance

Table 3 provides a view of the results in the context of the experimental methodology, depicting the allocation of users to groups and associated summary types. Focusing on the different summary settings the relative performance across the experimental queries in terms of DC, P, R and average time spent is shown.

The results show a slight increase in DC and R performance with summaries that provide novelty with additional context, $SumN_i$. For P, the baseline summary with a constant length, $SumB_c$, performs best. However, the margins of improvement are somewhat minimal.

Interestingly, the margin of difference in the time spent on $SumN_i$ compared to $SumN_c$ does not agree with what we might normally expect. The additional effort to digest a longer summary (e.g. $SumN_i$) we would presuppose to translate into more time spent compared to shorter summaries (e.g. $SumN_c$). However, the results show that is not necessarily the case and the times are instead very similar. A possible reason to explain the similarity could be that users may skim the longer summaries, glancing over familiar parts, content already seen, and instead focusing on the new parts. The baseline summaries follow a more expected pattern, though again the margin of difference is small. A further observation from the table is the similarity in time spent viewing summaries between $SumN_i$ and $SumN_c$, compared to the greater level of separation observed between $SumB_i$ and $SumB_c$. However, we cannot extract significant conclusions on the basis of task completion times, since intangibles, such as, user fatigue and individual differences among users might also be important.

An alternate view of the results is given in Table 4 where the measures are separated accordingly (DC, P, R) and the results presented in a form that permits easy evaluation of the original hypotheses. In this way, by evaluating the columns we gain insight into hypothesis 1, and the rows hypothesis 2.

If we consider the case of novelty versus baseline summaries, relating to hypothesis 1,

Group	Type	DC	P	R	Time (secs)
1 & 4	$SumB_i$	0.764	0.822	0.845	66
2 & 3	$SumB_c$	0.768	0.850	0.798	53
2 & 3	$SumN_i$	0.776	0.809	0.852	64
1 & 4	$SumN_c$	0.760	0.803	0.752	63

Table 3: Average performance across all queries for the different summary types based on techniques assigned to users.

DC:

Sum	*B*	*N*
i	0.764	0.776
c	0.768	0.760

P:

Sum	*B*	*N*
i	0.822	0.809
c	0.850	0.803

R:

Sum	*B*	*N*
i	0.845	0.852
c	0.798	0.752

Time:

Sum	*B*	*N*
i	66	63
c	53	64

Table 4: Isolating the summarisation techniques in relation to the performance measures.

then for DC and R we observe that the best performing summary is from among the novel approaches (N_i, DC=0.776 and R=0.852). For P, the inverse is the case where the best performing summary setting is from the baseline (B_c, P=0.850). However, as mentioned the margins of performance improvement are small and as such inconclusive. Further inspection of Table 4 shows that evidence for a constant versus increasing length summary, relating to hypothesis 2, to be inconclusive. Carrying out appropriate statistical tests (Chi-Squared test) we found no significance difference in the overall results for the different approaches.

7.2 Query level performance

If we consider results at a query level, then Figure 5 reports the performance for each query separately. In terms of DC and P then performance levels show a degree of alignment according to whether they contain novelty, or are from the baseline. On the whole there is a pattern of improvement over the first query, with performance levelling out for intermediate queries and a drop in performance for the final query. However, an exception to this pattern is DC for the baseline approaches in the second query seen by users, query 58 (Q58), where there is a drop in performance. For R, the different summary types share a similar performance profile

but with a greater spread in the range of performance levels. However, $SumN_c$, performs noticeably worse in R compared to all other approaches, particularly in query 78 (Q78).

Indeed, the poor performance in R for $SumN_c$ may be in part due to its summarisation strategy and a tendency to 'move away' from the relevance ranking in favor of new/novel information. This fact, combined with the fact that information from the previous summary is not presented, may result in users loosing the context of sentences in the summary. During decision making then, given a degree of indecision by users based on a lack of context, they may be more inclined to mark a document as irrelevant (and, thus, some relevant documents are wrongly classified as irrelevant). Therefore, recall is harmed. On the other hand, precision is not especially worse than the other approaches as it is not as likely that many irrelevant documents will be marked as relevant.

Comparing queries in terms of the average time spent, the first query takes the greatest amount of time, with a decrease in time spent on all other queries, Figure 6. Interestingly, despite spending less time, users perform no worse in making relevance decisions for the later queries. This may be attributed to learning effects as users become more efficient in completing experimental tasks. Beyond the second query there is little variation in the times for the remaining queries, which may suggest a threshold in task efficiency.

The degree of query topic difficulty, and the language and writing style of documents, are the main factors behind the fluctuation in the observed query level performance. A further contributing factor being a period of learning as users become familiar with the experimental task. This pattern may also be observed at a document level for queries, as users' refine their interpretations of relevance. The performance drop for the final query may be explained by an element of user fatigue.

7.3 Document level performance

We now provide some indication of performance at a document level, where possible effects due to document characteristics, as well as factors relating to how users have approached the experimental task, are more apparent. However, the results should be interpreted as tentative due to limitations in the experimental design. The experimental arrangement does not allow

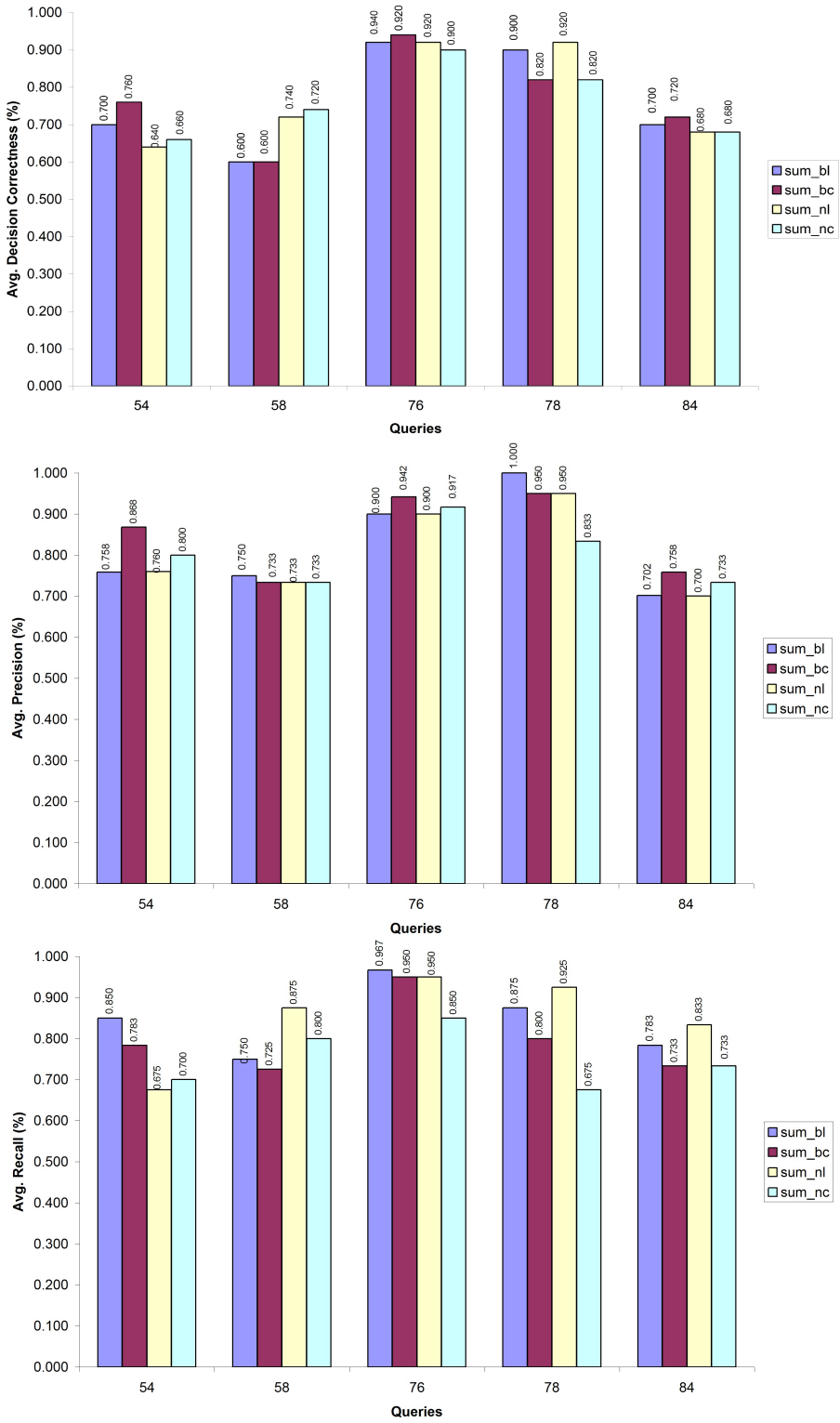

Figure 5: Average performance for individual queries for summary types based on techniques commonly seen by users.

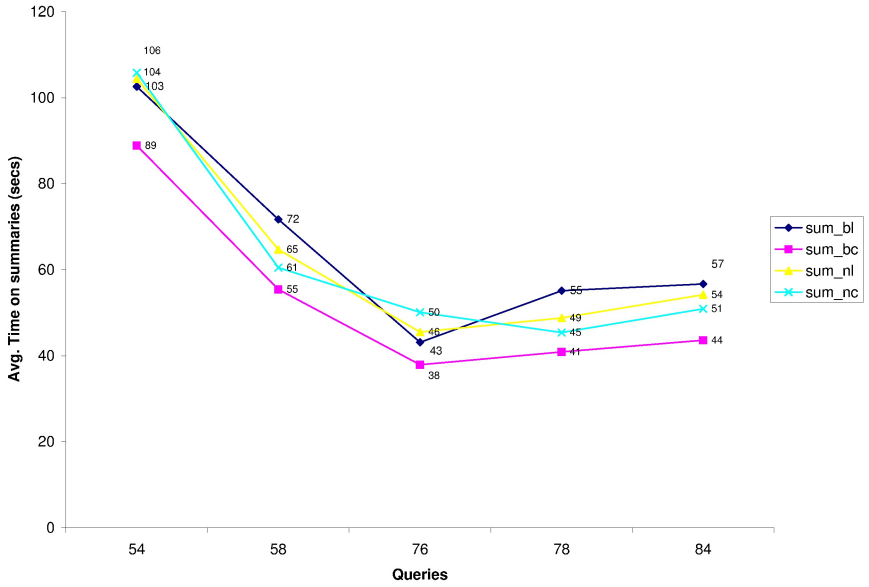

Figure 6: Average time spent viewing summaries for the experimental queries.

the extraction of definite conclusions at a document level.

To gain insight into users' performance we present the results of two experimental queries: Q58 and Q78. These two queries form opposite ends of the scale in terms of the observed performance results. The former being the worst performing query in the experiment, while the later being among the best performing queries. We focus on reporting DC results only, since this measure can account for both correct relevant and non-relevant decisions. Before we present results, it is worth recalling the mix of relevant and non-relevant documents for the two experimental queries. In Q58 there was a ratio of 6 relevant to 4 non-relevant documents and for Q78 an equal ratio, with 5 relevant and 5 non-relevant documents.

Figure 7 reports the DC levels for all documents in Q58, comparing the different types of summary and distinguishing both relevant and non-relevant documents. The results show high levels of DC for documents 5, 6, 9 and 10, and given such consistency in performance across users suggests there was little problem in correctly identifying these documents. There is a drop in performance for documents 2, 7 and 8. Indeed, for documents 2 and 8 most users consistently made incorrect decisions. This may be an indication that users have experienced difficulty in making decisions for these documents.

A factor that could contribute to a low performance for relevant documents is the presence

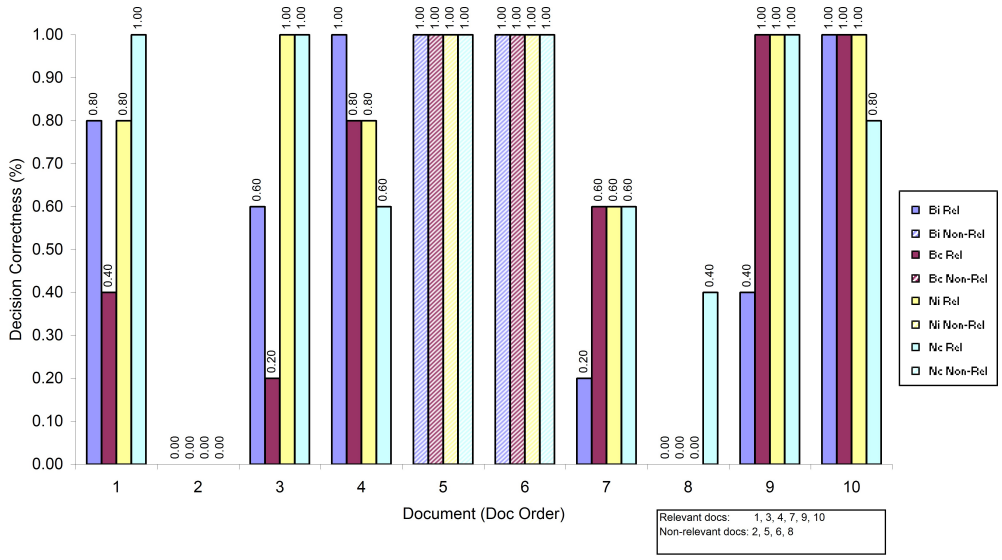

Figure 7: Decision correctness for all documents in Q58.

of few (TREC deemed) relevant sentences in summaries. For a non-relevant document the presence of sentences that suggest relevance, or are partially relevant, but in fact do not fulfil the full requirements of a relevant document, as deemed by the TREC assessors, could mislead users. This may explain the low levels of performance for the summaries of documents 2 and 8, both non-relevant documents. For the case of document 7, in the next section (Section 7.4) we shall analyse the composition of relevant document summaries to establish the proportion of relevant and non-relevant sentences.

Figure 8 shows the DC levels for all documents in Q78. Here, we can see that for most documents a high level of accuracy is achieved with the exception of documents 3, 5 and 9. Interestingly, all of these documents are relevant documents and as a result, for this query, it would seem that users appear to perform better with the non-relevant documents. This pattern is opposite to that in Q58, where users performed better with relevant documents. Achieving a greater level of accuracy in correctly identifying non-relevant documents could be attributed to a clearer distinction in a summary's content being off-topic.

Table 5 draws a comparison between users' performance with relevant and non-relevant documents for Q58 and Q78. Here, the average performance for all documents is reported according to the different summary types. Results from the table show, for Q58, despite the

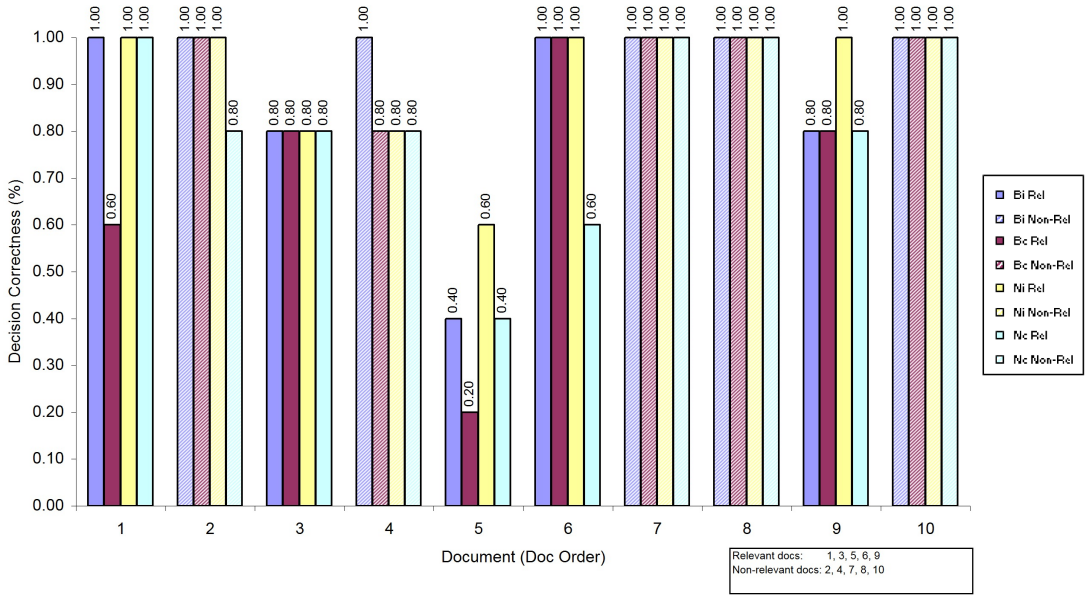

Figure 8: Decision correctness for all documents in Q78.

dip in performance for documents 7 and 8, seen in Figure 7, both novel approaches perform slightly better than the baselines. Also, that relevant documents marginally out-perform the non-relevant documents. For Q78, there is less distinction between novel and baseline summaries. However, increasing length summaries seem to fair better than the constant length summaries and, as previously mentioned, the improved performance with non-relevant documents is clearly evident.

Query	Rel docs only				NRel docs only			
	B_i	B_c	N_i	N_c	B_i	B_c	N_i	N_c
Q54	0.86	0.33	0.80	0.65	0.69	0.53	0.71	0.53
Q58	0.67	0.67	0.87	0.83	0.50	0.50	0.50	0.60
Q76	0.96	0.88	0.96	0.92	0.96	0.88	0.88	0.92
Q78	0.80	0.68	0.88	0.72	1.00	0.96	0.96	0.92
Q84	0.80	0.60	0.72	0.72	0.80	0.56	0.72	0.64

Table 5: Comparing average DC for different summary types based on all documents (Overall) and considering relevant (Rel) and non-relevant (NRel) documents separately.

Other factors that may influence users' performance at a document level include, the length of summaries and the time taken to complete the experimental tasks. Our initial intuition was that users would be more accurate in making decisions using longer summaries, since they would see more of the original document's content. Further, that greater decision

accuracy would be attained from longer viewing times. For both Q58 and Q78 there was little observed differences in performance among long and short summary lengths. Also, despite differences in summary viewing times among users, comparing hastier decisions to a longer time viewing summaries, there was again little performance variation.

7.4 Relevant sentences in summaries

We now report on the effects of relevant material in the experimental summaries, and whether having a greater proportion of relevant sentences has an influence on making correct decisions.

A further part of the test collection used for the experiment is relevance judgements at a sentence level. For relevant documents in the TREC Novelty track collection, all sentences are manually assessed and annotated as being either relevant or non-relevant. Using this listing of relevant sentences we were able to establish relevant sentences contained in the experimental summaries. Table 6 shows, for relevant documents in Q58 and Q78, the percentage of summaries containing relevant and non-relevant sentences, based on all summary levels combined. In addition, details of source documents: the document length as the number of sentences, and the number of contained relevant sentences are reported.

Apparent in Table 6 is the lack of relevant sentences in summaries where the percentage of relevant sentences in the text of a document is low. This effect is most clear for documents 3 and 9 in Q58 and appears to be independent of the summary type. As the number of relevant sentences increases in the text of a document both types of summary gain more relevant sentences. This can be seen for document 10 in Q58, where 47% of the summaries contain relevant sentences. Similar patterns are evident for relevant documents in Q78.

In terms of influence on performance, we can observe that the percentage of relevant sentences in summaries does impact on users' accuracy in DC. Combining the findings from the previous section, with insights gained from Table 6 shows that summaries that contain many relevant sentences out-perform those with low numbers of relevant sentences. In Q58, summaries for documents 2, 7 and 9 contain few relevant sentence for which users' achieve low levels of DC. By comparison, the remaining relevant documents in Q58, containing more relevant sentences and all perform better. A similar case is in Q78, comparing the low

performance with document 5 to the other relevant documents.

Query	Document				Summaries (% Rel)	
	DocId	Len (in sents)	Rel. Sents	% Rel.	$SumB_c$	$SumN_c$
58	1	40	8	20%	33%	20%
	3	42	2	5%	0%	0%
	4	39	6	15%	33%	25%
	7	62	2	3%	6%	11%
	9	81	9	11%	0%	11%
	10	53	22	42%	47%	47%
78	1	34	13	38%	44%	33%
	3	21	10	48%	67%	67%
	5	51	6	12%	25%	25%
	6	36	12	33%	67%	67%
	9	73	14	19%	67%	44%

Table 6: Document summaries for Q58 and Q76 (relevant documents only).

Figures 9 and 10 provide insight into the numbers of sentences that make up the reported percentage of summaries. The figures show, for each of the relevant documents, the number of relevant and non-relevant sentences at each summary level and evidence to explain the poor performance for certain documents. If we return to document 7 in Q58, highlighted in the previous section, there is only a single relevant sentence in the baseline summaries, and only a minor improvement of two relevant sentences for novel summaries. It therefore not unexpected that users made mistakes for this document.

We can also gauge the differences between summary types, evident beyond the first generic summary level. For Q58, there is no difference in the total number of relevant sentences for the query as a whole, whereas for Q78, there are 5 more relevant sentences for $SumB_c$ compared to $SumN_c$. However, despite containing fewer relevant sentences $SumN_c$ performed no worse than $SumB_c$. It would seem then, aside from the quantity of relevant sentences in summaries, that other factors, such as, the style of writing, technical details, and any assumed previous knowledge may have an impact on users decision accuracy.

7.5 Performance at different decision levels

It was an assumption of the study that users would make better decisions if shown more of the original document contents. As such, a study of the decision level, or "stopping point" was

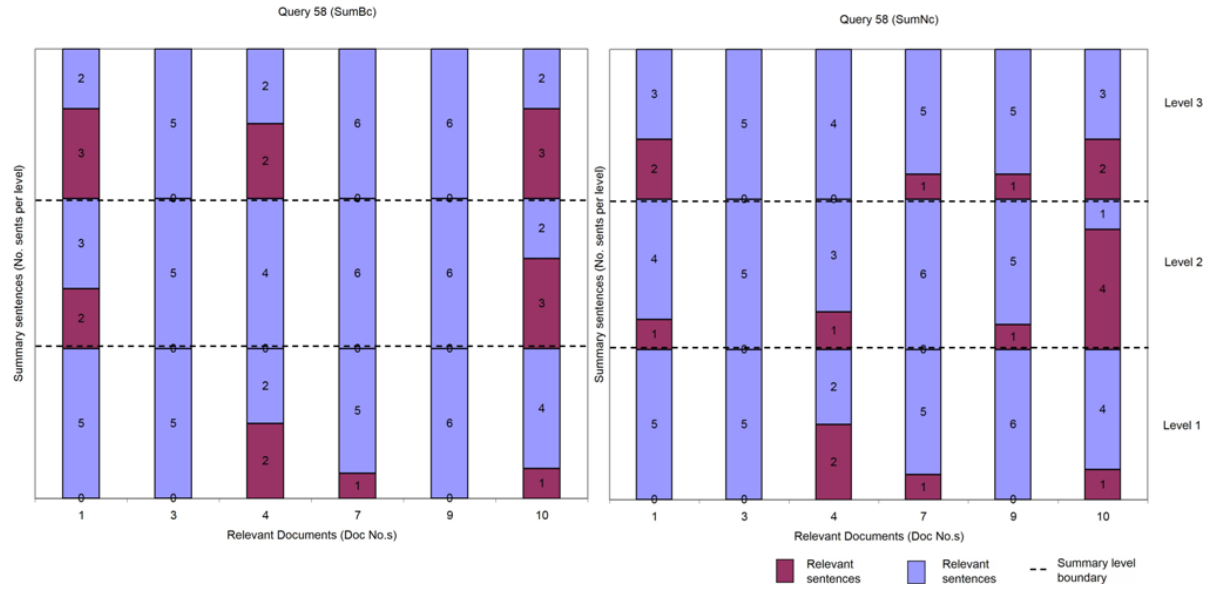

Figure 9: Relevant and non-relevant summary sentences for Query 58 (for $SumB_c$ and $SumN_c$).

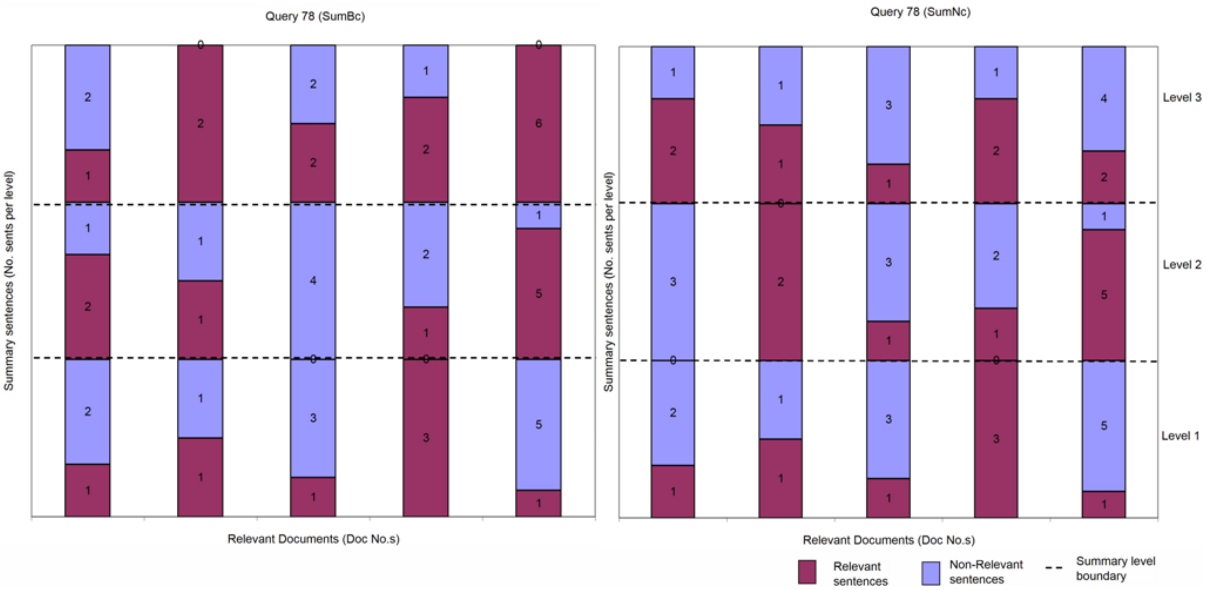

Figure 10: Relevant and non-relevant summary sentences for Query 78 (for $SumB_c$ and $SumN_c$).

not an objective in the paper. Nevertheless, we conducted a preliminary study (four users) to analyse whether this issue can have a significant effect. While in the previous section we reported the number of relevant sentences at the different levels of summary, here the investigation is concerned with the patterns of users' decisions given the freedom to select summary levels.

We now report the findings of a small additional experiment, similar to the one described, with the difference that users were able to make relevance decisions at any level. The purpose of this study was to investigate whether the original experimental design, to restrict users decisions until after all levels were seen, was valid. In addition, observe the utility of summary levels since users can select which summaries to based their decisions and determine the accuracy of their decisions. Our initial assumption, which underpinned the reason for restricting the decision level, was that shown more of the source document users would achieve a higher level of performance, compared to an early decision on the basis of seeing less of the source document.

In much the same experimental conditions as used previously, four users were given the same experimental task, to correctly identify relevant documents and assigned one of the test settings which were associated with the previous experimental groups.

Results from this new experiment indicate somewhat similar performance levels to those found in the previous experiment in terms of overall DC, P and R. An exception being in the time spent viewing summaries, which was reduced. Table 7 summarises the above finding, comparing the overall average DC, P, R and Time for both experiments. In the table, a fixed decision level refers to the previous experiment results, while a variable decision level refers to the setting described here, with the decision level being controlled by users. On the basis of these results it would seem that permitting users choice in the level of summary to make decisions does not improve the accuracy of making relevance decisions.

Decision level	DC	P	R	Time (Secs)
Fixed (Original Experiment)	0.767	0.821	0.812	61
Variable (New Experiment)	0.780	0.829	0.813	38

Table 7: Comparing the overall performance levels of experiments.

Level	User 101		User 106		User 111		User 116		Total	
1	28	(56%)	33	(66%)	16	(32%)	25	(50%)	102	(51.0%)
2	15	(30%)	13	(26%)	19	(38%)	12	(24%)	59	(29.5%)
3	7	(14%)	4	(8%)	15	(30%)	13	(26%)	39	(19.5%)

Table 8: Decisions made by users at the different summary levels.

Level	$SumB_i$		$SumB_c$		$SumN_i$		$SumN_c$	
1	30	(60%)	26	(52%)	23	(46%)	23	(46%)
2	**11**	**(22%)**	**12**	**(24%)**	**20**	**(40%)**	**16**	**(32%)**
3	**9**	**(18%)**	**12**	**(24%)**	**7**	**(14%)**	**11**	**(22%)**

Table 9: Decisions at the different summary levels according to summary type.

A further interesting insight is the utility of the summaries. For this purpose we assume those levels seen by users to be an indicator of the utility of the summaries. Table 8 presents the decisions made at the distinct summary levels by users. The table shows variation among users in the levels used to make decisions. However, a pattern most evident being that most users make fewer decisions with the lower levels (levels 2 and 3), with a large proportion of the decisions are made at level 1; in most cases half the of the total decisions. At the same time, as users do go beyond the first level, there is greater variation among users in decisions between levels 2 and 3. An alternative view of summary utility is given in Table 9. Comparing the levels of decision on the basis of summary type Table 9 focuses on the decisions made at level 2 and 3 summaries (level 1 is included only for completeness). The table shows a consistent trend in use for all summary types, with the majority of decisions being made at the first level. Interestingly, for novel summaries this pattern is less distinct as more decisions are made at level 2. Indeed, for novel summaries, the combined use of levels 2 and 3 is greater than at the first level of summary. Comparing summary use on the basis of length, then constant length summaries have slightly greater levels of access, compared to incremental length summaries. Interestingly, the greater frequencies of access of constant length summaries at level 3 (24% and 22%) may suggest that there is an increased loss of context in constant length level 2 summaries.

In terms of effectiveness in making relevance decisions at the distinct summary levels,

Level	$SumB_i$	$SumB_c$	$SumN_i$	$SumN_c$	Overall
1	0.860	0.750	0.625	0.917	0.799
2	**0.611**	**0.875**	**0.787**	**0.806**	**0.781**
3	**0.700**	**0.714**	**1.000**	**0.500**	**0.696**

Table 10: DC performance at the different summary levels according to summary type.

Table 10 provides a comparison of summary types on the basis of DC performance. Here, we can see that by isolating the summary types $SumN_i$ stands out as the best performing with the pattern of an increase in accuracy, with increasing level. However, this does not hold for $SumB_i$ summaries, and is indeed the inverse of the pattern observed for the overall picture (Table 9). Comparing the performance with novel summaries to the baselines then there is not much in the way of difference. In terms of performance at the levels, then users seem to perform best with level 1 summaries.

We now summarise the findings of the additional experiments reported in this section. The results suggest a preference among users to make decisions at the first summary level, however, in the case of novel summaries, a greater tendency to use further levels. Whilst achieving slight improvements in performance by going beyond the first level for novel summaries, on the whole, users are most accurate with the first summary. Therefore, despite the freedom to select summaries for decision making afforded by the new experimental setting this does not translating into greater levels of performance. Since users are traditionally reluctant to spend much effort to make relevance decisions and, thus, it is not surprising that our users decided often to assess the relevance of the documents right after seeing the first piece of material. Nevertheless, in this study our main aim was to determine which methods are effective to supply additional material to users when they want to spend more time reviewing a document's contents. With this in mind, we consider that our original experimental design to be valid. However, we recognise that in a more realistic deployment, that a setting more similar to the one described in this section would be more appropriate.

7.6 Discussions

We now discuss the findings in the experimental results in response to our initial intuitions held at the outset. The experimental hypotheses, which we describe in Section 3 can be briefly summarised as follows: (1) given the benefits of seeing more new information from a document, that novel summaries would out-perform summaries that contained redundant information; (2) also, based on viewing more a document, that longer summaries, including context, would perform better than short summaries.

In terms of novel query-biased summarisation, the results show no benefit of generating summaries that are novel over the baseline query-biased summaries. The inclusion of novelty in summaries, despite our initial intuitions, does not seem to be beneficial in terms of performance, or in terms of time savings. Key factors that are integral to the performance of the experimental systems and influence the results we observe are: the novelty detection algorithm adopted, and the quality of the baseline used.

If we recall how the experimental summaries were produced, a core part of the (query-biased) summary generation process was the ranking of sentences by relevance. Novelty detection was integrated as an additional feature. The baseline summaries then by definition contain a majority of relevant sentences, and, therefore, it is not unexpected that they perform well for the task we set our users. For relevant documents, the average number of relevant sentences in $SumB_c$ and in $SumN_c$ was 4 sentences. The total number of relevant sentences in all summaries combined, for all queries, in $SumB_c$ was 165 sentences, and in $SumN_c$ was 154 sentences.

An explanation for the lack of performance difference in experimental systems, despite the inclusion of novelty detection, could be due to a deficiency in the distinction of the summaries generated; that users' could not discriminate between the different types of summary, and perceived all summaries as being the same. This could also be a feature of the specific collection that comprise relatively short news documents. Therefore, given the baseline query-biased summaries are well suited to the experimental task, and difficult to improve upon, combined with the similarity in performance levels; we could surmise that the simple method of novelty detection we used, though proved to be no worse than more complex approaches

in (Allan et al., 2003), is not suitable for our purpose. A more complex method of detecting novelty, that is more selective of the sentences that make up a summary, is needed for the tasks we set our users.

Considering a constant length summary compared to an increasing length summary, the experimental results show that retaining contextual information in the summaries does not improve performance. Instead, it would seem that users are able to maintain the context of what has been previously seen, without explicit prompts in the content of a summary. Whilst there is little benefit in time savings between the two summarisation approaches, there is a benefit in savings of bandwidth for constant length summaries. This is an important finding from the point of view of mobile information access, where factors relating to costs of communication are more prominent. Benefits of a short summary then would include: reduced costs, both financially for pay-per-view content and in transmissions overheads; less navigation requirements in terms of scrolling and paging for content; and less cognitive effort to assimilate the information contained in a summary, due to a smaller amount of text to digest.

Finally, given the freedom to select summaries to base decisions, the results show a pattern of preference to make decisions at the first summary level. Users were also found to be accurate in their decisions using the first level. This supports the argument that the query-biased summaries, used for the generic first level summary and the baseline summaries, are difficult to improve on. Though again, this finding might be a consequence of the specific collections and topics used.

8 Conclusions and Future Work

Automatic text summarisation condenses a document, thereby reducing the need to refer to the full text. This can be seen as being beneficial given the problems associated with information overload. An effective way to produce a short summary maybe to include only novel information. However, producing a summary that only contains novel sentences (assuming we employ sentence extraction to build summaries) might imply a loss of context for the reader.

In this paper we considered summarisation with novelty detection, where information is not only condensed but also attempt is made to remove redundancy. Whilst the combination of summarisation paired with novelty detection is not a new concept, in the paper we concern ourselves with the mechanism of delivery, more specifically, to investigate if there a optimal strategy for showing summaries to users in response to the request to show me more.

We adopted two strategies to produce summaries that incorporate novelty in different ways: an incremental summary and a constant length summary. We compared the performance of groups of users with each of the summarisation strategies, and baseline query-biased summaries that did not include novelty. The aim was to establish whether a summary that contains only novel sentences provides sufficient basis to determine relevance of a document, or whether additional sentences need to be present to provide context.

Findings from the user study suggest that there is little difference in performance (DC, P, R and time) between the different summaries. That, the inclusion of novelty in summaries does not seem to offer benefits in terms of performance, or in terms of time savings. However, this may be due in part to the high levels of performance achieved with the baseline query-biased summaries, and the suitability of the novelty detection algorithm we adopted. In terms of a constant length summary compared to an increasing length summary, the experimental results show that retaining contextual information in the summaries does not improve performance. Instead, it would seem that users are able to maintain the context of what has been previously seen, without explicit prompts in the content of a summary. Whilst again there is little benefit in time savings between the two approaches to summary length, there is a benefit in savings of bandwidth for constant length summaries. This is an important finding from the point of view of mobile information access, where factors relating to costs of communication are more prominent. Finally, given the freedom to select summaries to make decisions, the results show both a pattern of preference, and high level of accuracy, for initial summaries.

If we revisit our initial objective to evaluate the usefulness of incorporating novelty detection in summarisation at a sentence level. On the basis of our results, it seems that users do not perceive major differences between approaches that account for novelty and those that do not. Therefore, the state of the art in sentence level novelty detection is not helpful in these

circumstances.

Extensions to the work we have presented include investigating the performance of a more refined approach to novelty detection beyond a simple count of new words. Also, it would be interesting to measure users' opinions on their confidence and perceived accuracy in making relevance decisions. In addition, investigating the use of different collections and topics could also be interesting. Finally, given the open issue of information overload, there remains the motivations to investigate strategies to assist in accessing information. Therefore, we intend to continue to explore the theme of 'show me more' to investigate other parameters of summarisation, such as, personalisation.

Acknowledgements

This work is supported by the EU Commission under the IST Project PErsonalised News content programminG (PENG) (IST-004597). More information about PENG can be found at http://www.peng-project.org/.

David E. Losada thanks the support from the "Ramn y Cajal" R&D programme (funded by FEDER & Ministerio de Edudacin y Ciencia) and project num. TIN2005-08521-C02-01.

9 References

Albers, M., & Kim, L. (2000). User web browsing characteristics using palm handhelds for information retrieval. *Proceedings of IEEE IPCC/ACM SIGDOC'00* (pp. 125–135). Piscataway, NJ, USA.

Allan, J. (2002). Introduction to topic detection and tracking (pp. 1–16). Norwell, MA, USA: Kluwer Academic Publishers.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic Detection and Tracking Pilot Study: Final Report. *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop* (pp. 194–218). Lansdowne, VA.

Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. *Proceedings of ACM SIGIR'03* (pp. 314–321). Toronto, Canada.

Bickmore, T., & Schilit, B. (1997). Digestor: "Device-independent Access to the World Wide Web". *Computer Networks and ISDN Systems*, *29*(8-13), 11075–1082.

Boguraev, B., Bellamy, R., & Swart, C. (2001). Summarisation Miniaturisation: Delivery of News to Hand-Helds. *Proceedings of Workshop on Automatic Summarization (NAACL'01)*. Pittsburgh, U.S.A.

Brandow, R., Mitze, K., & Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, *31*(5), 675–685.

Buchanan, G., Jones, M., & Marsden, G. (2002). Exploring Small Screen Digital Library Access with the Greenstone Digital Library. In M. Agosti, & C. Thanos (Eds.), *Proceedings of ACM ECDL'02*, Vol. 2458 (pp. 583–596). Lecture Notes in Computer Science, Rome, Italy: Springer-Verlag, Berlin.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. *Proceedings of 10th WWW Conference* (pp. 652–662). Hong Kong, China.

Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000). Power Browser: efficient web browsing for PDAs. *Proceedings of CHI 2000* (pp. 430–437). Amsterdam, Nederlands.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of ACM SIGIR'98* (pp. 335–336). Melbourne, Australia.

Corston-Oliver, S. (2001). Text compaction for display on very small screens. *Proceedings of Workshop on Automatic Summarization (NAACL'01)*. Pittsburgh, U.S.A.

Jones, M., Buchanan, G., & Mohd-Nasir, N. (1999a). Evaluation of WebTwig - a site outliner for handheld Web access. *Proceedings International Symposium on Handheld and Ubiqui-*

tous Computing, Vol. 1707 (pp. 343–345). Lecture Notes in Computer Science, Karlsrhue, Germany: Springer, Berlin.

Jones, M., Buchanan, G., & Thimbleby, H. (2003). Improving web search on small screen devices. *Interacting with Computers*, *15*(4), 479–495.

Jones, M., Marsden, G., Mohd-Nasir, N., & Boone, K. (1999b). Improving Web Interaction on Small Displays. *Proceedings of 8th WWW Conference* (pp. 156–157). Toronto, Canada.

Jones, S., Jones, M., & Deo, S. (2004). Using keyphrases as search result surrogates on small screen devices. *Personal Ubiquitous Computing*, *8*(1), 55–68.

Kim, L., & Albers, M. (2001). Web design issues when searching for information in a small screen display. *Proceedings of ACM SIGDOC'01* (pp. 193–200). Sante Fe, New Mexico, USA.

Li, X., & Croft, W. B. (2005). Novelty detection based on sentence level patterns. *Proceedings of ACM CIKM'05* (pp. 744–751). Bremen, Germany.

Loudon, G., Sacher, H., & Kew, L. (2002). Design Issues for Mobile Information Retrieval. *Proceedings of Workshop on Mobile Personal Information Retrieval (ACM SIGIR'02)* (pp. 3–14). Tampere, Finland.

MacKay, B., & Watters, C. (2003). The Impact of Migration of Data to Small Screens on Navigation. *IT & Society*, *1*(3), 90–101.

Mani, I. (1999). *Advances in Automatic Text Summarization.* Cambridge, MA, USA: MIT Press.

Otterbacher, J., Radev, D., & Kareem, O. (2006). News to Go: Hierachical Text Summarization for Mobile Devices. *Proceedings of ACM SIGIR'06* (pp. 589–596). Seattle, Washington, USA.

Paice, C. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, *26*(1), 171–186.

Radev, D., Kareem, O., & Otterbacher, J. (2005). Hierarchical Text Summarization for

WAP-Enabled Mobile Devices. *Proceedings of ACM SIGIR'05* (pp. 679–679). Salvador, Brazil.

Rush, J., Salvador, R., & Zamora, A. (1971). 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), 260–274.

Schiffman, B., & McKeown, K. (2005). Context and Learning in Novelty Detection. *Proceedings of ACL HLT-EMNLP'05* (pp. 716–723). Vancouver, British Columbia, Canada.

Sweeney, S., & Crestani, F. (2003). Supporting Searching on Small Screen Devices using Summarisation. *Proceedings of Mobile HCI'03 International Workshop*, Vol. 2954 (pp. 187–201). Lecture Notes in Computer Science, Udine, Italy: Springer, Berlin.

Sweeney, S., & Crestani, F. (2006). Effective search results summary size and device screen size: Is there a relationship? *Information Processing and Management*, *42*(4), 1056–1074.

Sweeney, S., Crestani, F., & Tombros, A. (2002). Mobile Delivery of News using Hierarchically Query-Biased Summaries. *Proceedings of ACM SAC'02* (pp. 634–639). Madrid, Spain.

Tombros, A., & Crestani, F. (2000). Users's perception of relevance of spoken documents. *Journal of the American Society for Information Science*, *51*(9), 929–939.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in Information Retrieval. *Proceedings of ACM SIGIR'98* (pp. 2–10). Melbourne, Australia.

Yang, C., & Wang, F. (2003). Automatic Summarization for Financial News Delivery on Mobile Devices. *Proceedings of 12th WWW Conference* (pp. 391–382). Budapest, Hungary.

Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. *Proceedings of ACM SIGIR'02* (pp. 81–88). Tampere, Finland.