# Multi-Document Summarization By Sentence Extraction

**Jade Goldstein*** **Vibhu Mittal†** **Jaime Carbonell*** **Mark Kantrowitz†**
*jade@cs.cmu.edu* *mittal@jprc.com* *jgc@cs.cmu.edu* *mkant@jprc.com*

*Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
U.S.A.

†Just Research
4616 Henry Street
Pittsburgh, PA 15213
U.S.A.

## Abstract

This paper discusses an sentence extraction approach to multi-document summarization that builds on single-document summarization methods by using additional, available information about the document set as a whole and the relationships between the documents. Multi-document summarization differs from single in that the issues of compression, speed, redundancy and passage selection are critical in the formation of useful summaries. Our approach addresses these issues by using domain-independent techniques based mainly on fast, statistical processing, a metric for reducing redundancy and maximizing diversity in the selected passages, and a modular framework to allow easy parameterization for different genres, corpora characteristics and user requirements.

## 1 Introduction

With the continuing growth of online information, it has become increasingly important to provide improved mechanisms to find and present textual information effectively. Conventional IR systems find and rank documents based on maximizing relevance to the user query (Salton, 1970; van Rijsbergen, 1979; Buckley, 1985; Salton, 1989). Some systems include sub-document relevance assessments and convey this information to the user via techniques such as *text tiling* (Hearst, 1997). More recently, single document summarization systems provide an automated generic abstract or a query-relevant summary (Tait, 1983; Kupiec et al., 1995; Paice, 1990; Klavans and Shaw, 1995; McKeown et al., 1995; Shaw, 1995; Aone et al., 1997; Boguraev and Kennedy, 1997; Hovy and Lin, 1997; Mitra et al., 1997; Teufel and Moens, 1997; Barzilay and Elhadad, 1997; Carbonell and Goldstein, 1998; Baldwin and Morton, 1998; Radev and McKeown, 1998; Strzalkowski et al., 1998). However, large-scale IR and summarization have not yet been truly integrated, and the functionality challenges on a summarization system are greater in a true IR or topic-detection context (Yang et al., 1998; Allan et al., 1998).

Consider the situation where the user issues a search query, for instance on a news topic, and the retrieval system finds hundreds of closely-ranked documents in response. Many of these documents are likely to repeat much the same information, while differing in certain parts. Summaries of the individual documents would help, but are likely to be very similar to each other, *unless the summarization system takes into account other summaries that have already been generated.* Multi-document summarization – capable of summarizing either complete documents sets, or single documents in the context of previously summarized ones – are likely to be essential in such situations. Ideally, multi-document summaries should contain the key shared relevant information among all the documents *only once*, plus other information unique to some of the individual documents that are directly relevant to the user's query.

Though many of the same techniques used in single-document summarization can also be used in multi-document summarization, there are at least four significant differences:

1. The degree of redundancy in information contained a group of topically-related articles is much higher than the degree of redundancy within an article, as each article is apt to describe the main point as well as necessary shared background. Hence *anti-redundancy* methods are more crucial.
2. A group of articles may contain a *temporal dimension,* typical in a stream of news reports about an unfolding event. Here later information may override earlier more tentative or incomplete accounts.
3. The *compression ratio* (i.e. the size of the summary with respect to the size of the document set) will typically be much smaller for collections of dozens or hundreds of topically related documents than for single document summaries. The SUMMAC evaluation (TIPSTER, 1998a) tested 10% compression summaries, but in our work summarizing 200-document clusters, we find that compression to the 1% or 0.1% level is required. Summarization becomes significantly more difficult when compression demands increase.
4. The co-reference problem in summarization presents even greater challenges for multi-document than for single-document summarization (Baldwin and Morton, 1998).

This paper discusses an approach to multi-document summarization that builds on previous work in single-document summarization by using additional, available information about the document set as a whole, the relationships between the documents, as well as individual documents.

## 2 Background and Related Work

Generating an effective summary requires the summarizer to *select*, *evaluate*, *order* and *aggregate* items of information according to their relevance to a particular subject or purpose. These tasks can either be approximated by IR techniques or done in greater depth with fuller natural language processing. Most previous work in summarization has attempted to deal with the issues by focusing more on a related, but simpler, problem. With *text-span deletion* the system attempts to delete "less important" spans of text from the original document; the text that remains is deemed a summary. Work on automated document summarization by text span extraction dates back at least to Luhn's work at IBM in the fifties (Luhn, 1958). Most of the work in sentence extraction applied statistical techniques (frequency analysis, variance analysis, etc.) to linguistic units such as tokens, names, anaphora, etc. (e.g., (Tait, 1983; Boguraev and Kennedy, 1997; Aone et al., 1997; Carbonell and Goldstein, 1998; Hovy and Lin, 1997; Mitra et al., 1997; Baldwin and Morton, 1998; Paice, 1990)). Other approaches include the utility of discourse structure (Marcu, 1997), the combination of information extraction and language generation (Klavans and Shaw, 1995; McKeown et al., 1995; Shaw, 1995; Radev and McKeown, 1998; McKeown et al., 1999), and using machine learning to find patterns in text (Teufel and Moens, 1997; Barzilay and Elhadad, 1997; Strzalkowski et al., 1998).

Several researchers have extended various aspects of the single document approaches to look at multi-document summarization (Radev and McKeown, 1998; Mani and Bloedern, 1997; TIPSTER, 1998b). These include comparing templates filled in by extracting information – using specialized, domain specific knowledge sources – from the document, and then generating natural language summaries from the templates (Radev and McKeown, 1998), comparing named-entities – extracted using specialized lists – between documents and selecting the most relevant section (TIPSTER, 1998b), finding co-reference chains in the document set to identify common sections of interest (TIPSTER, 1998b), or building activation networks of related lexical items (identity mappings, synonyms, hypernyms, etc.) to extract text spans from the document set (Mani and Bloedern, 1997). Our approach (Goldstein and Carbonell, 1998) differs from these in several ways (1) it attempts to use only domain-independent techniques, based mainly on fast, statistical processing, (2) it explicitly deals with the issue of reducing redundancy, and (3) most of the modules are parameterized, so that different genres or corpora characteristics can be taken into account easily.

## 3 Requirements for Multi-Document Summarization

There are two types of situations in which multi-document summarization would be useful: (1) the user is faced with a collection of dis-similar documents and wishes to assess the information landscape contained in the collection, or (2) there is a collection of topically-related documents, extracted from a larger more diverse collection as the result of a query, or a topically-cohesive cluster. In the first case, if the collection is large enough, it only makes sense to first cluster and categorize the documents (Yang et al., submitted), and then sample from, or summarize each cohesive cluster. Hence, a "summary" would constitute of a visualization of the information landscape, where features could be clusters or summaries thereof. In the second case, it is possible to build a synthetic textual summary containing the main point(s) of the topic, augmented with non-redundant background information and/or query-relevant elaborations. This is the focus of our work reported here, including the necessity to *eliminate redundancy* among the information content of multiple related documents.

Users' information seeking needs and goals vary tremendously. When a group of three people created a multi-document summarization of 10 articles about the Microsoft Trial from a given day, one summary focused on the details presented in court, one on an overall gist of the day's events, and the third on a high level view of the goals and outcome of the trial. Thus, an ideal multi-document summarization would be able to address the different levels of *detail*, which is difficult without natural language understanding. An interface for the summarization system needs to be able to permit the user to enter information seeking goals, via a query, a background interest profile and/or a relevance feedback mechanism.

Following is a list of requirements for multi-document summarization:

- *clustering*: The ability to cluster similar documents and passages to find related information.
- *coverage*: The ability to find and extract the main points across documents.
- *anti-redundancy*: The ability to minimize redundancy between passages in the summary.
- *summary cohesion criteria*: The ability to combine text passages in a useful manner for the reader. This may include:
  - *document ordering*: All text segments of highest ranking document, then all segments from the next highest ranking document, etc.
  - *news-story principle (rank ordering)*: present the most relevant and diverse information first so that the reader gets the maximal information content even if they stop reading the summary.

- *topic-cohesion*: Group together the passages by topic clustering using passage similarity criteria and present the information by the cluster centroid passage rank.
- *time line ordering*: Text passages ordered based on the occurrence of events in time.

- *coherence*: Summaries generated should be readable and relevant to the user.
- *context*: Include sufficient context so that the summary is understandable to the reader.
- *identification of source inconsistencies*: Articles often have errors (such as billion reported as million, etc.); multi-document summarization must be able to recognize and report source inconsistencies.
- *summary updates*: A new multi-document summary must take into account previous summaries in generating new summaries. In such cases, the system needs to be able to track and categorize events.
- effective user interfaces:
  - *Attributability*: The user needs to be able to easily access the source of a given passage. This could be the single document summary.
  - *Relationship*: The user needs to view related passages to the text passage shown, which can highlight source inconsistencies.
  - *Source Selection*: The user needs to be able to select or eliminate various sources. For example, the user may want to eliminate information from some less reliable foreign news reporting sources.
  - *Context*: The user needs to be able to zoom in on the context surrounding the chosen passages.
  - *Redirection*: The user should be able to highlight certain parts of the synthetic summary and give a command to the system indicating that these parts are to be weighted heavily and that other parts are to be given a lesser weight.

## 4 Types of Multi-Document Summarizers

In the previous section we discussed the requirements for a multi-document summarization system. Depending on a user's information seeking goals, the user may want to create summaries that contain primarily the common portions of the documents (their intersection) or an overview of the entire group of documents (a sampling of the space that the document span). A user may also want to have a highly readable summary, an overview of pointers (sentences or word lists) to further information, or a combination of the two. Following is a list of various methods of creating multi-document summaries by extraction:

1. *Summary from Common Sections of Documents*: Find the important relevant parts that the group of documents have in common (their intersection) and use that as a summary.
2. *Summary from Common Sections and Unique Sections of Documents*: Find the important relevant parts that the group of documents have in common and the relevant parts that are unique and use that as a summary.
3. *Centroid Document Summary*: Create a single document summary from the centroid document in the group.
4. *Centroid Document plus Outliers Summary*: Create a single document summary from the centroid document in the group and add some representation from outlier documents (passages or keyword extraction) to provide a fuller coverage of the document set. [1].
5. *Latest Document plus Outliers Summary*: Create a single document summary from the latest time stamped document in the group (most recent information) and add some representation of outlier documents to provide a fuller coverage of the document collection.
6. *Summary from Common Sections and Unique Sections of Documents with Time Weighting Factor*: Find the important relevant parts that the group of documents have in common and the relevant parts that are unique and weight all the information by the time sequence of the documents in which they appear and use the result as a summary. This allows the more recent, often updated information to be more likely to be included in the summary.

There are also much more complicated types of summary extracts which involve natural language processing and/or understanding. These types of summaries include: (1) differing points of view within the document collection, (2) updates of information within the document collection, (3) updates of information from the document collection with respect to an already provided summary, (4) the development of an event or subtopic of an event (e.g., death tolls) over time, and (5) a comparative development of an event.

Naturally, an ideal multi-document summary would include natural language generation to create cohesive readable summaries (Radev and McKeown, 1998; McKeown et al., 1999). Our focus is on fast, domain independent summaries, which is currently beyond the scope of natural language processing techniques.

## 5 System Design

In the previous sections we discussed the requirements and types of multi-document summarization systems. This section discusses our current implementation of

[1]This is similar to the approach of Textwise, whose multi-document summary consists of the most relevant paragraph and specialized word lists (TIPSTER, 1998a)

$$MMR\text{-}MD \stackrel{\text{def}}{=} Arg \max_{P_{ij} \in R \setminus S} \left[ \lambda(Sim_1(P_{ij}, Q, C_{ij})) - (1-\lambda) \max_{P_{nm} \in S} Sim_2(P_{ij}, P_{nm}, C, S)) \right]$$

$$Sim_1(P_{ij}, Q, C_{ij}, D_i, D) = w_1 * (P_{ij} \cdot Q) + w_2 * coverage(P_{ij}, C_{ij}) + w_3 * content(P_{ij}) + time\_sequence(D_i, D)$$

$$Sim_2(P_{ij}, P_{nm}, C, S, D_i) = w_a * (P_{ij} \cdot P_{nm}) + w_b * clusters\_selected(C_{ij}, S) + w_c * documents\_selected(D_i, S)$$

$$coverage(P_{ij}, C) = \sum_{k \in C_{ij}} w_k * |k|$$

$$content(P_{ij}) = \sum_{W \in P_{ij}} w_{type}(W)$$

$$time\_sequence(D_i, D) = \frac{timestamp(D_{maxtime}) - timestamp(D_i)}{timestamp(D_{maxtime}) - timestamp(D_{mintime})}$$

$$clusters\_selected(C_{ij}, S) = |C_{ij} \cap \bigcup_{v,w : P_{vw} \in S} C_{vw}|$$

$$documents\_selected(D_i, S) = \frac{1}{|D_i|} * \sum_w [P_{iw} \in S]$$

where
$Sim_1$ is the similarity metric for relevance ranking
$Sim_2$ is the anti-redundancy metric
D is a document collection
P is the passages from the documents in that collection (e.g., $P_{ij}$ is passage j from document $D_i$)
Q is a query or user profile; R = IR(D, P, Q, $\theta$), i.e., the ranked list of passages from documents retrieved by an IR system, given D, P, Q and a relevance threshold $\theta$, below which it will not retrieve passages ($\theta$ can be degree of match or number of passages)
S is the subset of passages in R already selected
$R \setminus S$ is the set difference, i.e, the set of as yet unselected passages in R
C is the set of passage clusters for the set of documents
$C_{vw}$ is the subset of clusters of C that contains passage $P_{vw}$
$C_v$ is the subset of clusters that contain passages from document $D_v$
$|k|$ is the number of passages in the individual cluster k $|C_{vw} \cap C_{ij}|$ is the number of clusters in the intersection of $C_{vw} and C_{ij}$ $w_i$ are weights for the terms, which can be optimized
W is a word in the passage $P_{ij}$
type is a particular type of word, e.g., city name
$|D_i|$ is the length of document $i$.

Figure 1: Definition of multi-document summarization algorithm - MMR-MD

a multi-document summarization system which is designed to produce summaries that emphasize "*relevant novelty*." Relevant novelty is a metric for minimizing redundancy and maximizing both relevance and diversity. A first approximation to measuring relevant novelty is to measure relevance and novelty independently and provide a linear combination as the metric. We call this linear combination "marginal relevance" – i.e., a text passage has high marginal relevance if it is both relevant to the query and useful for a summary, while having minimal similarity to previously selected passages. Using this metric one can maximize marginal relevance in retrieval and summarization, hence we label our method "maximal marginal relevance" (MMR) (Carbonell and Goldstein, 1998).

The Maximal Marginal Relevance Multi-Document (MMR-MD) metric is defined in Figure 1.

For MMR-MD we define $Sim_1$ and $Sim_2$ to cover some of the properties that we discussed in Section 3.[2]

For $Sim_1$, the first term is the cosine similarity metric for query and document. The second term computes a *coverage* score for the passage by whether the passage is in one or more clusters and the size of the cluster. The third term reflects the information content of the passage by taking into account both statistical and linguistic features for summary inclusion (such as query expansion, position of the passage in the document and presence/absence of named-entities in the passage). The final term indicates the temporal sequence of the document in the collection allowing for more recent information to have higher weights.

For $Sim_2$, the first term uses the cosine similarity metric to compute the similarity between the passage and previously selected passages. (This helps the system to minimize the possibility of including passages similar to ones already selected.) The second term penalizes passages that are part of clusters from which other passages have already been chosen. The third term penalizes documents from which passages have already been selected; however, the penalty is inversely proportional to document length, to allow the possibility of longer documents contributing more passages. These latter two terms allow for a fuller *coverage* of the clusters and documents.

Given the above definition, MMR-MD incrementally computes the standard relevance-ranked list – plus some additional scoring factors – when the parameter $\lambda$=1, and computes a maximal diversity ranking among the passages in the documents when $\lambda$=0. For intermediate values of $\lambda$ in the interval [0,1], a linear combination of both criteria is optimized. In order to sample the information space in the general vicinity of the query, small values of $\lambda$ can be used; to focus on multiple, potentially overlapping or reinforcing relevant passages, $\lambda$ can be set to a value closer to 1. We found that a particularly effective search strategy for document retrieval is to start with a small $\lambda$ (e.g., $\lambda$ = .3) in order to understand the information space in the region of the query, and then to focus on the most important parts using a reformulated query (possibly via relevance feedback) and a larger value of $\lambda$ (e.g., $\lambda$ = .7) (Carbonell and Goldstein, 1998).

Our multi-document summarizer works as follows:

- Segment the documents into passages, and index them using inverted indices (as used by the IR engine). Passages may be phrases, sentences, n-sentence chunks, or paragraphs.
- Identify the passages relevant to the query using cosine similarity with a threshold below which the passages are discarded.
- Apply the MMR-MD metric as defined above. Depending on the desired length of the summary, select a number of passages to compute passage *redundancy* using the cosine similarity metric and use the passage similarity scoring as a method of clustering passages. Users can select the number of passages or the amount of compression.
- Reassemble the selected passages into a summary document using one of the summary-cohesion criteria (see Section 3).

The results reported in this paper are based on the use of the SMART search engine (Buckley, 1985) to compute cosine similarities (with a SMART weighting of `lnn` for both queries and passages), stopwords eliminated from the indexed data and stemming turned on.

## 6 Discussion

The TIPSTER evaluation corpus provided several sets of topical clusters to which we applied MMR-MD summarization. As an example, consider a set of 200 apartheid-related news-wire documents from the Associated Press and the Wall Street Journal, spanning the period from 1988 to 1992. We used the TIPSTER provided topic description as the query. These 200 documents were on an average 31 sentences in length, with a total of 6115 sentences. We used the sentence as our summary unit. Generating a summary 10 sentences long resulted in a sentence compression ratio of 0.2% and a character compression of .3%, approximately two orders of magnitude different with compression ratios used in single document summarization. The results of summarizing this document set with a value of $\lambda$ set to 1 (effectively query relevance, but no MMR-MD) and $\lambda$ set to 0.3 (both query relevance and MMR-MD anti-redundancy) are shown in Figures 2 and 3 respectively. The summary in Figure 2 clearly illustrates the need for reducing redundancy and maximizing novel information.

Consider for instance, the summary shown in Figure 2. The fact that the ANC is fighting to overthrow the government is mentioned seven times (sentences #2,–#4,#6–#9), which constitutes 70% of the sentences in the summary. Furthermore, sentence #3 is an exact duplicate of

[2] $Sim_1$ and $Sim_2$ as previously defined in MMR for single-document summarization contained only the first term of each equation.

1. **WSJ910204-0176: 1** CAPE TOWN, South Africa – President F.W. de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.
2. **AP880803-0082: 25** Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
3. **AP880803-0080: 25** Three Canadian anti-apartheid groups issued a statement urging the government to sever diplomatic and economic links with South Africa and aid the African National Congress, the banned group fighting the white-dominated government in South Africa.
4. **AP880802-0165: 23** South Africa says the ANC, the main black group fighting to overthrow South Africa's white government, has seven major military bases in Angola, and the Pretoria government wants those bases closed down.
5. **AP880212-0060: 14** ANGOP quoted the Angolan statement as saying the main causes of confict in the region are South Africa's "illegal occupation" of Namibia, South African attacks against its black-ruled neighbors and its alleged creation of armed groups to carry out "terrorist activities" in those countries, and the denial of political rights to the black majority in South Africa.
6. **AP880823-0069: 17** The ANC is the main guerrilla group fighting to overthrow the South African government and end apartheid, the system of racial segregation in which South Africa's black majority has no vote in national affairs.
7. **AP880803-0158: 26** South Africa says the ANC, the main black group fighting to overthrow South Africa's white-led government, has seven major military bases in Angola, and it wants those bases closed down.
8. **AP880613-0126: 15** The ANC is fighting to topple the South African government and its policy of apartheid, under which the nation's 26 million blacks have no voice in national affairs and the 5 million whites control the economy and dominate government.
9. **AP880212-0060: 13** The African National Congress is the main rebel movement fighting South Africa's white-led government and SWAPO is a black guerrilla group fighting for independence for Namibia, which is administered by South Africa.
10. **WSJ870129-0051: 1** Secretary of State George Shultz, in a meeting with Oliver Tambo, head of the African National Congress, voiced concerns about Soviet influence on the black South African group and the ANC's use of violence in the struggle against apartheid.

Figure 2: Sample multi-document summary with $\lambda = 1$, news-story-principle ordering (rank order)

sentence #2, and sentence #7 is almost identical to sentence #4. In contrast, the summary in Figure 3, generated using MMR-MD with a value of $\lambda$ set to 0.3 shows significant improvements in eliminating redundancy. The fact that the ANC is fighting to overthrow the government is mentioned only twice (sentences #3,#7), and one of these sentences has additional information in it. The new summary retained only three of the sentences from the earlier summary.

Counting clearly distinct propositions in both cases, yields a 60% greater information content for the MMR-MD case, though both summaries are equivalent in length.

When these 200 documents were added to a set of 4 other topics of 200 documents, yielding a document-set with 1000 documents, the query relevant multi-document summarization system produced exactly the same results.

We are currently working on constructing datasets for experimental evaluations of multi-document summarization. In order to construct these data sets, we attempted to categorize user's information seeking goals for multi-document summarization (see Section 3). As can be seen in Figure 2, the standard IR technique of using a query to extract relevant passages is no longer sufficient for multi-document summarization due to redundancy. In addition, query relevant extractions cannot capture temporal sequencing. The data sets will allow us to measure the effects of these, and other features, on multi-document summarization quality.

Specially, we are constructing sets of 10 documents, which either contain a snapshot of an event from multiple sources or the unfoldment of an event over time. From these sets we are performing two types of experiments. In the first, we are examining how users put sentences into pre-defined clusters and how they create

1. **WSJ870129-0051 1** Secretary of State George Shultz, in a meeting with Oliver Tambo, head of the African National Congress, voiced concerns about Soviet influence on the black South African group and the ANC's use of violence in the struggle against apartheid.
2. **WSJ880422-0133 44** (See related story: "ANC: Apartheid's Foes – The Long Struggle: The ANC Is Banned, But It Is in the Hearts of a Nation's Blacks — In South Africa, the Group Survives Assassinations, Government Crackdowns — The Black, Green and Gold" – WSJ April 22, 1988)
3. **AP880803-0158 26** South Africa says the ANC, the main black group fighting to overthrow South Africa's white-led government, has seven major military bases in Angola, and it wants those bases closed down.
4. **AP880919-0052 5** But activist clergymen from South Africa said the pontiff should have spoken out more forcefully against their white-minority government's policies of apartheid, under which 26 million blacks have no say in national affairs.
5. **AP890821-0092 10** Besides ending the emergency and lifting bans on anti- apartheid groups and individual activists, the Harare summit's conditions included the removal of all troops from South Africa's black townships, releasing all political prisoners and ending political trials and executions, and a government commitment to free political discussion.
6. **WSJ900503-0041 11** Pretoria and the ANC remain far apart on their vision s for a post-apartheid South Africa: The ANC wants a simple one-man, one-vote majority rule system, while the government claims that will lead to black domination and insists on constitutional protection of the rights of minorities, including the whites.
7. **WSJ900807-0037 1** JOHANNESBURG, South Africa – The African National Congress suspended its 30-year armed struggle against the white minority government, clearing the way for the start of negotiations over a new constitution based on black-white power sharing.
8. **WSJ900924-0119 20** The African National Congress, South Africa's main black liberation group, forged its sanctions strategy as a means of pressuring the government to abandon white-minority rule.
9. **WSJ910702-0053 36** At a meeting in South Africa this week, the African National Congress, the major black group, is expected to take a tough line again st the white-run government.
10. **WSJ910204-0176 1** CAPE TOWN, South Africa – President F.W. de Klerk's proposal to repeal the major pillars of apartheid drew a generally positive response from black leaders, but African National Congress leader Nelson Mandela called on the international community to continue economic sanctions against South Africa until the government takes further steps.

Figure 3: Sample multi-document summary with $\lambda$ = 0.3, time-line ordering

sentence based multi-document summaries. The result will also serve as a gold standard for system generated summaries - do our systems pick the same summary sentences as humans and are they picking sentences from the same clusters as humans? The second type of experiment is designed to determine how users perceive the output summary quality. In this experiment, users are asked to rate the output sentences from the summarizer as good, okay or bad. For the okay or bad sentences, they are asked to provide a summary sentence from the document set that is "better", i.e., that makes a better set of sentences to represent the information content of the document set. We are comparing our proposed summarizer #6 in Section 4 to summarizer #1, the common portions of the document sets with no anti-redundancy and summarizer #3, single document summary of a centroid document using our single document summarizer (Goldstein et al., 1999).

## 7 Conclusions and Future Work

This paper presented a statistical method of generating extraction based multi-document summaries. It builds upon previous work in single-document summarization and takes into account some of the major differences between single-document and multi-document summarization: (i) the need to carefully eliminate redundant information from multiple documents, and achieve high compression ratios, (ii) take into account information about document and passage similarities, and weight different passages accordingly, and (iii) take temporal information into account.

Our approach differs from others in several ways: it is completely domain-independent, is based mainly on fast, statistical processing, it attempts to maximize the novelty of the information being selected, and different genres or corpora characteristics can be taken into account easily. Since our system is not based on the use of sophisticated natural language understanding or information extraction techniques, summaries lack co-reference resolution, passages may be disjoint from one another, and in some cases may have false implicature.

In future work, we will integrate work on multi-document summarization with work on clustering to provide summaries for clusters produced by topic detection and tracking. We also plan to investigate how to generate coherent temporally based event summaries. We will also investigate how users can effectively use multi-document summarization through interactive interfaces to browse and explore large document sets.

## References


James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

Chinatsu Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. 1997. A scalable summarization system using robust NLP. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73, Madrid, Spain.

Breck Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, Granada, Spain, June.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain.

Branimir Boguraev and Chris Kennedy. 1997. Salience based content characterization of text documents. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 2–9, Madrid, Spain.

Chris Buckley. 1985. Implementation of the SMART information retrieval system. Technical Report TR 85-686, Cornell University.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR-98*, Melbourne, Australia, August.

Jade Goldstein and Jaime Carbonell. 1998. The use of mmr and diversity-based reranking in document reranking and summarization. In *Proceedings of the 14th Twente Workshop on Language Technology in Multimedia Information Retrieval*, pages 152–166, Enschede, the Netherlands, December.

Jade Goldstein, Mark Kantrowitz, Vibhu O. Mittal, and Jaime Carbonell. 1999. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of SIGIR-99*, Berkeley, CA, August.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–65, March.

E. Hovy and C. Y. Lin. 1997. Automated text summarization in SUMMARIST. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 18–24, Madrid, Spain, July.

J. L. Klavans and J. Shaw. 1995. Lexical semantics in summarization. In *Proceedings of the First Annual Workshop of the IFIP Working Group FOR NLP and KR*, Nantes, France, April.

Julian M. Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual Int. ACM/SIGIR Conference on Research and Development in IR*, pages 68–73, Seattle, WA, July.

P. H. Luhn. 1958. Automatic creation of literature abstracts. *IBM Journal*, pages 159–165.

Inderjeet Mani and Eric Bloedern. 1997. Multi-document summarization by graph search and merging. In *Proceedings of AAAI-97*, pages 622–628. AAAI.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain.

Kathleen R. McKeown, Jacques Robin, and Karen Kukich. 1995. Designing and evaluating a new revision-based model for summary generation. *Info. Proc. and Management*, 31(5).

Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards Multidocument Summarization by Reformulation: Progress and Prospects. In *Proceedings of AAAI-99*, pages 453–460, Orlando, FL, July.

M. Mitra, A. Singhal, and C. Buckley. 1997. Automatic text summarization by paragraph extraction. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 31–36, Madrid, Spain, July.

Chris D. Paice. 1990. Constructing literature abstracts by computer: Techniques and prospects. *Info. Proc. and Management*, 26:171–186.

Dragomir Radev and Kathy McKeown. 1998. Generating natural language summaries from multiple online sources. *Compuutational Linguistics*.

Gerald Salton. 1970. Automatic processing of foreign language docuemnts. *Journal of American Society for Information Sciences*, 21:187–194.

Gerald Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

James Shaw. 1995. Conciseness through aggregation in text generation. In *Proceedings of 33rd Association for Computational Linguistics*, pages 329–331.

Tomek Strzalkowski, Jin Wang, and Bowden Wise. 1998. A robust practical text summarization system. In *AAAI Intelligent Text Summarization Workshop*, pages 26–30, Stanford, CA, March.

J. I. Tait. 1983. *Automatic Summarizing of English Texts*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

S. Teufel and M. Moens. 1997. Sentence extraction as a classification task. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 58–65, Madrid, Spain, July.

TIPSTER. 1998a. Tipster text phase III 18-month workshop notes, May. Fairfax, VA.

TIPSTER. 1998b. Tipster text phase III 24-month workshop notes, October. Baltimore, MD.

C.J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

Yiming Yang, Thomas Pierce, and Jaime G. Carbonell. 1998. A study on retrospective and on-line event detection. In *Proceedings of the 21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 28–36.

Yiming Yang, Jaime G. Carbonell, Ralf Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. (submitted). Learning approaches to topic detection and tracking. *IEEE EXPERT, Special Issue on Applications of Intelligent Information Retrieval*.