Summarisation and novelty in Mobile Information Access

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 Electrónica y Computacion
Universidad de Santiago de Compostela, Spain
dlosada@usc.es

ABSTRACT

The paper presents a user study which investigates the effects of incorporating novelty detection in automatic text summarisation. The motivation being the need to provide access to information that is tailored to small screen displays. Automatic text summarisation offers a means to deliver device-friendly content. An effective summary could be one that includes only new information. However, a consequence of focusing exclusively on novel parts may result in a loss of context, which may have an impact on the ability to correctly interpret the meaning of a summary given the source document. In the user study we compares two strategies to produce summaries that incorporate novelty in different ways; an incremental summary and a constant length summary. The aim is to establish whether a summary that contains only novel sentences provides sufficient basis to determine relevance of a document, or do we need to include additional sentences to provide context. Findings from the study seem to suggest that there is minimal difference in performance for the tasks we set our users. Therefore, for the case of mobile information access a summary that contains only novel information would be more appropriate.

1. INTRODUCTION

The continued development of mobile device technologies, their supporting infrastructures and associated services is important to meet the anytime, anywhere information access demands of today's users. The growing need to deliver information on request, in a form that can be readily and easily digested on the move, continues to be a challenge. This is despite improvements in device handsets with greater battery life, support for a greater range of applications with Java compatibility (J2ME[1]) and high speed 3G[2] network access. A key feature in providing access to information in a mobile context are the limitations on interaction, in particular the size of screen displays. While font rendering and screen resolutions can, and do improve, there remains an upper threshold dictated by what is legible to the human eye, given the inherent requirements on device sizes to be portable. Therefore, the design of content for mobile delivery remains an important factor.

Automatic summarisation can be employed to condense a document, presenting only the important parts of a full text thereby reducing the need to refer to the source document. Adopting such an approach removes the need to manually construct device friendly content, while offering means to tailor summaries to be informative or indicative [4]. The intended use, and therefore type of summary employed is an important characteristic, another is the length of the summary in relation to the display screen size. In terms of an optimal summary size, according to findings of previous work [11] it would appear that short summaries (7% of the document length) perform well for a range of display screen sizes.

Assuming then short summaries as the basis for length, are there other factors that could improve the effectiveness of summaries, particularly in light of the task of identifying items of interest, or relevant content? An effective way to produce a short summary could be to include only those parts that contain novel information. However, producing a summary that contains only novel sentences (assuming we employ an approach that uses sentence extraction) may result in a loss of context. Therefore, will the novel sentences alone provide sufficient basis to determine relevance, or do we need to include additional sentences to provide context? In this sense, we refer to context as the background, or more specifically to information previously digested from a source text. If we consider that 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. In contrast a summary based on novelty will contain only sentences that are both relevant and novel. Within a summary based on relevance there may be redundant information since sentences appearing later in the summary may repeat earlier concepts.

[1]The Java™2 Platform, Micro Edition (Java ME) provides a robust, flexible environment for applications running on consumer devices, such as mobile phones.

[2]3G refers to the third generation of developments in wireless technology, especially mobile communications, which support much higher data rates intended for applications other than voice.

In this paper we consider summarisation with novelty detection, where information is not only condensed but also an attempt is made to remove redundancy. We adopt the same strategy as we employed previously to produce query-biased summaries [10, 11, 12], with the difference that given an initial summary, subsequent summaries will not only be query-biased (presenting those sentences that are relevant to the query) but also take account of novelty by reflecting the history of previously seen summaries. The scenario that describes our experimental approach is as follows: given an interest in a topic describing a point of view, a user wishes to satisfy further interest by exploring a number of document summaries to identify relevant documents. We adopt two strategies to produce summaries that incorporate novelty in different ways; an incremental summary ($SumN_i$) and a constant length summary ($SumN_c$). We compare the performance of groups of users with each of the test systems to gain insight into the following research question. Will a summary that contains only novel sentences provide sufficient basis to determine relevance of a document, or do we need to include additional sentences in the summary to provide context?

The remainder of the paper is as follows. First, we outline briefly work related to novelty detection and how it can be combined with summarisation (section 2). We then describe the process of how we generated the novel summaries for our experiment (section 3). Next, we present details of the experiment we carried out (section 4), and some of the results we collected (section 5). Finally, we conclude the paper with a short discussion of the implications of our findings in combining summarisation with novelty and indicate directions for future work (section 6).

2. RELATED WORK

A large proportion of work in novelty detection has been carried out in Topic Detection and Tracking (TDT) for the purposes of new event, or first story detection [2, 1, 5, 7, 9, 15, 16]. Typically, work in this area applies TDT to news stories where the concern is event-based novelty detection. The emphasis then is on detecting overlaps in event coverage 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)[3]. In contrast to TDT, the Novelty track is concerned with topic-based novelty detection. Here, the focus is novelty detection at the 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. For the track participants are required to build a novel ranked list of 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 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 (we adopt a similar approach in this paper) to more complex techniques that use language models and Kullback-Leibler (KL) divergence with different smoothing strategies [3]. More recent approaches have investigated features in sentences, such as, various types of word combination patterns ranging from named entities and phrases, to other natural language structures.

[3]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.

The combination of summarisation paired with novelty detection is not a new concept. Early work combining query-relevance and information-novelty was in [6], where Maximal 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. For the purposes of 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, on previously seen summaries. In this way we provide the most relevant important parts of the document in response to the query first, any subsequent requests for more content, present only new information with respect to what has been already seen.

3. GENERATING THE NOVEL SUMMARIES

3.1 Query-biased summaries

A detailed description of the methods used to build the query-biased summaries can be found in [10, 11, 12, 13, 14]. The summarisation system employed in the experiment we report is similar to one described in [11]. The system uses a number of sentence extraction methods [8] that utilise information both from the documents of the collection and from the queries used.

The underlying 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 partial scores.

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, 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.

3.2 Summaries used in the experiment

We shall now focus on those parts concerning the integration of novelty in the summary generation process. To provide a point of reference for the rest of this section we first illustrate the complete range of summaries built for the user study. Figure 1 shows both the levels and types of summaries prepared. Reading in a vertical perspective the diagram can

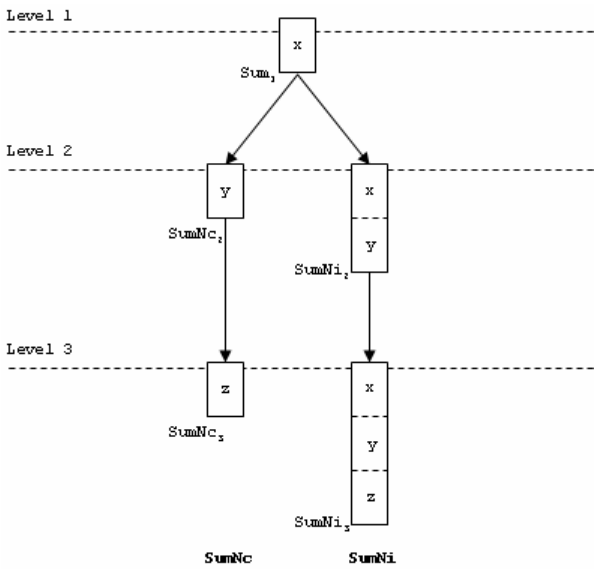

Figure 1: Illustrating the summary types built for the user study.

be divided along an imaginary central axis (beneath Sum_1) to show two approaches: one that combines novelty with constant length (left of centre), $SumN_c$; and the other that incorporates novelty with increasing length (right of centre), $SumN_i$. The horizontally dotted lines indicates increments in 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 for a sample document are given in Figure 2.

Key decisions made at the outset, which 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 performed well in similar tasks [11].

A further feature shown in the diagram (referring to Figure 1) is an indication of differences in how information content is presented. In the figure, x represents information gained from the summary at level 1. The contrasting methods of delivery are apparent then at levels 2 and 3. For $SumN_i$, levels 2 and 3 consists of the union of what was seen previously and the additional new information, whereas for $SumN_c$ only the new information is shown. The overall pattern then is that the same information is conveyed in both cases and only the method of delivery is varied.

3.3 Novel summaries

In a comparison of techniques to detect novelty at a sentence level, Alan et al. [3] found that simple word counting methods (e.g. *NewWords*) performed no worse than other methods tested; indeed performing best in the case where non-relevant sentences were present. This is the most realistic case when considering use in a real environment. We therefore make use of a similar approach to *NewWords* as our first attempt to take account of novelty when building summaries.

We shall now outline the process of building the summaries. We start from a ranked set of sentences, $s_{r_1}, s_{r_2}, ..., s_{r_n}$, obtained by the methods explained earlier (refer to section 3.1). This rank is used to produce an initial summary, Sum_1, (relevance-based only) whose length is l_1, determined from the original document length. The idea is that Sum_1 is the first summary presented to the user and, then, she/he can ask to see more information. The first method increases length (N_i) and increments the size of the next summary to be l_2, $= 2 * l_1$ producing a new summary where some 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 $s_{r_{l_1+1}}, s_{r_{l_1+2}}, ..., s_{r_n}$.

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. Taking the original relevance-based rank we first establish a list of sentences to form the history log of previously seen summary text and a set of candidate sentences, whose relevance score is greater than zero.

To compute the novelty scores for candidate sentences we generate a WordsSeen list from the history log. The WordsSeen list remains static and is not updated with new words identified as candidate sentences are evaluated. The 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. Weighting is applied 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 combined 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. It is at this stage that the generation process differs depending on the summary type. The difference in strategy is as below:

- **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

$SumN_{i_1} = x$, then $SumN_{i_2} = y$ and $SumN_{i_3} = z$.

The final stage of the process involves summary sentences being reordered as they occurred in the original document.

3.4 Sample summaries for a typical document

To illustrate the described process for building novel summaries we now provide an example for a typical document, e.g. APW19981020.1368. Table 1 shows the output of the summarisation processes; highlighting the difference between the summaries generated using the different settings for a sample document. For each distinct level there are the associated sentence identifiers, which are assigned during an initial phase of summarisation process. 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 generic (Sum_1, shown previously in Figure 1).

Figure 2 contains the summaries generated for the sample document. Annotations marking the type of summary, have been added for the purposes of reporting here. Also, for easy cross-referencing with Table 1, sentence identifiers have been included in the summary text.

Level	$SumN_i$	$SumN_c$
1	0,1,5	0,1,5
2	0,1,5,15,16,19	15,16,19
3	0,1,5,7,8,15,16,19,20	7,8,20

Table 1: Listings of summary sentence IDs for summaries of a typical document, e.g. APW19981020.1368.

Level 1: Sum_1(Generic)

UN warns that arrears issue will again threaten US vote, future

- UNITED NATIONS (AP) _ Washington's continued failure to pay its bills will again threaten its vote in the General Assembly next year and will lead to a backlash against enacting U.S.-demanded reforms, the United Nations warned. (0)
- Secretary-General Kofi Annan, who has been outspoken in the past week in criticizing the United States, said in a statement that the U.S. Congress and administration had reneged on personal promises to pay its bills this budget season. (1)
- But Congress failed to act on a separate spending bill concerning the dlrs 1.3 million the United Nations says the United States owes in back payments. (5)

Level 2: $SumNc_2$

UN warns that arrears issue will again threaten US vote, future

- The United States now accounts for two-thirds of the outstanding U.N. arrears. (15)
- The United Nations has managed to keep its operations going by borrowing money from a separate peacekeeping fund once the regular budget runs out, usually in September. (16)
- ``Where we stand today is that a large number of other member states are underwriting the United States' dues in the United Nations by agreeing to permit us to borrow from peacekeeping funds that are really owed to them,'' the official said. (19)

Level 3: $SumNc_3$

UN warns that arrears issue will again threaten US vote, future

- President Bill Clinton has threatened to veto the arrears bill because it contains a provision denying U.S. contributions to international family-planning organizations that advocate abortion rights. (7)
- In a related issue, Congress failed to allot any funding for the U.N. Population Fund _ a decision that will mean ``the unnecessary death and suffering of women who are deprived of the information and means to plan their families,'' the agency's executive director, Nafis Sadik said in a statement. (8)
- Annan has suggested asking the General Assembly to decide whether it wants to continue the practice, but the issue hasn't been placed on the assembly's agenda yet. (20)

Figure 2: Summary text for a typical document, e.g. APW19981020.1368 ($SumN_c$ only).

4. EXPERIMENTAL SETTINGS

4.1 The Test Collection

The documents used were taken from the AQUAINT collection used at the Novelty track, consisting of newswire stories from the New York Times (NYT), Associated Press Wire (APW) and Xinhua News Agency (XIE). Topics selected were used both as a data source and as a standard against which the users' relevance assessments were compared, enabling precision and recall figures to be calculated. For this last purpose the relevance assessments that are part of the collection and that were made by TREC assessors are used (refer to the discussion at the end of section 4.4).

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 4-5 relevant documents were present in each test set. The test collection then consisted of a total of 50 news articles.

4.2 Experimental Measures

The experimental measures used to assess the effectiveness of user relevance judgements 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 ($SumN_i$ and $SumN_c$). 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. *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 judgment was *decision-correctness* (DC), that is the user 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.

4.3 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.

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).

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. These 10 documents were represented as 5 documents summarised using technique which included novelty, $SumN$, and 5 summarised using a baseline technique that did not use novelty detection, $SumB$. The baseline summaries were in fact query-biased summaries. For each document there are three levels summary as, Sum_1, Sum_2, and Sum_3 (Figure 1).

The experiment was conducted in such a way that each user group experienced using the different system settings. The system configurations that were shown to users alternated so as to mix the different summary 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.

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 $SumB$).

4.4 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, users were informed that their performance scores would be penalised if they made mistakes. The information presented for each document was the automatically generated summaries.

Following an initial 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[4]) provided the necessary background to their 'information need' to allow users to make relevance judgements. For each query, an initial period was allowed to read and digest the query details. Following this, the first of the 10 documents were presented to users 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 sequence in which summaries were presented. Having seen summary $SumN_{c3}$ 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 complete, a 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 users to make 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 are therefore evaluating the best strategy for showing the user more.

5. RESULTS

We now report the results of the experiment described in the previous section. Table 3 provides a view of 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. Carrying out appropriate statistical tests (Chi-Squared test) we found no significance difference in the overall results for the different approaches.

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

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

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.

longer summary (e.g. $SumN_i$) we would expect 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 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.

If we consider results at a query level, then Table 4 shows 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, 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). 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. 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. An explanation for the fluctuation in observed query level performance could be 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. Other factors that may help explain variations in performance for queries being the degree of query topic difficulty, and the language and writing style of documents.

The average length of documents for the queries is shown in Table 5. The average length of a summary (for the generic Sum_1 at level 1) being 6 sentences. Table 6 provides some indication of the range in the lengths of summary used in the experiment.

In summary, the results from the user study suggest that there is little difference in performance (DC, P, R and in time spent viewing content) between novel summaries that include context ($SumN_i$) and those that contain only novel information ($SumN_c$). Since the same level of performance is achieved using both strategies then for the case of mobile information access, a novel constant length summary ($SumN_c$) is best. Therefore, for the point of view of mobile information access, given issues of bandwidth, we can concluded that an effective way to produce a short summary is to build one that includes only novel information. Other factors that support a short summary include: reduced transmissions costs, both financially for pay-per-view content and in bandwidth usage; less navigation requirements in terms of scrolling and paging; finally, less cognitive effort to assimilate the information contained in a summary due to a smaller amount of text to digest. However, the lack of improvement over the baseline does place doubt over the merit of building novel summaries and will require more investigation.

Query	Type	DC	P	R	Time (secs)
54	$SumB_i$	0.700	0.758	0.850	103
	$SumB_c$	0.760	0.868	0.783	89
	$SumN_i$	0.640	0.760	0.675	104
	$SumN_c$	0.660	0.800	0.700	106
58	$SumB_i$	0.600	0.750	0.750	72
	$SumB_c$	0.600	0.733	0.725	55
	$SumN_i$	0.720	0.733	0.875	65
	$SumN_c$	0.740	0.733	0.800	61
76	$SumB_i$	0.920	0.900	0.967	43
	$SumB_c$	0.940	0.942	0.950	38
	$SumN_i$	0.920	0.900	0.950	46
	$SumN_c$	0.900	0.917	0.850	50
78	$SumB_i$	0.900	1.000	0.875	55
	$SumB_c$	0.820	0.950	0.800	41
	$SumN_i$	0.920	0.950	0.925	49
	$SumN_c$	0.820	0.833	0.675	45
84	$SumB_i$	0.700	0.702	0.783	57
	$SumB_c$	0.720	0.758	0.733	44
	$SumN_i$	0.680	0.700	0.833	54
	$SumN_c$	0.680	0.733	0.733	51

Table 4: Average performance for individual queries for summary types based on techniques commonly seen by users.

	Q54	Q58	Q76	Q78	Q84
Avg. document length	55	47	54	42	49

Table 5: Average document length (in sentences) for queries in the experiment.

6. CONCLUSIONS AND FUTURE WORK

Automatic text summarisation is a potential solution to achieving device-friendly content for devices that have limited display screens. 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.

In this paper we considered summarisation with novelty de-

Summary Size	Count	% of total
Long (5 and 6 sentences)	254	63.5
Medium (3 and 4 sentences)	138	34.5
Short (2 sentences)	8	2
Total	400	-

Table 6: Range of summary lengths (in sentences) generated for the experiment.

tection, where information is not only condensed but also attempt is made to remove redundancy. We adopted two strategies to produce summaries that incorporate novelty in different ways; an incremental summary ($SumN_i$) and a constant length summary ($SumN_c$). We compared the performance of groups of users with each of the test systems. The aim was to establish whether a summary that contains only novel sentences provides sufficient basis to determine relevance of a document, or do we need to include additional sentences in the summary to provide context?

Findings from the user study suggest that there is little difference in performance (DC, P and R) between novel summaries that include context ($SumN_i$) and those that contain only novel information ($SumN_c$). Therefore, for mobile information access where issues of bandwidth and screen size are paramount then we can conclude that an effective way to produce a short summary is to build one that includes only novel information. However, the performance of the baseline summaries, for the task we set our users, questions the benefits of using novel summaries.

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. In addition, a further point of interest being to study the effects of permitting users to make decisions at any levels; to investigate summary level preference and if there is a corresponding impact on accuracy.

Acknowledgements

We would like to thank all the participants who took part in the user study. 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 "Ramón y Cajal" program, whose funds come from "Ministerio de Educación y Ciencia" and the FEDER program. His research is also partially supported by projects TIN2005-08521-C02-01 (from "Ministerio de Educación y Ciencia") and PGIDT03PXIC10501PN (from "Xunta de Galicia").

7. REFERENCES

[1] J. Allan, V. Lavrenko, and H. Jin. First story detection in TDT is hard. In *Proceedings of ACM CIKM'00*, pages 374–381, McLean, Virginia, USA, November 2000. ACM Press.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of ACM SIGIR'98*, pages 37–45, Melbourne, Australia, August 1998. ACM Press.

[3] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of ACM SIGIR'03*, pages 314–321, Toronto, Canada, July 2003. ACM Press.

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

[5] T. Brants and F. Chen. A system for new event detection. In *Proceedings of ACM SIGIR'03*, pages 330–337, Toronto, Canada, July 2003. ACM Press.

[6] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of ACM SIGIR'98*, pages 335–336, Melbourne, Australia, August 1998. ACM Press.

[7] M. Franz, A. Ittycheriah, J. S. McCarley, and T. Ward. First Story Detection: Combining Similarity and Novelty Based Approaches. Topic Detection and Tracking Workshop'01, 2001.

[8] C. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–Ű186, 1990.

[9] N. Stokes and J. Carthy. First Story Detection using a Composite Document Representation. In *Proceedings of ACL HLT'01*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[10] S. Sweeney and F. Crestani. Supporting Searching on Small Screen Devices using Summarisation. In *Proceedings of Mobile HCI'03 International Workshop*, pages 187–201, Udine, Italy, September 2003. Lecture Notes in Computer Science, volume 2954, Springer, Berlin.

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

[12] S. Sweeney, F. Crestani, and A. Tombros. Mobile Delivery of News using Hierarchically Query-Biased Summaries. In *Proceedings of ACM SAC'02*, pages 634–639, Madrid, Spain, March 2002.

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

[14] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of ACM SIGIR'98*, pages 2Ű–10, Melbourne, Australia, August 1998.

[15] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of ACM SIGIR'98*, pages 28–36, Melbourne, Australia, August 1998. ACM Press.

[16] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *Proceedings of ACM SIGKDD'02*, pages 688–693, Edmonton, Alberta, Canada, July 2002. ACM Press.