Summarisation and Novelty: An Experimental Investigation

Simon Sweeney[1], Fabio Crestani[1], and David E. Losada[2]

[1] Dept. Computer and Information Sciences
University of Strathclyde, Glasgow, Scotland, UK
{simon, fabioc}@cis.strath.ac.uk
[2] Depto. de Electrónica y Computacion
Universidad de Santiago de Compostela, Spain
dlosada@usc.es

1 Generating Novel Summaries

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.

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 detection, 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?

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. 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 [2]. Finally, we make use of a similar approach to *NewWords* in [1] as our first attempt to take account of novelty when building summaries. For a full description of the algorithm used to produce the experimental summaries please refer to an extended version of this paper to appear.

The starting point for generating our novel summaries is an initial seed summary, Sum_1, which is a query-biased summary. The query-biased summarisation

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 745–748, 2007.

system used to produce the summaries for the experiment was the same as that described in [3]. The length of this query-biased 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. 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. 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.

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. 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 emphasise 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.

2 Experimental Investigation

The documents used in the experiment were taken from the AQUAINT collection and consisted of newswire stories. A total of 5 randomly selected TREC queries and for each query, the 10 top-ranking documents were used as an input for

summary generation. The experimental measures to assess the effectiveness of user relevance judgements were the *time to complete the task*, *precision* (P), *recall* (R) and *decision-correctness* (DC). We define DC 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.

We recruited 20 users to form four experimental groups for the user study. Participants were recruited from members of staff and postgraduate students of the Department of CIS at the University of Strathclyde. 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 a technique which included novelty, $SumN$, and 5 summarised using a baseline technique that did not use novelty detection, $SumB$, which were query-biased summaries. For each document there are three levels of summary, Sum_1, Sum_2, and Sum_3.

The experimental procedure can be described as follows. Following an initial briefing users were presented with a list of 5 queries. The title and the description of each query (i.e., the 'title' and 'description' fields of the respective TREC topic[1]) 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 $SumN_{c3}$). 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. The process was repeated until all queries had been evaluated. Once all query tasks were completed a simple online questionnaire was given to the users. The key quantitative data of interest, user decisions and the individual summary timing data, were recorded in log files.

We now report results from the experiment. Due to restrictions of space we are unable to present a full analysis of all the data produced during the experimentation. Table 1 provides a view of the results in the context of the experimental methodology, depicting the allocation of users to groups and associated summary

Table 1. Average performance across all queries for the different summary types

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

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

types. 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. Appropriate statistical tests 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. A possible reason to explain the similarity in viewing times 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 $SumB_c$. It could be argued then, when we increase the size of the summary, using the baseline approach the user takes more time to digest it whereas, the increasing length summary reads better if it was constructed using novelty.

3 Conclusions and Future Work

In conclusion, 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 lack of improvement over the baseline does place doubt over the merit of building novel summaries and will require more investigation.

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.

References

1. 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.
2. 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.
3. 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.