Single-document and multi-document summary evaluation using Relative Utility

Dragomir R. Radev[1,2], Daniel Tam[1] and Güneş Erkan[1]
[1]Department of Electrical Engineering and Computer Science
[2]School of Information
{radev,dtam,gerkan}@umich.edu
University of Michigan, Ann Arbor MI 48109

Abstract

We present a series of experiments to demonstrate the validity of Relative Utility (RU) as a measure for evaluating extractive summarization systems. Like some other evaluation metrics, it compares sentence selection between machine and reference summarizers. Additionally, RU is applicable in both single-document and multi-document summarization, is extendable to arbitrary compression rates with no extra annotation effort, and takes into account both random system performance and interjudge agreement. RU also provides an option for penalizing summaries that include sentences with redundant information. Our results are based on the JHU summary corpus and indicate that Relative Utility is a reasonable, and often superior alternative to several common summary evaluation metrics. We also give a comparison of RU with some other well-known metrics with respect to the correlation with the human judgements on the DUC corpus.

1 Introduction

One major bottleneck in the development of text summarization systems is the absence of well-defined and standardized evaluation metrics. In this paper we will discuss Relative Utility (RU), a method for evaluating extractive summarizers, both single-document and multi-document. We will address some advantages of RU over existing co-selection metrics such as Precision, Recall, percent agreement, and Kappa. We will present some experiments performed on a large text corpus to discuss how RU is affected by interjudge agreement, compression rate (or summary length), and summarization method.

The main problem with traditional co-selection metrics (thus named because they measure the degree of overlap between the list of sentences selected by a judge and an automatically produced extract) such as Precision, Recall, and Percent Agreement for evaluating extractive summarizers is that human judges often disagree about which the top n% most important sentences in a document or cluster are and yet, there appears to be an implicit importance value for all sentences which is judge-independent. We base this observation on an experiment in which we asked three judges to give scores from 0 to 10 to each sentence in a multi-document cluster. Even though the relative rankings of the sentences based on the judge-assigned importance varies significantly from judge to judge, their absolute importance scores are highly correlated. We have measured the utility correlation for three judges on 3,932 sentences from 200 documents from the HK News corpus. The average pairwise Pearson correlation was 0.71, which is indicative of high agreement.

In the next section, we will formally introduce the Relative Utility method. The following two sections discuss our evaluation framework. Our goal was to understand what properties of multi-document extractive summaries make them hard to evaluate using co-selection metrics and how Relative Utility can be used to capture summaries in which equally important sentences are substituted for one another. Section 3 describes our experimental setup while Section 4 summarizes our results and our analysis of these results. In Section 5, we discuss the fact that the presence of a sentence in a multi-document summary may readjust the importance score of another sentence (e.g., when the two sentences are paraphrases of each other or when the included sentence subsumes the other sentences). We propose a variant of Relative Utility (RU with subsumption) which addresses this problem by giving only partial credit for redundant sentences that are included in a summary. Section 6 concludes our presentation by summarizing our conclusions and setting the agenda for future research.

2 The Relative Utility evaluation method

Extractive summarization is the process of identifying highly salient units (usually words, phrases, sentences, or paragraphs) within a cluster of documents. When a cluster consists of one document, the process is called single-document extractive summarization, otherwise the name is multi-document extractive summarization.

Extractive summarization is the only scalable and domain-independent

method for text summarization. It is used in a variety of systems (e.g., [Luh58],[JMBE98]).

One common class of evaluation metrics for extractive summaries based on text unit overlap includes Precision and Recall (P&R), Percent Agreement (PA), and Kappa. The generic name for this class of evaluation methods is *co-selection* as they measure to what extent an automatic extract overlaps with manual extracts.

Using metrics such as P&R or PA [JMBE98, GKMC99] to evaluate summaries creates the possibility that two equally good extracts are judged very differently.

Suppose that a manual summary contains sentences {1 2} from a document. Suppose also that two systems, A and B, produce summaries consisting of sentences {1 2} and {1 3}, respectively. Using P&R or PA, system A will be ranked much higher than system B. It is quite possible however, that for the purpose of summarization, sentences 2 and 3 are *equally* important, in which case the two systems should get the same score. It is known from the literature on summarization (e.g., [JMBE98]) that given a target summary length, judges often pick different sentences. We will call this observation the principle of Summary Sentence Substitutability (SSS).

The Relative Utility (RU) method [RJB00] allows ideal summaries to consist of sentences with variable membership. With RU, the ideal summary represents all sentences of the input document(s) with confidence values for their inclusion in the summary. It directly addresses the SSS problem because it allows for sentences in different summaries of the same input to be substituted (at a small cost) for one another.

For example, a document with five sentences {1 2 3 4 5} is represented as {1/10 2/9 3/9 4/2 5/4}. The second number in each pair indicates the degree to which the given sentence should be part of the summary according to a human judge. We call this number the *utility* of the sentence. Utility depends on the input documents and the judge. It does not depend on the summary length. In the example, the system that selects sentences {1 2} will not get a higher score than a system that chooses sentences {1 3} given that both summaries {1 2} and {1 3} carry the same number of utility points (10+9). Given that no other combination of two sentences carries a higher utility, both systems {1 2} and {1 3} produce optimal extracts at the given target length of two sentences.

In Relative Utility experiments, judges are asked to assign numerical scores to individual sentences from a single document or a cluster of related documents. A score of 10 indicates that a sentence is central to the topic of the cluster while a score of 0 marks a totally irrelevant sentence.

2.1 An example

S#	Text	J_1 util	J_2 util
2	The preliminary investigations showed that at this stage, human-to-human transmission of the H5N1 influenza A virus has not been proven and further investigations will be made to study this possibility, the Special Working Group on H5N1 announced today (Sunday)	9	8
3	The initial findings also showed that the four H5 cases did not share a common source, nor was the virus transmitted from one case to the others.	7	4
7	However, there is no cause for panic as available evidence does not suggest that the disease is widespread.	7	6
9	The WHO has been asked to alert vaccine production centres in the world in the case investigation to follow developments here with a view to preparing the necessary vaccines.	7	7
14	He said the Department would disseminate to doctors, medical professionals, colleges and health care workers available information about the H5 virus through letters and the Department of Health's homepage on the Internet (http:/www.info.gov.hk/dh/).	8	8

Figure 1: A 5-sentence extractive summary created from document D-19971207-001 (in cluster 398) by LDC Judge J_1.

S#	Text	J_1 util	J_2 util
11	To further enhance surveillance in Hong Kong, Dr Saw said, the Department of Health would extend surveillance coverage to all General Out-patient Clinics.	8	10
12	The Hospital Authority would also set up surveillance in public hospitals.	4	10
13	In the meantime, Dr Saw said, the Agriculture and Fisheries Department had also increased surveillance in poultry in collaboration with The University of Hong Kong.	6	10
19	Dr Saw advised members of the public that the best way to combat influenza infection was to build up body resistance by having a proper diet with adequate exercise and rest.	7	10
20	Good ventilation should be maintained to avoid the spread of respiratory tract infection.	8	10

Figure 2: A 5-sentence extractive summary created from document D-19971207-001 (in cluster 398) by LDC Judge J_2.

The following example illustrates an advantage that Relative Utility has over Precision/Recall. The two summaries shown in Figures 1 and 2 are 5-sentence extractive summaries created from the same document by two different judges. Because the two summaries are composed entirely of dif-

ferent sentences, the interjudge agreement as measured by Precision/Recall or Percent Agreement is 0, despite the fact that both summaries are reasonable.

Note that both judges gave each other's sentences fairly high utility scores, however. In fact, the interjudge agreement as measured by RU for this example is 0.76. RU agreement (see next section) is defined as the relative score that one judge would get given his own extract and the other judge's sentence judgements. For example, if judge 1 picks a single sentence in his extract and if the score that judge 2 gives to the same sentences is 8, and given that judge 2's top ranked sentence has a score of 10, then one can say that judge 1's score *relative* to judge 2 is 0.80 (or 8/10).

The 0.76 score is also markedly higher than the lowest possible score a summarizer could receive. Although not depicted in the example, a summarizer could have an RU agreement with judge J_1 as low as 0.14 and an agreement with judge J_2 as low as 0.38. In other words, given that interjudge agreement is significantly less than 1.0 but significantly more than the worst score possible, an automatic summarizer might score as low as .70 and still be almost as good as the judges themselves.

A related paper [DDM00] suggested that one problem with classic approaches to summary evaluation is that different collections of extracts rank differently when one ground truth (judgement) is substituted for another. In their experiments, recall for the same summary varied from 25% to 50% depending on what manual extract it was compared against. Our results strongly confirm Donaway et al.'s claims and suggest that RU is a viable evaluation alternative.

2.2 Defining Relative Utility

In this section, we will formally define Relative Utility (RU). To compute RU, a number of judges, N ($N \geq 1$), are asked to assign *utility scores* to all n sentences in a cluster of documents (which can consist of one or more documents). The top e sentences according to utility score are then called a sentence extract of size e (in the case of ties, some arbitrary but consistent mechanism is used to decide which sentences should be included in the summary). The formulas below assume that n is the number of sentences in a cluster of documents, e is the number of sentences in the desired extract, and N is the number of human judges providing utility scores.

We can then define the following metrics:

$$\vec{U_i} = \{u_{i,1}, u_{i,2}, ..., u_{i,n}\}$$

$$
\begin{aligned}
&= \text{sentence utility scores for judge } i \\
&\quad \text{for all } n \text{ sentences in the cluster} \\
\vec{U_i'} &= \{\delta_{i,1} \cdot u_{i,1}, \delta_{i,2} \cdot u_{i,2}, ..., \delta_{i,n} \cdot u_{i,n}\} \\
&= \text{extractive utility scores for judge } i
\end{aligned}
$$

In the formula for $\vec{U_i'}$, $\delta_{i,j}$ is the summary characteristic function for judge i and sentence j. It is equal to 1 for the e highest-utility sentences for a given judge, allowing us to adjust the summary size. For example, if $e = 2$, and $\vec{U_i} = \{10, 8, 9, 2, 4\}$, then $\delta_{i,1} = \delta_{i,3} = 1$ and $\delta_{i,2} = \delta_{i,4} = \delta_{i,5} = 0$. Note that $\sum_{j=1}^{n} \delta_{i,j} = e$.

We can now define some additional quantities:

$$
\begin{aligned}
U_i &= \sum_{j=1}^{n} u_{i,j} \\
&= \text{total self-utility for judge } i \\
U_i' &= \sum_{j=1}^{n} \delta_{i,j} \cdot u_{i,j} \\
&= \text{total extractive self-utility for judge } i \\
&\quad \text{(computed over all } n \text{ sentences)} \\
U_{i,k} &= \sum_{j=1}^{n} \delta_{i,j} \cdot u_{k,j} \\
&= \text{total extractive cross-utility for judges } i \text{ and } k \\
&\quad (i \neq k) \\
U_{i,avg} &= 1/(N-1) \cdot \sum_{k=1}^{N} U_{i,k} \quad \text{for } i \neq k \\
&= \text{(non-symmetric) judge utility for judge } i. \\
J &= U_{avg} = 1/N \cdot \sum_{i=1}^{N} U_{i,avg} \\
&= \text{interjudge performance} \\
&\quad \text{(average extractive cross-utility of all judges)} \\
U &= \sum_{j=1}^{n} \sum_{i=1}^{N} u_{i,j} \\
&= \text{total extractive utility for all judges.}
\end{aligned}
$$

$$
\begin{aligned}
U' &= \sum_{j=1}^{n} \varepsilon_j \cdot \sum_{i=1}^{N} u_{i,j} \\
&= \text{total utility for all judges}
\end{aligned}
$$

In the formula for U', ε_j (multi-judge summary characteristic function) is 1 for the top e sentences according to the sum of utility scores from all judges. U' is the maximum utility that any system can achieve at a given summary length e.

Note that $\sum_{j=1}^{n} \varepsilon_{i,j} = e$. Note also that $N = 1$ implies $U' = U'_1$ (single judge case).

A summarizer producing an extract of length e can be thought of as an additional judge. Its (non-normalized) RU will be computed as its performance against the human judges divided by the maximum possible performance. In other words, the ratio of the sum of its cross-utility with the totality of human judges and the maximum utility U' achievable at a given summary length e. As a result, a summary can be judged based on its utility *relative* to the maximum possible against the set of judges, hence the name of the method *RU*.

$$
\begin{aligned}
S &= \frac{\sum_{j=1}^{n} \delta_{s,j} \cdot \sum_{i=1}^{N} u_{i,j}}{U'} \\
&= \text{system performance } (\delta_{s,j} \text{ is equal to 1 for the} \\
&\quad \text{top } e \text{ sentences extracted by the system}).
\end{aligned}
$$

In the formula for S, $\sum_{i=1}^{N} u_{i,j}$ is the utility assigned by the totality of judges to a given sentence j extracted by the summarizer.

$$
\begin{aligned}
R &= 1/\binom{n}{e} \sum_{t=1}^{\binom{n}{e}} S_t \\
&= \text{random performance (computed over all } \binom{n}{e} \\
&\quad \text{possible extracts of length } e).
\end{aligned}
$$

R is practically a lower bound on S while J is the corresponding upper bound. In order to factor in the difficulty of a given cluster, one can normalize the system performance S between J and R:

$$
D = \frac{S - R}{J - R}
$$

$=$ normalized Relative Utility
(normalized system performance).

Assuming that $R \neq J$ (which is not unreasonable (!) and holds in practice), $D = 1$ only when $S = J$ (system is as good as the interjudge agreement) and $D = 0$ when $S = R$ (system is no better than random).

Reporting S values in the absence of corresponding J and R values is not very informative. Therefore, one should either report S, J, and R or report D alone.

When values for R and J are given as comparison, reporting S is sufficient. However, D should be used when R and J are ignored.

3 Experimental setup

We used the Hong Kong News summary corpus created at Johns Hopkins University in 2001. The original corpus consists of 18,146 aligned articles (on the document level) in plain text in English and Chinese without any markup.[1] We annotated the corpus with information about sentence and word boundaries for both English and Chinese, and part of speech and morphological information for articles in English only.

3.1 Clusters

The Linguistic Data Consortium (LDC) developed 40 queries that cover a variety of subjects ("narcotics rehabilitation", "natural disaster victims aided", "customs staff doing good job", etc.). Using an in-house information retrieval engine and human revision, documents highly relevant to the queries were obtained and the 10 most relevant (according to human assessors) were used to construct clusters. These 40 clusters of documents were used during the workshop for training and some specific evaluations. Figure 3 shows the first 20 queries that were used in our experiments.

The three human annotators from LDC judged each sentence within the 10 relevant documents in each cluster. They assigned each sentence a score on a scale from 0 to 10, expressing the importance of this sentence for the summary. This annotation allows us to compile human-generated 'ideal' summaries at different target lengths, and it is the basis for our different measures of sentence-based agreement, both between the human agreement and between the system and the human annotators. We can in fact, in

[1] http://www.ldc.upenn.edu

Group 125	Narcotics Rehabilitation
Group 241	Fire safety, building management concerns
Group 323	Battle against disc piracy
Group 551	Natural disaster victims aided
Group 112	Autumn and sports carnivals
Group 199	Intellectual Property Rights
Group 398	Flu results in Health Controls
Group 883	Public health concerns cause food-business closings
Group 1014	Traffic Safety Enforcement
Group 1197	Museums: exhibits/hours
Group 447	Housing (Amendment) Bill Brings Assorted Improvements
Group 827	Health education for youngsters
Group 885	Customs combats contraband/dutiable cigarette operations
Group 2	Meetings with foreign leaders
Group 46	Improving Employment Opportunities
Group 54	Illegal immigrants
Group 60	Customs staff doing good job.
Group 61	Permits for charitable fund raising
Group 62	Y2K readiness
Group 1018	Flower shows

Figure 3: 20 queries produced by the LDC.

addition to RU scores, produce any co-selection metric such as P/R and Kappa using the top ranked sentences.

Each query-based cluster contains 10 documents. Figure 4 shows the contents of cluster 125. The document IDs come from the HKNews corpus and indicate the year, month, day, and story number for each document.

```
<?xml version='1.0'?>
<CLUSTER LANG=''ENG''>
        <D DID=''D-20000408_011.e'' />
        <D DID=''D-19990927_011.e'' />
        <D DID=''D-19990425_009.e'' />
        <D DID=''D-19990218_009.e'' />
        <D DID=''D-19990829_012.e'' />
        <D DID=''D-19990729_008.e'' />
        <D DID=''D-19980430_016.e'' />
        <D DID=''D-19990211_009.e'' />
        <D DID=''D-19980306_007.e'' />
        <D DID=''D-19990802_006.e'' />
</CLUSTER>
```

Figure 4: Sample cluster.

3.2 Sentence utility judgements

All sentence utility scores given by the judges for a given cluster are represented in a so-called *sentjudge*. An example is shown in Figure 5. The total number of sentences in cluster 125 is 232. By convention, a 10% summary will contain 24 sentences (23.2 rounded up). (Note that in the case where a cluster contains a single document, sentjudges can used for single-document summarization).

While we have not studied the cost of acquiring such *sentjudges*, it appears to be comparable to that of generating human reference summaries for the other co-selection evaluation schemes. In the case where it is impossible to have human judges assign utility scores to each sentences, one could produce such judgement automatically from manual abstracts, which we discuss in Section 6.

DOC:SENT	JUDGE1	JUDGE2	JUDGE3	TOTAL
19980306_007:1	4	6	9	19
19980306_007:2	5	10	9	24
19980306_007:3	4	9	7	20
19980306_007:4	4	9	8	21
19980306_007:5	5	8	8	21
19980306_007:6	4	9	5	18
19980306_007:7	4	9	6	19
19980306_007:8	5	7	8	20
...				
20000408_011:13	1	5	3	9
20000408_011:14	6	4	2	12
20000408_011:15	2	6	6	14
...				

Figure 5: Sentjudge: sentence utilities as assigned by the judges.

3.3 Summarizers

For evaluation, we used two summarization systems that were available to us.

One summarizer that we used in the experiments is WEBSUMM [MB99]. It represents texts in terms of graphs where the nodes are occurrences of words or phrases and the edges are relations of repetition, synonymy, and co-reference. WEBSUMM assumes that nodes which are connected to many other nodes are likely to carry salient information, and it builds its summary correspondingly.

The second summarizer is the centroid-based summarizer MEAD [RJB00]. MEAD ranks sentences in a cluster of documents based on their positions

in a document and the cosine similarity between them and the sentence *centroid*, which is a pseudo-sentence (bag of words) that is closest to all sentences in the cluster. MEAD has a built-in facility which does not include in the summary sentences that are too similar (lexically) to the rest of the summary.

3.4 Extracts

An extract contains a list of the highest-scoring sentences that will be used in the summary. After the top sentences are picked, they are sorted in the order they appear.

We produced a large number of automatic extracts (at 10 target lengths using a number of algorithms of all 20 clusters and of all 18,146 documents in the corpus).

In the following example, we will evaluate one of the summarizers, MEAD using RU. Table 21 presents seven different 10% extracts produced from the same cluster (Cluster 125). An excerpt from the actual judgement scores is shown in Figure 5. As one can see, when all judges are taken into account, one sentence with high salience is sentence 2 from article 19980306_007 with a total utility score of 24. Given that MEAD includes that sentence in its 10% extract, it will get the maximum possible utility for this sentence. On the other hand, not all sentences extracted by MEAD have such a high utility. For example, sentence 3 from 19990802_006 which was also picked by MEAD only carries a utility of 15. If MEAD had picked a different sentence instead (e.g., sentence 2 from 20000408_011 with a utility of 28), its RU would be higher.

In this example, the total self-utility U_1 for judge 1 is 1218. The total self-utilities for judges 2 and 3 are 1380 and 1130, respectively. The values for extractive total utility U_i' for each of the three judges are 237, 218, and 224, respectively.

Table 1 shows the values for extractive cross-judge utility. The average, 0.73, is equal to the interjudge agreement J.

	Judge 1	Judge 2	Judge 3	Average
Judge 1	1.00	0.74	0.74	0.74
Judge 2	0.64	1.00	0.74	0.69
Judge 3	0.72	0.81	1.00	0.77

Table 1: Cross-judge utilities.

Using the formulas in the previous section, one can compute the value

for random performance, which is 0.57.

The performance of MEAD is 0.70 (compared to random = 0.57 and interjudge agreement = 0.73). When normalized, MEAD's performance is 0.80 on a scale from 0 to 1.

3.5 Comparing Relative Utility with P/R

Given an *ideal* extract E_1 consisting of e_1 sentences, one can measure how similar another extract E_2 including e_2 sentences is to it. Precision (P) is the ratio of sentences included in E_2 which are also included in E_1 while Recall (R) is the ratio of sentences included in E_2 to the total number e_1 of sentences in E_1. It can be trivially shown that if $e_1 = e_2$ and the two extracts have a sentences in common, $P = R = a/e$.

Percent agreement (PA) measures how many of the judges' decisions are shared amongst two judges. If d is the number of sentences in the input document (or cluster) that were not extracted by either judge and the input has n sentences, then PA is defined as $(a + d)/n$.

For example, suppose that two judges produce 10% extracts from a document containing 50 sentences. For example, if the same three sentences are extracted by both judges, then $P = R = 3/5 = 60\%; PA = (3 + 43)/50 = 92\%$. PA is known to significantly overestimate agreement (due to the inclusion of non-summary sentences in the evaluation) for both very short and very long extracts while P and R underestimate agreement (due to the Summary Sentence Substitutability principle).

We can now compare the RU values with these for Precision and Recall. Let's first look at judges 1 and 2. Out of 24 sentences, only four overlap between the two judges (19980306_007:2, 19990802_006:8, 19990802_006:9, and
19990829_012:2), or in other words, $P = R = 4/24 = .17$. (Note that when the two extracts are of the same length and the number of sentences that each of them includes is the same, Precision trivially equals Recall). Let's now look at judges 1 and 3. They overlap on only three sentences ($P = R = .13$). Similarly, $P = R = .17$ for judges 2 and 3.

Let's now turn to the performance of MEAD. MEAD has $P = R = 2/24 = .08$ with judge 1. The values for P and R are .13 and .17 when comparing MEAD with judge 2 and judge 3, respectively.

Such low numbers could indicate that it is impossible to reach consensus on extractive summaries. The numbers above are for multi-document extracts, although similar numbers hold for single-document extracts as well. For example, the average interjudge P/R for 10% extracts of each of the ten

	A	B	C	D	E	F	G	H	I	J
R	.648	.650	.652	.465	.626	.727	.508	.497	.644	.566
J	.715	.666	.859	.726	.876	.944	.909	.776	.710	.808

Table 2: Relative Utility - interjudge agreement (J) and random performance (R) for cluster 125, per document, 5% target length.

	A	B	C	D	E	F	G	H	I	J
R	.690	.685	.679	.523	.642	.741	.541	.553	.699	.595
J	.827	.730	.866	.828	.838	.913	.861	.876	.736	.874

Table 3: Relative Utility - interjudge agreement (J) and random performance (R) for cluster 125, per document, 20% target length.

single documents comprising cluster 125 is .22 for judges 1 and 2, .33 for judges 2 and 3, and .26 for judges 3 and 1.

Past work on evaluating extractive summaries [JMBE98, GKMC99] has indicated such low agreement for single-document extracts. We claim that Relative Utility is a better metric than P/R because it does not underestimate agreement in the case where multiple sentences are almost equally useful for an extract and the summarizer has to choose one over the other.

4 Experiments

We ran four experiments to compute Relative Utility values for a number of summarizers at ten summary lengths. We also produced Relative Utility values for a few baselines - lead-based and random summaries.

4.1 Single-document J/R values

In the experiments below, J is the upper bound. R is the lower bound on the performance of an extractive summarizer. Reasonable summarizers are expected to have Relative Utility S in the range between R and J. Note that occasionally (on a particular input and at a particular summary length) a summarizer can score worse than random or better than J. However, when averaging over a number of clusters, these outliers cancel out.

Tables 2, 3 and 4 show how single-document J and R vary by document within a cluster. The first table is for 5% extracts and the second one for 20% extracts.

Tables 12 and 13 show how J and R vary by compression rate in single and multi-document summaries. They also describe the performance of

	A	B	C	D	E	F	G	H	I	J
R	.74	.738	.724	.653	.695	.77	.647	.679	.764	.664
J	.836	.754	.878	.954	.91	.952	.919	.954	.811	.904

Table 4: Relative Utility - interjudge agreement (J) and random performance (R) for cluster 125, per document, 40% target length.

MEAD (S) and another single-document summarizer (WEBS). The value for LEAD is for a lead-based summarizer (that is, a summarizer that only includes the top n% of the sentences of a document or cluster).

4.2 Single-document RU evaluation

We computed J (interjudge agreement), R (random performance), S (system performance), and D (normalized system performance) over all 20 clusters (total = 200 documents). The results are presented in Table 5.

We explored different summarization technologies that work in both single- and multi-document mode. We included two baseline methods in our framework: random summaries (RANDOM, constructed from sentences picked at random from the source) and lead-based summaries (LEAD, produced from sentences appearing at the beginning of the text).

We should note the concept of a random summary produced by picking random sentences given a summary length is different from the idea of R as described above. To produce R, we average over all possible $\binom{n}{e}$ combinations of e sentences out of n where the random summary method produces only *one* such combination. It should be expected, over a large sample, that RANDOM extracts perform as poorly as R and our experiments show that such is indeed the case.

Random summaries should give a lower bound for the performance any system should have, while lead-based summaries give a nice and simple baseline that sometimes obtains very good performance for specific tasks [BMR95]. To provide a basis for comparison, we evaluated WEBSUMM in addition to MEAD.

The single-document results tables compare MEAD with WEBSUMM and the two baselines RANDOM and LEAD.

Several interesting observations can be made by looking at the data in Table 5. First, random performance is quite high although certainly beatable, as shown in Tables 2 and 3. Second, both the lower bound (R) and the upper bound (J) increase with summary length. The average value of R across all documents at the 5% target length is 0.598 while the average

value of J is 0.799. The corresponding values for the 20% target length are: $R = 0.635$ and $J = 0.835$. Third, even though the performances of MEAD and WEBSUMM (S) also increase with summary length, MEAD's normalized version (D) decreases slowly with summary length until the two summarizers score about the same on both S and D for longer summaries. Fourth, for summary lengths of 80% and above, R gets really close to J showing that reasonable summarization that significantly beats random at such summary lengths is quite difficult. Fifth, MEAD outperforms LEAD in lower compression rates. This last observation is very valuable given that some previous studies (e.g., [BMR95]) had indicated that lead-based extracts are at least as good as *more intelligent* extracts. The fact that a public-domain summarizer, not specifically trained for the particular type of documents used in this experiment can outperform LEAD indicates that even though the first few sentences in a document are indeed rather important, there are some other sentences, further down in a document whose utility exceeds that of the sentences in the lead extracts.

			MEAD		RANDOM		LEAD		WEBSUMM	
Percent	J	R	S	D	S	D	S	D	S	D
05	0.80	0.66	0.78	0.88	0.67	0.05	0.72	0.41	0.72	0.44
10	0.81	0.68	0.79	0.84	0.67	-0.02	0.73	0.42	0.73	0.44
20	0.83	0.71	0.79	0.68	0.71	0.01	0.77	0.52	0.76	0.43
30	0.85	0.74	0.81	0.64	0.75	0.10	0.80	0.55	0.79	0.44
40	0.87	0.76	0.83	0.63	0.77	0.03	0.83	0.64	0.82	0.51
50	0.89	0.79	0.85	0.61	0.79	0.01	0.86	0.63	0.85	0.55
60	0.92	0.83	0.88	0.59	0.83	0.02	0.89	0.63	0.87	0.42
70	0.94	0.86	0.91	0.58	0.87	0.08	0.92	0.69	0.90	0.48
80	0.96	0.91	0.93	0.45	0.91	0.05	0.94	0.66	0.93	0.36
90	0.98	0.96	0.97	0.37	0.96	0.04	0.98	0.68	0.97	0.53

Table 5: Single-document Relative Utility.

4.3 Single-document P/R evaluation

It is interesting to compare single-document RU with single-document P/R results. Table 6 shows how P/R varies by summarizer and summary length. For the lengths that make most sense in real life (5-30%), P/R agreement is quite low, both among judges and between systems and judges, whereas RU agreement is much higher.

	J0+J1	J1+J2	J2+J0	ALL JUDGE PAIRS	MEAD	RANDOM	LEAD	WEBSUMM
05	0.22	0.25	0.14	0.20	0.17	0.08	0.30	0.23
10	0.25	0.29	0.25	0.26	0.23	0.12	0.35	0.24
20	0.35	0.37	0.43	0.38	0.34	0.23	0.43	0.32
30	0.46	0.49	0.51	0.49	0.44	0.34	0.49	0.41
40	0.57	0.60	0.59	0.59	0.53	0.43	0.58	0.51
50	0.67	0.68	0.66	0.67	0.62	0.52	0.65	0.59
60	0.75	0.76	0.75	0.76	0.72	0.63	0.73	0.68
70	0.84	0.82	0.83	0.83	0.80	0.74	0.81	0.77
80	0.91	0.89	0.89	0.90	0.87	0.83	0.88	0.85
90	0.96	0.96	0.96	0.96	0.95	0.94	0.96	0.95

Table 6: Single-document Precision/Recall ($P = R$).

4.4 Single-document content-based evaluation

To further calibrate RU results, we compared them with a number of content-based measures [DDM00]. These include word-based cosine between two summaries, word overlap, bigram overlap, and LCS (longest common subsequence). These metrics are all based on the actual text of the extracts (unlike P/R/Kappa/RU, which are all computed on the sentence co-selection vectors). The content-based metrics are described in more detail in[RTS$^+$03].

To compute the content-based scores, we obtained manual abstracts at variable lengths. The comparative results are shown in Tables 7— 10.

Some interesting observations can be made from this comparison. First, the three content-based metrics rank LEAD ahead of both MEAD and WEBSUMM. Second, MEAD and WEBSUMM score approximately the same on all metrics with MEAD doing slightly better on the Word overlap, Bigram overlap, and Longest-common-subsequence measures and WEBSUMM on the cosine metric. Contrasting these findings with the results using RU, one can conclude that RU is somehow better able than the content-based measures in giving proper credit for substitutable sentences that are not lexically similar to the manual extracts.

Percent	LEAD	MEAD	RANDOM	WEBSUMM
10	0.55	0.46	0.31	0.52
20	0.65	0.61	0.47	0.60
30	0.70	0.70	0.60	0.68
40	0.79	0.78	0.69	0.77
50	0.84	0.83	0.75	0.82

Table 7: Similarity between Machine Extracts and Human Extracts. Measure: Cosine.

Percent	LEAD	MEAD	RANDOM	WEBSUMM
10	0.42	0.30	0.22	0.35
20	0.47	0.40	0.31	0.36
30	0.48	0.46	0.41	0.41
40	0.57	0.55	0.47	0.51
50	0.61	0.61	0.52	0.58

Table 8: Similarity between Machine Extracts and Human Extracts. Measure: Word overlap.

Percent	LEAD	MEAD	RANDOM	WEBSUMM
10	0.35	0.22	0.12	0.25
20	0.38	0.31	0.20	0.25
30	0.41	0.37	0.29	0.31
40	0.51	0.46	0.36	0.42
50	0.56	0.53	0.43	0.50

Table 9: Similarity between Machine Extracts and Human Extracts. Measure: Bigram overlap.

Percent	LEAD	MEAD	RANDOM	WEBSUMM
10	0.47	0.37	0.25	0.39
20	0.55	0.52	0.38	0.45
30	0.60	0.61	0.50	0.53
40	0.70	0.70	0.58	0.64
50	0.75	0.76	0.64	0.71

Table 10: Similarity between Machine Extracts and Human Extracts. Measure: longest-common-subsequence.

4.5 Multi-document RU evaluation

In this section, we provide multi-document RU results. Given that MEAD was the only multi-document summarizer available to us, in Table 11 we only include MEAD-specific results, in addition to the two baselines: RANDOM and LEAD.

As one can see from the table, multi-document RU is slightly lower than single-document RU. We believe that this can be explained by the fact that the distribution of scores by the same judge across different articles in the same cluster is not uniform. Some documents contain only a small number of high-utility sentences and contribute to the increase in RU for single-document vs. multi-document. In addition to RU, the lower bound (R) and the upper bound (J) are also slightly lower for multi-document extracts. As

a result, the normalized performance (D) is almost exactly the same in both cases.

			MEAD		RANDOM		LEAD	
Percent	J	R	S	D	S	D	S	D
05	0.76	0.64	0.73	0.81	0.63	-0.08	0.71	0.62
10	0.78	0.66	0.75	0.76	0.65	-0.01	0.71	0.47
20	0.81	0.69	0.78	0.74	0.71	0.15	0.76	0.55
30	0.83	0.72	0.79	0.65	0.72	0.01	0.79	0.67
40	0.85	0.74	0.81	0.62	0.74	-0.06	0.82	0.72
50	0.87	0.77	0.82	0.58	0.79	0.11	0.84	0.70
60	0.88	0.80	0.84	0.52	0.81	0.00	0.86	0.66
70	0.91	0.82	0.86	0.49	0.85	0.06	0.88	0.59
80	0.92	0.84	0.88	0.45	0.89	0.03	0.90	0.55
90	0.93	0.86	0.89	0.36	0.93	-0.04	0.91	0.52

Table 11: Multi-Document Relative Utility

	5	10	20	30	40	50	60	70	80	90
R	.66	.68	.71	.74	.76	.79	.83	.86	.91	.96
RANDOM	.67	.67	.71	.75	.77	.79	.83	.87	.91	.96
WEBSUMM	.72	.73	.76	.79	.82	.85	.87	.90	.93	.97
LEAD	.72	.73	.77	.80	.83	.86	.89	.92	.94	.98
MEAD	.78	.79	.79	.81	.83	.85	.88	.91	.93	.97
J	.80	.81	.83	.85	.87	.89	.92	.94	.96	.98

Table 12: (non-normalized) RU per summarizer and summary length (Single-document)

	5	10	20	30	40	50	60	70	80	90
R	.64	.66	.69	.72	.74	.77	.80	.82	.84	.86
RANDOM	.63	.65	.71	.72	.74	.79	.81	.85	.89	.93
LEAD	.71	.71	.76	.79	.82	.85	.87	.90	.93	.97
MEAD	.73	.75	.78	.79	.81	.82	.84	.86	.88	.89
J	.76	.78	.81	.83	.85	.87	.88	.91	.92	.93

Table 13: (non-normalized) RU per summarizer and summary length (Multi-document)

4.6 Multi-document content-based evaluation

We will now present a short summary of the multi-document content-based evaluation. In Table 14 we show a comparison between the performance of both MEAD and manual extracts (in this case, 50, 100, and 200 words

were chosen for pragmatic reasons - these are the lengths used in the DUC evaluation[DUC00]) when both methods are compared to manual abstracts. Except for the cosine measure, all other metrics show that MEAD's performance is quite comparable to human extracts.

LENGTH	COSINE		OVERLAP		BIGRAM		LCS	
	HUMAN	MEAD	HUMAN	MEAD	HUMAN	MEAD	HUMAN	MEAD
50	0.36	0.17	0.20	0.17	0.06	0.04	0.23	0.20
100	0.44	0.22	0.20	0.17	0.07	0.04	0.25	0.21
200	0.50	0.43	0.20	0.20	0.08	0.07	0.25	0.23

Table 14: (MEAD vs. MANUAL EXTRACTS) compared to MANUAL SUMMARIES

5 Relative Utility with Subsumption

One important property of multi-document summaries that unmodified RU does not address well is subsumption. Unlike sentence substitutability which exists between sentences that are equally worthy of inclusion in a summary but which may be very different in content, subsumption deals with pairs of sentences that have a significant amount of content overlap. In the extreme case, they could be paraphrases of each other or outright copies. It is not hard to realize that sentences with similar content are (a) likely to obtain similar utility scores independently of one another and (b) once one of them is included in a summary, the utility of the other sentence is automatically dropped.

We extended RU to deal with subsumption by introducing conditional sentence utility values [RJB00] which depend on the presence of other sentences in the summary.

Informational subsumption deals with the fact that the utility of a sentence may depend on the utility of other sentences already included in a summary. Two sentences may be almost identical in content and get the same utility scores from a judge and yet they should not be included in the summary at the same time.

Figure 6 shows an example. The sentence extracted from D-19990527-022 (S1) subsumes that from D-19980601-013 (S2), because S1 has the additional information that the hygiene facilities were provided by the "Provisional Regional Council". Since S1 contains all the information provided by S2, an extractive summary selecting both S1 and S2 should be penalized. This has been implemented as an option in our RU system.

RU penalizes summarizers that include subsumed sentences by reducing judge utility scores for those sentences by a parameter α. α takes a value from 0 to 1. When it is 1, subsumed sentences retain their original utility scores. When α is 0, the utility score is 0. The utility scores for sentences that subsume others (S1 in our example) are not modified. In general, the utility score of a subsumed sentence in an extract is reduced by the formula:

$U_{subsumed} = \alpha * U_{orig}$

In our experiment, information subsumption is identified by human judges. This is imaginably very time consuming. [ZOR03] studied methods to automatically identify subsumption, as well as other Cross-document Structural Relationships.

S#	Text	J_1 util	J_2 util	J_3 (of 10)
S1	Two to Four students studying in schools in the New Territories and outlying islands now have a chance to gain more environmental hygiene knowledge through visits to a number of Provisional Regional Council (Pro RC) 's hygiene facilities and participation in a lifeskill training camp during the summer holiday.	9	10	9
S2	Two to Four students studying in schools in the New Territories and outlying islands can now have a chance to gain more knowledge on their environmental hygiene facilities and at the same time take part in a challenging lifeskill training camp during this summer holiday.	5	6	9

Figure 6: An illustration of subsumption from documents D-19990527-022 and D-19980601-013 (in cluster 827)

We obtained subsumption data for 12 clusters and experimented with various α values. Note that since subsumption penalty is carried out for all utility scores, both J and R are recomputed. For example, it may no longer be possible to achieve a very high J if that would cause the inclusion of sentences that subsume one another. Compared with Table 13, where subsumed sentences are not penalized, MEAD and RANDOM both performed significantly better. Tables 15 – 17 illustrate the results of RU with subsumption for different values of α.

6 Experiments on DUC data

We have shown that relative utility gives higher interjudge agreement compared to other metrics. This is a strong evidence that relative utility corre-

			MEAD		RANDOM		LEAD	
Percent	J	R	S	D	S	D	S	D
10	0.77	0.63	0.79	1.47	0.68	0.55	0.68	0.61
20	0.80	0.66	0.81	1.18	0.72	0.55	0.74	0.69
30	0.82	0.69	0.82	1.13	0.74	0.39	0.79	0.88
40	0.84	0.71	0.84	1.26	0.74	0.36	0.82	1.15
50	0.86	0.74	0.86	1.40	0.78	0.42	0.84	1.25

Table 15: Multi-document Relative Utility with subsumption penalty 0.25.

			MEAD		RANDOM		LEAD	
Percent	J	R	S	D	S	D	S	D
10	0.78	0.62	0.86	1.89	0.75	0.92	0.66	0.28
20	0.80	0.64	0.88	1.46	0.76	0.77	0.74	0.62
30	0.83	0.67	0.88	1.42	0.78	0.67	0.80	0.84
40	0.85	0.70	0.90	1.52	0.77	0.57	0.83	0.99
50	0.86	0.73	0.92	1.65	0.80	0.55	0.85	1.01

Table 16: Multi-document Relative Utility with subsumption penalty 0.5.

			MEAD		RANDOM		LEAD	
Percent	J	R	S	D	S	D	S	D
10	0.84	0.65	1.03	1.90	0.90	1.13	0.64	0.23
20	0.82	0.63	0.97	1.60	0.82	0.96	0.74	0.57
30	0.84	0.66	0.97	1.69	0.84	0.81	0.82	0.83
40	0.86	0.69	1.00	1.76	0.81	0.73	0.85	0.98
50	0.88	0.72	1.01	1.94	0.82	0.75	0.87	0.98

Table 17: Multi-document Relative Utility with subsumption penalty 0.75.

lates better with human judgements. However, since the only summarizer systems available to us are MEAD and WEBSUMM, the results in Section 4 are not enough to conclude that relative utility is a better metric than the others in this sense. In this section, we compare relative utility with other metrics used in evaluating summaries.

6.1 Data sets and automatic sentence utility judgements

DUC data is perfectly suitable for our purpose since it includes many automatic summaries by different participant systems and human judge rankings of these systems. We used the generic multi-document summarization tasks of DUC 2003 and 2004 in our experiments. However, there is no manual utility scoring for each sentence in DUC, which is a burden against computing relative utility. To get sentence utilities, we applied the automatic sentence scoring algorithm described in [ROQT03]. In this method, manual

abstracts are used to score the sentences. Utility for a sentence is computed by looking at how similar the sentence is to the manual abstracts. We used cosine similarity for this purpose although the idea can be extended to any similarity metric.

6.2 Correlations of different metrics against human judgements

A total of 18 participant systems and 10 human summarizers ranked in DUC 2003, and 17 participant systems and 8 human summarizers in DUC 2004. Each human summarizer is also judged by other humans and placed in the ranking. Table 18 shows the Spearman rank order coefficients of DUC 2003 and 2004 multi-document summarization data between human rankings and different automatic content-based metrics. We also include BLEU [PRWZ01], which is a widely used evaluation metric among the machine translation community.

	Cosine	Word	Bigram	LCS	BLEU
DUC 2003	0.822	0.877	0.914	0.902	0.865
DUC 2004	0.754	0.878	0.803	0.839	0.804

Table 18: Spearman rank order correlation coefficients of DUC multi-document summarization data between human rankings and some automatic content-based evaluation metrics (in order: cosine, word overlap, bigram overlap, longest common subsequence, and BLEU).

Multi-document summaries are bounded by 100 words in DUC 2003 and 665 bytes in DUC 2004, which correspond to 2-5% of the document clusters. There are 30 document clusters in DUC 2003, and 50 clusters in DUC 2004, with 10 documents in each cluster. To get a better comparison with the content-based metrics, we also produced 2% extracts from the automatically created sentence utilities as well as 5%, 10%, and 20% extracts. Table 19 shows the the Spearman rank order coefficients between human rankings and different extractive evaluation metrics. RU gives higher correlation in all cases compared to P/R and Kappa. In comparison with the content-based metrics, RU correlates with human judgements as well as other metrics on DUC 2004. However, it is hard to say that this is the case for DUC 2003. There are at least three reasons for RU's worse performance. First of all, interjudge agreement in DUC 2003 is lower than it is in DUC 2004 (Table 20). Considering the judges are the same individuals as the manual summarizers, this may result in inconsistent rankings among differ-

ent judges. Second, our method to produce extracts from sentence utility scores is merely taking the sentences with the highest score. Since we do not consider information subsumption among the selected sentences, the extract may suffer from repeated information. This makes a crucial effect on human rankings, which are based on *coverage* of the summaries with respect to the manual summaries. Finally, our automatic sentence scoring algorithm is not as perfect as human scoring, which clearly effects the accuracy of RU. Last two reasons apply for the DUC 2004, too, which means that we could have even higher correlation if we had human sentence utility scores and used RU with subsumption.

	DUC 2003			DUC 2004		
Percent	P/R	Kappa	RU	P/R	Kappa	RU
2	0.664	0.663	0.718	0.780	0.782	0.826
5	0.737	0.743	0.761	0.844	0.844	0.882
10	0.723	0.726	0.753	0.827	0.827	0.868
20	0.795	0.789	0.801	0.812	0.789	0.845

Table 19: Spearman rank order correlation coefficients of DUC multi-document summarization data between human rankings and some automatic extractive evaluation metrics (in order: Precision/Recall, Kappa, and relative utility).

	Percentage			
	02	05	10	20
DUC 2003	0.647	0.705	0.743	0.796
DUC 2004	0.715	0.740	0.770	0.810

Table 20: Relative Utility - average interjudge agreement (J) for DUC multi-document summarization data.

7 Conclusions and Future work

Since interjudge agreement measured by Precision, Recall, and percent agreement are quite low for extractive summaries, it is practically impossible to write summarizers which are optimized for these measures. Relative Utility provides an intuitive mechanism which takes into account the fact that even though human judges may disagree on exactly which sentences belong in a summary, they tend to agree on the overall salience of each sentence. By

moving from binary decisions to variable-membership decisions, it is possible to catch that agreement and produce better summarizers.

Relative Utility has several additional advantages over P/R/PA. First, in a way similar to Kappa [SC88], it takes into account the difficulty of a problem by factoring in random and interjudge performance.

Second (and unlike Kappa), it can be used for evaluation at multiple compression rates (summary lengths). In one pass, judges assign salience scores to all sentences in a cluster (or in a single document). It is then possible to simulate extraction at a fixed compression rate by ranking sentences by utility. As a result, RU is a more informative measure of sentence salience than the alternative metrics.

Third, the RU method can be further expanded to allow sentences or paragraphs to exert negative reinforcement on one another, that is, allow for cases in which the inclusion of a given sentence makes another redundant and a system that includes both will be penalized more than a system which only includes one of the two "equivalent" sentences and another, perhaps less informative sentence.

In current work, we are investigating the connection between RU, subsumption and the taxonomy of cross-document relationships (such as paraphrase, follow-up, elaboration, etc.) set forth in Cross-Document Structure Theory (CST) [Rad00, ZBGR02].

The subsumption-based RU model will need further adjustment to address sentences which mutually increase their importance. For example, sentences with anaphoric expressions (e.g., "He then said...") will have a higher utility if the sentence containing the antecedent of the anaphora is also included.

Finally, we need to mention that the use of Relative Utility is not limited to the evaluation of sentence extracts. We will investigate its applicability to other evaluation tasks, such as ad-hoc retrieval and word sense disambiguation. One particularly promising area of application is in the evaluation of non-extractive summaries. In recent DUC conferences, abstractive summaries have been evaluated using model unit recall (also known as MLAC = mean length-adjusted coverage.) In this model, human reference summaries are split into atomic content pieces called model units. Example model units could be "Teachers went on strike in France" or "Two new SARS cases have been reported in Hong Kong". The current DUC evaluation measures recall only when the right model unit is included in the system summary. We will investigate assigning relative utility scores to model units in order to capture fact salience.

8 Acknowledgments

This work was partially supported by the National Science Foundation's Information Technology Research program (ITR) under grant IIS-0082884. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the participants, and do not necessarily reflect the views of the National Science Foundation.

We would also like to thank the CLAIR (Computational Linguistics And Information Retrieval) group at the University of Michigan and, more specifically, Adam Winkel, Sasha Blair-Goldensohn, Jahna Otterbacher, Naomi Daniel, and Timothy Allison for useful feedback.

References

[BMR95] Ron Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[DDM00] R.L. Donaway, K.W. Drummey, and L.A. Mather. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, pages 69–78. Association for Computational Linguistics, 30 April 2000.

[DUC00] *Proceedings of the Workshop on Text Summarization (DUC 2000)*, New Orleans, LA, 2000.

[GKMC99] Jade Goldstein, Mark Kantrowitz, Vibhu O. Mittal, and Jaime G. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *SIGIR 1999*, pages 121–128, Berkeley, California, 1999.

[JMBE98] Hongyan Jing, Kathleen McKeown, Regina Barzilay, and Michael Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 60–68, Standford (CA), USA, March 23-25 1998. The AAAI Press.

[Luh58] H.P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958.

[MB99] Inderjeet Mani and Eric Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67, 1999.

[PRWZ01] K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. Blue: A Method for Automatic Evaluation of Machine Translation. Research Report RC22176, IBM, 2001.

[Rad00] 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.

[RJB00] 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.

[ROQT03] Dragomir R. Radev, Jahna Otterbacher, Hong Qi, and Daniel Tam. Mead reducs: Michigan at duc 2003. In *Proceedings of DUC 2003*, Edmonton, AB, Canada, 2003.

[RTS+03] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. Evaluation challenges in large-scale multi-document summarization: the mead project. In *Proceedings of ACL 2003*, Sapporo, Japan, 2003.

[SC88] Sidney Siegel and N. John Jr. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition, 1988.

[ZBGR02] Zhu Zhang, Sasha Blair-Goldensohn, and Dragomir Radev. Towards CST-enhanced summarization. AAAI 2002, August 2002.

[ZOR03] Zhu Zhang, Jahna Otterbacher, and Dragomir R. Radev. Learning cross-document structural relationships using boosting. In *Proceedings of ACM CIKM 2003*, New Orleans, LA, November 2003.

MEAD	LEAD	RANDOM	JUDGE1	JUDGE2	JUDGE3	ALLJUDGES
19980306_007:2	19980306_007:1	19980306_007:4	**19980306_007:2**	19980306_007:1	19980306_007:15	**19980306_007:2**
19980306_007:15	**19980306_007:2**	19980306_007:6	19980306_007:3	**19980306_007:2**	19980306_007:17	19980306_007:15
19980306_007:26	19980430_016:1	19980306_007:19	19980306_007:4	19980306_007:18	19980430_016:1	19980430_016:13
19980306_007:27	19980430_016:2	19980306_007:22	19980306_007:6	19990425_009:1	19980430_016:2	19980430_016:16
19980430_016:17	19990211_009:1	19980430_016:1	19980306_007:7	19990425_009:2	19980430_016:13	19990425_009:1
19980430_016:20	19990211_009:2	19980430_016:3	19980306_007:9	19990729_008:12	19980430_016:14	19990425_009:2
19980430_016:38	19990218_009:1	19980430_016:20	19980306_007:11	19990802_006:2	19980430_016:16	19990425_009:3
19990211_009:2	19990218_009:2	19980430_016:24	19980306_007:12	19990802_006:6	19980430_016:17	19990425_009:7
19990211_009:4	19990218_009:3	19980430_016:42	19980306_007:13	19990802_006:8	19980430_016:19	19990425_009:8
19990211_009:6	19990425_009:1	19990218_009:14	19990425_009:7	19990802_006:9	19990211_009:3	19990729_008:8
19990218_009:4	19990425_009:2	19990425_009:18	19990425_009:10	19990802_006:13	19990218_009:2	19990802_006:8
19990425_009:2	19990425_009:3	19990729_008:4	19990802_006:7	19990802_006:16	19990218_009:4	19990802_006:9
19990425_009:6	19990729_008:1	19990729_008:13	19990802_006:8	19990829_012:1	19990425_009:1	19990802_006:10
19990425_009:7	19990729_008:2	19990802_006:19	19990802_006:9	19990829_012:2	19990425_009:3	19990802_006:13
19990425_009:9	19990802_006:1	19990802_006:23	19990802_006:10	19990927_011:1	19990425_009:8	19990802_006:16
19990425_009:13	19990802_006:2	19990829_012:16	19990829_012:2	19990927_011:2	19990425_009:12	19990829_012:2
19990729_008:3	19990829_012:1	19990927_011:11	19990829_012:5	19990927_011:10	19990729_008:8	19990829_012:6
19990729_008:8	19990829_012:2	19990927_011:14	19990829_012:6	19990927_011:11	19990802_006:13	19990829_012:13
19990729_008:13	19990927_011:1	19990927_011:18	19990829_012:12	19990927_011:12	19990829_012:2	19990927_011:11
19990802_006:3	19990927_011:2	19990927_011:21	19990829_012:13	19990927_011:13	19990829_012:6	19990927_011:12
19990802_006:16	19990927_011:3	19990927_011:26	19990927_011:4	19990927_011:18	19990829_012:13	20000408_011:1
19990802_006:17	20000408_011:1	20000408_011:15	19990927_011:5	19990927_011:20	19990927_011:14	20000408_011:2
19990829_012:7	20000408_011:2	20000408_011:20	19990927_011:6	19990927_011:21	20000408_011:13	20000408_011:4
19990927_011:9	20000408_011:3	20000408_011:21	20000408_011:2	20000408_011:1	20000408_011:15	20000408_011:5

Table 21: Seven 10% extracts (document-id:sentence-id) produced from the same cluster. Note: order within a column is not relevant.