# A Text-Extraction Based Summarizer

## Tomek Strzalkowski, Gees C. Stein and G. Bowden Wise

GE Corporate Research & Development
1 Research Circle
Niskayuna, NY 12309, USA

## Abstract

We present an automated method of generating human-readable summaries from a variety of text documents including newspaper articles, business reports, government documents, even broadcast news transcripts. Our approach exploits an empirical observation that much of the written text display certain regularities of organization and style, which we call the Discourse Macro Structure (DMS). A summary is therefore created to reflect the components of a given DMS. In order to produce a coherent and readable summary we select continuous, well-formed passages from the source document and assemble them into a mini-document within a DMS template. In this paper we describe an automated summarizer that can generate both short indicative abstracts, useful for quick scanning of a list of documents, as well as longer informative digests that can serve as surrogates for the full text. The summarizer can assist the users of an information retrieval system in assessing the quality of the results returned from a search, preparing reports and memos for their customers, and even building more effective search queries.

## Introduction

A good summarization tool can be of enormous help for those who have to process large amounts of documents. In information retrieval one would benefit greatly from having content-indicative quick-read summaries supplied along with the titles returned from search. Similarly, application areas like routing, news on demand, market intelligence and topic tracking would benefit from a good summarization tool.

Perhaps the most difficult problem in designing an automatic text summarization is to define what a summary is, and how to tell a summary from a non-summary, or a good summary from a bad one. The answer depends in part upon who the summary is intended for, and in part upon what it is meant to achieve, which in large measure precludes any objective evaluation. A good summary should at least be a good reflection of the original document while being considerably shorter than the original thus saving the reader valuable reading time.

In this paper we describe an automatic way to generate summaries from text-only documents. The summarizer we developed can create general and topical indicative summaries, and also topical informative summaries. Our approach is domain-independent and takes advantage of certain organization regularities that were observed in news-type documents. The system participated in a third-party evaluation program and turned out to be one of the top-performing summarizers. Especially the quality/length ratio was very good since our summaries tend to be very short (10% of the original length).

The summarizer is still undergoing improvement and expansion in order to be able to summarize a wide variety of documents. It is also used successfully as a tool to solve different problems, like information retrieval and topic tracking.

## Task Description and Related Work

For most of us, a summary is a brief synopsis of the content of a larger document, an abstract recounting the main points while suppressing most details. One purpose of having a summary is to quickly learn some facts, and decide what you want to do with the entire story. Depending on how they are meant to be used one can distinguish between two kinds of summaries. *Indicative* summaries are not a replacement for the original text but are meant to be a good reflection of the kind of information that can be found in the original document. *Informative* summaries can be used as a replacement of the original document and should contain the main facts of the document. Independent of their usage summaries can be classified as *general* summaries or *topical* summaries. A general summary addresses the main points of the document ignoring unrelated issues. A topical sum-

mary will report the main issues relevant to a certain topic, which might have little to do with the main topic of the document. Both summaries might give very different impressions of the same document. In this paper we describe a summarizer that summarizes one document, text only, at a time. It is capable of producing both topical and generic indicative summaries, and topical informative summaries.

Our early inspiration, and a benchmark, have been the Quick Read Summaries, posted daily off the front page of New York Times on-line edition (http://www.nytimes.com). These summaries, produced manually by NYT staff, are assembled out of passages, sentences, and sometimes sentence fragments taken from the main article with very few, if any, editorial adjustments. The effect is a collection of perfectly coherent tidbits of news: the who, the what, and when, but perhaps not why. Indeed, these summaries leave out most of the details, and cannot serve as surrogates for the full article. Yet, they allow the reader to learn some basic facts, and then to choose which stories to open.

This kind of summarization, where appropriate passages are extracted from the original text, is very efficient, and arguably effective, because it doesn't require generation of any new text, and thus lowers the risk of misinterpretation. It is also relatively easier to automate, because we only need to identify the suitable passages among the other text, a task that can be accomplished via shallow NLP and statistical techniques. Nonetheless, there are a number of serious problems to overcome before an acceptable quality summarizer can be built. For one, quantitative methods alone are generally too weak to deal adequately with the complexities of natural language text. For example, one popular approach to automated abstract generation has been to select key sentences from the original text using statistical and linguistic cues, perform some cosmetic adjustments in order to restore cohesiveness, and then output the result as a single passage, e.g., (Luhn 1958) (Paice 1990) (Brandow, Mitze, & Rau 1995) (Kupiec, Pedersen, & Chen 1995). The main advantage of this approach is that it can be applied to almost any kind of text. The main problem is that it hardly ever produces an intelligible summary: the resulting passage often lacks coherence, is hard to understand, sometimes misleading, and may be just plain incomprehensible. In fact, some studies show (cf. (Brandow, Mitze, & Rau 1995)) that simply selecting the first paragraph from a document tends to produce better summaries than a sentence-based algorithm.

A far more difficult, but arguably more "human-like" method to summarize text (with the possible exception of editorial staff of some well-known dailies) is to comprehend it in its entirety, and then write a summary "in your own words." What this amounts to, computationally, is a full linguistic analysis to extract key text components from which a summary could be built. One previously explored approach, e.g., (Ono, Sumita, & Miike 1994) (McKeown & Radev 1995), was to extract discourse structure elements and then generate the summary within this structure. In another approach, e.g., (DeJong 1982) (Lehnert 1981) pre-defined summary templates were filled with text elements obtained using information extraction techniques. Marcu (Marcu 1997a) uses rhetorical structure analysis to guide the selection of text segments for the summary; similarly Teufel and Moens (Teufel & Moens 1997) analyze argumentative structure of discourse to extract appropriate sentences. While these approaches can produce very good results, they are yet to be demonstrated in a practical system applied to a reasonable size domain. The main difficulty is the lack of an efficient and reliable method of computing the required discourse structure.

## Our Approach

The approach we adopted in our work falls somewhere between simple sentence extraction and text-understanding, although philosophically we are closer to NYT cut-and-paste editors. We overcome the shortcomings of sentence-based summarization by working on paragraph level instead. Our summarizer is based on taking advantage of paragraph segmentation and the underlying Discourse Macro Structure of News texts. Both will be discussed below.

### Paragraphs

Paragraphs are generally self-contained units, more so than single sentences, they usually address a single thought or issue, and their relationships with the surrounding text are somewhat easier to trace. This notion has been explored by Cornell's group (Salton *et al.* 1994) to design a summarizer that traces inter-paragraph relationships and selects the "best connected" paragraphs for the summary. Like in Cornell's system, our summaries are made up of paragraphs taken out of the original text. In addition, in order to obtain more coherent summaries, we impose some fundamental discourse constraints on the generation process, but avoid a full discourse analysis.

We would like to note at this point that the summarization algorithm, as described in detail later,

does not explicitly depend on nor indeed require input text that is pre-segmented into paragraphs. In general, any length passages can be used, although this choice will impact the complexity of the solution. Lifting well-defined paragraphs from a document and then recombining them into a summary is relatively more straightforward than recombining other text units. For texts where there is no structure at all, as in a closed-captioned stream in broadcast television, there are several ways to create artificial segments. The simplest would be to use fixed word-count passages. Or, content-based segmentation techniques may be applicable, e.g., Hearst's Text-Tiling (Hearst 1997).

On the other hand, we may argue that essentially any length *segments* of text can be used so long as one could figure out a way to reconnect them into paragraph-like passages even if their boundaries were somewhat off. This is actually not unlike dealing with the texts with very fine grained paragraphs, as is often the case with news-wire articles. For such texts, in order to obtain an appropriate level of chunking, some paragraphs need to be reconnected into longer passages. This may be achieved by tracking co-references and other text cohesiveness devices, and their choice will depend upon the initial segmentation we work up from.

## Discourse Macro Structure of a Text

It has been observed, eg., (Rino & Scott 1994), (Weissberg & Buker 1990), that certain types of texts, such as news articles, technical reports, research papers, etc., conform to a set of style and organization constraints, called the Discourse Macro Structure (DMS) which help the author to achieve a desired communication effect. For instance, both physics papers and abstracts align closely with the *Introduction-Methodology-Results-Discussion-Conclusion* macro structure. It is likely that other scientific and technical texts will also conform to this or similar structure, since this is exactly the structure suggested in technical writing guidebooks, e.g. (Weissberg & Buker 1990). One observation to make here is that perhaps a proper summary or an abstract should reflect the DMS of the original document. On the other hand, we need to note that a summary can be given a different DMS, and this choice would reflect our interpretation of the original text. A scientific paper, for example, can be treated as a piece of news, and serve as a basis of an un-scientific summary.

News reports tend to be built hierarchically out of components which fall roughly into one of the two categories: the *What-Is-The-News* category, and the optional *Background* category. The *Background*, if present, supplies the context necessary to understand the central story, or to make a follow-up story self-contained. The Background section is optional: when the background is common knowledge or is implied in the main news section, it can, and usually is omitted. The *What-Is-The-News* section covers the new developments and the new facts that make the news. This organization is often reflected in the summary, as illustrated in the example below from NYT 10/15/97, where the highlighted portion provides the background for the main news:

SPIES JUST WOULDN'T COME IN FROM COLD WAR, FILES SHOW

*Terry Squillacote was a Pentagon lawyer who hated her job. Kurt Stand was a union leader with an aging beatnik's slouch. Jim Clark was a lonely private investigator.* [A 200-page affidavit filed last week by] the Federal Bureau of Investigation says the three were out-of-work spies for East Germany. And after that state withered away, it says, they desperately reached out for anyone who might want them as secret agents.

In this example, the two passages are non-consecutive paragraphs in the original text; the string in the square brackets at the opening of the second passage has been omitted in the summary. Here the human summarizer's actions appear relatively straightforward, and it would not be difficult to propose an algorithmic method to do the same. This may go as follows:

1. Choose a DMS template for the summary; e.g., Background+News.

2. Select appropriate passages from the original text and fill the DMS template.

3. Assemble the summary in the desired order; delete extraneous words.

It is worth noting here that the background-context passage is critical for understanding of this summary, but as such provides essentially no relevant information except for the names of the people involved. Incidentally, this is precisely the information required to make the summary self-contained, if for no other reason than to supply the antecedents to the anaphors in the main passage (*the three, they*).

## The Algorithm

The summarizer can work in two modes: generic and topical. In the generic mode, it simply summarizes the main points of the original document.

In the topical mode, it takes a user supplied statement of interest, a topic, and derives a summary related to this topic. A topical summary is thus usually different from the generic summary of the same document. The summarizer can produce both indicative and informative summaries. An indicative summary, typically 5-10% of the original text, is when there is just enough material retained from the original document to indicate its content. An informative summary, on the other hand, typically 20-30% of the text, retains all the relevant facts that a user may need from the original document, that is, it serves as a condensed surrogate, a digest.

The process of assembling DMS components into a summary depends upon the complexity of the discourse structure itself. For news or even for scientific texts, it may be just a matter of concatenating components together with a little of "cohesiveness glue", which may include deleting some obstructing sentences, expanding acronyms, adjusting verb forms, etc. In a highly specialized domain (e.g., court rulings) the final assembly may be guided by a very detailed pattern or a script that conforms to specific style and content requirements.

Below we present a 10-step algorithm for generating summaries of news-like texts. This is the algorithm underlying our current summarizer. The reader may notice that there is no explicit provision for dealing with DMS structures here. Indeed, the basic Background+News summary pattern has been tightly integrated into the passage selection and weighting process. This obviously streamlines the summarization process, but it also reflects the notion that news-style summarization is in many ways basic and subsumes other more complex summarization requirements.

## The Generalized Summarization Algorithm

s0: Segment text into passages. Use any available handles, including indentation, SGML, empty lines, sentence ends, etc. If no paragraph or sentence structure is available, use approximately equal size chunks.

s1: Build a paragraph-search query out of the content words, phrases and other terms found in the title, a user-supplied topic description (if available), as well as the terms occurring frequently in the text.

s2: Reconnect adjacent passages that display strong cohesiveness by one-way *background links*, using handles such as outgoing anaphors and other backward references. A background link from passage $N+1$ to passage $N$ means that if passage $N+1$

is selected for a summary, passage $N$ must also be selected. Link consecutive passages until all references are covered.

s3: Score all passages, including the linked groups with respect to the paragraph-search query. Assign a point for each co-occurring term. The goal is to maximize the overlap, so multiple occurrences of the same term do not increase the score.

s4: Normalize passage scores by their length, taking into account the desired target length of the summary. The goal is to keep summary length as close to the target length as possible. The weighting formula is designed so that small deviations from the target length are acceptable, but large deviations will rapidly decrease the passage score. The exact formulation of this scheme depends upon the desired tradeoff between summary length and content. The following is the basic formula for scoring passage $P$ of length $l$ against the passage-search query $Q$ and the target summary length of $t$, as used in current version of our summarizer:

$$NormScore(P,Q) = \frac{RawScore(P,Q)}{\sqrt{\frac{|l-t|}{t}+1}}$$

where:

$$RawScore(P,Q) = \sum_{q \in Q} weight(q,P) + prem(P)$$

with sum over unique content terms $q$, and

$$weight(q,P) = \begin{cases} 1 & \text{if } q \in P \\ 0 & \text{otherwise} \end{cases}$$

with $prem(P)$ as a cummulative non-content based score premium (cf s7).

s5: Discard all passages with length in excess of 1.5 times the target length. This reduces the number of passage combinations the summarizer has to consider, thus improving its efficiency. The decision whether to use this condition depends upon our tolerance to length variability. In extreme cases, to prevent obtaining empty summaries, the summarizer will default to the first paragraph of the original text.

s6: Combine passages into groups of 2 or more based on their content, composition and length. The goal is to maximize the score, while keeping the length as close to the target length as possible.

Any combination of passages is allowed, including non-consecutive passages, although the original ordering of passages is retained. If a passage attached to another through a background link is included into a group, the other passage must also be included, and this rule is applied recursively. We need to note that the background links work only one way: a passage which is a background for another passage, may stand on its own if selected into a candidate summary.

S7: Recalculate scores for all newly created groups. This is necessary, and cannot be obtained as a sum of scores because of possible term repetitions. Again, discard any passage groups longer than 1.5 times the target length. Add premium scores to groups based on the inverse degree of text discontinuity measured as a total amount of elided text material between the passages within a group. Add other premiums as applicable.

S8: Rank passage groups by score. All groups become candidate summaries.

S9: Repeat steps S6 through S8 until there is no change in top-scoring passage group through 2 consecutive iterations. Select the top scoring passage or passage group as the final summary.

## Implementation and some Examples

The summarizer has been implemented in C++ with a Java interface as a demonstration system, primarily for news summarization. At this time it can run in both batch and interactive modes under Solaris, and it can also be accessed via Web using a Java compatible browser. Below, we present a few example summaries. For an easy orientation paragraphs are numbered in order they appear in the original text.

TITLE: Mrs. Clinton Says U.S. Needs 'Ways That Value Families'
SUMMARY TYPE: indicative
TARGET LENGTH: 5%
TOPIC: none

(6) The United States, Mrs. Clinton said, must become "a nation that doesn't just talk about family values but acts in ways that values families."

SUMMARY TYPE: indicative
TARGET LENGTH: 15%
TOPIC: Hidden cameras used in news reporting

(4) Roone Arledge, the president of ABC News, defended the methods used to report the segment and said ABC would appeal the verdict.

(5) "They could never contest the truth" of the broadcast, Arledge said. "These people were doing awful things in these stores."

(6) Wednesday's verdict was only the second time punitive damages had been meted out by a jury in a hidden-camera case. It was the first time punitive damages had been awarded against producers of such a segment, said Neville L. Johnson, a lawyer in Los Angeles who has filed numerous hidden-camera cases against the major networks.

(7) Many journalists argue that hidden cameras and other undercover reporting techniques have long been necessary tools for exposing vital issues of public policy and health. But many media experts say television producers have overused them in recent years in a push to create splashy shows and bolster ratings. The jurors, those experts added, may have been lashing out at what they perceived as undisciplined and overly aggressive news organizations.

TITLE: U.S. Buyer of Russian Uranium Said to Put Profits Before Security
SUMMARY TYPE: informative
TARGET LENGTH: 25%
TOPIC: nuclear nonproliferation

(1) In a postscript to the Cold War, the American government-owned corporation that is charged with reselling much of Russia's military stockpile of uranium as civilian nuclear reactor fuel turned down repeated requests this year to buy material sufficient to build 400 Hiroshima-size bombs.

(2) The incident raises the question of whether the corporation, the U.S. Enrichment Corp., put its own financial interest ahead of the national-security goal of preventing weapons-grade uranium from falling into the hands of terrorists or rogue states.

(7) The corporation has thus far taken delivery from Russia of reactor fuel derived from 13 tons of bomb-grade uranium. "The nonproliferation objectives of the agreement are being achieved," a spokesman for the Enrichment Corp. said.

(8) But since the beginning of the program, skeptics have questioned the wisdom of designating the Enrichment Corp. as Washington's "executive agent" in managing the deal with Russia's Ministry of Atomic Energy, or MINATOM.

(19) Domenici, chairman of the energy subcommittee of the Senate Appropriations Committee, which is shepherding the privatization plan through Congress, was never informed of the offer by the administration. After learning of the rebuff to the Russians, he wrote to Curtis asking that the Enrichment Corp. "be immediately replaced as executive agent" and warning that "under no circumstances should the sale of the USEC proceed until this matter is resolved." Once Domenici entered the fray, the administration changed its tune.

(20) Curtis sent a letter to Domenici stating that all the problems blocking acceptance of the extra six tons had been solved. People close to the administration said that the Enrichment Corp. has now been advised to buy the full 18-ton shipment in 1997. Moreover, Curtis quickly convened a new committee to monitor the Enrichment Corp. for signs of foot-dragging.

## Evaluation

Our program has been tested on a variety of news-like documents, including Associated Press newswire messages, articles from the New York Times, The Wall Street Journal, Financial Times, San Jose Mercury, as well as documents from the Federal Register, and the Congressional Record. The summarizer is domain independent, and it can be easily adapted to most European languages. It is also

very robust: we used it to derive summaries of thousands of documents returned by an information retrieval system. Early results from these evaluations indicate that the summaries generated using our DMS method offer an excellent tradeoff between time/length and accuracy. Our summaries tend to be shorter and contain less extraneous material than those obtained using different methods. This is further confirmed by the favorable responses we received from the users.

Thus far there has been only one systematic multisite evaluation of summarization approaches, conducted in early 1998, organized by U.S. DARPA[1] in the tradition of Message Understanding Conferences (MUC) (DAR 1993) and Text Retrieval Conferences (TREC) (Harman 1997a), which have proven successful in stimulating research in their respective areas: information extraction and information retrieval. The summarization evaluation focused on content representativeness of indicative summaries and comprehensiveness of informative summaries. Other factors affecting the quality of summaries, such as brevity, readability, and usefulness were evaluated indirectly, as parameters of the main scores. For more details see (Firmin & Sundheim 1998).

The indicative summaries were scored for relevance to pre-selected topics and compared to the classification of respective full documents. In this evaluation, a summary was considered successful if it preserved the original document's relevance or non-relevance to a topic. Moreover, the recall and precision scores were normalized by the length of the summary (in words) relative to the length of the original document, as well as by the clock time taken by the evaluators to reach their topic relevance decisions. The first normalization measured the degree of content compression provided by the summaries, while the second normalization was intended to gauge their readability. The results showed a strong correlation between these two measures, which may indicate that readability was in fact equated with meaningfulness, that is, hard to read summaries were quickly judged non-relevant.

For all the participants the best summaries scored better than the fixed-length summaries. When normalized for length our summarizer had the highest score for best summaries and took the second place for fixed-length summaries. The F-scores for indicative topical summaries (best and fixed-length) were very close for all participants. Apparently it is easier to generate a topical summary then a general summary. Normalizing for length did move our score

up, but again, there was no significant difference between participants.

The informative (topical) summaries were scored for their ability to provide answers to *who, what, when, how*, etc. questions about the topics. These questions were unknown to the developers, so systems could not directly extract facts to satisfy them. Again, scores were normalized for summary length, but no time normalization was used. This evaluation was done on a significantly smaller scale than for the indicative summaries, simply because scoring for question answering was more time consuming for the human judges than categorization decisions. This evaluation could probably be recast as categorization problem, if we only assumed that the questions in the test were the topics, and that a summary needs to be relevant to multiple topics.

Informative summaries were generated using the same general algorithm with two modifications. First, the expected summary length was set at 30% of the original, following an observation by the conference organizers while evaluating human generated summaries. Second, since the completeness of an informative summary was judged on the basis of it containing satisfactory answers to questions which were not part of the topic specification, we added extra scores to passages containing possible answers: proper names (who, where) and numerics (when, how much). Finally, we note that the test data used for evaluation, while generally of news-like genre, varied greatly in content, style and the subject matter, therefore domain-independence was critical.

Again our summarizer performed quite well, although the results are less significant since the experiment was carried out on such a small scale. The results were separated out for three different queries. For two queries the system was very close to the top performing system, and for the third query the system had an F-score of about 0.61 versus 0.77 for the best system.

In general we are quite pleased with the summarizer performance, especially since our system was not trained on the kind of texts that we had to summarize.

## Related Work and Future Work

The current summarizer is still undergoing improvement and adaptation in order to be able to summarize more than a single text news document at a time. At the same time we are investigating how summarization can be used in related but different problems. Both will be described below.

## A better and more flexible summarizer

Currently our summarizer is especially tuned for English one-document text-only news summarization. While we are still working on improving this, we also want the system to be able to summarize a wider variety of documents. Many challenges remain, including summarization of non-news documents, multimodal documents (such as web pages), foreign language documents and (small or large) groups of documents covering one or more topics.

Typically, a user needs summarization the most when dealing with a large number of documents. Therefore, the next logical step is to summarize more than one documents at a time. At the moment we are focusing on multi-document (cross-document) summarization of English text-only news documents. Just as for single-document summarization, multi-document summarization can be generic or topical and indicative or informative. Other factors that will influence the types of summary are the number of documents (a large versus a small set) and the variety of topics discussed by the documents (are the documents closely related or can they cover very different topics). Presentation of a multi-document offers a wide variety of choices. One could create one large text summary that gives an overview of all the main issues mentioned in all summaries. Or perhaps give different short summaries for similar documents. If the number of documents is very large it might be best to create nested summaries with high-level descriptions and the possibility to 'zoom in' on a subgroup with a more specific summary. A user will probably want to have the ability to trace information in a summary back to its original document; source information should be a part of the summary. If one views summarization in the context of tracking a topic, the main goal of the summary might be to show the new information every next document contains, while not repeating information already mentioned in previous documents. Another type of summary might highlight the similarities documents have (e.g., all these documents are on protection of endangered species) and pointing out the differences they have (e.g., one on bald eagles, some on bengal tigers,..). As one can see, there are many questions to be answered and the answers depend partially on the task environment the summarizer will be used in.

Currently we are focussing on summarizing a small set of text-only documents (around 20) all on a similar topic. The summary will reflect the main points/topics discussed by the documents. Topics discussed by more than one document should only be mentioned once in the summary together with its different sources. When generating the summary we want to ensure coherence by placing related topic close to each other. The main issues we are addressing is the detection of similar information in order to avoid repetition in the summary and the detection of related information in order to generated a coherent summary. This work is right now in progress. Our next step will be summarizing large amounts of similar information.

## Applying summarization to different problems

Information retrieval (IR) is a task of selecting documents from a database in response to a user's query, and ranking these documents according to relevance. Currently we are investigating the usage of summarization in order to build (either automatically or with the help of the user) more effective information need statements for an automated document search system. The premise is quite simple: use the initial user's statement of information need to sample the database for documents, summarize the returned documents topically, then add selected summaries to the initial statement to make it richer and more specific. Adding appropriate summaries can be either done by the user who reads the summaries or automatically. Both approaches are described in our other paper appearing in this volume.

The task of tracking a topic consists of identifying those information segments in a information stream that are relevant to a certain topic. Topic tracking is one of the three main tasks in the TDT (Topic Detection and Tracking) tasks that we hope to use our summarizer for. The information stream consists of news, either from a tv broadcast or a radio broadcast. Speech from these programs has been recognized by a state-of-the-art automatic speech recognition system and also transcribed by human transcriptionists. A topic is defined implicitly by a set of training stories that are given to be on this topic. The basic idea behind our approach is simple. We use the training stories to create a set of keywords (the query). Since we process continuous news the input is not segmented into paragraphs or any other meaningful text unit. Before applying our summarizer each story is divided into equal word-size segments. We summarize every story using our query, and use similarity of the summary to the query to decide whether a story is on topic or not. We are still in the process of refining our system and hope to have our first results soon. Initial results suggest that this is a viable approach. It is encouraging to notice that the absence of a paragraph structure does not prevent the system from generating useful

summaries.

## CONCLUSIONS

We have developed a method to derive quick-read summaries from news-like texts using a number of shallow NLP techniques and simple quantitative methods. In our approach, a summary is assembled out of passages extracted from the original text, based on a pre-determined Background-News discourse template. The result is a very efficient, robust, and portable summarizer that can be applied to a variety of tasks. These include brief indicative summaries, both generic and topical, as well as longer informative digests. Our method has been shown to produce summaries that offer an excellent tradeoff between text reduction and content preservation, as indicated by the results of the government-sponsored formal evaluation.

The present version of the summarizer can handle most written texts with well-defined paragraph structure. While the algorithm is primarily tuned to newspaper-like articles, we believe it can produce news-style summaries for other factual texts, as long as their rhetorical structures are reasonably linear, and no prescribed stylistic organization is expected. For such cases a more advanced discourse analysis will be required along with more elaborate DMS templates.

We used the summarizer to build effective search topics for an information retrieval system. This has been demonstrated to produce dramatic performance improvements in TREC evaluations. We believe that this topic expansion approach will also prove useful in searching very large databases where obtaining a full index may be impractical or impossible, and accurate sampling will become critical.

Our future development plans will focus on improving the quality of the summaries by implementing additional passage scoring functions. Further plans include handling more complex DMS's, and adaptation of the summarizer to texts other than news, as well as to texts written in foreign languages. We plan further experiments with topic expansion with the goal of achieving a full automation of the process while retaining the performance gains.

## Acknowledgements

## References

Brandow, R.; Mitze, K.; and Rau, L. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management* 31(5):675–686.

DARPA. 1993. *Proceedings of the 5th Message Understanding Conference*, San Francisco, CA: Morgan Kaufman Publishers.

DARPA. 1996. *Tipster Text Phase 2: 24 month Conference*, Morgan-Kaufmann.

DeJong, G. G. 1982. An overview of the frump system. In Lehnert, W., and Ringle, M., eds., *Strategies for Natural Language Processing*. Lawrence Erlbaum, Hillsdale, NJ.

Firmin, T., and Sundheim, B. 1998. Tipster/summac evaluation analysis. In *Tipster Phase III 18-month Workshop*.

Harman, D., ed. 1997a. *The 5th Text Retrieval Conference (TREC-5)*, number 500–253. National Institute of Standards and Technology.

Hearst, M. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Kupiec, J.; Pedersen, J.; and Chen, F. 1995. A trainable document summarizer. In *Conference of the ACM Special Interes Group on Information Retrieval (SIGIR)*, 68–73.

Lehnert, W. 1981. Plots units and narrative summarization. *Cognitive Science* 4:293–331.

Luhn, H. 1958. The automatic creation of literature abstracts. *IBM Journal* 159–165.

Marcu, D. 1997a. From discourse structures to text summaries. In *Proceedings of the ACL Workshop on Intelligent, Scallable Text Summarization*, 82–88.

Marcu, D. 1997b. The rhetorical parsing of natural language texts. In *Proceedings of 35th Annual Meetings of the ACL*, 96–103.

McKeown, K., and Radev, D. 1995. Generating summaries of multiple news articles. In *Proceedings of the 8th Annual ACM SIGIR Conference on R&D in IR*.

Ono, K.; Sumita, K.; and Miike, S. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the International Conference on Computational Linguisitcs (COLING-94)*, 344–348.

Paice, C. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management* 26(1):171–186.

Rino, L., and Scott, D. 1994. Content selection in summary generation. In *Third International Conference on the Cognitive Science of Natural Language Processing*.

Salton, G.; Allan, J.; Buckley, C.; and Singhal, A. 1994. Automatic analysis, theme generation, and summarization of machine readable texts. *Science* 264:1412–1426.

Strzalkowski, T., and Wang, J. 1996. A self-learning universal concept spotter. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-96)*, 931–936.

Teufel, S., and Moens, M. 1997. Sentence extraction as classification task. In *Proceedings of the ACL Workshop on Intelligent, Scallable Text Summarization*.

Weissberg, R., and Buker, S. 1990. *Writing up Research: Experimental Research Report Writing for Student of English*. Prentice Hall, Inc.