Evaluating Summarisation Technologies: A Task Oriented Approach

Paul McLellan[1], Anastasios Tombros[1], Joemon Jose[1], Iadh Ounis[1],
and Miles Whitehead[2]

[1] Department of Computing Science, University of Glasgow, Glasgow G12 8RZ, U.K.
{mclellap, tombrosa, jj, ounis}@dcs.gla.ac.uk
[2] Research and Standards Group, Reuters Group Plc,
85 Fleet Street, London EC4P 4AJ, U.K.
miles.whitehead@reuters.com

Abstract. This paper presents a novel task-oriented approach for the evaluation of automatic text summarisation systems. Evaluation of systems has traditionally been a troublesome area in summarisation research. We propose a scheme that evaluates three existing systems by determining their relative effectiveness in an interactive search task, under conditions that approximate the intended use of the systems.

1. Introduction

Information available in a digital form has rapidly grown over the past few years. This explosive growth of information is at odds with the concerns of modern society, since time is an extremely valuable commodity. The vast quantity of available information causes users to be presented with more documents than they have time to read. Consequently, there is a demand for new innovations and technology designed to alleviate the problem.

Automatic Text Summarisation (ATS) systems offer a means of tackling the problem of information overload. They are capable of reducing the size and complexity of documents providing a concise representation which can be absorbed at a single glance. The utility of ATS systems in the context of Digital Libraries (DL) has been previously emphasised both for textual [8] and multimedia data [2]; in the context of the present paper we shall limit our discussion to textual data. Assessing the performance of automatic text summarisation systems is the issue that forms the primary focus of this paper.

Information Retrieval (IR) systems are designed to facilitate the location of relevant documents from potentially vast collections. This is particularly useful given the current proliferation of information. However, it is acknowledged that no retrieval system is infallible. The upshot of this is that any set of results *will* contain some non-relevant documents. This does not present a problem as long as the user is able to easily distinguish between relevant and non-relevant information.

Typical IR systems usually present little more than the title of each document that is retrieved. This is rarely enough information with which to make an accurate assessment of a document's relevance. Consequently, time is wasted accessing the source text of documents that may be of little or no interest. Document summaries can provide a solution to this problem.

By presenting users with a brief summary in addition to the document title, it is possible to increase the accuracy of a user's relevance assessment [16]. This is only feasible if the summary being displayed is of a suitable standard since a poor summary can be more damaging than no summary at all. It is therefore essential to have some method of judging the quality of a document summary.

It is inherently difficult to measure the quality of a summary. Simply stating that a particular summary is 'good' is of little use, as any measure of performance must be comparable to those of other summaries. Otherwise it becomes impossible to make direct comparisons between summaries or determine which is most suitable for a specific task. This paper presents a task-oriented approach to the evaluation of document summaries. The method described focuses on the use of summaries as an aid to browsing, providing a measure of performance that is both relevant to the application in question, and comparable to those of other summaries. Therefore, the main contribution of this paper is the proposal of the task-based evaluation.

The research detailed in this paper was initiated by Reuters desire to make use of document summaries in order to enhance the extensive news and information service they provide. Three different summarisation systems are being considered, and evaluated based on their effectiveness in assisting users to determine the utility of textual documents for an interactive search task quickly, and accurately. Such a task would be similar to one that users of the summarisation systems at Reuters would have to carry out. The focus of this research is to develop a task-oriented method for assessing the relative suitability of each of the three systems, and not to decide on the 'best candidate'.

2. Automatic Text Summarisation

Perhaps the major motivation for ATS is that summarisation provides a means of reducing the size and complexity of a document, while retaining the significant information of the original. Kupiec et al. [7] states that "document extracts consisting of roughly 20% of the original can be as informative as the full text". The ability to condense a document to this degree carries with it certain benefits. A representation such as this provides a concise description which can be absorbed in considerably less time than the original. Also, in the age of electronic documents there are added advantages to be had from summarisation. It is often a costly process to access the full text of documents since they may be located on a remote system. However, document summaries offer a more manageable format that can be quickly transmitted and easily stored.

The purpose for which a summary is designed can have a great bearing on its format and content. This is a concept which [13] define as the *intent* of a summary. They suggest that summaries are either indicative or informative in nature. An *indicative* *summary* contains just enough information to allow users to judge the relevance of the associated document. Indicative summaries are typically intended as an aid to browsing, and enable users to decide whether or not the full text is worth viewing. An i*nformative summary* on the other hand, is intended to act as a surrogate for the original document. It presents all the significant information contained within the full text, effectively removing any need to view the source document. For the purpose of this report we shall concentrate on the indicative function of summaries.

Traditionally summaries are manually produced; indeed it can be argued that the only perfect document summary is one written by its author. However, the ongoing ‘information revolution’ has created increasing demand for systems capable of fast, effective document summarisation. This has resulted in a great deal of research and subsequent development of a number of different ATS systems.

The goal of any ATS system is to produce an abstract style representation of a document or, in some cases, many documents. To achieve this goal ATS systems generally employ one of two methods, language generation or sentence extraction. Language generation involves a great degree of document processing and computation. It works by carefully analysing the source text to identify the key points of a document. A sentence is then generated in natural language to present each of these points. These sentences are then combined to form the document summary. More information regarding language generation systems can be found in [10], [11].

The three ATS systems being evaluated in this report all use a sentence extraction approach. Such methods proceed by selecting sentences from the original document [7], [12], [16]. Sentence extraction involves assigning importance scores to sentences based on term frequency and other characteristics of the document. A predefined number of these top-scoring sentences then form the summary. Despite the limitations of textual cohesion, balance and coverage, it is possible to produce domain independent summaries that are indicative [12].

3. Summary Evaluation

In order to compare the quality of summaries produced by different ATS systems it is important to have some form of standard evaluation. Hongyan *et al*. [5] suggests that one of the main failings in the field of automatic summarisation is the lack of just such a methodology. Many developers adopt non-standard techniques that are only suitable for their particular implementation making direct comparison across systems impossible. However, this does not imply that there are *no* common evaluation techniques, only that those which exist are not always applied.

Evaluation of ATS systems can be either *intrinsic* or *extrinsic* [4]. An intrinsic method is one which measures the overall quality of a system whereas extrinsic methods evaluate a system’s performance in relation to specific tasks.

3.1 Intrinsic Evaluation

Intrinsic evaluation typically involves comparing automatic summaries with a pre-prepared ‘ideal’ summary for the document. This target summary can be generated by a professional abstractor or by merging summaries produced by several human subjects. The latter approach can reduce the subjectiveness of the final abstract as it represents the majority opinion of several people rather than the views of an individual.

Comparisons with the ‘ideal’ summary are made in terms of precision and recall measures. Precision can be defined as ‘give me *only* significant information’. This means that the automatic summary should not contain any points that are not expressed in the target summary. Precision for an automatic summary A, is given by the number of 'correct' sentences in A divided by the total number of sentences in A.

Recall, on the other hand, equates to ‘give me *all* the significant information’. The automatic summary should present every point contained in the ‘ideal’ summary. Recall for an automatic summary A, is given by the number of 'correct' sentences in A divided by the total number of 'correct' sentences in the ideal summary.

The precision and recall values calculated for different summaries can be compared to assess their relative performance.

There is one fundamental problem with this method of evaluation. It relies on the assumption that any document has only one ‘ideal’ summary. However, it is acknowledged that a single document may have any number of acceptable abstracts, and also that judges' perception of an ideal summary may change over time [16]. Despite these facts intrinsic evaluation is still widely used.

3.2 Extrinsic Evaluation

Extrinsic, or task-oriented evaluation is designed to assess the performance of a summarisation system with respect to a particular task. The precise nature of the task involved is largely dependent on the intent of the summary being assessed. However, this type of evaluation usually involves some form of information retrieval or news analysis task.

The difficulty with task-oriented evaluation lies in establishing that a subject’s performance was directly influenced by the experimental condition being assessed. For example, an individual might be extremely skilled at the task involved and would therefore display a higher level of performance than less proficient subjects. However, this increased performance may not be attributed to the experimental condition, and consequently any conclusions that were made might be invalid. The problem can be solved through careful control of all confounding variables involved in the evaluation. This ensures that any factors which are not being evaluated are counterbalanced to limit their effect on the experimental results.

We opted for an extrinsic evaluation of the three summarisation systems, because we believe that it best captures the essence of summary utility. By measuring the effectiveness of a system in an end-user, operational environment, under conditions that approximate the ones in which the system will be used, we believe that one can draw more useful conclusions than by means of an intrinsic evaluation.

4. System Architecture

A special web-based application was developed to support the evaluation of the different summarisation systems. Construction of this application was motivated by the physical distribution of the test subjects involved in the evaluation (i.e. Glasgow and London). This made it impossible to observe subjects during testing and also meant that evaluation could not be conducted at a fixed location.

The on-line interface that was developed allowed subjects to participate in the evaluation from remote locations while enforcing all necessary experimental conditions. The system was also responsible for recording the results for each test subject. Detailed knowledge regarding the application's design and implementation is not essential for the purpose of the paper. However, for completeness the following section provides a brief overview of the system's architecture.

The system is comprised of four main components, the XML parser, the IR system, the document repository and the web application itself.

The interaction between these four modules is illustrated in Figure 1. The Reuters Research & Standards Group (RSG) supplied all necessary test data (articles and summaries in XML format). There was no direct involvement with the production of document summaries and consequently no details can be given regarding the exact process involved.

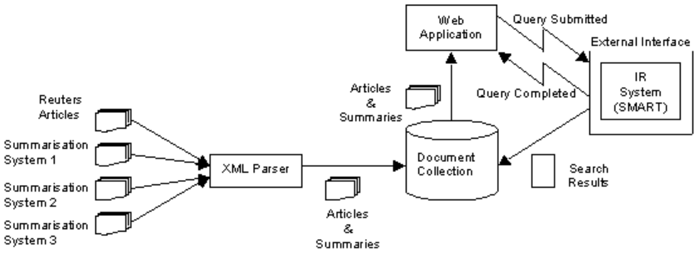


Figure 1. System architecture

The SMART retrieval system [14] formed the basis of the IR component. This was required to allow subjects to search the document collection. The purpose of this search facility will become clear when we come to look at the experimental design later in the paper.

Finally, the web application provided the front end for the evaluation system handling all interaction with the test subjects. It guided subjects through the experimental process and recorded details of their activity for subsequent analysis. Figure 2 provides a screenshot of the application showing the set of results for a sample query and a summary of one of the documents.

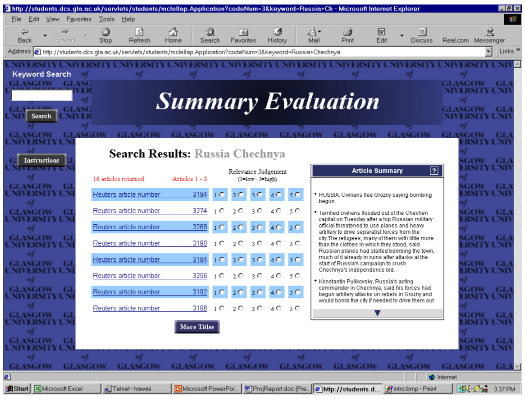
5. Experimental Design

Although the purpose of the evaluation was identified earlier in the paper, a more complete specification is given here in order to clarify the key aims of the experiment. Three text summarisation technologies were being evaluated, all of which generate summaries by extracting a subset of sentences from the source text. A collection of a thousand news stories was been summarised by each system to provide test data for the evaluation. The aim of the experiment was to obtain performance measures for each of the technologies so that a comparison could be made and the most suitable system selected.

Measuring the quality of a summary is difficult, as it tends to be an extremely subjective issue. It was suggested earlier that the content of the 'ideal' summary is dependant upon the purpose of that summary. Therefore, in order to properly assess the quality of a summary it is essential to be clear about what the summary is expected to do.

For the purpose of this experiment it was assumed that a document summary is intended to provide enough information to allow a user to judge the relevance of the related article. It is not expected to remove the need to view that article.

Given the above assumption, an 'ideal' summary would be one which allows the user to correctly judge the document's relevance to a particular information need. It should be pointed out here that a correct judgement, in this context, is one which is the same as that made after viewing the complete text. The best analogy for this is a user searching a digital library for relevant material to help them complete a specific task. The user does not expect to glean all the necessary information from the synopsis of digitally available material. His intention is simply to decide whether or not it is worthwhile viewing the rest of the data in question. If, after reading the material he has chosen he discovers that it was not relevant to his task, then clearly the summary of the material was misleading. This is an important point and one which has particular significance to the way in which summary quality is measured in our study.



http://students.dcs.gla.ac.uk/servlets/students/mclellap.Application?codeNum=3&keyword=Russia+Chechnya

Summary Evaluation

Search Results: Russia Chechnya

16 articles returned	Articles 1 - 8	Relevance Judgement (1=low - 5=high)
Reuters article number	3194	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3274	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3268	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3190	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3184	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3256	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3192	1 ○ 2 ○ 3 ○ 4 ○ 5 ○
Reuters article number	3186	1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Article Summary ?

- RUSSIA: Civilians flee Grozny saying bombing begun.
- Terrified civilians flooded out of the Chechen capital on Tuesday after a top Russian military official threatened to use planes and heavy artillery to drive separatist forces from the city.The refugees, many of them with little more than the clothes in which they stood, said Russian planes had started bombing the town, much of it already in ruins after attacks at the start of Russia's campaign to crush Chechnya's independence bid.
- Konstantin Pulikovsky, Russia's acting commander in Chechnya, said his forces had begun artillery attacks on rebels in Grozny and would bomb the city if needed to drive them out.

Figure 2. User interface

In order to obtain meaningful results, it was important that the summaries were evaluated in terms of the context in which they would be used. It was therefore essential that the summary intent identified here was also reflected in the format of the experiment. For this reason an extrinsic, task-oriented approach was adopted for evaluation.

5.1 Relevant Literature

There were a few previous studies which yielded useful information for this experiment. The first of these is an MSc thesis by Anastasios Tombros [16]. In it, the author states that "The minimal function that any useful summary should provide is being indicative of the source's content, hence helping a reader to decide whether looking at the whole document will be worthwhile". This echoes the assumption made earlier concerning the perceived role of the summary in the evaluation.

Another study, [9], presents a study of 'indicativity'. This is defined as "the ability of a catalog field to indicate the utility of a document to a searcher for a given problem …". In their experiment, subjects were presented first with titles, abstracts, index terms and the index terms that matched those of their query. They were then told to assess the usefulness of 20 documents using a simple three-point scale. Subjects were later shown the full texts of these documents and asked to make the same assessments. An indicativity rating was obtained by measuring "the fraction of evaluations made on the basis of the information in a field that were the same as those made on the basis of the full text …".

In a similar experiment, involving the incremental presentation of document fields (titles, abstracts, full text), Saracevic [15] found that "it seems to be easier … to recognise non-relevance from the shorter formats than relevance". This implies that if a document is initially judged to be non-relevant then it is less likely that the subject's opinion will change when presented with more information. This is an interesting point which should be considered when analysing the results of this experiment.

One final report of interest is [3]. This paper was concerned with relevance judgement using magnitude estimation techniques. They found that the order in which documents were presented to users could effect the relevance judgements made. They discovered that "where documents are presented to judges in a high to low rank, they will consistently underestimate the significance of documents at the higher end. In a low to high situation, there is overestimation of documents, particularly at the low to middle range". They therefore recommended that documents should always be presented in a random order.

5.2 Methodology

Subjects. The experiment employed a randomised block design, with test subjects divided into two distinct groups. The first of these was composed of individuals from Glasgow University. Subjects were chosen from a range of faculties to avoid having purely 'science minded' participants. The second group contained members of the

Reuters research department and editorial staff. It was felt that this mix of backgrounds would provide different perspectives on the evaluation.

A total of 24 subjects took part in the experiment, 12 from each group. Within these groups subjects were randomly assigned to one of three experimental conditions, relating to each of the summarisation systems being evaluated. Subjects were evenly, and randomly, distributed giving a total of 4 people from each group for each set of summaries.

Experimental Procedure. We are concerned that our subjects should be placed in a simulated work task situation in which their information needs would evolve, in just the same dynamic manner as such needs might be observed to evolve in the subjects' real working environment [1], [6]. To this end, we generated the following work task simulation:

> ‘Imagine you have been set the task of writing a short report. The subject of the report is to be the history of the Russian-Chechen conflict. This report should provide details of the key stages in the conflict and its causes. In order to help you complete this task, you are provided with a collection of Reuters news stories and a simple search facility.’

It should be noted that participants were not expected to produce a report as part of the experiment. The purpose of this definition was merely to provide subjects with a basis for their relevance assessments. The choice of topic was determined by the nature of the document collection. The collection represented 1000 arbitrarily selected articles from a single day. The restricted number and nature of the articles made it necessary to focus on a subject matter that would guarantee a substantial number of documents for ranking. This is also the main reason for which we did not allow subjects to choose a topic of personal interest.

Subjects were then asked to define a query which they felt would provide them with documents relevant to the given task. Participants were free to select any terms for their query, the only restriction being that they must retrieve at least 8 documents. If this was not the case then subjects were advised to refine their query and try again. However, it was found that no participant retrieved less than 16 articles during the experiment. Allowing an individual to determine the bounds of their own search was designed to help them formulate a clearer sense of relevance. It was felt that the act of selecting query terms compelled subjects to think about exactly what it was they were searching for.

All queries were submitted to SMART and the top 16 documents returned. In accordance with the findings of [3] these documents were presented to subjects in random order. The subjects’ task was to judge the relevance of each of these documents. Relevance assessments were made in terms of a simple five-point scale ranging from non-relevance (1) to complete relevance (5). This scale offered more freedom than a binary judgement system.

The experiment comprised two separate stages. Firstly, subjects were shown a brief summary of each of the 16 documents. They were then asked to assess the relevance of each document based on the information contained within that summary. In the second phase, subjects were presented with the same 16 documents and asked to assess their relevance again. This time however, their relevance judgements were made based on the full text of each document.

In both stages documents were displayed as a list of article numbers rather than titles to ensure that a subject's judgement was based solely on the information contained within the summary / full text. Also, the order in which the documents were presented was different in each stage. This was designed to reduce the chance of subjects basing their relevance assessment on judgements made in the previous stage.

Once the experiment had been successfully completed, subjects were solicited for any feedback they had regarding the evaluation process.

6. Experimental Measures

Three measures were used to quantify the relative performance of each system and allow direct comparisons to be made. It was essential to establish a baseline for these measurements, the ground truth. In this case, the ground truth was taken to be a subject's relevance assessment based on the complete text of a document. The rationale for this choice is that once a user has seen the full text there is no more information to be presented, and so a more accurate judgement cannot be made. Therefore, if a judgement changes after having seen the full text, it is clear that the summary did not contain enough information to make a correct judgement. The extent of the change provides an indication of just how misleading the summary was.

6.1 Indicativity

This measurement is based upon that presented by [9]. In terms of the experiment, it is concerned with the fraction of judgements made based on an article's summary that remain unchanged once the full text has been seen. The assumption is that if a subject's assessment does not alter when presented with the complete document, then the summary must have provided a suitable impression of the document's content.

In order to calculate an indicativity rating, it is necessary to find the number of documents for which a subject's relevance judgement remains constant i.e. those documents that were awarded the same score for both stage one and stage two. This number is then divided by the total number of documents that were evaluated by the subject. The resulting value provides the indicativity rating for that summary.

This is a rather simplistic measurement as no significance is attached to the extent by which a subject's initial judgement changes. The only concern here is that their opinion has changed at all.

6.2 Average Variance

This measurement can be seen as being complementary to that of indicativity, as the focus is on those documents where the subject's relevance assessment changes after being presented with the full text instead of those for which the values remain the same. The important factor here is the average amount by which the two judgements differ. It is assumed that the greater the average variation, the more inaccurate the summary must be. If the subject considers a document to be irrelevant in the first phase and yet completely relevant in the second, then the summary provided must be insubstantial. It is unimportant whether this variation is positive or negative as long as there is some difference between the two judgements.

6.3 Positivity

Positivity is neither as important, nor as scientifically sound as either indicativity or average variance. It represents a speculative measure designed to identify any trends that might exist in the relevance adjustments of a subject. For example, if the score assigned to a document increases once the full text has been revealed, then this would constitute a ‘positive’ change. On the other hand, if the score decreases then the change would be a ‘negative’ one.

If subjects are found to be consistently misjudging document relevance in one or other direction then this might suggest something about the nature of the summaries involved i.e. whether they are positively or negatively misleading.

7. Results & Analysis

The results for each of the experimental measures will now be presented, along with a discussion of the relative performance of each system. The overall scores for all 24 subjects will be discussed before the results of the two test groups (i.e. University and Reuters) are contrasted.

7.1 Indicativity

Firstly we shall consider the combined results of both groups. An overall indicativity rating for each of the three systems was calculated by averaging the results of all the subjects assigned to that system. These values are presented in Table 1 (column 2).

Table 1. Indicativity ratings

	Overall	**University**	**Reuters**
System 1	0.578	0.547	0.609
System 2	0.695	0.75	0.641
System 3	0.563	0.516	0.609

System 2 was found to have the highest overall indicativity rating with a score of 0.695. This value equates to almost 70% of documents being correctly assessed (i.e. the assessment did not change after the source text had been read) based on the summaries provided. There is only a marginal difference between the ratings of Systems 1 and 3, with System 1 perfroming slightly better. However, the difference involved is negligible, just 0.015. This is less than a single document.

We shall now look at the indicativity ratings obtained for the University group. These values are presented in Table 1 (column 3). The table clearly shows that System 2 has the highest indicativity of the three with a score of 0.75. There is a difference of over 0.2 between System 2 and the next highest score. This value constitutes a fifth of all documents that were ranked, almost thirteen extra articles. However, Systems 1 and 3 exhibit similar performance levels, with roughly half of all relevance judgements being correct. This is still an acceptable level of quality.

If we now compare these results with those of the Reuters group (column 4) we can see a certain degree of concurrency. Once again, System 2 was found to have the highest score, 0.641, which is less than that for the University group. The total variation between the three systems was also considerably smaller than that for the University group, with a difference of little over 0.03. This suggests that the relative indicativity of all systems was similar.

Despite the different indicativity ratings obtained by the two test groups, they do both point toward the same system. It is therefore reasonable to conclude that System 2 offers the clearest indication of document relevance, with the highest overall score. It is also apparent from these results that Systems 1 and 3 exhibit only a small variation in performance, if indeed any at all.

7.2 Average Variance

Table 2 (column 2) shows the combined variance results for all test subjects. As for indicativity, these overall scores represent an average of the results for every subject using that particular system.

These values offer an indication of how inaccurate, or misleading, the three sets of summaries were. Each of the scores shown represents the subjects' average error in relevance assessment. This error is in terms of the five-point scale which was employed during testing. The 'best' system is therefore that which shows the least variance, System 2 in this case. The average error in judgement was only 0.445. Once again, we see that Systems 1 and 3 have lower performance levels.

The results for the University group are presented in Table 2 (column 3). Again we see System 2 displaying the lowest average variance. Subjects who used this system were able to judge the relevance of the corresponding document with an average error of just 0.422. The variance values for the other systems are considerably greater with average errors of 0.625 and 0.656. These values imply that both of these systems presented more misleading representations than that offered by System two.

Table 2. Variance scores

	Overall	**University**	**Reuters**
System 1	0562	0.625	0.578
System 2	0.445	0.422	0.469
System 3	0.578	0.656	0.5

Now let us consider the variance scores that were calculated for the Reuters group. Table 2 (column 4) shows that once again there was some parity between the test groups. The overall variation between the systems is minimal, roughly 0.03. Again, this is in contrast to the results of the University group which suggested a marked difference in performance. The average variance has decreased for all systems (compared to the University scores), but System 2 was still found to have the lowest score and hence the best performance. One could therefore conclude that System 2 provides the most accurate representation of document content.

7.3 Positivity

This measure was intended to identify any trends in the relevance adjustments of test subjects. As such, there is no 'best' result for positivity, it is merely designed to supplement the findings of the previous measures. The positivity scores for both groups are presented in Table 3 to highlight the relationship between them.

These results indicate a certain consistency for Systems 1 and 2, in that both groups made the majority of adjustments in the same direction. It can be seen that subjects who were assigned to System 1 tended to overestimate the relevance of documents, resulting in negative adjustments when the source text was revealed. System 2, on the other hand, had the opposite effect, with subjects typically judging articles to be less relevant than they really were.

Table 3. Positivity scores

	University	**Reuters**
System 1	-1	-1
System 2	4	5
System 3	-15	7

The results for System 3 present a different picture. There is a severe disparity between the two scores which raises certain doubts about the validity of the measurement. The implication here is that the 'direction' of a subject's relevance adjustments may not be directly influenced by the quality of the summary. If a summary is inaccurate, then an incorrect relevance assessment will be made. However, the nature of this error in judgement is more likely to be a result of the subjective opinion of the individual rather than the summary itself. Overall, the findings for positivity are somewhat inconclusive.

8. Conclusions

The main contribution of the research reported in this paper is the proposal of a task-oriented framework for the evaluation of summarisation systems. Evaluation of summarisation systems has been a troublesome area. Our work takes the view that a task-oriented methodology, that evaluates the effectiveness of a summarisation system in an operational environment, can provide an accurate measure of evaluation. In our methodology we took care to ensure that the operational environment would simulate the actual environment under which the systems would be used as closely as possible.

There are also some conclusions concerning the comparison of the three systems that can be drawn from our results. Firstly, there is an indication that System 2 displays the highest performance level of the three. The combination of high indicativity and low variance offers a balanced summary quality.

Secondly, Systems 1 and 3 displayed similar levels of performance. The most notable difference between these systems was found in their positivity ratings. System 1 showed a negative trend for both test groups, whereas System 3 received radically different scores for each. This may indicate a subtle deficiency in the summaries generated by the third system, although no definite conclusions can be drawn.

Although these conclusions seem logical based on the results obtained, the reader should be aware that statistical testing has not confirmed their significance. There are two important factors that may have contributed to this. Firstly, the number of test subjects involved in the evaluation was limited, only eight per experimental condition. This is undoubtedly the most likely explanation for the lack of significance shown, and was largely due to the difficulty involved in finding willing test subjects. Secondly, it is possible that the chosen subject matter for the evaluation (Russian-Chechen conflict) was also a contributing factor. The topic was determined by the restricted nature of the test collection rather than any personal interest of the participant. Consequently, the subject's lack of interest in the topic may have affected their concentration or judgement of relevance. It is our intention to apply our proposed evaluation framework on a larger and more relevant document collection, and also in an actual user context.

To conclude, we believe that the evaluation methodology proposed in this paper does capture the utility of a summarisation system in an operational environment that approximates the summaries' intended use.

References

1. Borlund, P. and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225-250.
2. Christel, M.G. Winkler, D.B. Taylor, C.R. (1997). Multimedia abstractions for a digital video library. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pp. 21-29. Philadelphia, PA, USA.
3. Eisenberg, M., Barry, C. (1988). Order Effects: A Study of the Possible Influence of Presentation Order on User Judgements of Document Relevance. *Journal of the American Society for Information Science,* 39(5):293-300.
4. Hand, T. F. (1997). A Proposal for Task-based Evaluation of Text Summarization Systems. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarisation*, pp. 31-38. Madrid, Spain.
5. Hongyan, J., Barzilay, R., McKeown, C. Elhadad, M. (1998). Summarization Evaluation Methods: Experiments and Analysis. In *AAAI Spring Symposium on Intelligent Text Summarization, Technical Report SS-98-06 of the AAAI*, pp. 60-68. Stanford, CA, USA.
6. Jose, J.M. Furner, J., Harper, D.J. (1998). Spatial querying for image retrieval. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 232-240. Melbourne, Australia.
7. Kupiec, J., Pederson, J., Chen, F. (1995). A Trainable Document Summarizer. *Proceedings of the 18th Annual ACM SIGIR Conference*, pp. 68-73. Seattle, WA, USA.

8. Lopez, M.J.M. Rodriguez, M.B. Hidalgo, J.M.G. (1999). Using and evaluating user directed summaries to improve information access. In *Proceedings of the 3rd European Conference on Digital Libraries*, pp. 198-214. Paris, France.

9. Marcus, R. S., Kugel, P., Benenfeld, A. R. (1978). Catalog Information and Text as Indicators of Relevance. *Journal of the American Society for Information Science*, 29:15-30.

10. McKeown, K. Robin, J. Kukich, K. (1995). Generating Concise Natural Language Summaries. *Information Processing and Management* 31(5):703-745.

11. McKeown, K. Radev, D. R. (1995). Generating Summaries from Multiple News Articles. *Proceedings of the 18th Annual ACM SIGIR Conference*, pp. 74-82. Seattle, WA, USA.

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

13. Rush, J., Salvador, R., Zamora, A. (1971). Automatic abstracting and indexing: 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.

14. Salton, G. (1971). *The SMART retrieval system - experiments in automatic document processing*. Prentice-Hall, Englewood Cliffs, NJ.

15. Saracevic, T. (1969). Comparative Effects of Titles, Abstracts, and Full Texts on Relevance Judgements. *Proceedings of American Society for Information Science*, 6:293-299.

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