A Proposal for Task-based Evaluation of Text Summarization Systems

Thérèse Firmin Hand
Department of Defense
9800 Savage Rd
Ft Meade, MD 20755-6000, USA
tfirmin@romulus.ncsc.mil

Abstract

Evaluation is a key part of any research and development effort, but the goals and focus of evaluations are often narrow in scope, addressing a specific algorithm or technique, or analyzing a single result All of the evaluation work done to date on text summarization systems has been by the *developers* of individual systems, usually to study and improve sentence selection criteria Under TIPSTER III, DARPA is sponsoring a task-based evaluation of multiple text summarization systems This focus of this evaluation will be on *user* needs, and the feasibility of applying summarization technology to a variety of tasks

1 Introduction

The explosion of on-line textual material and the advances in text processing technology have provided an important opportunity for broad application of text summarization systems Numerous techniques for deriving summaries from full text documents have already been implemented, and there are several commercial summarization products available The summaries generated by these systems are potentially useful in a variety of settings In 1997, the US Government will begin a Defense Advanced Research Projects Agency (DARPA)-sponsored program under the TIPSTER umbrella to evaluate full text summarization systems to provide feedback to researchers and commercial institutions on the utility of various approaches to specific summarization tasks TIPSTER, discussed in more detail later, is a DARPA initiative with participation from multiple US government agencies and research and commercial institutions to push the state of the art in text processing technologies

2 Concepts of Text Summarization

Automatic summaries are usually described in terms of certain key features which relate to the concepts of intent, focus, and coverage

- *Intent* describes the potential use of the summary, either indicative or informative Indicative summaries, used in this context, provide just enough information to judge the relevancy of the full text Informative or substantive summaries serve as substitutes for the full documents, retaining all important details
- *Focus* refers to the scope of the summary, either generic or user-directed A generic summary is based on the main concept(s) of a document, whereas a user- or goal-directed summary is based on the topic of interest indicated by the recipient of the summary
- *Coverage* indicates whether the summary is based on a single document or multiple documents

Much of the historical work in automatic text summarization has been geared towards the creation of indicative, generic summaries of single documents For example, the work of Luhn (1958), Edmundson (1969), Johnson *et al* (1993) and Brandow *et al* (1995) all generated this type of summary, although their approaches have included different combinations of statistical and linguistic techniques Luhn (1958) considered frequency of word occurrence within a document and the position of the word in a sentence, Edmundson (1969) looked at cue words, title and heading words, and structural indicators, Johnson *et al* (1993) used indicator phrases, and Brandow *et al* (1995) applied sentence weighting using signature word selection Most of these approaches claim some degree of domain independence, however they have been tested only on a specific type of data, such as newspaper articles (Brandow *et al* 1995) or technical literature (Edmundson 1969)

More recently, the scope of research has expanded to include informative, user-directed, and multi-document summaries Reimer and Hahn (1988), Maybury (1993), and McKeown and Radev (1995) used knowledge-based approaches to generate informative summaries that can serve as substitutes for the original document

The expansion in focus to include user-directed summaries has been influenced by research in information retrieval community on passage-based retrieval, as in the work of Knaus *et al* (1996) Also, advances in statistical learning algorithms, such as those implemented by Kupiec *et al* (1995) and Aone *et al* (1997) have combined generic summaries and user-customization, allowing the user to affect the content of the summaries by manipulating sentence extraction features

The potential for multi-document summarization as proposed by the work of Strzalkowski (1996) and Mani and Bloedorn (1997) is based in part on advances in information retrieval and information extraction performance

3 Previous Evaluations

During the course of their development, most of the above systems were subject to some form of evaluation Many of these evaluations relied on the presence of a human-generated target abstract, or the notion of a single 'best' abstract, although there is fairly uniform acceptance of the belief that any number of acceptable abstracts could effectively represent the content of a single document Human-generated abstracts attempt to capture the central concept(s) of a document using the terminology of the document, along the lines of a generic summary The comparisons made between the human-generated versus machine-generated summaries were intended primarily for the developers' own benefit, and evaluate the technology itself, rather than the utility of the technology for a given task Other evaluations did focus on specific tasks and potential uses of automatic summaries, but only with respect to a single system and a limited document set

Many different techniques were attempted in the area of intrinsic or developer-oriented evaluations, which judge the quality of summaries Edmundson (1969) compared sentence selection in the automatic abstracts to the target abstracts, and also performed a subjective evaluation of the content Johnson *et al* (1993) proposed matching a template of manually generated key concepts with the concepts included in the abstract, and performed one sample abstract evaluation Paice and Jones (1993) used a set of statistics to determine if the summary effectively captured the focal concepts, the non-focal concepts, and conclusions Using a strictly statistical measure, Kupiec *et al* (1995) calculated the percentage of sentence matches and partial matches between their automatic summary and a manually generated abstract The main problem with this type of evaluation is its reliance on the notion of a single 'correct' abstract Since many different representations of a document can form an effective summary, this is an inappropriate measure

In extrinsic or task-oriented evaluations, the information retrieval notion of relevancy of a document to a specific topic is the common measure for summarization testing Miike *et al* (1994) analyzed key sentence coverage and also recorded timing and precision/recall statistics to make relevance decisions based on summaries for a domain-specific summarizer Brandow *et al* (1995) had news analysts compare the summaries generated using statistical and natural language processing (NLP) techniques to summaries using the initial sentences (called the "lead summaries") of the document Brandow *et al* (1995) discovered that in general, experienced news analysts felt that the lead summaries were more acceptable than the summaries created using sophisticated NLP techniques Mani and Bloedorn (1997) generated similar precision/recall and timing measures for an information retrieval experiment using a graph search and matching technique and

Task	Intent	Focus	Coverage	Evaluation Decision	Quantitative measures
Categorization	Indicative	Generic	Single document	appropriate category	time accuracy
Adhoc	Indicative	User-directed	Single document	relevant to topic	time accuracy

TABLE 1. Proposed Evaluation

learned that their summaries were effective enough to support accurate retrieval.

4 Proposed Evaluation

Full text summarization is a major task in TIPSTER Phase III. TIPSTER Phase I sponsored research in information extraction and information retrieval, and supported the Message Understanding Conferences (MUC) and Text REtrieval Conferences (TREC) for evaluating extraction and retrieval performance, respectively (Merchant, 1993). TIPSTER Phase II concentrated on defining a common architecture to facilitate integration of the two technologies. TIPSTER Phase III continues to advance research in extraction and retrieval, and adds text summarization in both the research and formal evaluation arenas (Merchant, 1996). This proposed evaluation will be a formal, large scale, multiple task, multiple system evaluation independent from any single approach or methodology.

As outlined in Table 1, the proposed evaluation for text summarization will be task-based, judging the utility of a summary to a particular task. It will be an evaluation for users, determining fitness for a particular purpose, versus an evaluation strictly for developers. It is not intended to pick the best systems, but to understand some of the issues involved in building summarization systems and evaluating them. It will provide an environment whereby systems will be judged independently on their applicability to a given task.

We will begin with at least two tasks for the first evaluation, following the MUC and TREC examples of testing along multiple dimensions. We hope this will avoid any redirection of research efforts based on relative performance on any given task.

Additional tasks will be added in subsequent years to evaluate other aspects of text summaries. These tasks will also reflect continued maturation of the technology.

4.1 Goals

Automatic text summarization systems lend themselves to many tasks. An informative summary may be used as the basis for executive decisions. An indicative summary may be used as an initial indicator of relevance prior to reviewing the full text of a document (and possibly eliminating the need to view that full text). Summaries (used in place of full text documents) may also be used to improve precision in information retrieval systems, since users would be searching only the content-relevant words or phrases within a document (Brandow *et al.*, 1995). For this initial evaluation, we will concentrate on tasks that appear to offer the possibility of near term payoff for users. We attempted to devise tasks that model the real world activities of information analysts and consumers of large quantities of text. These tasks were designed based on interviews with users who spend a majority of their workday searching through volumes of on-line text for information relevant to their area of interest.

We will begin with tasks that address the focus (generic or user-directed) of the summaries. The first task, categorization, will evaluate generic summaries, and the other, adhoc retrieval, will evaluate user-directed summaries, as described below

4.1.1 Task 1 - Categorization

While information routing systems are becoming prevalent in many work environments, there is still a role in many such places for a central review authority to scan and distribute all incoming documents based on their content, essentially performing a manual routing task These reviewers deal both with a broad topic base and with data from multiple sources They must browse a document quickly to determine the key concepts, and forward that document to the appropriate individual

A related task involves scanning a large set of documents that has been selected using an extremely broad indicator or concept A user will browse through this data and categorize it according to various parameters For example, on the World-Wide-Web (WWW), information seekers frequently enter short, broad queries that return hundreds or even thousands of documents The user must determine which documents represent the greatest potential for providing information of interest

Integrating text summarization into each of the above scenarios, the user would be presented a generic summary in lieu of the full text, from which he or she will make a categorization decision

The evaluation task will simulate the manual routing scenario described above The goal will be to decide quickly whether or not a document contains information about any of a limited number of topic areas The document will be limited to a single topic

Selections from the TREC test collections of query topics and documents will be used as the data for the evaluation We will select a minimum of five distinct topics, approximately 200 documents per topic At least two of the topics will be entity-based (i e based on the MUC categories of person, location, and organization) The topics will be related at a very broad level The document set provided will be that returned as a result of five simple queries to a commonly used information retrieval system, which should provide an adequate mix of shorter and longer documents The resulting documents will be randomly mixed The TREC test collections are described in detail in Harman (1993)

Only the documents will be provided to the evaluation participants Summarization systems developed by the participants will automatically generate a generic summary of each document There will not be any constraints on the format of the summary All summaries submitted by the participants will be combined by the evaluation organizers into a single group and randomly mixed

The full text of the document and the lead sentences of the document (up to the specified cutoff length) will be used as baselines The summaries provided by the participants, the baseline lead summaries, and the full text documents will be mixed together, resulting in N+2 versions of a single document, where N is the number of evaluation participants This document set will be randomly divided among the assessors Assessors for the evaluation will be professional information analysts Each assessor will read a summary or document and categorize it into one of the five topic areas that were selected by the organizers, or 'none of the above', which can be considered a sixth category No assessor will read more than one version (summary or full text) of a single document The assessor's decision-making process will be timed The assessor will then move on to the next document of summary

In addition to the TREC relevance judgments, a minimum of two additional assessors will read all of the full text documents to establish a ground truth relevance decision for each

The assessors will be timed, and their categorization decisions will be compared to the ground truth assessments This methodology will assure that the assessors' own categorization performance can be measured along with the performance of the summarization systems

4.1.2 Task 2 - Adhoc retrieval

Both the volume of data available on-line and the prevalence of information retrieval engines have created an immediate application for implementing a text summarization filter as a back end to an information retrieval engine, whereby the user could quickly and accurately judge the relevancy of documents returned as a result of a query The user's query has direct bearing on the content of the documents returned

Applying text summarization to the above scenario, the user would be presented a summary based on the query (a user-directed summary), instead of the full text, from which he or she will make a relevance assessment

The second evaluation task will simulate the adhoc retrieval scenario described above The goal will be to decide the relevancy of a retrieved document by looking only at the user-directed summary that has been generated by the system under evaluation

The TREC collection will also provide the common test data used for this task in the same proportions as for the categorization tasks, five hand-selected topics and approximately 200 documents for each topic The document set provided will be that returned as a result of five queries to a commonly used information retrieval system In this case, both the topics and documents will be provided to the participants Summarization systems developed by the participants will then automatically generate a summary using the topic as the indication of user interest The full text and a keyword-in-context (KWIC) list will be used as baselines

Assessors will work with one topic at a time All summaries received from the participants for a given topic, along with the full text and the KWIC summaries will be combined into a single group, randomly mixed, and divided among the assessors Each assessor will review a topic, then read each summary or document and judge whether or not it is relevant to the topic at hand The assessor will then move on to the next topic No assessor will read more that one representation of a single document

In addition to the TREC relevance judgments, a minimum of two additional assessors will read all of the full text documents to establish a ground truth relevance decision

4.2 Evaluation Criteria

Both evaluations highlight the acceptability of a summary for a given task, with the assumption that there is not a single 'correct' summary The main purpose will be to determine if the evaluator would make the same decision if given the full text, and how much longer it would take to make that decision The ideal outcome would be that the decision could be made with the same accuracy in shorter time, given the document summary For each task, we will record the time required to make each decision, and the actual decision The decision for each evaluator will then be compared to the relevance decision for the baselines Analysis of the results will include consideration of the effects of summary length on the time taken to make the relevance decision as well as its effects on decision accuracy

Quantitative measures

- Categorization/Relevance Decisions

Determining relevance to a given topic is an inherently subjective activity We intend to mitigate this by using a sound statistical model to determine the appropriate number of summaries to evaluate, and by structuring the evaluation in such a way as to avoid bias of any single assessor As previously discussed, we will establish low-end and high-end baselines and use multiple assessors to create ground truth decisions

- Time Required

The time required to make a relevance or categorization decision using a summary will be recorded and compared with the time required to make the same decision using the full text

- Summary Length

In previous studies, 20-30% of full document length was often used as optimal cutoff length for informative summaries, with the supposition that indicative summaries would require far less information ((Brandow *et al* , 1995) and (Kupiec *et al* , 1995)) For the initial evaluation, which will use indicative summaries only, a document cutoff length will be established at 10% of the original document length Any summary exceeding that margin will be truncated

Qualitative measures

- User Preference

Evaluators will be asked to indicate whether they prefer the full text or the summary as a basis for decision-making In addition to this qualitative

Task	Intent	Focus	Coverage	Evaluation Goal	Quantitative measures
Index summaries for information retrieval	Indicative	Generic	Single document	Improve IR precision	Precision Recall
Summarize across documents	Indicative or Informative	Generic or User-directed	Multiple document	Reduce information processing load	Time Accuracy
Executive decision making	Informative	Generic or User-directed	Single or multi-document	Include all relevant information	Key concept matching Formatted questions

TABLE 2. Future Evaluations

assessment, the evaluator will be encouraged to provide feedback as to why the summary was or was not acceptable for a given task This feedback will then be made available for system developers It could also provide a basis for subsequent evaluations

5 Future Direction of Evaluation

This initial evaluation will address only a limited number of issues involving automatic text summarization technology As we gain more experience working with these systems and integrating them into a user's work flow, the scope of the evaluations will necessarily grow and change Some additional features and tasks to be addressed potentially in future evaluations have already been identified, including cohesiveness of a summary, optimal length of a summary, and multi-document summaries Selected tasks are outlined in Table 2 and described briefly below

5.1 Tasks and Measures

We are addressing two information retrieval types of tasks during the first evaluation, however, potential applications go beyond this limited scope One of the frequently mentioned uses of a text summary is as a substitute for the document during the indexing process of an information retrieval system The notion is that indexing based on summaries would result in more results retrievals because only the key concepts and content-bearing words would have been indexed This idea could be evaluated using standard precision and recall information retrieval measures

Summarizing across multiple documents is another extremely useful application While single document summaries are expected to provide improved efficiency for the end-user, much of the information reviewed from one summary to the next will be redundant Automatically generated summaries could result in even larger efficiency gains and productivity improvements by distilling the information from multiple documents into a single summary An evaluation of this type of summary would be much more complex, possibly comparing at a phrase-matching or key concept level the combined factual information included in a single summary with manually identified key information in individual documents The evaluation would verify that the relevant aspects of key facts across documents have been successfully identified and combined in the resulting summary

A third application could focus on a decision-making task based on an informative summary An evaluation of this type of summary could include filling out a template indicating key concepts in a document, similar to the Paice and Jones (1993) and Johnson *et al* (1993) evaluations, possibly augmented by a question/answer measure based on the full text and the summary

5.2 Data

Newspaper articles, such as those which will be used for the first evaluation, represent only a small portion of the type of information available online A useful, effective summarizer should be able to accept text in a variety of formats With each subsequent evaluation, new sources of data will be added These new sources could be news feeds or web pages They will tend to be less formatted, vary greatly in length, and cover multiple topics At some point, we hope to introduce documents in languages other than English for summarization either into their native language or into English

6 Acknowledgments

The author is grateful to Donna Harman and Beth Sundheim for their support and assistance in designing the evaluation

The views expressed in this paper are those of the author and do not necessarily reflect the views of the Department of Defense or any of its agencies

References

Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen 1997 A Scalable Summarization System Using Robust NLP In *Proceedings of ACL-97*, Madrid, Spain, July To appear

Ronald Brandow, Karl Mitze, and Lisa F Rau 1995 Automatic Condensation of Electronic Publications by Sentence Selection *Information Processing and Management*, 31(5) 675-685

Kenneth W Church and Lisa F Rau 1995 Commercial Applications of Natural Language Processing *Communications of the ACM*, 38(11) 71-79

H P Edmundson 1969 New Methods in Automatic Abstracting *Journal of the ACM*, 16(2) 264-285

Brigette Endres-Niggemeyer, Jerry Hobbs, and Karen Sparck Jones 1993 Summarizing Text for Intelligent Communication In *Dagstuhl Seminar Report*, IBFI GmbH, Schloss Dagstuhl, Wadern, Germany

J R Galliers and Karen Sparck Jones 1993 Evaluating Natural Language Processing Systems *University of Cambridge Computer Laboratory Technical Report No 291*, Computer Laboratory, University of Cambridge

Donna Harman 1993 Overview of the First Text REtrieval Conference (TREC-1) In *TREC-2 Proceedings*, Gaithersburg, Maryland

Donna Harman 1996 Overview of the Fourth Text REtrieval Conference (TREC-4) In *The Fourth Text REtrieval Conference (TREC-4)*, pages 1-24, Gaithersburg, Maryland, 1995

F C Johnson, C D Paice, W J Black, and A P Neal 1993 The application of linguistic processing to automatic abstract generation *Journal of Document and Text Management*, 1(3) 215-241

Daniel Knaus, Elke Mittendorf, Peter Schauble, and Páraic Sheridan 1996 Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System In *The Fourth Text REtrieval Conference (TREC-4)*, pages 233-238, Gaithersburg, Maryland, 1995

Julian Kupiec, Jan Pedersen, and Francine Chen 1995 A Trainable Document Summarizer *SIGIR '95*, pages 68-73, Seattle, Washington, 1995

H P Luhn 1958 The Automatic Creation of Literature Abstracts *IBM Journal*, pages 159-165

Inderjeet Mani and Eric Bloedorn 1997 Multidocument Summarization by Graph Search and Matching In *Proceedings of AAAI-97*, Providence Rhode Island, 1997 To appear

Mark T Maybury 1993 Automated Event Summarization Techniques In *Dagstuhl Seminar Report*, pages 100-108, IBFI GmbH, Schloss Dagstuhl, Wadern, Germany

Kathleen McKeown and Dragomir R Radev 1995 Generating Summaries of Multiple News Articles *SIGIR '95*, pages 74-82, Seattle, Washington

Roberta Merchant 1993 Tipster Program Overview In *Tipster Text Program*, pages 1-2, Fredericksburg, Virginia

Roberta Merchant 1996 TIPSTER Phase III In *TIPSTER Text Phase III Kickoff Workshop*, Columbia, Maryland, October

Andrew H Morris, George M Kasper, and Dennis A Adams 1992 The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance *Information Systems Research* 3 1, pages 17-35

Seiji Miike, Etsuo Itoh, Kenji Ono, and Kazuo Sumita 1994 A Full-Text Retrieval System with a Dynamic Abstract Generation Function *SIGIR '94*, pages 152-161, Seattle, Washington

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

Chris D Paice and Paul A Jones 1993 The Identification of Important Concepts in Highly Structured Technical Papers *SIGIR '93*, pages 69-77

G J Rath, A Resnick, and T R Savage 1961 The Formation of Abstract by the Selection of Sentences *American Documentation*, pages 139-143

U Reimer and U Hahn 1988 Text Condensation as a Knowledge Base Abstraction *IEEE Conference on AI Applications*, pages 338-344

Tomek Strzalkowski 1996 Robust Natural Language Processing and User-Guided Concept Discovery for Information Retrieval, Extraction, and Summarization Tipster Phase III In *TIPSTER Text Phase III Kickoff Workshop*, Columbia, Maryland, October

Beth Sundheim 1995 Overview of Results of the MUC-6 Evaluation In *Sixth Message Understanding Conference (MUC-6)*, pages 13-31, Columbia, Maryland

Sarah Taylor 1996 TIPSTER Text Program Overview In *TIPSTER Text Phase II*, Tysons Corner, Virginia