Sub-event based multi-document summarization

Naomi Daniel, Dragomir Radev

University of Michigan
340 West Hall
Ann Arbor, MI 48109
(734) 763-2285

ndaniel@umich.edu, radev@umich.edu

1. **Paper ID:**
2. **Keywords:** Multiple-document summarization, inter-judge agreement, relative utility, manual summarization, automatic summarization, sub-event-based summarization, event-based summarization, topic based summarization.
3. **Contact Author:** Naomi Daniel
4. **Under consideration for other conferences (specify)?** No
5. **Abstract**

The production of accurate and complete multiple-document summaries is challenged by the complexity of judging the usefulness of information to the user. We experimented with two new methods for summary creation, comparing our results with well-known baselines and with MEAD. Our aim was to determine whether identifying sub-events in a news topic could help us capture essential information to produce better summaries. We used six methods to create multi-document summaries and then compared them to find which method was the most successful. In two experiments, we used human judges to determine the relative utility of sentences, as related to either a news topic or its sub-events. We then compared three summaries created from this data, with three automatically created summaries. We examined the use of inter-judge agreement and a relative utility metric that accounts for the complexity of determining sentence quality in relation to a topic. Though this paper focuses on creating manual summaries, we hope to use what we discovered to improve automatic multi-document summarization through MEAD.

Sub-event based multi-document summarization

Paper ID:

Abstract

The production of accurate and complete multiple-document summaries is challenged by the complexity of judging the usefulness of information to the user. Our aim is to determine whether identifying sub-events in a news topic could help us capture essential information to produce better summaries. We used six methods to create multi-document summaries and then compared them to find which method was the most successful. In two experiments, we used human judges to determine the relative utility of sentences, as related to either a news topic or its sub-events. We then compared summaries created from this data, with three automatically created summaries. We examine the use of inter-judge agreement and a relative utility metric that accounts for the complexity of determining sentence quality in relation to a topic. Though this paper focuses on creating manual summaries, we hope to use what we discovered in future work on automatic multi-document summarization.

1. Introduction

Multiple articles on a particular topic tend to contain redundant information, as well as information that is unique to each article. For instance, different news sources covering the same topic may take different angles, or new information may become available in a later report. So, while all the articles are related to the larger topic, each article may be associated with any of several sub-events. We wanted to find a way to capture the unique sub-event information that is characteristic in multiple-document coverage of a single topic. We predicted that breaking documents down to their sub-events and capturing those sentences in each sub-event with the highest utility would produce an accurate, thorough, and diverse multi-document summary.

To determine whether sub-event summaries would produce useful multiple-document summaries, we compared six methods of summarization to see which produces the best summaries. The methods included three automatic and three manual methods of producing summaries. We used inter-judge agreement and relative utility to capture and measure subtleties in determining sentence relevance. We created multi-document summaries using both a sub-event-based approach and a topic-based approach. Generally, we expected to find that the manual summaries performed better than the automatic summaries. Our intent was to gather preliminary information on the use of sub-events to improve automatic multi-document summarization, as well as to use topic-based manual summaries to improve the MEAD summarizer. These manual summarization techniques were studied with the aim of improving automatic multi-document summarization techniques.

2. Related Work

Much work has preceded and informed this paper. Allan et al's work on summarizing novelty recognizes that news topics consist of a series of events – what we call "sub-events," to distinguish the difference between a news topic and its sub-events. Their method uses an algorithm to identify "novel" sentences, rather than the use of human judges. This work differs in that it does not take relative utility into account, a concept which we feel has great bearing on what is a complex problem, instead considering sentences simply either "on-topic" or "off-topic" (Allan et al., 2001a, Allan et al., 2001b). Goldstein (1999) uses Maximal Marginal Relevance (MMR) to identify "novel" information to improve query answering results, as well as applying this method to multiple-document summarization. Success in the use of inter-judge agreement has led us to pursue the use of the current evaluation methods. However, this experiment differs from prior work in that we use judges to evaluate the relevance of sentences to sub-events, rather than to evaluate summaries (Radev et al., 2000).

3. Article Corpus

Our study involves two experiments carried out by different judges on one corpus of news articles. The article corpus was selected from a cluster of eleven articles used in a previous experiment, describing the 1991 plane crash of Gulf Air flight 072, from which we chose a corpus of five news articles, containing a total of 159 sentences. All the articles cover a single news event, the plane crash and its aftermath. The articles were gathered on the web from sources reporting on the event as it unfolded, and come from various news agencies, such as ABC News, Fox News, and the BBC. All the articles give some discussion of the events leading up to and following the crash, with particular articles focusing on areas of special interest, such as the toll on Egypt, from where many of the passengers had come. The article titles in Table 1, below, illustrate the range of sub-events that are covered under the crash topic.

Article ID	Source	Date	Headline
30	BBC	24-Aug-00	Bodies recovered from Gulf Air crash
41	Fox News	24-Aug-00	Egyptians Suffer Second Air Tragedy in a Year
81	USA Today	24-Aug-00	One American among 143 dead in crash
87	ABC News	25-Aug-00	Prayers for victims of Bahrain crash
97	Fox News	25-Aug-00	Did Pilot Error Cause Gulf Air Crash?

Table 1. Corpus article characteristics.

4. Experiment 1: Sub-Event Analysis

We completed two experiments to gather human data for our summarization work. The first experiment involved having human judges analyze the sentences in our corpus for degree of salience to a series of sub-events comprising the topic. The second experiment used different groups of human subjects to examine the relevance of sentences in the same corpus to the overall news topic as a whole. Here, we will discuss the sub-event experiment. We will touch on the second, topic-based, experiment later in the paper.

4.1 Description of Sub-Event User Study

The goal of this experiment was to study the effectiveness of breaking a news topic down into sub-events, in order to capture not simply salience, but also diversity (Goldstein, 1998).

The sub-events were chosen to cover all of the material in the reports, and represent the most significant aspects of the news topic. For the Gulf Air crash, the sub-events we identified were:

1. The plane takes off
2. Something goes wrong
3. The plane crashes
4. Rescue and recovery effort
5. Gulf Air releases information
6. Government agencies react
7. Friends, relatives and nations mourn
8. Black box(es) are searched for
9. Black box(es) are recovered
10. Black box(es) are sent for analysis

We instructed judges to rank the degree of sentence relevance to each sub-event. Figure 1 shows an example of the sub-event forms the judges completed. Judges were instructed to use a scale, such that a score of ten indicated that the sentence was critical to the sub-event, and a score of 0 indicated that the sentence was irrelevant. Thus, the judges processed five 159-word documents ten times, once pertaining to each sub-event. This experiment produced 1590 data points for each judge, which were analyzed according to the methods described in the next section.

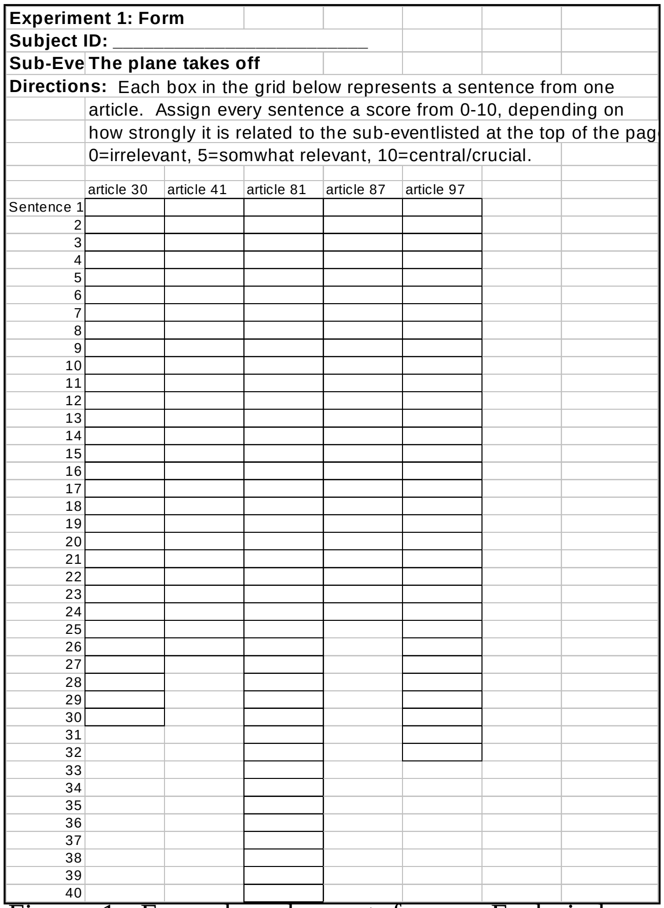
We used the data on the relevance of the sentences to the sub-events to calculate inter-judge agreement. In this manner, we determined which sentences had the overall highest relevance to each sub-event. We used this ranking to produce summaries at different levels of compression.

5. Methods for Producing Summaries

To gather data about the effectiveness of breaking news topics down into their sub-events for creating summaries, we utilized data from human judges, upon which we manually performed three algorithms. These algorithms and their application are described in detail below. However, in the future we anticipate using Topic Detection and Tracking technology (Allen, 1998) to group sentences by sub-event and apply these algorithms automatically.

	Sub-Event 1			Sub-Event 2			Sub-Event 3		
	Judge 1	Judge 2	Judge 3	Judge 1	Judge 2	Judge 3	Judge 1	Judge 2	Judge 3
Article 30, Sentence 1	1	0	0	5	0	5	8	8	10
2	1	0	0	7	4	7	**10**	**10**	**10**
3	4	0	0	**10**	**10**	**10**	10	5	7
4	1	0	3	5	0	2	8	0	2
5	**0**	**0**	**0**	3	0	0	5	0	2
6	**0**	**0**	**0**	3	0	0	6	0	2
7	**0**	**0**	**0**	3	0	0	6	0	2
8	**0**	**0**	**0**	3	4	2	**10**	**10**	**10**
9	0	0	2	**0**	**0**	**0**	8	0	0
10	**0**	**0**	**0**	3	0	0	6	0	2

Table 2. First ten sentences of article 30, shown with scores given by three judges for three sub-events. Judges often disagree on the degree of sentence relevance. Some sentences are used in more than one sub-event.



Experiment 1: Form

Subject ID: ___________________________

Sub-Eve The plane takes off

Directions: Each box in the grid below represents a sentence from one article. Assign every sentence a score from 0-10, depending on how strongly it is related to the sub-eventlisted at the top of the pag 0=irrelevant, 5=somwhat relevant, 10=central/crucial.

	article 30	article 41	article 81	article 87	article 97
Sentence 1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					

Figure 1: Example sub-event form - Each judge filled out ten such forms, one for each sub-event.

5.1 Sub-Event-Based Algorithms

Using the judges' data, we calculated the inter-judge agreement on sub-topic sentence relevance. We then ranked the sentences using three different algorithms, according to the utility scores for each sub-topic, to create multi-document summaries. From this data, we created summary extracts using three algorithms, as follows:

- Algorithm 1) Highest score anywhere - from highest score anywhere from any sub-event, inter-judges scores.
- Algorithm 2) Sum of all scores - from highest score in all sub-events combined, inter-judge scores.
- Algorithm 3) Sub-Event Round Robin - from highest score in each sub-event, inter-judges scores.

Algorithm 1 - Highest Score Anywhere (HSA): This algorithm was produced by summing the data across all judges to produce a total inter-judge score and keeping sub-events distinct, to see the inter-judge utility scores given to sub-events. We ordered the sentences by ranking these scores in descending order and omitting duplicates, to produce the ten and twenty percent extracts. For example, with data from seven judges on ten sub-events, the highest possible score for each sentence was seventy. Thus seventy was the highest score.

In the case that there was a tie between sentences, we ordered them by sub-event number (first sub-event first and tenth sub-event last).

Algorithm 2 - Sum of All Scores (SAS): This algorithm was produced by summing the data across all judges to produce a total inter-judge score, and combining events so that we could see the utility scores given across sub-events. We ordered the sentences by ranking these cross-event inter-judge utility scores in descending order and omitting duplicates, to produce the ten and twenty percent extracts.

Algorithm 3 – Sub-Event Round Robin (SRR): This algorithm was produced by summing the data across all judges to produce a total inter-judge score and keeping sub-events distinct, to see the inter-judge utility scores given to sub-events. We ordered the sentences by ranking the inter-judge utility scores in descending order within each sub-event. We then chose the top sentence from each sub-event (one through ten), the second highest sentence from each sub-event, and so on, omitting duplicates, until we had produced the ten and twenty percent extracts.

In this manner, we created thirty-six sub-event-based summary extracts – six clusters, three algorithms, two compression rates – which we then analyzed.

The Sum of All Scores algorithm most closely replicates a topic-based summary by combining the ten sub-event scores into one pan-topic score for each sentence. As our model extract was produced using human data on topic-based sentence relevance, the Sum of All Scores algorithm is the sub-event algorithm that most closely matched our model extract. In contrast, the Highest Score Anywhere algorithm maintains the structure of the sub-event breakdown, preferring the highest score in any sub-event. Likewise, the Round Robin algorithm maintains the sub-event breakdown, but rather than preferring the highest score in any event, it selects the highest score from each sub-event, serially; this algorithm most closely resembles the Lead-based automatic summarizer.

5.2 Automatic Multi-Doc Summaries

The three automatic summarization methods that we used in our comparison have already been established. We compared our manual summaries to these established automatic multiple-document summarization methods: Centroid-based (MEAD), Lead-based and Random.

MEAD: First, we produced summaries using the MEAD system. MEAD works by producing word cluster centroids, creating summaries by sentence extraction (Radev et al, 2002).

Lead-Based: We also produced summaries by the Lead-based, or “round robin,” method. This method involves assigning the highest score to the first sentence in each article, then the second sentence in each article, and so on.

Random: Our third automatic summarization method involved generating a summary by randomly selecting the sentences for inclusion.

6. Metric

We used Relative Utility as our metric. Relative utility (RU) is a metric by which sentence relevance can be measured. It allows us to distinguish the degree of importance between sentences, providing a more flexible model for evaluating sentence utility than precision, recall, or sentence overlap (Radev et al., 2000). Studies involving sentence extraction have often been predicated upon determining the usefulness of sentences as either useful or non-useful (Allan et al. 2001b). However, determining the usefulness of sentences is more complex than a simple relevant/irrelevant binary categorization can account for.

Another advantage of the relative utility metric is that, although human judges have a low level of agreement on which sentences belong in a summary, they tend to agree on how important sentences are to a topic or event; thus, relative utility makes it "possible to catch that agreement and produce better summarizers" (Radev et al., 2002)

We asked human subjects to assign a score to each sentence in a corpus of articles. The score reflects the subject's perception of a sentence's relevance to the topic it describes. The scale our judges were instructed to use ranged from zero to ten. A score of zero indicated that the sentence was irrelevant to the overall cluster, whereas a score of ten indicated that the sentence was crucial to the understanding of the cluster.

Relative utility is determined by applying a calculation for inter-judge agreement. Inter-judge agreement is determined by first calculating total utility by adding together the utility scores given to each sentence by each judge. For example, if a judge assigns a score of 8 to sentence 2 from a document and a 9 to sentence 5 from the document and these are the two highest-scoring sentences in the document, then no 2-sentence summary will be able to achieve a higher score. In the best case, a 2-sentence summary will get a relative utility of 1 (= 17/17) if it includes the same two sentences as the judge (or any other pair of sentences totaling the same number of points). If the summary includes sentences totaling 15 points, its relative utility will go down to 15/17. The relative utility method (Radev et al. 2000, Radev et al. 2002) extends this idea to include multiple reference judges and summaries of an arbitrary length.

Manual summaries, it is expected, will set the upper bound for the ranking of our summaries. We expect automatic summaries to fare below manual summaries, with the random automatic summaries setting the lower bound.

7. Extract Creation

Summaries can be created by abstracting or extracting [Mani, 2001]. For purposes of comparison with and improvement of MEAD, an extractive summarizer, we used an extractive method to create all six summary types: sum of all scores, highest score anywhere, round robin, MEAD, lead-based, and random.

7.1 Extracts

We used the data on sentence relevance from our judges to create summaries from the simple extraction of sentences by our judges, using the MEAD summarizer. The MEAD summarizer produces summaries in the form of sentence extracts (Radev et al, 2002). These extracts come out in the form illustrated in Figure 2.

The extract shows the sentence order of the finished summary, then the article and sentence number indicating from where the sentence was drawn. Notice that “S ORDER” refers to the rank of each sentence (1=best, 16=worst), “DID” refers to the document ID, and “SNO” refers to the sentence number. So, the first entry in Figure 2, “<S ORDER="1" DID="41" SNO="2" />,” can be read as ‘the best ranked sentence in this extract comes from document number 41, sentence 2.’

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE EXTRACT SYSTEM
"/clair/tools/mead/dtd/extract.dtd">

<EXTRACT QID="GA3N1" COMPRESSION="10"
SYSTEM="/clair/tools/mead/programs/hkmead.pl
Centroid 1 Position 1 Length 9" LANG="ENG">
<S ORDER="1" DID="41" SNO="2" />
<S ORDER="2" DID="41" SNO="26" />
<S ORDER="3" DID="81" SNO="2" />
<S ORDER="4" DID="81" SNO="7" />
<S ORDER="5" DID="81" SNO="17" />
<S ORDER="6" DID="81" SNO="27" />
<S ORDER="7" DID="81" SNO="44" />
<S ORDER="8" DID="87" SNO="2" />
<S ORDER="9" DID="87" SNO="3" />
<S ORDER="10" DID="87" SNO="5" />
<S ORDER="11" DID="87" SNO="10" />
<S ORDER="12" DID="87" SNO="16" />
<S ORDER="13" DID="87" SNO="20" />
<S ORDER="14" DID="87" SNO="22" />
<S ORDER="15" DID="87" SNO="24" />
<S ORDER="16" DID="97" SNO="3" />
```

Figure 2. Sample extract.

7.2 Clusters

Each of the summarization methods was employed at both ten and twenty percent compression rates. We used the summaries thus produced to consider how compression rates could influence the effectiveness of the six summarization methods. We additionally looked at varying combinations of the five articles, such that we examined the corpus in six clusters, as shown in Figure 3. We selected these article combinations to maximize the diversity of sources in each cluster, and to achieve a variable number of articles in a cluster.

8. Comparison

In order to evaluate a summary, it must be compared against an ideal summary. This ideal summary, or ground truth, is the model against which the system extracts are compared. The data gathered from this experiment was used to create a model summary against which to compare the sub-event summaries created in the first experiment. The topic-based data gathered in the second experiment set the standard against which the six other summarization creation methods were compared. Here, we discuss the second experiment, which focuses on the topic of the corpus as a whole.

Combination 1) articles 30 + 41 + 81 + 87 + 97

Combination 2) articles 30 + 41 + 81

Combination 3) articles 41 + 81 + 87

Combination 4) articles 81 + 87 + 97

Combination 5) articles 87 + 97 + 30

Combination 6) articles 97+ 30 + 41

Figure 3. Article clusters.

8.1 Description of Topic-Based User Study

The topic-based experiment used a different group of judges, who examined the relevance of sentences to the news topic as a whole (the Gulf Air crash), utilizing the same document corpus as was utilized in the first experiment. This experiment mirrored the first in its application of relative utility scores. However, unlike the first experiment, these judges were asked to consider the topic covered by the articles as a whole, rather than as sub-events. The relative utility rankings assigned to sentences by the judges reflect this whole-topic orientation.

We expect human judges to produce summaries that are superior to automatic summaries. Topic-based summary creation is not a unique task (as is sub-event summarization), therefore, we used the summaries created from this data as a metric against which to measure our sub-event-based summaries.

8.2 Upper and Lower Bounds

Determining upper and lower bounds enables us to conclude where each method ranks in relation to the other. Random summaries were used as a baseline – creating a lower bound for our summaries. We assumed that our manual summaries should be the upper bound, more successful than the automatic summarizers as a consequence of reflecting human judgement in sentence selection. MEAD summaries should fall in the middle.

9. Results

Some of our results met our expectations, while others surprised us (see Table 3). The Sum of All Scores manual algorithm produces the best summaries at the twenty percent compression rate. At the ten percent compression rate, data shows Lead-based summaries performing best, with the Sum of All Scores algorithm following right behind. MEAD scores in the mid-range as expected, for both compression rates, just behind the Sub-Event Round Robin Algorithm. In contrast, the random method leads in low scores, with the Highest Score Anywhere algorithm scoring only slightly higher. Random sets the lower bound. Here, we discuss the details of our findings and their significance in more detail.

9.1 Manual Algorithms

Compared to MEAD, both the Sum of All Scores and Sub-Event Round Robin algorithms performed better, while the Highest Score Anywhere algorithm performed less well. This result is reasonable, based upon the characteristics of the algorithms. Algorithm 2 (SAS), the best performer among the manual summaries, used the sum of all scores across events and judges; thus, it tapped into which sentences were most popular overall. Algorithm 3 (SRR), also better than MEAD, used a round robin technique, which, similarly to the Lead-based results, tapped into the pyramid quality of news journalism. Algorithm 1 (HSA), poorest performer of the manual summaries, used the highest score in any event by inter-judge score; its weakness was in negating both the benefits of the pyramid structure of news journalism, reflected in the judges' sentence rankings (captured by SRR), as well as the popularity of sentences across events (captured by SAS).

9.2 Compression Rate

For extracts at the ten percent compression rate, Lead-based sets the upper, and random the lower, bound. However, the Sum of All Scores algorithm performed better at the twenty percent compression rate, beating Lead-based for best summaries. Each method produced better summaries overall at ten percent compression rate, except for Algorithm 2, which performed better at the twenty percent compression rate.

	10%						20%					
	HSA	**SAS**	**SRR**	**MEAD**	**Lead**	**Rand**	**HSA**	**SAS**	**SRR**	**MEAD**	**Lead**	**Rand**
Cluster 1	0.641	0.686	0.717	0.617	**0.795**	0.613	0.542	**0.745**	0.683	0.621	0.722	0.527
Cluster 2	0.629	0.739	0.716	0.629	**0.8**	0.425	0.637	**0.786**	0.659	0.623	0.741	0.465
Cluster 3	0.568	0.698	0.544	0.672	**0.701**	0.339	0.572	**0.735**	0.631	0.647	0.629	0.483
Cluster 4	0.406	0.669	0.651	0.662	**0.714**	0.533	0.539	0.722	0.596	0.653	**0.738**	0.537
Cluster 5	0.646	0.675	0.698	0.604	**0.797**	0.538	0.598	0.739	0.733	0.631	**0.749**	0.529
Cluster 6	0.622	0.698	0.693	0.595	**0.88**	0.467	0.623	0.762	0.717	0.552	**0.817**	0.555
Average=	0.5853	0.6942	0.6698	0.6298	**0.7812**	0.4858	0.5852	**0.7482**	0.6698	0.6212	0.7327	0.516

Table 3. Results: Best performing algorithm at each cluster/compression rate shown in bold. Relative utility is 0 when there is no match between the model and the system extract; it is 1 when the two are identical.

9.3 Lead-Based Summaries

We were surprised to find Lead-based summaries producing better summaries than MEAD, as earlier work in this area stresses results showing that "intelligent" summarizers, such as MEAD, are expected to produce higher quality summaries than the simplistic algorithm demonstrated by Lead-based sentence extraction (Radev et al, 2002). This result may be explained by the pyramid structure of news journalism, which essentially pre-ranks document sentences in order of importance, in order to convey the most critical information first. As our corpus was comprised entirely of news articles, this effect could be exaggerated in our results. As expected, the Random summarizer set the lower bound.

9.4 Manual Summaries and MEAD

Most significantly, among the mid-range performers, the data demonstrates what we expected to find: Two of the three new sub-event-based algorithms (SRR and SAS) perform better than MEAD. Using human subjects and inter-judge agreement metrics allowed us to illustrate the effectiveness of conceptualizing document summarization as a complex task. Complexities in determining sentence relevance revolve around the user's information need, changes in report coverage over time, and the need for timely, accurate, and succinct information. Identifying sub-events in news topic coverage is one method that we have shown can be utilized to help create better summaries.

9.5 Sample Summaries

The plane's flight recorder has been recovered, said a government official early on Thursday, and a search was continuing for the cockpit voice recorder. At the same news conference, Gulf Air CEO Sheikh Ahmed bin Saif al-Nahayan said the plane's black boxes, recovered from shallow waters, had not been opened yet. Divers will begin a search for the jet's cockpit voice and data recorders at first light, Bahraini civil defence chief Brigadier Abdul-Rahman bin Rashid Al-Khalifa said. "It will be sent for examination and interpretation in Europe or America." "The black box flight data recorder cannot be opened in Bahrain, where it is currently under guard with civil aviation affairs," a Gulf Air statement read. The probe into why Gulf Air Flight GF072 crashed into the sea off Bahrain Wednesday turned to possible pilot error Friday, despite the airline's insistence that the pilot was not at fault. More than 130 bodies are reported to have been recovered after a Gulf Air jet carrying 143 people crashed into the Gulf off Bahrain on Wednesday. Bahraini Information Ministry official Said Al-Bably said one of its engines had caught fire. He also said U.S. investigators were on their way to the crash site and the investigation was expected to start immediately. Airline officials said later 137 bodies - including up to 30 children - had been pulled from the water. The plane was reportedly moving too fast and was not at the right altitude to land. Gulf Air officials said there was no fire and other witnesses have said they did not see flames. "Up till now we have not found any survivors," said Abdul-Rahman bin Rashed al-Khalifa, administration director of Bahrain's Civil Defence. The Bahraini authorities launched a major rescue operation, helped by the US Navy's Fifth Fleet, based in Bahrain. Sixty-three Egyptians were on board the Airbus A320, which crashed into shallow Persian Gulf waters Wednesday night after circling and trying to land in Bahrain. Bahrain's state television quoted witnesses of the crash who described seeing a fire in one of the aircraft's engines.

Figure 4. Sum of All Scores Summary, Cluster 6, 10%

Cairo airport officials said the plane left the Egyptian capital at 1625 local time (1325 GMT). Bahraini Information Ministry official Said Al-Bably said one of its engines had caught fire. The Airbus A320 - flight GF072 - crashed shortly before coming into land in Bahrain after a three-long flight from Cairo. Airline officials said later 137 bodies - including up to 30 children – had been pulled from the water. Flight 072 crashed in shallow water near shore and Ali Ahmedi, a spokesman and an acting vice president for Gulf Air, has said the pilot gave no indication to air traffic controllers that there were any problems in the plane. The Emir of Bahrain, Sheikh Hamad bin Issa al-Khalifa, has announced that a commission will be set up to establish what brought the plane down. Weeping relatives of passengers meanwhile pleaded with policemen ringing the airport outside the capital Manama. Divers will begin a search for the jet's cockpit voice and data recorders at first light, Bahraini civil defence chief Brigadier Abdul-Rahman bin Rashid Al-Khalifa said. Both of the plane"s "black boxes" _ the flight data and voice cockpit recorders _ were to be shipped abroad for data recovery but aviation experts had not finalized plans on Friday, Gulf Air said. It came down a little under three hours later. "It U-turned and tried to land, then in 15 seconds it went sharply down into the sea and there was a huge fire," he said. The plane crashed in shallow waters about five kilometres (three miles) from Bahrain airport. The Bahraini coastguard and marine were joined in the rescue effort by three US Navy helicopters, two US destroyers, small boats and an ocean-going tug with a crane. At the same news conference, Gulf Air CEO Sheikh Ahmed bin Saif al-Nahayan said the plane's black boxes, recovered from shallow waters, had not been opened yet. The Bahraini authorities launched a major rescue operation, helped by the US Navy's Fifth Fleet, based in Bahrain. Distraught relatives also gathered at Cairo airport, demanding information. The plane's flight recorder has been recovered, said a government official early on Thursday, and a search was continuing for the cockpit voice recorder.

Figure 5. Sub-Event Round Robin Summary, Cluster 5, 10%.

CAIRO, Egypt -- The crash of a Gulf Air flight that killed 143 people in Bahrain is a disturbing deja vu for Egyptians: It is the second plane crash within a year to devastate this Arab country. Egypt, which lacks the oil wealth of the Gulf and has an economy struggling to revive from decades of socialist stagnation, has a long tradition of sending workers to the Gulf to fill everything from skilled to menial jobs. MINA SALMAN PORT, Bahrain (AP) -- A man's black shoe, a plastic sandal and bits of yellow foam padding bobbed Thursday in the waters off this tiny island nation, where families were burying loved ones a day after Gulf Air Flight 072 crashed, killing all 143 aboard. Ali Ahmedi, a spokesman and an acting vice president for Gulf Air, said it was too early to speculate on what caused the plane to crash as it circled the airport before coming in to land. A few recognizable pieces of the Gulf Air Airbus 320 protruded from the water: a ripped tail wing with the airline's black, red and gold logo, skin of the fuselage with the letters 'LF AIR' above the surface. They included 64 Egyptians, 36 Bahrainis, 12 Saudi Arabians, nine Palestinians, six from the United Arab Emirates, three Chinese, two British and one each from the United States, Canada, Oman, Kuwait, Sudan, Australia, Oman, the Philippines, Poland, India and Morocco. MANAMA, Bahrain (AP) _ Three bodies wrapped in cloth, one the size of a small child, were lain before the faithful in the Grand Mosque Friday during a specialprayer for the dead in honor of the 143 victims of the Gulf Air crash. Bahrain"s Prime Minister Sheik Khalifa bin Salman Al Khalifa and other top officials stood side-by-side with 2,000 Muslims reciting funeral prayers before the bodies, which were among the 107 adults and 36 children killed in Wednesday"s air disaster, said Information Ministry spokesman Syed el-Bably. At dawn Friday, the divers began searching for "diplomatic cargo" being carried by a U.S. government courier, according to Cdr. Jeff Gradeck, spokesman for the U.S. Navy"s 5th Fleet, which is based in Bahrain. Flight 072 crashed in shallow water near shore and Ali Ahmedi, a spokesman and an acting vice president for Gulf Air, has said the pilot gave no indication to air traffic controllers that there were any problems in the plane.

Figure 6. MEAD Summary, Cluster 3, 10%.

More than 130 bodies are reported to have been recovered after a Gulf Air jet carrying 143 people crashed into the Gulf off Bahrain on Wednesday. The Airbus A320 - flight GF072 - crashed shortly before coming into land in Bahrain after a three-long flight from Cairo. Bahraini Information Ministry official Said Al-Bably said one of its engines had caught fire. CAIRO, Egypt -- The crash of a Gulf Air flight that killed 143 people in Bahrain is a disturbing deja vu for Egyptians: It is the second plane crash within a year to devastate this Arab country. Sixty-three Egyptians were on board the Airbus A320, which crashed into shallow Persian Gulf waters Wednesday night after circling and trying to land in Bahrain. On Oct. 31, 1999, a plane carrying 217 mostly Egyptian passengers crashed into the Atlantic Ocean off Massachusetts. The probe into why Gulf Air Flight GF072 crashed into the sea off Bahrain Wednesday turned to possible pilot error Friday, despite the airline's insistence that the pilot was not at fault. According to reports on CNN and local newspapers in Bahrain Friday, the control tower was concerned with the velocity and altitude of the plane and had discussed these concerns with the pilot as he circled the airport in his first approach to land. The plane was reportedly moving too fast and was not at the right altitude to land.

Figure 7. Lead-based Summary, Cluster 6, 10%.

10. Conclusions and Future Work

We believe that SAS performed better at the twenty percent compression rate as a result of two characteristics: as the sum of scores across sub-events, this algorithm preferred both sentences that received higher scores, as well as sentences which were highly ranked most frequently. Therefore, it is weighted toward those sentences that carry information essential to several sub-events. Because of these sentences' relevance to more than one sub-event, they are most likely to be important to the majority of readers, regardless of the user's particular information task. This can also be interpreted as a sort of popularity weighting, with those sentences getting the most and best scores from judges producing the most useful summaries. The patterns uncovered by this result should be leveraged for future improvements to automatic summarizers.

We conclude that multi-document summarization is improved by two specific elements. Firstly, taking into account varying degrees of relevance, as opposed to a polarized relevant/non-relevant metric. Secondly, recognizing the sub-events that comprise a single news event are essential, as relevance is particular to the aims of a reader's aims or interests.

Accounting for these characteristics improves the quality of multi-document summaries. These findings suggest that the information characteristics captured by sub-event based, manual sub-event summarization can and should be incorporated into MEAD to improve automatic multi-document summarization.

Our current work uses manual sub-event recognition. However, having shown that sub-events are useful for producing good summaries, further work in this area will involve leveraging these qualities by automating the sub-event identification process. Topic detection and tracking (TDT) technology can be used to help automatically identify sub-events. By integrating these findings into the MEAD summarizer, we hope to improve automatic summarization of multiple documents through the use of sub-event recognition.

11. References

[Allan et al, 2001a] James Allan and Rahul Gupta and Vikas Khandelwal. “Temporal summaries of news topics.” Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

[Allan et al, 2001b] James Allan and Rahul Gupta and Vikas Khandelwal. “Topic models for summarizing novelty.” ARDA Workshop on Language Modeling and Information Retrieval. Pittsburgh, Pennsylvania.

[Allan et al, 1998] "On-line New Event Detection and Tracking." Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne, Australia.

[Goldstein, 1998] Jade Goldstein & Jamie Carbonell, 1998. “The use of MMR, diversity-based reranking for reordering documents and producing summaries.” Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne, Australia.

[Goldstein, 1999] Jade Goldstein, 1999. “Automatic text summarization of multiple documents.” Thesis proposal. Carnegie Melon University.

[Mani, 2001] Inderjeet Mani, 2001. “Automatic summarization.” Natural Language Processing, ed. Ruslan Mitkov. Philadelphia, PA: John Bejamins Publishing.

[Radev et al, 2000] Dragomir Radev and Hongyan Jing and Malgorzata Budzikowska. “Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies.” ANLP/NAACL Workshop on Summarization. Seattle, WA.

[Radev et al, 2001] Dragomir Radev and Sasha Blair-Goldensohn and Zhu Zhang. "Experiments in single and multi-document summarization using MEAD." First Document Understanding Conference. New Orleans, LA.

[Radev et al, 2002] Dragomir Radev , Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Celebi, Hong Qi, Daniu Liu, and Elliott Drabek. "Evaluation Challenges in large-scale multi-document summarization: the MEAD project." Submitted to SIGIR 2002.