Creating and Evaluating Multi-Document Sentence Extract Summaries

Jade Goldstein* Vibhu Mittal†‡ Jaime Carbonell* Jamie Callan*
jade@cs.cmu.edu mittal@cs.cmu.edu jgc@cs.cmu.edu callan@cs.cmu.edu

*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 passage extraction approaches to multi-document summarization that use available information about the document set as a whole and the relationships between the documents to build on single document summarization methodology. 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, as well as the user's goals in creating the summary. 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. We examined how humans create multi-document summaries as well as the characteristics of such summaries and use these summaries to evaluate the performance of various multi-document summarization algorithms.

1. INTRODUCTION

With the continuing rapid expansion of online information, it has become increasingly important to provide improved mechanisms to find and present textual information effectively. Conventional information retreival systems including modern search engines find and rank documents based on maximizing relevance to the user query [22, 5, 23], yet these systems still require users to read the documents to locate the relevant sections of text for their information seeking goals. 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 [29, 1].

‡Current address: Xerox PARC, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA. e-mail: vibhu@mittal.net

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 – is 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.

Given the enormous amounts of information that is accessible, good quality multi-document summaries are needed to "save" the user from the time consuming task of reading relevant documents. Multi-document summaries can be used for a variety of purposes including: (1) to locate the sections of text pertinent to a users' information seeking goals, which can be browsing or finding specific answers to questions, (2) to indicate the content of a document collection, and (3) to provide updates to "known" information (a particular summary or stored representation of what an user has previously seen).

Though many of the same techniques used in single-document summarization can also be used in multi-document summarization, there are at least five significant differences: (1) *anti-redundancy* methods are needed since the degree of redundancy as previously mentioned is significantly higher in a group of topically related articles that in an individual article as each article tends to describe the main point as well as necessary shared background, (2) The group of articles may contain a *temporal dimension*, typical in a stream of news reports about an unfolding event, in which case later information may override earlier incomplete reports, (3) the summary size required by the user will typically be much smaller for collections of topically related documents than for single documents requiring a lower *compression factor* (i.e. the size of the summary with respect to the size of the document set), thereby requiring a far more careful selection of passages, (4) the *co-reference* issue presents a greater challenge when entities and facts occur across documents than in a single-document situation, and (5) the user interface will need to address the users' information seeking goals by allowing rapid effective interaction with the summary such as for the purposes of viewing context of a passage within the summary, view related information to the summary passages including the original document and/or single document summaries, and create new related summaries.

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. The results of a study are reported in which the characteristics of human-generated multi-document summaries are examined and the summaries are applied as a "gold standard" for evaluating our multi-document summarization system.

2. 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 [12]. 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. [27, 19, 9, 18, 2]. Other approaches include the utility of discourse structure [14], the combination of information extraction and language generation [11, 17, 24, 21, 16], and using machine learning to find patterns in text [28, 4, 26].

Several researchers have extended various aspects of the single document approaches to look at multi-document summarization [13, 21, 3, 7, 15]. 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 [21], comparing named-entities – extracted using specialized lists – between documents and selecting the most relevant section [15], finding co-reference chains in the document set to identify common sections of interest [3], or building activation networks of related lexical items (identity mappings, synonyms, hypernyms, etc.) to extract text spans from the document set [13]. Recently, several approaches have focused [7, 25, 20] on using fast, statistical processing and dealing with the issues of redundancy.

Our approach incorporates the basic concept of the above statistical approaches - fast processing and anti-redundancy measures as well as operating on parameterized weighted modules to allow text extract summaries of various types depending on the users' information seeking goals.

3. MULTI-DOC SUMMARIZATION

Users' information seeking needs and goals vary tremendously. When people are asked to create multi-document summaries from a group of articles, the summaries vary significantly. People seem to apply various metrics including what information is most important and interesting to them (which is also based on their previous knowledge of the event) and provide various levels of detail on specific points. For example, 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. At a minimum, the 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 (which can contain references to the users "knowledge" through summaries or other mechanisms) and/or a relevance feedback mechanism.

Following are some requirements for a multi-document summarization system:

- *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 ordering the passages by rank, by date, etc.
- *quality*: Summaries generated should be readable and relevant as well as contain sufficient context so that the points are understandable to the reader.
- *identification of source inconsistencies*: Articles often have errors (such as billion reported as million, etc.) or differing information (such as closing prices of stock, number of deaths); 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*: Where the user can interact with the summary by accessing the sources of a passage, viewing related passages to the passage shown, eliminating sources of information from the summary, viewing context of passages in the summary, and create new summaries based on passages of the summary.

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

$$Sim_1(P_{ij}, Q, C_{ij}, D_i, D) = w_1 * (P_{ij} \cdot Q) + w_2 * coverage(P_{ij}, C_{ij}) + w_3 * content(P_{ij}) + w_4 * time\_sequence(D_i, D)$$

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

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

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

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

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

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

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, θ), i.e., the ranked list of passages from documents retrieved by an IR system, given D, P, Q and a relevance threshold θ, below which it will not retrieve passages (θ 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

4. TYPES OF 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

[1] This is similar to the approach of Textwise, which constructs multi-document summaries consisting of the most relevant paragraph and specialized word lists [15]

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 [21, 16]. 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 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) [6].

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 based on 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 λ=1, and computes a maximal diversity ranking among the passages in the documents when λ=0. For intermediate values of λ 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 λ can be used; to focus on multiple, potentially overlapping or reinforcing relevant passages, λ 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 λ (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 λ (e.g., $\lambda = .7$) [6].

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, as well as specify summary types (weighting parameters) to produce the types of summaries mentioned in Section 4.
- 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 [5] 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. REDUNDANCY EXAMPLE

To motivate the need for anti-redundancy measures, consider the following output from our summarizer not using anti-redundancy measures shown in: Figure 2 for a 10 document set spanning 3 days on the January 2000 Norway Rail crash. Sentences 1 and 2 are near duplicates, Sentences 4 and 5 are also near duplicates, Sentence 9 is contained in Sentence 10, Sentence 8 is contained in Sentence 6 and Sentence 3 contains similar information to that of Sentence 7. Thus nearly 50% of the information is "useless". In contrast, the summary in Figure 3, generated using MMR-MD with a value of λ set to 0.3 shows significant improvements in eliminating redundancy. The new summary retains only one sentence from the original summary although the majority of the information in the original summary is contained in the new summary.

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

1. **10 1** Norway's train drivers on Thursday began a boycott of a line where two trains crashed this week, killing at least 16 people, after a driver apparently passed a red stop signal.
2. **9 1** Norway's train drivers on Thursday began a boycott of a line where two trains crashed this week, killing about 20 people, after a driver apparently passed a red stop signal.
3. **5 1** ASTA, Norway (Reuters) - Norwegian rescuers on Wednesday recovered bodies from the burned-out wreck of two trains in which up to 33 people, including schoolchildren, were feared killed in a head-on collision.
4. **8 1** ASTA, Norway (Reuters) - Norwegian rail controllers tried to telephone two train drivers to tell them to halt before a head-on collision that killed 20 to 30 people but had a wrong list of numbers, a television report said Wednesday.
5. **6 1** ASTA, Norway (Reuters) - Norwegian rail controllers tried to telephone two train drivers to tell them to halt before a collision that killed up to 33 people but had an incorrect list of numbers, a television report said Wednesday.
6. **3 6** If the death toll is as high as feared it will pass Norway's most recent comparable crash, when 27 people died further north on the same line in 1975, and be worse than Europe's last large rail accident, in which 31 people died near London's Paddington station in October.
7. **4 1** ASTA, Norway (Reuters) - Children on a shopping trip on the last day of the Christmas holiday were feared to be among 33 people believed to have died in a head-on collision between two trains in Norway, police said on Wednesday.
8. **4 19** If the death toll is as high as feared it will be Norway's worst rail crash since 1975, when 27 people died in an accident further north on the same line.
9. **4 22** Officials said the line lacked some modern safety controls used on other lines in Norway, including a system to prevent trains from driving through red stop signs.
10. **3 16** Officials said it was too early to speculate on what went wrong but the line lacked some modern safety controls used on other lines in Norway, including a system to prevent trains from driving past red stop signs

Figure 2: Sample multi-document summary with $\lambda = 1$ (no anti-redundancy), rank order: Sentence Number, Document Number, Sentence Number in Document, Sentence

7. DATA SETS: PROPERTIES

An ideal multi-document summary must contain the relevant information to fulfill a user's information seeking goals, as well as eliminate irrelevant and redundant information. A first step in creating such summaries is to identify how well a multi-document text summarizer can extract what people perceive as key information and to evaluate types of data sets that reflect 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. Our constructed data sets will allow us to measure the effects of these, and other features, on multi-document summarization quality.

Specifically, we constructed a database of human generated multi-document sentence extract summaries as well as assessor-marked subtopics for each sentence in each article. This database consists of 25 sets of 10 newswire articles from news sources taken primarily from Yahoo categories. The sets reflect four types of article clusters, (1) a snapshot of an event from multiple sources (e.g. the first report of an airline crash), (2) a snapshot from the same source (the first 10 articles from the same source on the airline crash spanning possibly a few days), (3) the unfoldment of an event over time (e.g. updates on the airline crash spanning a few weeks or months) or (4) a similar event in multiple locations (e.g. the millennium flu bug).

Three assessors assigned sentences in the articles to provided subtopics (on average 16) for the events. They also selected the ten most informative sentences for the collection of ten articles and the ten most informative sentences for a specific query (information seeking goal) for the articles. They also selected the three most representative articles for the entire set of articles.

In general, the "flavor" of the multi-document summary depended on the type analyzed. A snapshot has many redundant sentences and would generally have fewer lead sentences, and possibly more consecutive sentences.

For the most important topic of the provided topics, assessors had 56% agreement and of the three most important topics of the topic set, assessors had 62% agreement. For the three most representative articles of the document set, assessors had 42% agreement on the most representative article and 67% agreement on the articles selected as the three most representative articles. All articles were presented in their ordered time sequence of article appearance (although assessors were allowed to work with them in any order) and the majority of articles selected as the most representative articles were in the latter half of the data sets, supporting our summarizer algorithm's use of an additional weighting for documents with a more recent time stamp (this does not appear to be the case for the sentences selected from the articles).

We also examined human generated multi-document summaries for three specific queries, in which there was no limit on the number of sentences extracted. We compared these *no limit* summaries (with a sentence average of 41) to the fixed 10 sentence summaries (see Table 1) as well as characteristics of single document summaries for the newswire genre from our previous work [8]. For multi-document 10 sentence summaries, the assessors used on average 1.3 first

1. 2 25 "I heard a terrible crash...(and) thought at first that we had collided with an elk," Jeanette Haug, 23, told Norway's NTB news agency.
2. **3 1** ASTA, Norway (Reuters) - Norwegian rescue workers will start the search on Wednesday through the burned-out wrecks of two trains in which up to 33 people are feared to have died in a head-on collision.
3. **3 13** Rescuers did not try to enter the trains after firefighters doused the blaze, fearing possible explosions and saying the charred carriages were still dangerously hot despite freezing temperatures outside.
4. **5 7** Flags flew at half mast at railway stations around Norway after what could be the nation's worst rail crash, surpassing an 1975 accident in which 27 died farther north on the same line.
5. **5 11** "We have seen more dead bodies inside the trains" beyond the seven known dead, Ove Osgjelten, police rescue chief, told Reuters at the site.
6. **6 21** Police say that 67 people of the 100 aboard the two trains survived the accident, some with severe burns, leaving 33 feared dead.
7. **8 30** At least one 12-year-old girl on a shopping trip was feared killed on the northbound train but local schools reported that several others feared missing were safe.
8. **9 8** Police say a total of 19 people have now been reported missing, giving a guide to the likely number of dead, but down from early estimates of up to 33 killed.
9. **10 1** Norway's train drivers on Thursday began a boycott of a line where two trains crashed this week, killing at least 16 people, after a driver apparently passed a red stop signal.
10. **10 28** One television report said the controllers in nearby Hamar saw a crash was imminent and tried to warn the drivers but had the wrong list of phone numbers.

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

sentences of the available 10, compared to the no-limit summaries in which an average of 2.5 were used.

Our collected data will serve as a gold standard for system generated summaries - do our systems pick similar summary sentences to humans and are they picking sentences from the same clusters as humans? The next section will address the answer to the first question by describing our evaluation method for comparing human generated summaries to the system generated ones.

8. EVALUATION

Sparc Jones & Galliers define two types of summary evaluations: (i) intrinsic, measuring a system's quality, and (ii) extrinsic, measuring a system's performance in a given task [10]. Automatically produced summaries by text extraction can often result in a reasonable summary. However, this summary may fall short of an *optimal* summary, i.e., a readable, useful, intelligible, appropriate length summary from which the information that the user is seeking can be extracted. Thus extrinsic evaluations are important for determining the ultimate utility of summaries.

Our current evaluation is intrinsic - we will evaluate how similar our summaries are to the "gold standard" described in Section 3. We compute the similarity between two sentences and instead of using this as a redundancy penalty as it is used in Maximal Marginal Relevance (see Section 5) and in Radev's Cross Sentence Information Subsumption (CSIS) [20], we use this to score the machine generated sentences with respect to the human generated ones.

We are currently using cosine similarity as our similarity metric. Our scoring algorithm functions as follows:

1. Calculate a score for each summarizer generated sentence with respect to each human generated sentence using cosine similarity.
2. Perform N passes (where N is the number of sentences in the output summary) through the system, one for each sentence in the output summary, removing the highest scoring sentence pair.
3. Compute a score for the summarizer generated summary by averaging the scores for the extracted sentence pairs.
4. Compute a final score for the summarizer generated summary by averaging over the number of human generated summaries.

We used this scoring method to score our summarizer against our human generated summaries for both query-relevant and overall document content (generic) summaries. For our summarizer, we used three types of output summaries, (1) concatenate all the documents and perform single document summarization with no anti-redundancy measures (one of our baselines), (2) create a 10 sentence single document summary from the highest ranking human selection centroid document, and (3) use MMR-MD with anti-redundancy measures. We also compared our results to the agreement among human judges. The results are shown in Table 2.

There was not much difference in the scores between the summarization methods, although upon examination of the individual summaries, there is clear evidence of redundant information (as shown in Figure 2 compared to Figure 3). We hypothesize that the scoring does not reflect this partly due to the fact that certain techniques are good at retrieving certain types of information, i.e., producing particular types

Property	Query-Relevant Multi-Doc	Generic Multi-Doc	No-Limit Multi-Doc	Human $\Rightarrow$ Extracted Single Doc
number of doc sets (docs)	15 (150)	15 (150)	3 (30)	(2250)
avg sent/doc set (doc)	293 (29)	293 (29)	342 (34)	(26)
Summary Features				
% of doc set (doc) length	3%	3%	12%	(20%)
incl. 1st sentence(s)	13%	14%	25%	69%
Summary Composition				
single sentences	84%	88%	48%	-
2 consecutive sents	13%	10%	26%	-
3 consecutive sents	2%	2%	16%	-
>= 4 consecutive sents	1%	0%	10%	-

Table 1: Summary Data Comparison

Type	Human Comparison	Centroid Document Single-Doc Summaries	MMR-MD Baseline $\lambda = 1.0$	MMR-MD $\lambda = 0.6$	MMR-MD $\lambda = 0.3$
query-relevant	0.36	0.29	0.28	0.29	0.28
generic	0.33	0.29	0.30	0.30	0.27

Table 2: Summarizer Type Results: Similarity Score

of summaries and we include a variety of data in our data collection, such as a single days events, updates on events as well as similar events in multiple locations. For example, in the case where one wants to retrieve a broad coverage of how the flu affected different geographic locations, one would tend to use a λ close to 1.0 because one would be less concerned with the redundant information such as over crowded hospitals. However, a λ close to 0.3 tends to eliminate the redundancy in a summary for a collection of articles on a day's event from different news sources. Another main reason for the lack of difference in summarizer performance] is that our similarity score is not sufficiently fine tuned to distinguish certain summary quality characteristics such as the level of redundancy or whether the selected summary sentences have covered the points in the summary. We will need to develop further our multi-document summarization scoring methods to truly distinguish summary quality.

9. CONCLUSIONS AND FUTURE WORK

This paper presented a statistical method of generating various types of extraction based multi-document summaries. Our system builds upon previous work in single-document summarization - taking into account some of the major issues arising in multi-document summarization: (i) the need to carefully eliminate redundant information from multiple documents, and achieve high compression ratios, (ii) information about document and passage similarities, and weighting different passages accordingly, and (iii) the importance of temporal information.

Our approach is mainly domain-independent and based on fast, statistical processing, maximizing the novelty of the information being selected, as well as allowing different genres or corpora characteristics to 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.

We have illustrated the importance in eliminating redundant information from multi-document summaries and shown that genre characteristics such as the importance of the lead sentence for newswire stories does not hold in the same manner for multi-document summaries. Furthermore, we have shown that our summarizer performance comes very close to the similarity between human assessors, indicating that perhaps it is generating reasonable summaries. We plan to develop improved measures for summary similarity and quality as well as test summary quality by specifically asking people to rate the chosen summary sentences.

In future work, we will integrate multi-document summarization with document 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. In addition, we will examine how to construct interactive interfaces so that users can effectively use multi-document summarization to browse and explore large document sets.

10. REFERENCES

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

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

June 1998.

[3] Breck Baldwin, Thomas S. Morton, and Amit Bagga. Overview of the University of Pennsylvania's Tipster Report. In *TIPSTER Text Phase III Proceedings October 96-October 98*, pages 151–162, 1999. Omnipress, Inc.

[4] Regina Barzilay and Michael Elhadad. 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, 1997.

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

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

[7] Jade Goldstein and Jaime Carbonell. Summarization: Using MMR for Diversity-Based Reranking and Evaluating Summaries. In *TIPSTER Text Phase III Proceedings October 96-October 98*, pages 181–196, 1999. Omnipress, Inc.

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

[9] Eduard Hovy and Chin-Yew Lin. Automated text summarization in SUMMARIST. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 18–24, Madrid, Spain, July 1997.

[10] Karen Sparc Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems: an Analysis and Review.* Springer, New York, 1996.

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

[12] Hans P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, pages 159–165, 1958.

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

[14] Daniel Marcu. 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, 1997.

[15] Mary McKenna and Elizabeth Liddy. Multiple and Single Document Summarization Using DR-LINK. In *TIPSTER Text Phase III Proceedings October 96-October 98*, pages 215–222, 1999. Omnipress, Inc.

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

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

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

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

[20] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *ANLP/NAACL 2000 Workshop*, pages 21–29, April 2000.

[21] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple online sources. *Compuutational Linguistics*, 24(3), 1998.

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

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

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

[25] Gee C. Stein, Tomek Strzalkowsi, and G. Bowden Wise. Summarizing multiple documents using text extraction and interactive clustering. In *Proceedings of the Conference Pacific Association for Computation Linguistics*, pages 200–208, August 1999.

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

[27] J. I. Tait. *Automatic Summarizing of English Texts.* PhD thesis, University of Cambridge, Cambridge, UK, 1983.

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

[29] Yiming Yang, Thomas Pierce, and Jaime G. Carbonell. 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, 1998.