Summarization Evaluation: An Overview

Inderjeet MANI
The MITRE Corporation, W640
11493 Sunset Hills Road
Reston, VA 20190-5214, USA
imani@mitre.org

Abstract

This paper provides an overview of different methods for evaluating automatic summarization systems. The challenges in evaluating summaries are characterized. Both intrinsic and extrinsic approaches are discussed. Methods for assessing informativeness and coherence are described. The advantages and disadvantages of specific methods are assessed, along with criteria for choosing among them. The paper concludes with some suggestions for future directions.

Keywords: *Summarization Evaluation, Intrinsic, Extrinsic, Informativeness, Coherence.*

1 Introduction

The goal of automatic summarization[1] is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs [19]. Summaries can be user-focused (or topic-focused, or query-focused), i.e., tailored to the requirements of a particular user or group of users, or else, they can be 'generic', i.e., aimed at a particular-usually broad-readership community. They can take the form of an extract, i.e., a summary consisting entirely of material copied from the input, or an abstract, i.e., a summary at least some of whose material is not present in the input (see [20], [19] for a detailed introduction to the field).

Evaluation has long been of interest to automatic summarization, with extensive evaluations being carried out as early as the 1960's, e.g., [7]. While there has not been a great deal of consensus on evaluation issues, a fairly clear picture emerges from studying the evaluation methods explored so far. This short paper will briefly characterize and assess these methods.

[1] This work has been funded by DARPA's Translingual Information Detection, Extraction, and Summarization (TIDES) research program, under contract number DAA-B07-99-C-C201 and and ARPA Order H049.

There are several serious challenges in evaluating summaries, which makes summarization evaluation a very interesting problem:

- Summarization involves a machine producing output that results in natural language communication. In cases where the output is an answer to a question, there may be a correct answer, but in other cases it is hard to arrive at a notion of what the correct output is. There is always the possibility of a system generating a good summary that is quite different from any human summary used as an approximation to the correct output. (Similar problems occur with machine translation, speech synthesis, and other such output technologies.)
- Since humans may be required to judge the system's output, this may greatly increase the expense of an evaluation. An evaluation which could use a scoring program instead of human judgments is preferable, since it is easily repeatable.
- Summarization involves compression, so it is important to be able to evaluate summaries at different compression rates. This increases the scale and complexity of the evaluation.
- Since summarization involves presenting information in a manner sensitive to a user's or application's needs, these factors need to be taken into account. This in turn complicates the design of an evaluation.

Methods for evaluating text summarization (and, indeed, natural language processing systems) can be broadly classified into two categories [36]. The first, an *intrinsic* evaluation, tests the summarization system in of itself. The second, an *extrinsic* evaluation, tests the summarization based on how it affects the completion of some other task. Intrinsic evaluations have assessed mainly the coherence and informativeness of summaries. Extrinsic evaluations, on the other hand, have tested the impact of summarization on tasks like relevance assessment, reading comprehension, etc.

2 Intrinsic Methods

2.1 Criteria

2.1.1 Summary Coherence

Clearly, one aspect of a summary is how it reads. This can be assessed by having humans grade summaries for coherence based on specific criteria.

Summaries which are extracts may be extracted out of context, in which case coherence problems may occur, like dangling anaphors and gaps in the rhetorical structure of the summary. Readability criteria are often tied to these problems. For example, [25] had subjects grade readability of summaries based on the presence of dangling anaphors, lack of preservation of the integrity of structured environments like lists or tables, 'choppiness' of the text, presence of tautologous statements such as "Predicting the future is difficult", etc. Abstracts, like extracts, can also be incoherent, especially when natural language generation is used. [33] had judges grade the acceptability of abstracts produced by cut-and-paste operations on the source text, based on general readability criteria such as: good spelling and grammar, clear indication of the topic of the source document, impersonal style, conciseness, readability and understandability, acronyms being presented with expansions, etc.

When subjects can grade or rank summary sentences for coherence, the scores for the summary sentences can be compared with the scores for reference summaries, with the scores for the source sentences, or with the scores for other summarization systems. Automatic scoring has a limited role to play here, given that tools such as readability measurements (based on word and sentence length) and grammar and style checkers result in extremely coarse assessments [18].

2.1.2 Summary Informativeness

Informativeness aims at assessing the summary's information content. As a summary of a source becomes shorter, there is less information from the source that can be preserved in the summary. Therefore, one measure of the informativeness of a summary is to assess how much information from the source is preserved in the summary. Another measure is how much information from a reference summary is covered by information in the system summary. In other words, as in the case of coherence, comparisons can be made between system summaries, the source, reference summaries, and scores for other summarization systems. However, while subjective grading can be used for informativeness, informativeness is more amenable than coherence to automatic scoring.

2.1.3 Coherence versus Informativeness

Both these dimensions are valuable for summarization. They are somewhat orthogonal dimensions; it is certainly possible to have a well-written but atrocious summary. They are also related. [18], in evaluating a summarizer that merged and truncated sentences from an extract, found an inverse relation between informativeness and coherence. In the discussion below, I will focus mainly on informativeness, as it presents more interesting intellectual challenges for evaluation.

2.2 Comparison against reference output

2.2.1 Introduction

The idea of a reference summary, against which machine output can be compared, is a very natural one. The classic evaluation of [7] had humans evaluate machine summaries by comparing them to human-created abstracts (on a 5-point scale of similarity). Even if one could collect multiple reference summaries of a document, there is always the possibility of a system generating a summary that is quite different from any of the reference summaries, but which is still an informative and coherent summary. This is especially true in the case of generated abstracts. In other words, the set of reference summaries will necessarily be *incomplete*.

However, once a subject is given specific instructions on how to construct a summary, the space of possible summaries is constrained to some extent by these instructions. The human summary may be supplied by a document's author, or by a subject asked to construct an abstract or extract. When a subject is used to construct a reference extract, she may be required to judge every sentence on a boolean or multi-point scale for summary-worthiness. Alternatively, the subject may be given a compression rate, and told to pick the top 25% of the sentences in the document. Or, the subject may be told to rank all the sentences in the document, or rank the top 20% of the sentences. Further, the subject may be required to extract clauses or paragraphs, or arbitrary-sized passages, instead of sentences. As might be expected, these different compression instructions can make a considerable difference in what gets extracted.

2.2.2 Agreement among reference summaries

Previous studies, most of which have focused on extracts, have shown evidence of low agreement among humans as to which sentences are good summary sentences. In an early paper involving humans extracting 20-sentences from each of 10 Scientific American articles, [31] showed that different judges may produce different extracts of the same source (with all

6 subjects agreeing on an average of only 1.6 sentences per extract), and that a judge given the same document again eight weeks later may produce a substantially different extract (agreeing on their previous sentence selections a little over half the time). They found there was a lot more variability among the human subjects than among the machine summaries and very little agreement between human and machine selections. [35] obtained results in a similar vein: 2 subjects showed only 46% overlap in their extracts when asked to extract at least 5 paragraphs from each of 50 articles from Funk and Wagnall's encyclopedia. However, there is also evidence that judges may agree more on the most important sentences to include [12], [21].

2.2.3 Automatic Scoring

The comparison between summaries is best carried out by humans, but it can also be computed automatically. A variety of different measures can be used, based on studies by [6], [30]:

- Sentence Recall measures how many of the reference summary sentences the machine summary contains (likewise, Precision can also be used). This measure is mainly applicable to machine summaries that are sentence extracts, not abstracts, though an alignment of abstracts with source sentences can be carried out, e.g., [14], [13], [22]. However, such a measure isn't sensitive enough to distinguish between many possible summaries, and summaries which are quite different in content will have to be given the same scores.

- An alternative to using Sentence Recall is Sentence Rank, where a summary is specified in terms of a ranking of sentences in terms of summary-worthiness. The sentence rankings of the machine summary and the reference summary can then be compared by using a correlation measure. However, Sentence Rank is a somewhat unnatural way of getting a human to summarize, and this class of measure is again applicable to extract summaries, not abstracts.

- Utility-based measures [30] are based on a more fine-grained approach to judging summary-worthiness of sentences, rather than just boolean judgments. The advantage here is more precise measurements of informativeness; however, people do vary in their ability to discriminate.

- Content-Based measures based on similarity of vocabulary [34] are oriented in principle towards both extracts and abstracts. One of the virtues of such content-based measures is that the number of sentences involved can, if desired, be ignored in the similarity computation. The disadvantage of these measures is that they do not discriminate very well between summaries that involve differences in meaning expressed by negation, word ordering differences, etc. However, such measures can be useful for comparing extracts, or comparing summaries with abstracts that have a high degree of cut-and-paste relationship with the source, where these problems may not manifest themselves as much. They are also useful for comparing more fragmentary summaries, such as lists of phrases.

2.3 Comparison against summarization input

2.3.1 Introduction

One type of evaluation involves providing both the summary and the source to subjects, asking them to determine summary informativeness in the context of the source, e.g., [4]. Many of the automatic content-based measures discussed above can also be used to compare a summary against a source document. Such a comparison becomes more meaningful when the system's output is compared against semantic information extracted by a human from the source.

2.3.2 Semantic Methods

It is possible to mark up the meaning of each sentence in a text, with a subjective grading of to what extent the summary covers the propositions in the source [38]. However, such an effort can be prohibitively costly in terms of the extent of human annotation required and the delicacy of the grading involved. A more practical evaluation method which is semantic in nature is found in the information extraction approach of [29]. To evaluate their summarizer, which produced abstracts of documents on crop agriculture, they characterized each text in terms of its focal concepts (e.g., leaf canopy development and barley) and non-focal concepts. They then measured the summary's coverage of these concepts in terms of Correct, Incorrect, and Missing. A more discourse-sensitive approach is found in [25], who had subjects grade the summary's coverage of rhetorical links in the text.

2.3.3 Surface Methods

Instead of representing concepts at a deep level, it is possible to judge whether the key ideas in the source (identified by underlining source passages) are covered in the summary. A variant of this was carried out in the TIPSTER SUMMAC text summarization evaluation Q&A (Question and Answer) task [17], where passages in the source were marked up based on criteria of relevance to a topic. Given a document and a

topic, an accurate topic-focused summary of the document would contains answers to questions that covered 'obligatory' information that needed to be present in any document judged relevant to the topic. For example, for a topic concerning "Prison overcrowding", a topic-related question would be "What is the name of each correction facility where the reported overcrowding exists?". A document relevant to prison overcrowding might contain an answer to such a question; a summary could be compared against such an answer in the document. Given the source document with answers marked up, and the summary of the source, humans judged the summary as Correct (contained the answer to the question, along with sufficient context), Partially Correct (insufficient context), or Missing (did not contain the answer) for any given question. Based on this, the authors discovered an informativeness ratio of accuracy to compression of about 1.5.

A somewhat shallower approach, which does not require marking up each document by hand, is to rely on keywords associated with the source. In the evaluation by [33], judges were presented with their system's generated abstracts and 5 lists of keywords, each list being obtained from keywords provided with the source article in the publication journal. The judge had to associate the right list of keywords with the abstract. Since the keywords were indicative of the content of the source article, if the abstract could be related to the right keywords, the abstract would be viewed as covering key concepts in the source document.

2.4 Comparing Human and Automatic Scoring

When an automatic scorer is used in place of human judgments, it is important to validate the scoring scheme. One way to do this is to compare the performance of the automatic scorer with that of a human; if the two sets of scores are closely correlated, the automatic scoring may be used as a substitute for human scoring in the task at hand. This was in fact the strategy taken in the SUMMAC Q&A task above, where an automatic scoring program was developed that measured the overlap between summaries and answers in the answer key using four different Content-based vocabulary overlap measures. The correlation between the summary participants' scores on each of the four overlap measures and each of the scores assigned by the human was very strong. [15] and [18] have reused the SUMMAC Q&A data along with automatic scoring.

3 Extrinsic Methods

3.1 Introduction

The idea of an extrinsic summarization evaluation is to determine the effect of summarization on some other task. A variety of different tasks can be considered:

- If the summary involves instructions of some kind, it is possible to measure the efficiency in executing the instructions.
- One can examine the summary's usefulness with respect to some information need or goal, such as finding documents relevant to one's need from a large collection, routing documents, producing an effective report or presentation using a summary, etc.
- It is also possible to assess the impact of a summarizer on the system in which it is embedded, e.g., how much does summarization help the question answering system? Another possibility is to measure the amount of effort required to post-edit the summary output to bring it to some acceptable, task-dependent state.

The variety of tasks to which summarization can be applied is in fact very large, and will expand as computer technology continues to evolve. I discuss a few selected ones to convey an idea of the type of evaluation carried out.

3.2 Relevance Assessment

In the task of relevance assessment, a subject is presented with a document and a topic, and asked to determine the relevance of the document to the topic. The influence of summarization on accuracy and time in the task is then studied. There have been numerous extrinsic evaluations involving this task paradigm [37], [12], [4], [16], [17], of which the TIPSTER SUMMAC evaluation [17] was the most large-scale, developer-independent evaluation.

Although the SUMMAC evaluation included the intrinsic Q&A evaluation component described above, its main focus was on an extrinsic evaluation, based on tasks which modeled real-world activities typically carried out by information analysts in the U.S. Government. In the ad hoc task, the focus was on indicative topic-focused summaries. (Indicative summaries [3] aim only at providing a reference function for selecting documents for more in-depth reading.) This task relates to the real-world activity of an analyst conducting full-text searches using an information retrieval system to quickly determine the relevance of a retrieved document. Given a document (which could be a summary or a full-text source - the subject was not told which), and a topic description, the human subject was asked to determine whether the document was relevant to the topic. In the categorization task, the evaluation sought to find out whether a generic summary could effectively present enough information to allow an analyst to quickly and correctly categorize a document.

Here the topic was not known to the summarization system. Given a document, which could be a generic summary or a full-text source (the subject was not told which), the human subject would choose a single category out of five categories (each of which had an associated topic description) to which the document was relevant, or else choose "none of the above".

The accuracy of the subject's relevance assessment decision was measured in terms of 'ground-truth' judgments of the full-text source relevance, which were separately obtained from the TREC collection [9]. Thus, an indicative summary would be 'accurate' if it accurately reflected the relevance or irrelevance of the corresponding source.

In meeting the evaluation goals, the main question to be answered was whether summarization saved time in relevance assessment, without impairing accuracy. The results from relevance assessment by 51 subjects of the output of 16 summarization systems showed that summaries at relatively low compression rates (generic summaries at 10% of source length, and topic-related summaries as short as 17% of source length) reduced relevance assessment time by 40% (for generic summaries) to 50% (for topic-related summaries), with no statistically significant degradation in accuracy. Systems that performed most accurately in the production of topic-related summaries used statistical methods involving term frequency and co-occurrence statistics, and vocabulary overlap comparisons between text passages, and extracted similar sets of sentences. However, in the absence of a topic, these statistical methods did not provide any additional leverage: in the case of generic summaries, the systems (which relied in this case on inclusion of the first sentence of the source) were indistinguishable in accuracy.

While the SUMMAC results are definitive and can be expected to stand the test of time, and have influenced other evaluations (e.g., [27], as well as the body of Q&A work cited above), it had several shortcomings:

- The evaluation required that the source documents be relatively short since they have to be read in a reasonable amount of time. However, if the documents are too short, there is no need to summarize them!

- Reliance on pre-existing data for relevance judgments constrained the genre of documents available, in this case to newswire texts. Scientific texts, editorials, etc. might challenge the summarizers to a greater extent. In addition to general issues of corpus selection such as size, heterogeneity, annotation standards, etc., adequate document lengths and compression rates of summaries are critical.

- The SUMMAC tasks did not cover the full spectrum of single-document summarization; the tasks could be carried out just with extracts, rather than abstracts. (Although participants could have submitted abstracts, none did.)

3.3 Reading Comprehension Tasks

In reading comprehension tasks, the human first reads full sources or summaries assembled from one or more documents. The human then answers a multiple-choice test. The system then automatically scores the answers, measuring the percentage of correct answers. Thus, a human's comprehension based on the summary can be objectively compared with that based on the source. The reasoning here is that if reading a summary allows a human to answer questions as accurately as he would reading the source, the summary is highly informative.

[26] carried out on an extrinsic evaluation of the impact of summarization in a task of question-answering. The authors picked four Graduate Management Admission Test (GMAT) reading comprehension exercises. The exercises were multiple-choice, with a single answer to be selected from answers shown alongside each question. The authors measured how many of the answers the subjects got correct under different conditions, including a full-text condition (where the subjects were shown the original passages), an extract condition (where the subjects were shown automatically generated generic extracts from the passages), an abstract condition (where the subjects were shown a generic human abstract of about 25% compression created by a professional abstractor instructed to create informative abstracts), and a control no-text condition (where the subjects had to pick the correct answer just from seeing the questions, without seeing the passages themselves). However, the number of exercises in this study was very small, and the exercises being summarized tended to be very short.

[11] measured informativeness in terms of the extent to which humans could reconstruct the essential information in a document by reading a summary of it. In a pilot experiment, they had subjects recreate the source text based on reading the summary, the source document, and also based on guessing alone. The group shown the source text were able to do so in just 10 keystrokes (one can view this as an upper bound of performance for a summary); the summary group needed 150 keystrokes; and those given no prior information needed 1,100 keystrokes (this could be viewed as a lower bound).

Another variety of Reading Comprehension evaluation is found in the evaluation of the SUMGEN summarizer [23]. SUMGEN summarizes output logs from a battle simulation, and news from templates of key information created by annotators from business news

sources describing joint ventures. In their evaluation, subjects filled in answers to questions for both sources and summaries. For the battle simulation, subjects were asked to fill in the names, participants, time and duration of all missions that appeared in the texts. For the business news application, they were asked to fill in the type, partners, and status of all joint ventures mentioned in the text.

All in all, reading comprehension exercises present a promising avenue for summarization evaluation, and it is expected that more work using such methods will take place in the future, e.g., for Multi-Document Summarization evaluation. Clearly, there may be many pre-existing materials found in various pedagogical settings that may be leveraged here.

3.4 Extrinsic compared to Intrinsic

The choice of an intrinsic or extrinsic method depends very much on the goals of the developers, funders, and consumers of the summarization technology. In general, at early stages of the technology cycle, intrinsic evaluations are recommended, with an emphasis on evaluation of summarization components; as the technology matures, more situated, task-based tests of the system as a whole involving 'real' users become more important.

Extrinsic evaluations have the advantage of assessing the utility of summarization in a task, so they can be of tremendous practical value to a funder or consumer of summarization technology. However, they are less useful to developers in terms of offering feedback as to how they might improve their systems. This can be somewhat alleviated when a system's performance in an intrinsic evaluation predicts performance in an extrinsic one. Also, developers need repeatable, less expensive, automatically-scorable evaluations. This can be achieved by developing annotated corpora which can provide reference data to support intrinsic evaluations, whether in the form of a representation of the input source document, as in the Q&A annotated mini-corpus, or in the form of reference summaries linked to their sources [13], [5], [22], [8], etc.

4 Presentation Strategies

Unfortunately, very little work has been done in this area. [24] evaluated the summarization component of their Broadcast News Navigator, which performs an analysis of news video along with the closed-captioned text. The summaries used lists of topic terms and proper names, and single sentence extracts from the closed-captioned text, along with key frames extracted from the video, in various combinations. On a relevance assessment task, they found that the mixed-media presentation strategies saved time without loss of accuracy compared to viewing the video.

Human factors studies, which address issues like standard ways of presenting different tools, use of particular colors, iconology, presentation techniques, and types of display, are also useful here. However, the methods for carrying out systematic evaluations of interactive systems are not yet well understood. Often, there is a real need to separate the evaluation of the summarization component itself from the evaluation of the interface in which it is used. After all, good summarizers may live in bad interfaces, and vice versa.

5 Component-level and Situated Tests

Components used in different phases of summarization, including analysis of input text, its transformation into a reduced form, and synthesis of output summaries can each be independently evaluated using various metrics. For example, [18] evaluates algorithms to extract information, truncate and merge sentences, and resolve dangling anaphors. Further, it is also possible to measure the impact of including one or more components on the overall summarization system, and as mentioned earlier, the impact of the summarization system on an overall system in which it is embedded.

Once a system is mature enough to have end-users, it is useful to assess the impact of the summarization tool on these end-users. This can involve assessing the tool's frequency of use, use of basic and advanced features of the system, ability to customize the output for particular needs, the perceived complexity of the user interface, etc. Typical metrics include time and cost to task completion, quality of solution, and user satisfaction, as well as extensibility and portability (e.g., based on coverage of a test set with or without adding new knowledge [32]). It is also useful to carry out regression testing to determine stability of performance of successive system versions on unseen test sets. In the special case of summarization systems, one could also examine whether particular menu items in the user interface are appropriate, whether the ability to edit summaries is supported, whether the various compression rates work as advertised, whether people can be easily trained to the use the summarizer, etc. Some work in this direction is found in [28], which examines user interface issues involved in the beta-test use of DimSum, a summarizer from SRA Corporation [1].

6 Conclusions and Future Directions

Evaluation offers many advantages to automatic summarization: as has been the case with language understanding technologies, it can foster the creation of reusable resources and infrastructure; it creates an

environment for comparison and replication of results; and it introduces an element of competition to produce better results [10]. At present, evaluation methods for summarization require more research, particularly in developing cost-effective user-centered evaluations and repeatable evaluations. Summarization corpora can play a valuable role in this process.

In recent years, new areas such as multi-document summarization and multi-lingual summarization have assumed increasing importance, posing new requirements for evaluations (see [2] for a roadmap for future summarization evaluations associated with the Document Understanding Conference). New applications for summarization, such as question-answering, condensation and navigation of book-length materials, summaries for hand-held devices, etc., will create new opportunities as well as challenges for summarization evaluation.

References

[1] Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. 1997. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*. MIT Press, pp. 71-80.

[2] Baldwin, N., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D., Sparck Jones, K., Sundheim, B., Teufel, S., Weischedel, R., and White, M. 2000. An Evaluation Roadmap for Summarization Research. www-nlpir.nist.gov/projects/duc/roadmapping.html

[3] Borko, H. and Bernier, C. 1975. Abstracting Concepts and Methods. Academic Press.

[4] Brandow, R., Mitze, K., and Rau, L. 1994. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5), pp. 675-685. Reprinted in Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 293-303.

[5] CMP-LG Annotated Corpus. www.itl.nist.gov/div894/894.02/related_projects/tipster_summac/index.html.

[6] Donaway, R. L., Drummey, K. W., and Mather, L. A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. *Proceedings of the Workshop on Automatic Summarization*, pp. 69-78.

[7] Edmundson, H.P. 1969. New methods in automatic abstracting. *Journal of The Association for Computing Machinery*, 16(2), pp. 264-285. Reprinted in Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 21-42.

[8] Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 121-128.

[9] Harman, D.K. and Voorhees, E.M. 1996. The fifth text retrieval conference (trec-5). *National Institute of Standards and Technology NIST SP 500-238*.

[10] Hirschman, L. and Mani, I. 2001. Evaluation. In Mitkov, R. (ed.), *Handbook of Computational Linguistics*. Oxford University Press, to appear.

[11] Hovy, E., and Marcu, D. 1998. COLING-ACL'98 Tutorial on Text Summarization. www.isi.edu/ marcu/coling-acl98-tutorial.html.

[12] Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. 1998. Summarization evaluation methods: Experiments and analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization, Spring 1998, Technical Report, AAAI, 1998*.

[13] Jing, H. and McKeown, K. 1999. The Decomposition of Human-Written Summary Sentences. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 129-136.

[14] Kupiec, J., Pedersen, F., and Chen, F. 1995. A trainable document summarizer. *Proceedings of the 18th ACM SIGIR Conference (SIGIR'95)*, pp. 68-73. Reprinted in Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 55-60.

[15] Lin, C-Y. 1999. Training a selection function for extraction. *Proceedings of the the Eighteenth International Conference on Information and Knowledge Management (CIKM'99)*, pp. 1-8.

[16] Mani, I. and Bloedorn, E. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1, pp. 35-67.

[17] Mani, I., Firmin, T., House, D., Chrzanowski, M., Klein, G., Hirschman, L., Sundheim, B., and Obrst, L. 1998. "The TIPSTER SUMMAC Text Summarization Evaluation: Final Report". MITRE Technical Report MTR 98W0000138. McLean, VA: The MITRE Corporation.

[18] Mani, I., Gates, B., and Bloedorn, E. 1999. Improving Summaries by Revising Them. *Proceedings of the 37th Annual Meeting of the ACL*, pp. 558-565.

[19] Mani, I. and Maybury, M., eds. 1999. *Advances in Automatic Text Summarization.* MIT Press.

[20] Mani, I. 2001. *Automatic Summarization.* John Benjamins (to appear).

[21] Marcu, D. 1999. "Discourse trees are good indicators of importance in text". In Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 123-136.

[22] Marcu, D. 1999. The automatic construction of large-scale corpora for summarization research. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 137-144.

[23] Maybury, M. 1995. Generating Summaries from Event Data. *Information Processing and Management, 31,5, pp. 735-751.*

[24] Merlino, A. and Maybury, M. "An empirical study of the optimal presentation of multimedia summaries of broadcast news". In Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 391-401.

[25] Minel, J-L., Nugier, S., and Piat, G. 1997. How to appreciate the quality of automatic text summarization. In Mani, I. and Maybury, M., eds., *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 25-30.

[26] Morris, A., Kasper, G., and Adams, D. 1992. The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. *Information Systems Research,* 3(1), pp. 17-35. Reprinted in Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 305-323.

[27] NTCIR Text Summarization Challenge. 2001. galaga.jaist.ac.jp:8000/tsc

[28] Okurowski, M.E., Wilson, W., Urbina, J., Taylor, T., Clark, R. C., and Krapcho, F. 2000. "Text summarization in use: lessons learned from real world deployment and evaluation." In Proceedings of the Workshop on Automatic Summarization, pp. 49-58.

[29] Paice, C.D. and Jones, P. A. 1993. "The Identification of Important Concepts in Highly Structured Technical Papers." In Proceedings of the 16th International Conference on Research and Development in Information Retrieval (SIGIR'93), pp. 69-78. New York: Association for Computing Machinery.

[30] Radev, D. R., Jing, H., and Budzikowska, M. 2000. "Summarization of multiple documents: clustering, sentence extraction, and evaluation". In Proceedings of the Workshop on Automatic Summarization, pp. 21-30.

[31] Rath, G.J., Resnick, A., and Savage, T.R. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2), pp. 139-143. Reprinted in Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 287-292.

[32] Robin, J. 1994. Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design and implementation. Ph.D. Thesis, Columbia University.

[33] Saggion, H. and Lapalme, G. 2000. Concept Identification and Presentation in the Context of Technical Text Summarization. *Proceedings of the Workshop on Automatic Summarization*, pp. 1-10.

[34] Salton, G. and McGill, M. J. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

[35] Salton, G., Singhal, A., Mitra, M., and Buckley, C. 1997. Automatic Text Structuring and Summarization. *Information Processing and Management*, 33(2), 193-207. Reprinted in Mani, I., and Maybury, M., eds., *Advances in Automatic Text Summarization*, MIT Press, pp. 341-355.

[36] Sparck-Jones, K., and Galliers, J. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review.* Lecture Notes in Artificial Intelligence 1083. Springer-Verlag.

[37] Tombros, A., and Sanderson, M. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st ACM SIGIR Conference (SIGIR'98)*, pp. 2-10.

[38] van Dijk, T. A. 1979. "Recalling and Summarizing Complex Discourse". In W. Burchart and K. Hulker (eds.), Text Processing, 49-93. Berlin:Walter de Gruyter.