

Automatic Text Summarization by Paragraph Extraction

Mandar Mitra[†], Amit Singhal[‡], Chris Buckley^{††}

[†]Department of Computer Science, Cornell University, mitra@cs.cornell.edu

[‡]AT&T Labs Research, singhal@research.att.com

^{††}Sabir Research, Inc., chrisb@sabir.com

Abstract

Over the years, the amount of information available electronically has grown manifold. There is an increasing demand for automatic methods for text summarization. Domain-independent techniques for automatic summarization by paragraph extraction have been proposed in (Salton et al., 1994, Salton et al., 1996b). In this study, we attempt to evaluate these methods by comparing the automatically generated extracts to ones generated by humans. In view of the fact that extracts generated by two humans for the same article are surprisingly dissimilar, the performance of the automatic methods is satisfactory. Even though this observation calls into question the feasibility of producing perfect summaries by extraction, given the unavailability of other effective domain-independent summarization tools, we believe that this is a reasonable, though imperfect, alternative.

1 Introduction

As the amount of textual information available electronically grows rapidly, it becomes more difficult for a user to cope with all the text that is potentially of interest. Automatic text summarization methods are therefore becoming increasingly important. Consider the process by which a human accomplishes this task. Usually, the following steps are involved (Brandow et al., 1995)

- 1 understanding the content of the document,
- 2 identifying the most important pieces of information contained in it,
- 3 writing up this information

Given the variety of available information, it would be useful to have domain-independent, automatic techniques for doing this. However, automating the first and third steps for unconstrained texts is currently beyond the state

*This study was supported in part by the National Science Foundation under grant IRI-9300124

of the art (Brandow et al., 1995). Thus, the process of automatic summary generation generally reduces to the task of *extraction*, i.e., we use heuristics based upon a detailed statistical analysis of word occurrence to identify the text-pieces (sentences, paragraphs, etc.) that are likely to be most important, and concatenate the selected pieces together to form the final extract¹ (Luhn, 1958, Earl, 1970).

Techniques for sentence extraction have been proposed in (Brandow et al., 1995, Luhn, 1958, Paice, 1900, Kupiec et al., 1995). In (Salton et al., 1994, Salton et al., 1996b), the paragraph is chosen as the unit of extraction. It was expected that since a paragraph provides more context, the problems of readability and coherence that were seen in the summaries generated by sentence extraction would be, at least partially, ameliorated. Various properties of the extracts generated by different paragraph selection algorithms were observed in previous studies. In this study, we intend to do a more detailed evaluation of these different algorithms.

The remainder of the paper is organized as follows: section 2 briefly introduces text relationship maps, which constitute the main tool used in our extraction schemes, and outlines the paragraph selection algorithms, section 3 describes the experiments we conducted in order to evaluate these algorithms, section 4 discusses the evaluation method we adopted and the results of our experiments, finally, section 5 concludes the study.

2 Background

2.1 Text Relationship Maps

Usually, in information retrieval, each text or text excerpt is represented by a vector of weighted terms of the form $D_i = (d_{i1}, d_{i2}, \dots, d_{in})$ where d_{ik} represents an importance weight for term T_k attached to document D_i . The terms attached to documents for content representation purposes may be words or phrases derived from the document texts by an automatic indexing procedure, and the term weights are computed by taking into account the occurrence characteristics of

¹Henceforth, the term *summary* is used in this sense of a representative extract.

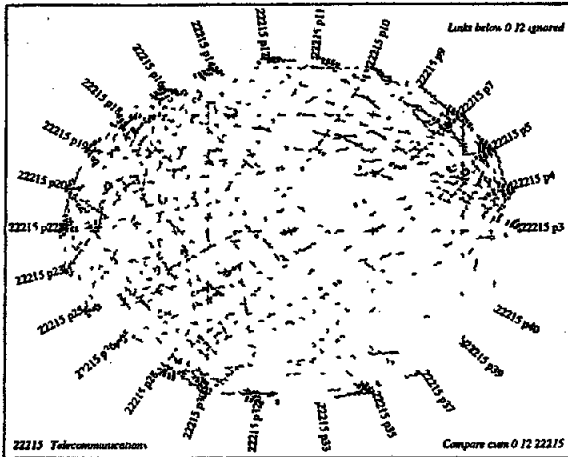


Figure 1 Text Relationship Map for article *Telecommunications*

the terms in the individual documents and the document collection as a whole (Salton and McGill, 1983) Assuming that every text or text excerpt is represented in vector form as a set of weighted terms, it is possible to compute pairwise similarity coefficients, showing the similarity between pairs of texts, based on coincidences in the term assignments to the respective items. Typically, the vector similarity might be computed as the inner product between corresponding vector elements, that is, $Sim(D_i, D_j) = \sum_{k=1}^i d_{ik} d_{jk}$, and the similarity function might be normalized to lie between 0 for disjoint vectors and 1 for completely identical vectors (Salton, 1989) The Smart information retrieval system (Salton, 1971) is based on these principles and is used in our experiments

In order to decide which paragraphs of a document are most useful for text summarization, we first want to determine how the paragraphs are related to each other. This task is accomplished using a *text relationship map*. A text relationship map is a graphical representation of textual structure, in which paragraphs (in general, pieces of text) are represented by nodes on a graph and related paragraphs are linked by edges (Salton and Allan, 1993) Nodes are joined by links based on a numerical similarity computed for each pair of texts using information retrieval techniques described above. Typically, a threshold value is selected, and all pairs of paragraphs whose similarity exceeds the threshold are connected by links. Since the similarity between two text vectors is based upon the vocabulary overlap between the corresponding texts, if the similarity between two vectors is large enough (above a threshold) to be regarded as non-random, we can say that the vocabulary matches between the corresponding texts are meaningful, and the two texts are "semantically related" (Salton et al., 1997)

Figure 1 shows a typical text relationship map. The paragraphs of the article *Telecommunications* (from the

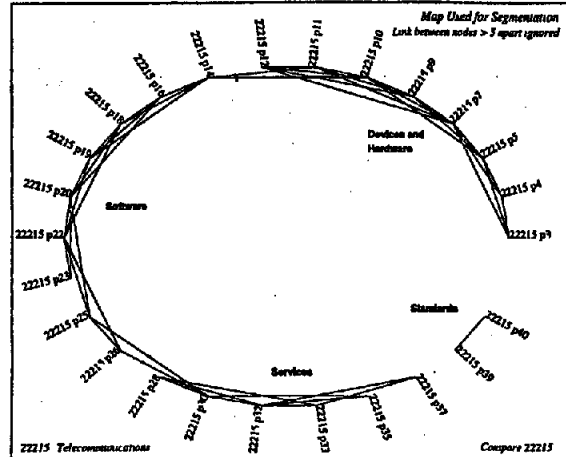


Figure 2 Text Segmentation for article *Telecommunications*

Funk and Wagnalls Encyclopedia (Funk and Wagnalls, 1979)) are denoted by nodes. Paragraphs which are sufficiently similar are joined by a link. The similarity threshold used in this map is 0.12. Important conclusions about text structure can be drawn from a text relationship map. For example, the importance of a paragraph within the text is likely to be related to the number of links incident on the corresponding node. The map can be used to identify related passages covering particular topic areas. It also provides information about the homogeneity of the text under consideration. When the map is well connected and has many cross-links between paragraphs, and direct links between adjacent paragraphs, one expects a unified, homogeneous treatment of the topic (Salton et al., 1996b)

A text relationship map maybe used to decompose a document into *segments* (Salton et al., 1996a). A segment is a contiguous piece of text that is linked internally, but largely disconnected from the adjacent text (Hearst and Plaunt, 1993). Segments are our (automatic) approximation to sectioning when a text does not have well defined sections (as is the case with numerous articles on the web these days). Consider Figure 2, for example. It shows the relationship map for the article on *Telecommunications* at a similarity threshold of 0.12 with links between distant paragraphs (paragraphs that are more than five apart) deleted. Paragraphs 3 to 12 are linked to each other, but there are few links connecting them to other nearby paragraphs. This suggests that these paragraphs deal with one topic, and the topic shifts from paragraph 12 to 14. Thus, paragraphs 3 to 12 form a segment. On reading the text, we find that they, in fact, deal with the *devices and hardware* used in telecommunications, and the topic shifts from paragraph 14 to a discussion of the *software* used in telecommunications.² Similarly, paragraphs 28

²Paragraph 13 is actually the heading for the Software sec-

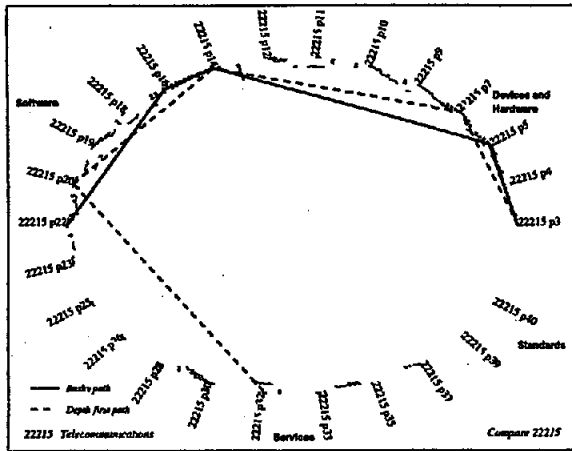


Figure 3 Global bushy and depth-first paths for article *Telecommunications*

to 35 form a segment, and this segment describes the *public telecommunication services* like electronic-mail. Paragraphs 39 and 40 form the last segment on *standards in telecommunication*. For the algorithm used to automatically generate segments for a document, see (Salton et al, 1996b, Salton et al, 1996a)

2.2 Text Traversal

We now come to the problem of generating summaries by selecting paragraphs of the document for inclusion. This could be accomplished by automatically identifying the important paragraphs on the map and traversing the selected nodes in text order to construct an extract, or *path*. Various criteria may be used to associate importance with paragraphs, giving rise to different paths. In this study, we evaluate four types of paths.

Bushy path

The *bushiness* of a node on a map is defined as the number of links connecting it to other nodes on the map. Since a highly bushy node (paragraph) is related to a number of other nodes, it has an overlapping vocabulary with many other paragraphs and is likely to discuss topics covered in several paragraphs. Such paragraphs are good overview paragraphs and are desirable in a summary, and therefore are good candidates for extraction. A global bushy path is constructed out of the n most bushy nodes on the map, where n is the targeted number of paragraphs in the summary. These nodes are arranged in chronological order, *i.e.*, the order in which they appear in the original document, to form the summary.

Depth-first path

The nodes on a bushy path are connected to a number of other paragraphs, but not necessarily to each other. Since heading paragraphs are not full-text and are not available in all domains, we do not leverage their presence in our summarization algorithms.

20% Global Bushy Path

Para 3 Telecommunications, broadly speaking, the process of transmitting information in an electronic form between any two devices by using any kind of transmission line. More specifically, however, telecommunications refers to the process.

Para 5 The devices used in telecommunications can be computers, terminals (devices that transmit and receive information), and peripheral equipment such as printers (see Computer, and see Office Systems). The transmission line used.

Para 14 Among the different kinds of software are terminal-emulation, file-transfer, host, and network software. Terminal-emulation software makes it possible for a device to perform the same functions as a terminal. File-transfer software is.

Para 16 Three major categories of telecommunication applications can be discussed here: host-terminal, file-transfer, and computer-network communications.

Para 22 In file-transfer communications, two devices are connected: either two computers, two terminals, or a computer and a terminal. One device then transmits an entire data or program file to the other device. For example, a person.

Table 1 Text for global bushy path for article *Telecommunications*

Therefore, while they may provide comprehensive coverage of an article, they may not form a very coherent extract, and the readability of the summary might be poor. To avoid this problem, we use the following strategy to build depth-first paths: start at an important node — the first node or a highly bushy node are typical choices — and visit the next most similar node at each step. Note that, only the paragraphs that follow the current one in text order are candidates for the next step. Since each paragraph is similar to the next one on the path, abrupt transitions in subject matter should be eliminated, and the extract should be a coherent one. However, since the subject matter of the paragraphs on the path is dictated to some extent by the contents of the first paragraph, all aspects of the article may not be covered by a depth-first path (Salton and Singhal, 1995, Salton et al, 1996b).

Segmented bushy path

Some articles contain segments dealing with a specialized topic. The paragraphs in such a segment would be well connected to each other, but poorly connected to other paragraphs. A bushy path would not include these paragraphs, and would thereby completely exclude an aspect of the subject matter covered in the article. A segmented bushy path attempts to remedy this problem. It is obtained by constructing bushy paths individually for each segment and concatenating them in text order. At least one paragraph is selected from each segment. The remainder of the extract is formed by picking more bushy nodes from each segment in proportion to its length. Since all segments are represented in the extract, this algorithm should, in principle, enhance the comprehensiveness of the extract (Salton et al, 1996b).

20% Global Depth-First Path

Para 3 Telecommunications, broadly speaking, the process of transmitting information in an electronic form between any two devices by using any kind of transmission line. More specifically, however, telecommunications refers to the process

Para 7 Each telecommunications device uses hardware, which connects a device to the transmission line, and software, which makes it possible for a device to transmit information

Para 14 Among the different kinds of software are terminal-emulation, file-transfer, host, and network software. Terminal-emulation software makes it possible for a device to perform the same functions as a terminal. File-transfer software is

Para 20 Finally, most host computers can communicate properly with only one kind of terminal. To communicate with such computers, terminal-emulation software is installed on a computer to make the linkage succeed

Para 32 An information-retrieval service leases time on a host computer to terminals, so that these terminals are able to retrieve information from the host computer. An example is CompuServe Information Services. To gain access to

Table 2 Text for global depth-first path for article *Telecommunications*

Augmented segmented bushy path

Typically authors introduce a new topic (for example a "Section") in the first few paragraphs that discuss the topic in the text. If proper sectioning information were available for all documents, a reasonable summarization scheme might be to select the first paragraph from each Section. A segmented bushy path might skip the less bushy introductory paragraph of a segment in favor of a

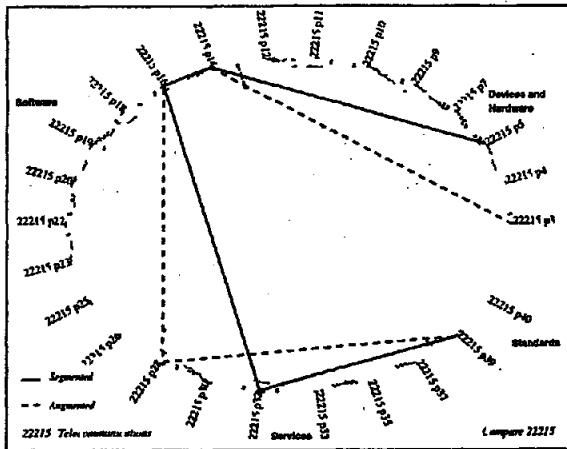


Figure 4 Segmented bushy and augmented segmented bushy paths for article *Telecommunications*

more bushy paragraph which is somewhere in the middle of the segment. This is quite detrimental to the readability of the summary. To remedy this problem, we define the augmented segmented bushy path which always picks the introductory paragraph from a segment, and other bushy

20% Segmented Bushy Path

Para 5 The devices used in telecommunications can be computers, terminals (devices that transmit and receive information), and peripheral equipment such as printers (see Computer, and see Office Systems). The transmission line used

Para 14 Among the different kinds of software are terminal-emulation, file-transfer, host, and network software. Terminal-emulation software makes it possible for a device to perform the same functions as a terminal. File-transfer software is

Para 16 Three major categories of telecommunication applications can be discussed here: host-terminal, file-transfer, and computer-network communications

Para 32 An information-retrieval service leases time on a host computer to terminals, so that these terminals are able to retrieve information from the host computer. An example is CompuServe Information Services. To gain access to

Para 39 Certain telecommunication methods have become standard in the telecommunications industry as a whole, because if two devices use different standards they are unable to communicate properly. Standards are developed in

Table 3 Text for segmented bushy path for article *Telecommunications*

paragraphs based upon the length requirements of the summary

Figure 3 shows a 20% global bushy path and a global depth-first path constructed for the article on telecommunications. The corresponding texts for these paths are shown in Tables 1 and 2. Note that the bushy path does not include any material from the last two segments (on telecommunication services and standards). The depth first path misses out the segment on standards. On the other hand, the segmented bushy path (see Figure 4 and Table 3) does include a paragraph from each of the last two segments and is more indicative of the contents of the article than either of the global paths. But the segmented bushy path picks paragraphs from the middle of a segment, for example paragraph 5 in the first segment and paragraph 32 in the segment on telecommunications services. Presenting a paragraph from a topic without introducing the topic is once again detrimental to the readability of the summary. This could be fixed by augmenting the segmented bushy paths by forcing them to select the introductory paragraph from every segment. The augmented segmented bushy path for this article (see Table 4) is actually a very good indicative summary for the article

3 Experiment

Several automatic extraction schemes, including the above, have been proposed earlier (Salton et al, 1996b, Salton et al, 1996a). General features of the extracts produced by these different algorithms have been noted, based on manually examining some of the extracts. However, objective evaluation of these algorithms has always been problematic. In (Salton and Singhal, 1995), an attempt was made to evaluate the summaries based on

20% Augmented Segmented Bushy Path

Para 3 Telecommunications, broadly speaking, the process of transmitting information in an electronic form between any two devices by using any kind of transmission line. More specifically, however, telecommunications refers to the process

Para 14 Among the different kinds of software are terminal-emulation, file-transfer, host, and network software. Terminal-emulation software makes it possible for a device to perform the same functions as a terminal. File-transfer software is

Para 16 Three major categories of telecommunication applications can be discussed here: host-terminal, file-transfer, and computer-network communications

Para 28 Public telecommunication services are a relatively recent development in telecommunications. The four kinds of services are network, information-retrieval, electronic-mail, and bulletin-board services

Para 39 Certain telecommunication methods have become standard in the telecommunications industry as a whole because if two devices use different standards they are unable to communicate properly. Standards are developed in

Table 4 Text for augmented segmented bushy path for article *Telecommunications*

ranked retrieval. Since relevance judgments were not available for passages or extracts, the available relevance judgments for full documents were extrapolated to the extracts. However, the portion of a document that is relevant to a query may well get left out of a passage, and so, results obtained from such an evaluation are unreliable.

Since the goal of our summarization schemes is to automate a process that has traditionally been done manually, a comparison of automatically generated extracts with those produced by humans would provide a reasonable evaluation of these methods. We assume that a human would be able to identify the most important paragraphs in an article most effectively. If the set of paragraphs selected by an automatic extraction method has a high overlap with the human-generated extract, the automatic method should be regarded as effective. Thus, our evaluation method takes the following form: a user submits a document to the system for summarization, in one case, the system presents a summary generated by another person, in the other, it produces an automatically generated extract. The user compares the two summaries — manual and automatic — to his/her own notion of an ideal extract. To evaluate the automatic methods, we compare the user's 'satisfaction' in the two cases. Such an evaluation methodology has its shortcomings, for example it does not account for the readability aspect of a summary, it also ignores the fact that user satisfaction is related to whether a user has seen the full-article or not. Unfortunately, given the lack of a good testbed for evaluating automatic summarization, it is the best we can do.

Fifty articles were selected from the Funk and Wagnalls Encyclopedia (Funk and Wagnalls, 1979). For each article, two extracts were constructed manually. One of these extracts was used as the manual summary. The

other one, which then becomes a user's (ideal) summary, is used as the oracle to compare the performance of the manual summary and an automatic summary. The following instructions were given to those who constructed the manual extracts:

Please read through the articles. Determine which n paragraphs are the most important for summarizing this article. $n = \text{MAX}(5, 1/5\text{th the total number of paragraphs (round to the next higher number for fractions)})$. Mark the paragraphs which you chose.

The resulting database of 100 manual summaries (two for each of the fifty articles) was used in the final evaluation of the automatic methods. Summaries were then automatically generated for the articles, using each of the four methods described above. In each case, the automatic and manual extracts had the same number of paragraphs³.

In manual summarization by paragraph extraction, there are certain paragraphs in a text that certainly belong in a summary extract, but then there are many paragraphs whose importance is subjectively judged by the individual doing the extraction. To reduce the effect of the arbitrariness introduced by individual's subjective notions, for very short articles, we asked our subjects to extract at least five paragraphs, hoping that the intersection of the two manual summaries will indeed yield the most important paragraphs in an article. The articles used in our evaluation had anywhere between thirteen and forty eight content paragraphs. The current implementation of the Smart system also considers the section headings, etc as individual paragraphs. Such paragraphs were marked as non-content and were ignored in the summarization process.

4 Results and Discussion

The following scenario was assumed for evaluation of the automatic summaries:

- A user walks up to the system and presents an article for summarization.
- In the first case, the system asks another human to do the summarization and presents it to the user. The user compares this summary to his/her own notion of an ideal summary.
- In the second case, the system automatically generates a summary and returns it to the user. The user again compares this summary to his/her own notion of an ideal summary.
- The user satisfaction in the above two cases is measured by the "degree of overlap" between the summary presented by the system and the user's notion of an ideal summary.

³Different users could count paragraphs differently. Thus, for a few articles, the lengths of the two manually generated summaries were different. In such cases, the automatic procedures took the average of these two lengths as the target length for the extract.

If the user's satisfaction is about the same in the above two cases, then our automatic summarization schemes are summarizing as well as a human would summarize by paragraph extraction

For each automatic summarization algorithm, four quantities were computed

- 1 *Optimistic evaluation* Since the two manual extracts for an article are different, the amount of overlap between an automatic and a manual extract depends on which manual extract is selected for comparison. The optimistic evaluation for an algorithm is done by selecting the manual extract with which the automatic extract has a higher overlap, and measuring this overlap. This is the same as using the human whose notion of an ideal extract is closer to the automatic extract as our user
- 2 *Pessimistic evaluation* Analogously, a pessimistic evaluation is done by selecting the manual extract with which the automatic extract has a lower overlap. This is the same as using the human whose notion of an ideal extract is more dissimilar to the automatic extract as our user. Thus, in some sense, is the worst case scenario
- 3 *Intersection* For each article, an intersection of the two manually generated summaries is computed. The fact that the paragraphs in this intersection were deemed important by both the readers suggests that they may, in fact, be the most important paragraphs in the article. We compute the percentage of these paragraphs that is included in the automatic extract
- 4 *Union* We also calculate the percentage of automatically selected paragraphs that is selected by at least one of the two users. This is, in some sense, a precision measure, since it provides us with a sense of how often an automatically selected paragraph is potentially important

In our experimentation, we observed that many subjects tend to select paragraph 3 in the summaries. This is because this paragraph is the first content paragraph in an article and tends to be a dictionary-style definition for the article. For example, for article 15930 (*Monopoly*), this paragraph reads

Monopoly, economic situation in which there is only a single seller or producer of a commodity or a service. For a monopoly to be effective, there must be no practical substitutes for the product or service sold, and no serious threat of the entry of a competitor into the market. This enables the seller to control the price

Such dictionary-style definitions are generally liked by readers and thus are usually included in a summary by our subjects

In general, in written texts, the first content paragraph tends to be an introductory paragraph and is a good starting paragraph for summarization. For the encyclopedia articles, we use this information and we always include paragraph 3 in the bushy and the depth-first summaries

This paragraph might be missed by the segmented bushy paths but is recaptured by the augmented segmented bushy paths. In case such collection specific information is not available, we use the first paragraph with a reasonable number of links to the rest of the paragraphs as the introductory paragraph (Salton and Singhal, 1995)

Table 5 shows the overlap for the two manual extracts, and the different evaluation measures averaged over all fifty articles, for the bushy, depth-first, segmented bushy, and augmented segmented bushy extracts. In addition to using these four methods, extracts were also generated for the articles by selecting the required number of paragraphs at random. To eliminate any advantage that the bushy, depth-first, and augmented segmented bushy extracts might have due to the presence of the introductory paragraph, paragraph 3 is always included in the random paths. The evaluation results for these random extracts are also shown in the table. Random selection of paragraphs serves as the weakest possible baseline. If an algorithm does not perform noticeably better than a random extract, then it is certainly doing a poor job of summarization. Also, Brandow, Mitze, and Rau found in (Brandow et al., 1995) that simply selecting the first few sentences (the lead sentences) produced the most acceptable summaries. To test their findings in our environment, we also selected the first 20% paragraphs of an article and used it as yet another automatic summary

Manual Extracts

The most unexpected result of our experiment was the low level of agreement between the two human subjects. The overlap between the two manual extracts is only 46% on an average, i.e., an extract generated by one person is likely to cover 46% of the information that is regarded as most important by another person. This ratio suggests that two humans disagree on more than half the paragraphs that they consider to be critical. In addition, as indicated above, the first paragraph of these encyclopedia articles is a general introduction to the article and is often selected by both subjects — in 50% of the cases in which the intersection between the two users' extracts is a single paragraph, this paragraph is the first one. This increases the chances of overlap between the two manual extracts. If we exclude this special paragraph from the article, the overlap figures for two humans will be even worse

The lack of consensus between users on which paragraphs are important can be explained as follows. On a first reading, users earmarked certain paragraphs as important. Some of these paragraphs were then eliminated, in order to reduce the extract to the stipulated size. Often, the choice between which paragraphs to keep and which to exclude was a difficult one, and in such situations, some arbitrariness is bound to creep in. This casts some shadows on the utility of automatic text summarization by text extraction. It is possible that the user satisfaction might be higher in reality when the true user does not read the portion of an article not presented to him/her by the summarization system and does not get an opportunity to form his/her own ideal view of an extract

| Overlap between manual extracts: 46% | | | | |
|--------------------------------------|----------------|-----------------|------------------|-----------|
| Algorithm | Optimistic (%) | Pessimistic (%) | Intersection (%) | Union (%) |
| Global bushy | 45 60 | 30 74 | 47 33 | 55 16 |
| Global depth-first | 43 98 | 27 76 | 42 33 | 52 48 |
| Segmented bushy | 45 48 | 26 37 | 38 17 | 52 95 |
| Augmented seg bushy | 46 66 | 27 59 | 41 83 | 55 44 |
| Random | 39 16 | 22 07 | 38 47 | 44 24 |
| Initial (Lead) | 47 99 | 29 50 | 50 00 | 55 97 |

Table 5 Evaluation measures for automatic extraction methods

Automatic Extracts

Table 5 indicates that global bushy paths and augmented segmented bushy paths produce the best extracts among the four paths considered in this study. 55% of the paragraphs selected by the process were considered important by at least one user. Optimistically speaking, a global bushy or an augmented segmented bushy path may be expected to agree approximately 46% with a user. This number is at par with the agreement between two humans (45.81%). This result is reassuring in terms of the method's viability for generating good extracts, since the scheme performs as well as a human.

About 47% of the paragraphs deemed important by both users are included in the bushy extract for an article. This figure is somewhat disheartening. We expected a better coverage of these vital paragraphs by our extracts. A further study of these paragraphs might reveal some properties that users look for in a paragraph to decide its importance. It might then be possible to automate this selection process. We also identified the articles for which the intersection of the two user summaries is a single paragraph. For 78% of these articles, this paragraph was included in the bushy path.

Segmented bushy paths perform worse than expected. This is because the first paragraph of an article is very often selected by users, and segmented bushy paths occasionally omit this paragraph. This results in a decrease in the overlap between automatic and manual extracts. In contrast, the other paths are *guaranteed* to include the first paragraph, and perform better. But, in general, the performance of segmented bushy paths was satisfactory (45.48% overlap with the user in the optimistic method). Similarly, the performance of the depth-first path was also satisfactory. All paths achieved the minimum requirement of performing significantly better than a random extract.

But more interestingly, we observe that extracts produced by selecting the first few paragraphs of the articles also performed comparably to the best paragraph extraction scheme. Admittedly, our evaluation methodology lacks the evaluation of the readability aspect of a summary which was one of the main motivations of moving from a sentence-based extraction strategy to paragraph-based extraction. With very high chances, the lead summary will outperform all other automatic summaries in terms of readability. We believe this because automatic

summaries are a forced concatenation of paragraphs distributed all across a document, whereas a lead summary is a nicely coherent sequence of paragraphs, as written by the author. Overall, the lead summaries are comparable to the best summarization strategy and could be more readable than all other summaries. This truth is rather discouraging for the feasibility of automatic summarization by text extraction but agrees with the observations in (Brandow et al., 1995). News reports, used in (Brandow et al., 1995), frequently contain a leading paragraph that summarizes the story contained in the rest of the report. Likewise, in the encyclopedia articles used in this study, the first paragraphs usually define the topic, and provide a general outline about it.

To sum up

- The good news is that interpreted in light of the fact that the overlap between the two manual extracts is, on an average, 46%, and given the enormous reduction in the amount of resources required⁴, our results indicate that automatic methods for extraction compare very favorably with manual extraction.
- But the bad news is that a summary formed by extracting the initial paragraphs of an article is as good as the best automatic summary and might just be more readable from a user's perspective. This brings into question the overall utility of automatic text summarization by text (sentence or paragraph) extraction.

It is possible that the nature of the articles used in this study (encyclopedia articles) and in (Brandow et al., 1995) (news articles) have a structure that yields very good summaries, simply by extracting the initial part of an article. It will be interesting to see if observations from this study and from (Brandow et al., 1995) carry over to other, more non-encyclopedia like and non-news like domains (for example legal documents or U.S. Patents). In our studies with text summarization (by text extraction), we have always felt a very strong need for a good evaluation test-bed. Lack of good objective evaluation techniques for text summarization has always been the biggest problem in all our work, and has consistently

⁴The system took about 15 minutes to generate 3 summaries for each of 50 articles. A human would require about 10 minutes to produce a summary for a typical article from this set.

discouraged more experimentation and exploration of interesting research possibilities (like the one mentioned above regarding articles from other domains)

5 Conclusion

In this study, we have tried to evaluate automatic summarization methods proposed earlier. If a good testbed for evaluating summaries were available, the evaluation methodology adopted in this study could be improved, but we believe it is the best we can currently do. Under our evaluation scheme, the four extraction algorithms examined perform comparably, but they produced significantly better extracts than a random selection of paragraphs. The absolute performance figures are not high, but given the low overlap between two human-generated extracts, they are eminently satisfactory. However, this wide variation between users brings us to the question of whether summarization by automatic extraction is feasible. If humans are unable to agree on which paragraphs best represent an article, it is unreasonable to expect an automatic procedure to identify the best extract, whatever that might be. We also find that presenting the user with the initial part of an article is as good as employing any "intelligent" text extraction scheme. In summary, automatic summarization by extraction is admittedly an imperfect method. However, at the moment, it does appear to be the only domain-independent technique which performs reasonably.

Acknowledgments

We are deeply indebted to (late) Professor Gerard Salton for all his guidance during the initial stages of this work. Without the invaluable advice and support of Professor Salton, this work would not have been possible.

We thank Nawaaz Ahmed, David Fielding, Nicholas Howe, S. Ravikumar, Cynthia Robinson, and Divakar Vishwanath for generating extracts for the articles used in the evaluation process.

References

- R Brandow, K Mitze, and L F Rau, Automatic Condensation of Electronic Publications by Sentence Selection, *Information Processing and Management*, 31(5), 675-685, 1995
- L Earl, Experiments in Automatic Extracting and Indexing, *Information Storage and Retrieval*, 6 4, 313-334, October 1970
- Funk and Wagnalls New Encyclopedia, Funk and Wagnalls, New York, 1979
- M A Hearst and C Plaunt, Subtopic Structuring for Full-Length Document Access, Proc SIGIR '93, 59-68, *Association for Computing Machinery*, New York, November 1993
- J Kupiec, J Pedersen, and F Chen, A Trainable Document Summarizer, Proc SIGIR '95, 68-73, *Association for Computing Machinery*, New York, July 1995
- H P Luhn, The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, 2(2), 159-165, 1958
- C D Paice, Constructing Literature Abstracts by Computer Techniques and Prospects, *Information Processing and Management*, 26(1), 171-186, 1990
- G Salton, ed, *The SMART Retrieval System — Experiments in Automatic Document Processing*, Prentice Hall Inc, NJ, 1971
- G Salton and M McGill *Introduction to Modern Information Retrieval* McGraw Hill Book Co, New York, 1983
- G Salton *Automatic text processing—the transformation, analysis and retrieval of information by computer* Addison-Wesley Publishing Co, Reading, MA, 1989
- G Salton and J Allan, Selective Text Utilization and Text Traversal, Proc Hypertext-93, *Association for Computing Machinery*, New York, November 1993, 131-144
- G Salton, J Allan, C Buckley, and A Singhal, Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts, *Science* 264, 1421-1426, 3 June 1994
- G Salton, J Allan, and A Singhal, Automatic Text Decomposition and Structuring, *Information Processing and Management*, 32(2), 127-138, 1996
- G Salton and A Singhal, Selective Text Traversal, Technical Report, TR 95-1549, Department of Computer Science, Cornell University, Ithaca, NY, September, 1995
- G Salton, A Singhal, C Buckley, and M Mitra, Automatic Text Decomposition Using Text Segments and Text Themes *Hypertext '96*, The Seventh ACM Conference on Hypertext, Association for Computing Machinery, New York, 53-65
- G Salton, A Singhal, M Mitra, and C Buckley, Automatic Text Structuring and Summarization, *Information Processing and Management*, 33(2), 193-208, 1997