Experiments in Single and Multi-Document Summarization Using MEAD

Dragomir R. Radev
School of Information and
Department of EECS
University of Michigan
Ann Arbor, MI 48109

Sasha Blair-Goldensohn
School of Information
University of Michigan
Ann Arbor, MI 48109

Zhu Zhang
School of Information
University of Michigan
Ann Arbor, MI 48109

ABSTRACT

In this paper, we describe four experiments in text summarization. The first experiment involves the automatic creation of 120 multi-document summaries and 308 single-document summaries from a set of 30 clusters of related documents. We present official results from a multi-site manual evaluation of the quality of the summaries. The second experiment is about the identification by human subjects of *cross-document structural relationships* such as *identity*, *paraphrase*, *elaboration*, and *fulfillment*. The third experiment focuses on a particular cross-document structural relationship, namely *subsumption*. The last experiment asks human judges to determine which of the input articles in a given cluster were used to produce individual sentences of a manual summary. We present numerical evaluations of all four experiments. All automatic summaries have been produced by MEAD, a flexible summarization system under development at the University of Michigan.

1. INTRODUCTION

The University of Michigan's summarization system, named MEAD, was initially developed to produce multi-document extractive summaries. The main idea behind MEAD is the use of the centroid-based feature [7] which identifies sentences that are highly relevant to an entire cluster of related documents. Version 2.0 of MEAD was developed in 2001 and addresses DUC-specific constraints such as absolute summary length, very short summaries, as well as the requirement to produce both single-document and multi-document summaries.

In this paper we present a brief description of the MEAD system, including two deployed web-based applications: NewsInEssence and WebInEssence. We then turn to the version of MEAD as used in DUC'2001, focusing on the results of the evaluation. We then briefly describe three user studies which were undertaken in 2001 with the goal of understanding how information provenance, cross-document subsumption, and the identification of cross-document structural relationships can be used in the production of better multi-document summaries.

2. MEAD: A CENTROID-BASED SUMMARIZER

MEAD is based on work described in [7]. It is based on sentence extraction. For each sentence in a cluster of related documents, MEAD computes three features and uses a linear combination of the three to determine what sentences are most salient. The three features used are centroid score, position, and overlap with first sentence (which may happen to be the title of a document).

The input to MEAD is plain text and a compression rate. MEAD uses the LT-POS software, developed at the University of Edinburgh [3] to mark sentence boundaries automatically. For each sentence S_i, MEAD then computes three values:

- the centroid score C_i [7] which is a measure of the centrality of a sentence to the overall topic of a cluster (or document in the case of a single-document cluster),
- the position score P_i which decreases linearly as the sentence gets farther from the beginning of a document, and
- the overlap-with-first score F_i which is the inner product of the TF*IDF-weighted vector representations of a given sentence and the first sentence (or title, if there is one) of the document: $(|S_i|.|S_1|)$.

All three features are normalized in the range 0–1. The overall score for $S_i, V(S_i) = w_c * C_i + w_p * P_i + w_f * F_i$ is a linear combination of the three features. For this paper, only the combination of weights $w_c = 3 or 4; w_p = 2; w_f = 1$ is used. The value of 3 for w_c is used to produce shorter multi-document summaries (50- and 100- word summaries). For 200- and 400-word summaries, w_c was set to 4. A trainable version of MEAD was subsequently developed and will be briefly mentioned in the conclusion of this paper.

MEAD discards sentences that are too similar to other sentences. A parameter, *redundancymax* is used to decide which sentences are too similar (based on a cosine similarity). For the DUC experiments, a similarity threshold of .7 was used for multi-document summaries. That value was raised to .95 for single-document summaries. Any sentence that is not discarded due to high similarity and which gets a high score (within the specified compression rate) is included in the summary.

In addition to the above values, MEAD uses a large number of other parameters that can be set by the user. Some of them need to be mentioned. The first one, *shortestsentmin* indicates the minimum sentence length (in words) that will be included in a summary. The second parameter, *shortestfirstsentmin* specifies the minimal length in words of the first sentence to be included in a summary. The *defaultidf* value indicates what IDF should be given to words that are seen for the first time and for which an IDF value is therefore not known. The values for these three parameters that were used in the evaluation are 9, 15, and 5, respectively.

Another parameter, *wordutilpower* specifies the power to which number of words in sentence will be raised before doing score-per-word division - examples: set to 0 to have scores divided by 1, so all sentences scored the same regardless of length; set to .5 to have bias toward longer sentences; set to 1 to divide by the number of words in the sentence, so sentence score is inversely proportional to length; set to 1.5 to have bias toward shorter sentences. In general, lower settings favor longer sentences, higher ones favor shorter sentences.

A version of MEAD was used in the development of two Web-based summarizers - WebInEssence and NewsInEssence. WebInEssence[6] works on arbitrary web pages while NewsInEssence[5] specializes on clusters of related news stories extracted in real time from Web sources.

3. RESULTS FROM DUC 2001

We produced 120 multi-document summaries from the 30 clusters provided by DUC (30 clusters * 4 compression rates: 50-word, 100-word, 200-word, and 400-word summaries) as well as 308 single-document summaries. Some sample summaries are included in Figure 1. These four summaries are from the same cluster, DUC cluster d05.

We present our results as given to us by the DUC evaluators. Table **??** includes our performance on 11 criteria:

The following two Tables: 2 and 3 show MEAD's performance on three of the criteria: overall peer grammaticality, overall peer cohesion, and overall peer organization.

System	Grammaticality	Cohesion	Organization	Total
O	554	422	452	1428
MEAD	558	390	424	1372
Q	556	381	397	1334
R	569	439	478	1486
S	547	367	380	1294
T	534	410	460	1404
V	553	455	479	1487
W	521	393	409	1323
X	535	375	413	1323
Y	481	369	403	1253
Z	493	334	363	1190

Table 2: Single-document evaluation

4. A STUDY OF CROSS-DOCUMENT STRUCTURAL RELATIONSHIPS

In this section, we present an experiment in which subjects were asked to analyze a set of documents using a set of proposed relationships from Cross-Document Structure Theory (CST) [4]. We then present the experimental results and consider the implications for further work in CST.

System	Grammaticality	Cohesion	Organization	Total
L	432	212	220	864
M	382	235	259	876
N	423	232	258	913
O	439	249	270	958
MEAD	426	224	252	902
R	418	250	284	952
S	418	220	233	871
T	407	271	303	981
U	380	152	129	661
W	363	172	148	683
Y	284	201	205	690
Z	380	209	225	814

Table 3: Multi-document evaluation

CST proposes a taxonomy of the informational relationships between documents in clusters of related documents. Some of the relationships are direct descendents of these used in SUMMONS [8] except that in CST, these relationships are domain-independent. CST posits that by identifying these cross-document "links", one can produce superior multi-document summaries.

The concept of using CST for multi-document summaries relates to the that of using Rhetorical Structure Theory (RST) [1] for single-document summarization [2]. However, while Marcu relied on "cue phrases" in implementing algorithms to discover the valid RST "trees" for a single document, such a technique is not very plausible for discovering CST "links" between documents. For instance, the "cue phrase" "*although* **statementX**, **statementY**" might indicate the RST relationship "concession" in some circumstances. Marcu is able to use these phrases for guidance because of the conventions of writing and the valid assumptions that authors tend to write documents using certain rhetorical techniques.

However, in the case of multiple documents and CST inter-document relationships (links), we cannot expect to encounter a reliable analog to the cue phrase. This is because separate documents, even when they are related to a common topic, are not (generally) written with an overarching structure in mind. Particularly in the case of news article clusters, we are most often looking at articles which are written by different authors working from partially overlapping information as it becomes available. So, except in cases of explicit citation, we cannot expect to find a static phrase in one document which reliably indicates a particular relationship to some phrase in another document.

Nonetheless, with the proliferation of available online news sources, it becomes increasingly attractive to be able to map the inter-document relationships proposed by CST in an automated fashion. As argued in [4], being able to produce a set of CST arcs which map between a set of documents in a news cluster would enable multi-document summarization which was not only generally superior, in terms of reduced redundancy and other generally desirable features, but also summaries tailored to individual preferences.

How, then, to approach the problem of discovering CST relationships in a set of documents? We present an exploratory experiment, in which human subjects were asked to find these relationships over a multi-document news cluster. It is our hope that the results of this experiment will be an early step in the eventual development of automated CST parsing techniques.

50 words Mad cow disease, or bovine spongiform encephalopathy, or BSE, was diagnosed only in 1986. THE CONDITION known in cattle as ’mad cow disease’, spongiform encephalopathies, has been found in Britain’s sparsely-scattered antelope population, the government has admitted. He believes that BSE can trigger human brain disease.

100 words ”Mad cow disease” has killed 10,000 cattle, restricted the export market for Britain’s cattle industry and raised fears about the safety of eating beef. Mad cow disease, or bovine spongiform encephalopathy, or BSE, was diagnosed only in 1986. THE CONDITION known in cattle as ’mad cow disease’, spongiform encephalopathies, has been found in Britain’s sparsely-scattered antelope population, the government has admitted. He believes that BSE can trigger human brain disease. ’Our worst predictions are coming true,’ he said. The German government yesterday announced the launch of a new research project to examine whether the cattle disease bovine spongiform encephalopathy can be transmitted to human beings.

200 words Mad cow disease, or bovine spongiform encephalopathy, or BSE, was diagnosed only in 1986. ”Mad cow disease” has killed 10,000 cattle, restricted the export market for Britain’s cattle industry and raised fears about the safety of eating beef. Some experts believe that cattle contracted the disease as a result of eating food contaminated with the remains of sheep infected with a BSE-like disease called scrapie. THE CONDITION known in cattle as ’mad cow disease’, spongiform encephalopathies, has been found in Britain’s sparsely-scattered antelope population, the government has admitted. GOVERNMENT veterinary and health experts were yesterday putting out reassuring messages about bovine spongiform encephalopathy, or ’mad cow’ disease, in the face of growing public anxiety. Dr Kenneth Calman, the government’s chief medical officer, yesterday repeated the official advice that beef can be eaten safely: ’There is no scientific evidence of a causal link between BSE in cattle and CJD in humans.’ The epidemic of bovine spongiform encephalopathy or ’mad cow’ disease- which has killed more than 100,000 animals in the UK- is causing a new wave of public concern. The German government yesterday announced the launch of a new research project to examine whether the cattle disease bovine spongiform encephalopathy can be transmitted to human beings.

400 words Mad cow disease, or bovine spongiform encephalopathy, or BSE, was diagnosed only in 1986. ”Mad cow disease” has killed 10,000 cattle, restricted the export market for Britain’s cattle industry and raised fears about the safety of eating beef. ”Mad cow” disease, an enigmatic nervous disorder that has killed thousands of cattle in Britain, is causing trade friction in Europe and is threatening the $3.7-billion British beef industry. Some experts believe that cattle contracted the disease as a result of eating food contaminated with the remains of sheep infected with a BSE-like disease called scrapie. THE CONDITION known in cattle as ’mad cow disease’, spongiform encephalopathies, has been found in Britain’s sparsely-scattered antelope population, the government has admitted. GOVERNMENT veterinary and health experts were yesterday putting out reassuring messages about bovine spongiform encephalopathy, or ’mad cow’ disease, in the face of growing public anxiety. Dr Kenneth Calman, the government’s chief medical officer, yesterday repeated the official advice that beef can be eaten safely: ’There is no scientific evidence of a causal link between BSE in cattle and CJD in humans.’ Both BSE and CJD are caused by mysterious particles of infectious protein called prions. The epidemic of bovine spongiform encephalopathy or ’mad cow’ disease- which has killed more than 100,000 animals in the UK- is causing a new wave of public concern. He believes that BSE can trigger human brain disease. One argument put forward by the health department is that CJD has such a long incubation period- typically 10 to 20 years- that clinical symptoms would not yet have appeared, even if BSE had triggered any cases of CJD. Scientists trying to understand the epidemic face an unusual problem: BSE, scrapie and CJD are caused by a bizarre, infectious agent, the prion, which does not follow the normal rules of microbiology. Language: English Article Type:CSO [Article by Nigel Hawkes, Science Editor: ”Zoo Antelope Catch Mad Cow Disease”] [Text] Scientists at London zoo have discovered that a strain of ”mad cow disease” affecting a type of antelope can be transmitted much more easily than was thought. The German government yesterday announced the launch of a new research project to examine whether the cattle disease bovine spongiform encephalopathy can be transmitted to human beings. Several German scientists have expressed concern that BSE- popularly known as ’mad cow disease’ because of the way it debilitates the brains of cattle -may be transmissible to humans who eat contaminated beef or take medicines made with ingredients from contaminated animals.

Figure 1: Sample multi-document summaries produced by MEAD at four compression rates

Metric	Avg. all peers (L-Z)	Avg. MEAD	StDev
Overall peer grammaticality	3.53	3.58	0.72
Overall peer cohesion	2.30	2.50	1.19
Overall peer organization	2.46	2.79	1.21
Unmarked peer units (PUs) that ought to be in model in place of something there	0.28	0.27	0.86
Unmarked PUs that don’t deserve to be in the model, but related to the subject	2.79	2.70	1.59
Unmarked PUs that are unrelated to the subject of the model	0.40	0.29	0.93
Number of model units (MU)	8.80	8.69	6.28
Number of peer units (PU)	5.90	5.72	4.69
Number of unique PUs marked expressing some of the same content as one or more MUs	2.96	3.31	2.79
Number of peer units marked for this MU	0.35	0.40	0.82
Extent which marked PUs express meaning of the current MU	0.61	0.73	1.30

Table 1: Comparison of MEAD with the rest of DUC participants

4.1 The experiment

The experiment which we conducted required subjects to read a set of news articles and write down the inter-document relationships which they observed. Specifically, the articles were on the subject of an airplane crash of a flight from Egypt to Bahrain in August, 2000. They were written by several different news organizations and retrieved from online news web sites in the days following the accident. The cluster contained eight articles in total. Six of the articles focus generally on the crash and its direct aftermath; one mentions the crash while focusing on the history of the particular model of jet plane involved; and one focuses on the toll of the crash in Egypt, where many passengers were from.

The subjects, eight graduate students and one professor at the University of Michigan, were given the articles and a set of instructions. The instructions specified five sets of article pairs comprised by random pairings of the eight articles mentioned above. Each article was included in at least one pair; no article was included in more than two pairs. For each pair, the subjects were instructed to first read the articles carefully. They were then instructed to look for and note down any occurrences of relationships like those in Figure 2. (Subjects were also provided with the examples shown in Figure 2 to illustrate each relationship type.) It was stated in the instructions that the relationships comprised only a "proposed" list, and not to be considered exhaustive. Subjects were invited to make up new relationship types if they observed cross-document relationships which did not correspond to those in Figure 2.

Although subjects were given examples of the proposed relationships at the sentence level, the instructions also explicitly stated that it was possible for a relationship to hold with one or both text spans being more than one sentence long. There was no provision for subjects to mark relationships with one or both text spans less than a full sentence in length. Subjects were instructed not to note down examples of these relationships across spans within a single document. Also, subjects were instructed that it was possible for more than one relationship to exist across the same pair of text spans, and to note down as many relationships as they observed for each pair of text spans.

No guidelines were given to subjects about how many relationships to identify per article pair. Rather, they were simply instructed to "continue writing down relations until you are reasonably certain that no further interesting relationships hold" for a given document pair.

4.2 Results

A summary of the raw results of the experiment are shown in Table 4.

Table 5 indicates the total number of links observed per article pair. Articles 41 and 47 are the articles mentioned above which focus on the airplane model and Egyptian perspective, respectively.

Table 6 describes the sentence pairs for which judges noted relationships. The total number of sentence pairs for all five article pairs assigned was 4579, which is $\sum_{i=1}^{5} n_i \times m_i$, where i is the number of the article pair, n is the number of sentences in the first article in the pair, and m is the number of sentences in the second article in the pair. Of course, by combining sentences to form longer text spans, a hugely larger number of text-span pairs are possible. Therefore, the other numbers in Table 6 should be carefully understood.

Articles	Total CST Relationships Identified
7 and 63	92
81 and 87	100
30 and 97	76
41 and 81	31
30 and 47	110

Table 5: Total Identifications of CST Relationships by Article Pair

In the second and third rows, the numbers of "sentence pairs" listed speaks of distinct sentence pairs for which either one or multiple judges observed a relationship, respectively. That is, if one judge observed relationship X between sentences 1-2 of document A and sentence 2 of document B, this would count as two sentence pairs. However, if the identical observation was made save that the first text span was limited to sentence 1, this would count as one sentence pair in the context of Table 6. Furthermore, in the context of Table 6, counting a pair as being observed to have a relationship by multiple judges, it is not necessary that a) the relationship observed be the same one OR b) the judges have marked a relationship for the exact same text spans. For example:

- Judge John identified relationship X between Doc A/Sents 1-2 and Doc B/Sent 2
- Judge Kyle identified relationship Y between Doc A/Sent 1 and Doc B/Sent 2

In the context of Table 6, this equates to one sentence pair identified by multiple judges (A/1-B/2), and one sentence pair identified by a single judge (A/2-B/2).

Judges Finding a Relationship	**Number of Sentence Pairs**
No Judges	4,291
One Judge	200
Multiple Judges	88

Table 6: Sentence Pairs by Number of Judges Marking a CST Relationship

As can be seen in Table 6, there are 88 sentence pairs (as just defined) for which multiple judges identify at least one CST relationship. Table 7 describes the breakdown of these 88 pairs in terms of inter-judge agreement. Although subjects were permitted to mark more than one relation per sentence pair, they are counted as "in agreement" here if at least one of the relations they mark agrees with one of the relations marked by another judge.

- Judge Frank identifies relationships X and Y for a given sentence pair
- Judge Horace identifies (only) relationship X for the same pair

In Table 7, these judges would be counted as agreeing.

Relationship	Description	Span 1 (S1)	Span 2 (S2)
Identity	The same text appears in more than one location	Tony Blair was elected for (Repetition) a second term today.	Tony Blair was elected for a second term today.
Equivalence (Paraphrase)	Two text spans have the same information content	Derek Bell is experiencing a resurgence in his career.	Derek Bell is having a “comeback year.”
Translation	Same information content in different languages	Shouts of “Viva la revolucion!” echoed through the night.	The rebels could be heard shouting, “Long live the revolution”.
Subsumption	S1 contains all information in S2, plus additional information not in S2	With 3 wins this year, Green Bay has the best record in the NFL.	Green Bay has 3 wins this year.
Contradiction	Conflicting information	There were 122 people on the downed plane.	126 people were aboard the plane.
Historical Background	S1 gives historical context to information in S2	This was the fourth time a member of the Royal Family has gotten divorced.	The Duke of Windsor was divorced from the Duchess of Windsor yesterday.
Citation	S1 explicitly cites document S2	Prince Albert then went on to say, “I never gamble.”	An earlier article quoted Prince Albert as saying “I never gamble.”
Modality	S1 presents a qualified version of the information in S2, e.g., using “allegedly”	Sean “Puffy” Combs is reported to own several multimillion dollar estates.	Puffy owns four multimillion dollar homes in the New York area.
Attribution	S1 presents an attributed version of information in S2, e.g. using “According to CNN,”	According to a top Bush advisor, the President was alarmed at the news.	The President was alarmed to hear of his daughter’s low grades.
Summary	S1 summarizes S2.	The Mets won the Title in seven games.	After a grueling first six games, the Mets came from behind tonight to take the Title.
Follow-up	S1 presents additional information which has happened since S2	102 casualties have been reported in the earthquake region.	So far, no casualties from the quake have been confirmed.
Indirect speech	S1 indirectly quotes something which was directly quoted in S2	Mr. Cuban then gave the crowd his personal guarantee of free Chalupas.	“I’ll personally guarantee free Chalupas,” Mr. Cuban announced to the crowd.
Elaboration / Refinement	S1 elaborates or provides details of some information given more generally in S2	50% of students are under 25; 20% are between 26 and 30; the rest are over 30.	Most students at the University are under 30.
Fulfillment	S1 asserts the occurrence of an event predicted in S2	After traveling to Austria Thursday, Mr. Green returned home to New York.	Mr. Green will go to Austria Thursday.
Description	S1 describes an entity mentioned in S2	Greenfield, a retired general and father of two, has declined to comment.	Mr. Greenfield appeared in court yesterday.
Reader Profile	S1 and S2 provide similar information written for a different audience.	The Durian, a fruit used in Asian cuisine, has a strong smell.	The dish is usually made with Durian.
Change of perspective	The same entity presents a differing opinion or presents a fact in a different light.	Giuliani criticized the Officer’s Union as “too demanding” in contract talks.	Giuliani praised the Officer’s Union, which provides legal aid and advice to members.

Figure 2: Proposed CST relationships and examples

Relationship Type	Subject										
	1	2	3	4	5	6	7	8	9	**Sum**	**Avg**
Identity	1	0	0	2	1	1	1	0	1	9	.78
Equivalence	8	2	2	36	5	7	5	4	1	70	7.78
Translation	0	0	0	0	0	0	0	0	0	0	0.00
Subsumption	16	3	2	7	3	1	3	3	0	39	4.22
Contradiction	4	4	0	7	4	5	0	4	1	31	3.22
Historical Background	35	3	0	4	1	0	1	0	0	44	4.89
Citation	0	0	0	0	0	0	0	0	0	0	0.00
Modality	0	1	0	0	0	0	0	1	0	2	0.22
Attribution	0	0	1	8	4	0	2	0	0	15	1.67
Summary	1	0	0	0	0	0	0	0	0	1	0.11
Follow-up	8	6	2	13	4	4	2	3	0	42	4.67
Indirect speech	1	1	1	0	1	0	0	2	0	6	0.67
Elaboration / Refinement	6	15	2	22	17	9	5	3	6	85	9.44
Fulfillment	1	0	0	1	2	0	0	0	0	4	0.44
Description	44	10	0	5	5	0	0	0	0	64	7.11
Reader Profile	0	0	0	0	1	0	0	0	0	1	0.11
Change of Perspective	0	0	0	0	0	0	0	1	0	1	0.11
	125	45	16	105	48	27	19	21	9	415	45.44

Table 4: Identifications of CST Relationships by Subject and Type

Discrete Relationship Types Observed	Judges in Agreement	Sentences
Only one	All	16
More than one	At least two	35
More than one	None	37

Table 7: Judge Agreement on Relationship Types among Sentence Pairs Linked by Multiple Judges

4.3 Observations

Because our data comes from observations about (a subset of) a single news cluster, it would clearly be premature to make conclusions about the natural frequencies of these relationships based on the data in Table 5. Nonetheless, we can at least speculate that human subjects are capable of identifying some subset of these relationships when reading articles from this news cluster.

On average, subjects identified approximately 45 occurrences of the proposed relationships per article. Interestingly, some relationships were identified much more frequently than others. The relationships "Elaboration/Refinement," "Equivalence," and "Description" were identified most frequently. Other relationships, such as "Translation," "Citation," "Summary," "Reader Profile" and "Change of Perspective" were observed never or only by one subject. Although subjects were encouraged in the study instructions to name new relationships, none did so.

As noted above, we need more data before we can say if the lack of identifications for these unobserved / rarely observed relationships is because of a true lack of frequency or some other factor. For instance, some of the proposed relationship names, like "modality," may not be intuitive enough for judges to feel comfortable identifying them, even though examples were given.

However, the most encouraging data concerns the relatively high level of overlap when multiple judges made an observation for a sentence. In 51 of 88 cases where more than one judge marked a sentence pair, at least two judges concurred about at least one relationship holding for the pair. Although approximately two-thirds of the marked pairs were marked by only one judge, the overall data sparseness (in comparison to the number of possible sentence pairs, only about 1/100th of pairs were marked) makes this ratio less discouraging.

Further analysis of the data is still needed. The level of judge agreement would seem to indicate that at least some of the proposed CST relationships are recognizable with a suitable degree of correspondence by humans. Before attempting to build automated means of detecting CST hierarchies for a document cluster, a better understanding of which relationships can be empirically demonstrated must be found.

Another key step is to gather further data. In order to do so, an automated markup tool in the style of Alembic Workbench or SEE would be extremely helpful. Not only is there a great deal of transcription (and associated possibilities for error) involved in running this experiment on paper, but a number of subjects expressed the belief that an automated tool like this would allow them to provide better and more consistent data.

5. TWO MORE USER STUDIES

We will now briefly note two additional experiments in progress. The first one deals with cross-document subsumption while the second one is about information provenance.

5.1 Cross-document subsumption

In this experiment, we asked five paid judges to find all pairs of sentences in a given cluster of documents such that one of the sentences in the pair subsumes the second one. Subsumption is just one of many cross-document structural relationships that were discussed in the previous section. We chose it for further analysis as it appears to be most closely related to generic multi-document summarization. The main idea is that if sentence S_1 subsumes sentence S_2, then S_2 need not be included in the summary if S_1 is to be included. For a detailed discussion of cross-document subsumption, refer to [7].

The five judges were given a subset of queries from the Johns Hopkins Workshop corpus (see the final Section of this paper). The documents in each query-induced cluster of relevant documents are from the Hong Kong News corpus distributed by the Linguistic Data Consortium. A total of 12 clusters (consisting of 10 articles each) were given to two judges each. Table 8 indicates the number of subsumptions found for each cluster. These numbers were computed by John Blitzer from Cornell University who is currently performing further analysis of the subsumption data.

Cluster number	**Associated query**	**Number of subsumptions**
112	Autumn and sports carnivals	434
125	Narcotics Rehabilitation	49
199	Intellectual Property Rights	111
241	Fire safety, building management concerns	15
323	Battle against disc piracy	258
398	Flu results in Health Controls	52
447	Housing (Amendment) Bill Brings Assorted Improvements	103
551	Natural disaster victims aided	649
827	Health education for youngsters	142
883	Public health concerns cause food-business closings	90
1014	Traffic Safety Enforcement	323
1197	Museums: exhibits/hours	228

Table 8: Subsumptions identified

5.2 Information provenance

In this experiment, we wanted to verify two hypotheses: (1) that information in a multi-document summary can be traced back to the article (or articles) which were used to produce it and (2) that human subjects can reach high levels of agreement in determining information provenance.

The participants (six in all) were presented with a cluster of ten news articles and four 400-word multi-document summaries (by 4 human assessors). The summaries were not necessarily produced through sentence extraction. They exist at various compression rates, however (50, 100, 200, 400 words). Only the 400-word summaries are used in this experiment. The topic of the news cluster is about day care issues in the U.S. The summaries and source articles were provided as part of the DUC training data.

Each summary contains a certain number of sentences. The information in each sentence is supposed to come from a certain article or articles in the cluster. The task of the participants is to identify the information source for each sentence in the summaries. If two articles provide overlapped or even totally redundant information for the same sentence in summary, both should be identified.

Summarizing large amount of information is a highly intelligent human behavior. The participants were also asked to report any patterns they noticed about how human generating summaries.

Preliminary results. Some preliminary results are shown in Table 9.

Here we are interested in how many sentences in the final summaries each article contributes to. Notice that the contribution of articles is not evenly distributed, according to the statistics in the table above.

Features of the articles. In Table 10, we summarize some features of the documents in the news cluster as we try to find ways to correlate them with the summarization process.

Table 11 summarizes the level of interjudge agreement among the six judges out of 103 sentences.

High (5 more more judges agree)	76
Medium (3 or 4 judges agree)	22
Low (no more than 2 judges agree)	5

Table 11: Interjudge agreement

For experiments of this kind, it is very important to measure the degree of agreement among human judges. According to the criteria and statistics in the table above, it is reasonable to pursue this experiment further.

If the human abstractor uses a strategy similar to sentence extraction, judges in our experiment tend to have very high agreement among them; on the contrary, if the human summarizer regenerates the summary totally according to his own understanding of the news cluster, our judges usually have trouble agreeing with each other.

Some other interesting observations were made by the participants:

- Article #7 is only partially relevant to the topic.
- Some human assessors tend to sequence their paragraphs in a way that each paragraph corresponds to one or two source articles; others have a more integrated style, using many documents for each paragraph in the summary.
- Summaries produced by different human assessors tend to focus on different subsets of the document cluster.

6. CONCLUSION

We presented work done at the University of Michigan as part of the DUC evaluation, 2001. We described our summarizer, MEAD, which is based on centroid-based features. We included the results of our participation in DUC. We also presented some results from three preliminary experiments. These results are likely to be used in the development of next year's release of MEAD. The first step toward the new version was already taken at the summer workshop, held at Johns Hopkins University.

6.1 The Johns Hopkins summer workshop

During the summer of 2001, a team of 10 researchers from 6 countries met together at the Center for Language and Speech processing at Johns Hopkins University (www.clsp.jhu.edu) and worked together for eight weeks (preliminary work was done in advance) to achieve the following goals:

- develop a public-domain trainable summarization system by rewriting MEAD from scratch and including in the new architecture a module that allows salience decisions to be made based on cross-document relationships such as the ones posited by CST,
- develop a summarization evaluation system,

Article	Subject A	Subject B	Subject C	Subject D	Subject E	Subject F	Avg. Contribution
1	6	12	11	9	10	8	9.33
2	11	11	14	8	11	7	10.33
3	14	18	17	16	19	18	17.00
4	17	15	18	9	14	17	15.00
5	5	3	9	4	4	7	5.33
6	11	11	14	10	10	10	11.00
7	11	11	11	6	10	3	8.67
8	11	12	23	7	11	10	12.33
9	5	6	6	4	6	6	5.50
10	8	6	7	5	7	7	6.67

Table 9: Information provenance

Article	Length in words	Length in sentences	Number of proper nouns	Occurrence of "daycare"	Early/ Late
1	1171	54	10	1	E
2	1080	45	50	0	E
3	1670	84	85	0	E
4	817	34	10	0	M
5	460	19	41	0	M
6	1245	58	45	0	M
7	1002	42	30	0	M
8	1369	63	68	0	L
9	1799	82	55	0	L
10	382	10	12	0	L

Table 10: Article features

- develop a large annotated corpus for further research in text summarization, and
- perform a meta-evaluation of a large variety of evaluation metrics including co-selection (precision, recall, F-measure, Kappa), content-based metrics (cosine, binary cosine, longest common subsequence, and word overlap), relative utility, and relevance preservation.

The resulting system is quite robust: it was used to produce several hundred million summaries of different compression rates (10 in total), two languages (English and Chinese), generic and query-based, both single- and multi-document summaries. The evaluation was carried out on 10 summarization systems (including a trainable version of MEAD) in a variety of settings.

All the goals of the meeting were achieved. During the fall of 2001, the summarizer (MEAD), the evaluation toolkit, and the annotations to the corpus will be released to the community. The corpus itself is being made available by the LDC.

7. REFERENCES

[1] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[2] Daniel Marcu. From Discourse Structures to Text Summaries. In *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 11 1997.

[3] Andrei Mikheev. Document centered approach for text normalization. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, 2000.

[4] Dragomir Radev. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *Proceedings, 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, October 2000.

[5] Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Interactive, domain-independent identification and summarization of topically related news articles. In *5th European Conference on Research and Advanced Technology for Digital Libraries*, Darmstadt, Germany, 2001.

[6] Dragomir R. Radev, Weiguo Fan, and Zhu Zhang. NewsInEssence: A personalized web-based multi-document summarization and recommendation system. In *NAACL Workshop on Automatic Summarization*, Pittsburgh, PA, 2001.

[7] 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 Workshop on Summarization*, Seattle, WA, April 2000.

[8] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, September 1998.