MEAD ReDUCs: Michigan at DUC 2003

Dragomir R. Radev[1,2], **Jahna Otterbacher**[1], **Hong Qi**[1], **and Daniel Tam**[2]
[1]School of Information, [2]Department of EECS
University of Michigan
{radev,jahna,hqi,dtam}@umich.edu

Abstract

We present the results of Michigan's participation in DUC 2003. Using mean length-adjusted coverage, we obtained the best score of all systems on task 4 - question-focused summaries.

1 Introduction

The year 2003 marked the third time the University of Michigan's CLAIR (Computational Linguistics And Information Retrieval) group participated in the DUC evaluation. We entered our system, MEAD (Radev et al., 2001; Radev et al., 2002), in three of the four tasks (2, 3, and 4). We did not participate in Task 1 ("Very short summaries"). The performance of our system (System 23 in the official result sheet) was quite good - using MLAC (mean length-adjusted coverage) as the primary evaluation metric, we ranked first of nine teams on Task 4 ("Short summaries in response to a question", fourth out of 16 teams on task 2 ("Short summaries focused by events", and fifth out of 11 teams on task 3 ("Short summaries focused by viewpoints").

In this report we will describe our general approach to the different tasks, paying particular attention to the ways in which we adapted our existing extractive summarizer, MEAD, to perform different tasks.

2 The DUC 2003 evaluation

The DUC 2003 evaluation included a new corpus and a new set of tasks. The corpus included three sets of documents (30 from the TREC Ad Hoc task, 30 from TDT, and 30 from the TREC Novelty task). The goal was to produce short summaries (10 words for Task 1 and 100 words for Tasks 2, 3, and 4). Unlike the two previous evaluations, DUC 2003 doesn't evaluate single-document summaries.

The goal of Task 1 is to produce a 10 word summary (headline) from the 30 TDT and 30 TREC clusters. Entries are evaluated for coverage and usefulness[1].

In Task 2, the input includes the 30 TDT clusters. The event-focused summaries are evaluated for quality and length-adjusted coverage. Length-adjusted coverage is based on the concept of *model units* which are either short phrases or entire sentences and which carry the most salient content of a document. The output of the automatic (peer) summarizers is decomposed into *peer units* (PU) which are then compared to the model units (MU) using traditional information retrieval measures such as precision and recall. A DUC-specific measure is the length-adjusted coverage (recall) which combines MU recall and peer length into one formula.

Task 3 deals with viewpoint-focused summaries. A viewpoint is a natural language string no longer than one sentence. Systems participating in Task 3 are evaluated using quality and length-adjusted coverage.

Finally, Task 4 deals with question-focused summaries. They are evaluated using the same measures as Tasks 2 and 3.

3 Evaluation measures

Here follows the full list of metrics used in the DUC 2003 official evaluation results.

Count of quality questions with non-0 answers: Simply the number of the 12 quality questions with scores > 0

Mean of the quality question scores: Average of the 12 scores each with value 0–3

Number of peer units: Number of rough sentences in the peer

Number of marked peer units: Number of peer units that the assessor felt expressed at least some of the meaning of the model

[1] `http://duc.nist.gov`

Number of unmarked peer units: Number of peer units that the assessor felt did not express any of the meaning of the model

Number of model units: The number of roughly elementary discourse units (e.g., clauses etc) in the model

Mean coverage: As indicated in the protocol, the assessor judges the coverage by the peer summary of each unit in the model. This is the mean of those coverage scores.

Median coverage: median of the per-model-unit coverage scores

Sample std of coverage scores sample standard deviation of the per-model-unit coverage scores

The following three length-adjusted measures are the analogous statistics but for a coverage score that emphasizes brevity as well as coverage. The coverage scores will be rewarded if the system produces a summary shorter than the predefined target length; penalized otherwise.

Mean length-adjusted coverage

Median length-adjusted coverage

Sample std of adjusted coverage scores

For each peer summary, the official evaluation results also include the evaluator's judgements for 12 questions (Table 1).

Q1: About how many gross capitalization errors are there?
Q2: About how many sentences have incorrect word order?
Q3:About how many times does the subject fail to agree in number with the verb?
Q4: About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) - causing the sentence to be ungrammatical, unclear, or misleading?
Q5: About many times are unrelated fragments joined into one sentence?
Q6: About how many times are articles (a, an, the) missing or used incorrectly?
Q7: About how many pronouns are there whose antecedents are incorrect, unclear, missing, or come only later?
Q8: For about how many nouns is it impossible to determine clearly who or what they refer to?
Q9: About how times should a noun or noun phrase have been replaced with a pronoun?
Q10: About how many dangling conjunctions are there ('and', 'however'...)?
Q11: About many instances of unnecessarily repeated information are there?
Q12: About how many sentences strike you as being in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship, or just don't fit in topically with neighboring sentences?

Figure 1: The twelve *quality* questions used in DUC.

4 Our system

We used the latest version of the MEAD system[2] augmented with a number of new rerankers. For a detailed discussion of MEAD, we refer the reader to (Radev et al., 2001). Suffice it to say that MEAD is an extractive summarization environment based on a three-step architecture. During the first step *the feature extractor*, each

[2] `http://www.summarization.com`

```
Centroid 1 Position 1 LengthCutoff 9 SimWithFirst 2
QueryTitleWordOverlap 1 QueryDescCosineNoIDF 0.5

mmr-reranker-word.pl 0.5 MEAD-cosine enidf
```

Figure 2: Sample MEAD policy.

sentence in the input document (or cluster of documents) is converted into a feature vector using features such as Position, Centroid, Length, OverlapWithFirst, etc. Second, the feature vector is converted to a scalar value using the *combiner*. At the last stage known as the *reranker*, the scores for sentences included in related pairs are adjusted upwards or downwards based on the type of relation between the sentences in the pair. Generally speaking, a *negative* relation exists between sentences that overlap in content (e.g., sentence pairs exhibiting subsumption or paraphrase) and therefore the presence of one of them in the summary should suppress the other one, while sentence pairs are related *positively* if the presence of one of them requires the presence of the other (e.g., due to an anaphoric relationship between them). The third stage of the MEAD architecture is based on custom *rerankers* which adjust the sentence scores assigned by the first and second stages. We used several rerankers in our experiments. Some of them (e.g., Maximal Marginal Relevance, MMR), are based on work by others (Carbonell and Goldstein, 1998) while others are based on our CST theory (Radev, 2000).

A MEAD policy is a combination of three components: (a) the command lines for all features, (b) the formula for converting the feature vector to a scalar, and (c) the command line for the reranker. A sample policy might be the one shown in Figure 2. This example indicates the four features used (Centroid, Position, LengthCutoff, and SimWithFirst), their relative weights (except for LengthCutoff where the number 9 indicates the threshold for selecting a sentence based on length), and the reranker (in this example, word-based MMR with a similarity threshold computed as the cosine between two sentences).

4.1 Converting manual abstracts to sentjudge files

To evaluate performance during training and testing, we used primarily the Relative Utility metric (RU) (Radev et al., 2000). RU allows for summaries to be automatically evaluated at different compression rates using a single set of judgement values on a scale of 0 (unimportant) to 10 (central) for each sentence in the input set. For example, if we want to evaluate an automatically produced extract using relative utility, one would need to have a *sentjudge* file for the same input documents. This technique was very successful in the presence of actual sentence judgements (Radev et al., 2003). In DUC we didn't have such judgements but instead we produced them automatically from the manual abstracts provided by the DUC evalua-

tors.

For our purpose, we compared each sentence against the multi-document abstract to get a reasonable approximation for sentence utility scores, and thus produced a set of automatic sentjudge files for each cluster using the manual abstracts available on DUC 2003 web site.

```
Storms blew ships onto rocks off the Shetlands
and in the Sumburgh-Fair Isle channel. Off
northwest Spain, fog sent a ship onto rocks,
and Spanish and Romanian freighters sank in
heavy seas.  Hurricanes sank a gold cargo in
1857 and a Buenos Aires steamer in 1915.
Hurricane Bertha destroyed freighter Corazon.
The Derbyshire vanished in a typhoon. Storms
caused the sinking of ships in the North Sea,
off Cornwall, in an Algerian port, at Drakes
Bay, the ferry Estonia in the Baltic, a
Liberian freighter, and a Korean fishing boat.
A North Atlantic gale damaged German freighter
Yarrowanga. A storm-damaged Saudi tanker was
left ablaze in the Strait of Hormuz.
```

Figure 3: The D448.M.100.N.H multi-document abstract.

```
<DOCSENT DID="LA122589-0032">
<BODY>
<TEXT>
<S SNO="1" PAR="1" RSNT="1">Two seamen were
missing after a Romanian freighter sank off
Spain's northwest coast Sunday, Spanish rescue
services said.</S>
<S SNO="2" PAR="1" RSNT="1">A spokesman said
the Tololovem went down after being smashed
by high waves off Cape Finisterre.</S>
</TEXT>
</BODY>
</DOCSENT>
```

Figure 4: A sample docsent: LA122589-0032.docsent.

Figure 3 shows the multi-document abstract over cluster D448 by judge H. A document in MEAD docsent format from the cluster D448, LA122589-0032, is shown is Figure 4. For each sentence in this docsent file, we compared it against the manual abstract and computed the utility score for the sentjudge file. Figure 5 shows a piece of the sentjudge file for cluster D448.

4.2 Training

For training, we used the DUC 2002 corpus and evaluations. We ran MEAD using a large number of policies and chose the one that achieved the highest RU on DUC 2002. The policy shown in Figure 2 was used for Task 4. For Task 3, we used the same policy minus the last feature on the list (QueryDescCosineNoIDF) and for Task 2, we used the same policy as for Task 3 except for the QueryTitleWordOverlap feature.

```
<SENT-JUDGE QID="448">
...
 <S DID="LA122589-0032" SNO="1" PAR="1" RSNT="1">
    <JUDGE N="H" UTIL="3.2651" />
    ...
  </S>
  <S DID="LA122589-0032" SNO="2" PAR="1" RSNT="1">
    <JUDGE N="H" UTIL="0.7186" />
 ...
  </S>
...
</SENT-JUDGE>
```

Figure 5: D448.sentjudge automatically generated from the corresponding manual abstract. In this example one can see that the first sentence of document LA122589-0032 is more related to the manual abstract than the next sentence.

5 Results

In this section we will present two sets of results. First, the official numbers and then some results using relative utility on non-DUC data. Due to a lack of space, we omitted some experimental results with various policies. These results will be added to a longer version of this paper.

5.1 Official results

The official mean length-adjusted coverage (MLAC) results on Tasks 2, 3, and 4 are shown in Table 1. The letters A–J represent the manual summaries while the numbers 2–6 stand for five different baseline systems. The rest of the numbers (10–26) correspond to the different peer systems which entered the evaluation. Our system (number 23) scored higher than all peer systems on Task 4 (MLAC=0.1367) and finished in the top five on Tasks 2 (MLAC=0.1756) and 3 (MLAC=0.1122).

Tables 2, 3, and 4 show the official P/R results.

On task 4, our system got the best score of all systems on questions 6, 7, 9, and 10. It was in the top three on questions 8, 11, and 12.

On task 2, our system achieved its best performance (tied for first) on question 6. It was also in the top three on questions 4, 7, and 12.

On task 3, we finished among the top 3 systems on questions 2, 6, 7, and 8. Tables 9–11 appendix contains the full official results of Tasks 2, 3, and 4.

5.2 Non-DUC data

The following tables show our performance on some non-DUC clusters. The clusters are described in Table 5 while the overall performance is shown in Tables 6, 7 and 8. The results for Task 3 are omitted due to a lack of space.

Summarizer	Task 2	Task 3	Task 4
2	0.0906	0.0790	-
3	0.1168	0.1014	-
4	-	-	0.1132
5	-	-	0.1259
6	0.1824	-	-
10	0.1452	0.1283	0.1038
11	0.1490	0.1152	-
12	0.1436	-	-
13	0.1890	0.0848	0.1073
14	0.1747	-	0.1342
15	0.0552	0.0713	-
16	0.1792	0.1206	0.1214
17	0.0984	0.1106	0.0844
18	0.1519	0.1225	-
19	0.0996	-	0.0479
20	0.1655	0.1233	0.0850
21	0.1258	0.0985	-
22	0.1779	0.1282	0.1279
23	0.1756	0.1122	0.1367
26	0.1429	-	-
A	0.3170	0.2209	0.2002
B	0.3189	0.2462	0.3227
C	0.3262	0.2551	0.2889
D	0.3508	0.2878	0.2272
E	0.2799	0.2086	0.2450
F	0.2540	0.2012	0.1805
G	0.2811	0.2048	0.2199
H	0.3609	0.1933	0.2810
I	0.2944	0.1956	0.2901
J	0.2552	0.1787	0.2311

Table 1: Mean length-adjusted coverage (MLAC), tasks 2, 3, and 4.

Summarizer	Precision	Recall	Precision Rank	Recall Rank
6	0.7965	0.2941	4	2
10	0.8442	0.2124	2	12
11	0.7945	0.1895	7	15
12	0.7177	0.2908	10	3
13	0.7959	0.2549	5	7
14	0.7604	0.2386	9	8
15	0.2735	0.2190	16	11
16	0.7905	0.2712	8	4
17	0.6701	0.2124	14	13
18	0.6810	0.2582	12	5
19	0.6737	0.2092	13	14
20	0.8667	0.2549	1	6
21	0.6869	0.2222	11	10
22	0.8361	0.1667	3	16
23	0.7955	0.2288	6	9
26	0.6601	0.3301	15	1

Table 2: Precision/Recall in task 2.

Summarizer	Precision	Recall	Prec. Rank	Rec. Rank
10	0.736842105	0.180645161	2	10
11	0.731343284	0.158064516	3	11
13	0.663551402	0.229032258	7	2
15	0.265560166	0.206451613	11	4
16	0.71559633	0.251612903	4	1
17	0.638297872	0.193548387	8	9
18	0.558558559	0.2	10	7
20	0.708333333	0.219354839	5	3
21	0.567567568	0.203225806	9	6
22	0.75308642	0.196774194	1	8
23	0.670212766	0.203225806	6	5

Table 3: Precision/Recall in task 3.

Summarizer	Precision	Recall	Prec. Rank	Rec. Rank
10	0.695652174	0.182509506	2	8
13	0.548387097	0.19391635	6	5
14	0.632478632	0.281368821	4	2
16	0.672413793	0.305882353	3	1
17	0.542553191	0.19391635	7	6
19	0.290322581	0.136882129	9	9
20	0.485436893	0.190114068	8	7
22	0.613207547	0.247148289	5	3
23	0.833333333	0.209125475	1	4

Table 4: Precision/Recall in task 4.

6 Acknowledgments

This work was partially supported by the National Science Foundation's Information Technology Research program (ITR) under grant IIS-0082884.

References

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April.

Dragomir Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multi-document summarization using MEAD. In *First Document Understanding Conference*, New Orleans, LA, September.

Dragomir R. Radev, Michael Topper, and Adam Winkel. 2002. Multi document centroid-based text summarization. In *ACL Demo Session*, Philadelphia, PA.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale multi-document summarization. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, July.

Dragomir Radev. 2000. 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.

Data	Task 2 RU	Task 2 Norm. RU	Task 3 RU	Task 3 Norm. RU	Task 4 RU	Task 4 Norm. RU
DUC02	0.4438	0.4506	0.5159	0.6307	0.5160	0.6307
Gulfair	0.5040	0.0136	0.6700	0.4046	0.6700	0.4046
HKNews	0.7255	0.6436	0.8179	1.6841	0.8179	1.6841
All	0.6547	0.7569	0.74582	1.357	0.74582	1.357

Table 6: Average performances over the three tasks

Cluster	Ave. Judge Performance	Expected Random	Leadbased RU	Leadbased Norm RU	RU	Norm RU
d061j	0.8175	0.4447	0.4039	-0.1239	0.5820	0.3684
d062j	0.6680	0.1424	0.1522	-0.0040	0.3602	0.4144
d063j	0.8813	0.3064	0.4708	0.3263	0.6438	0.5869
d064j	0.3986	0.1366	0.3199	0.3717	0.5776	1.6836
d065j	0.6611	0.3218	0.2640	-0.1783	0.0556	-0.8004
Gulfair	0.9090	0.4984	0.1461	-0.8905	0.5040	0.0136
1014	0.8000	0.5792	0.7889	0.9497	0.7778	0.8994
112	0.7570	0.6680	0.5314	-2.2001	0.6791	0.1257
1197	0.7125	0.5059	0.7222	1.0835	0.8083	1.4638
125	0.8083	0.5356	0.7998	1.3882	0.7500	0.7861
199	0.8213	0.6633	0.6722	0.0604	0.8944	1.4629
241	0.3890	0.5110	0.5270	0.2274	0.6055	NA
323	0.6000	0.6099	0.6576	2.4493	0.7167	NA
398	0.6833	0.5961	0.6681	0.6511	0.8833	3.2928
447	0.7924	0.6425	0.6565	0.0229	0.9208	1.8573
46	0.6454	0.6606	0.7556	NA	0.7741	NA
551	0.8000	0.6366	0.7222	0.5424	0.2333	-2.4684
60	0.8606	0.6376	0.6867	0.2285	0.6694	0.1427
61	0.6840	0.5285	0.6028	0.3792	0.6139	0.5490
62	0.7498	0.6133	0.7863	1.0444	0.5976	-0.1150
827	0.8769	0.6822	0.6390	-0.2555	0.8788	1.0100
883	0.7015	0.6459	0.4833	-1.5944	0.7977	2.7287
885	0.7250	0.6643	0.7444	1.0000	0.7333	1.1373

Table 7: MEAD and Leadbased Performance on Task 2

Cluster	Ave. Judge Performance	Expected Random	Leadbased RU	Leadbased Norm RU	RU	Norm RU
d061j	0.8217	0.4465	0.4039	-0.1239	0.7406	0.7839
d062j	0.6680	0.1424	0.1522	-0.0040	0.3602	0.4144
d063j	0.8813	0.3064	0.4708	0.3263	0.7375	0.7499
d064j	0.3986	0.1366	0.3199	0.3717	0.5776	1.6836
d065j	0.6611	0.3248	0.2640	-0.1783	0.1639	-0.4784
Gulfair	0.9244	0.4970	0.1461	-0.8905	0.6700	0.4046
1014	0.8000	0.5792	0.7889	0.9497	0.7556	0.7987
112	0.7570	0.6680	0.5314	-2.2001	0.8739	2.3140
1197	0.7125	0.5059	0.7222	1.0835	0.8667	1.7461
125	0.8083	0.5356	0.7998	1.3882	0.9176	1.3972
199	0.8213	0.6633	0.6722	0.0604	0.7370	0.4668
241	0.4469	0.5183	0.5270	0.2274	0.7179	NA
323	0.6000	0.6099	0.6576	2.4493	0.7167	NA
398	0.6833	0.5961	0.6681	0.6511	0.9333	3.8661
447	0.8057	0.6498	0.6565	0.0229	0.8516	1.2947
46	0.6454	0.6606	0.7556	NA	0.7556	NA
551	0.8000	0.6366	0.7222	0.5424	0.8333	1.2040
60	0.8606	0.6376	0.6867	0.2285	0.8077	0.7626
61	0.6840	0.5285	0.6028	0.3792	0.9014	2.3977
62	0.7498	0.6133	0.7863	1.0444	0.6925	0.5799
827	0.8769	0.6822	0.6390	-0.2555	0.9131	1.1857
883	0.7015	0.6459	0.4833	-1.5944	0.8310	3.3279
885	0.7250	0.6643	0.7444	1.0000	0.8000	2.2358

Table 8: MEAD and Leadbased Performance on Task 4

Peer summarizer code (baseline[1-5], manual[A-J], system submission[6-26])	Count of quality questions with non-0 answers	Mean number of quality questions with non-0 answers	Q1 (0 = 0, 1 = 1-5, 2 = 6-10, 3 = 11 or more)	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Number of peer units	Number of marked peer units	Number of unmarked peer units	Number of model units	Mean coverage	Median coverage	Sample std of coverage scores	Mean length-adjusted coverage	Median length-adjusted coverage	Sample std of adjusted coverage scores
2	1.33	0.77	0.70	0.03	0.03	0.10	0.20	0.00	0.03	0.10	0.00	0.00	0.07	0.07	4.10	2.13	1.97	10.20	0.16	0.01	0.26	0.09	0.00	0.15
3	2.90	1.03	1.00	0.10	0.03	0.13	0.23	0.00	0.07	0.17	0.10	0.00	0.63	0.50	3.17	2.30	0.87	10.20	0.19	0.03	0.30	0.12	0.02	0.18
6	1.40	0.82	0.00	0.00	0.07	0.07	0.10	0.00	0.07	0.20	0.07	0.00	0.37	0.50	3.77	3.00	0.77	10.20	0.28	0.14	0.35	0.18	0.09	0.23
10	1.83	0.90	0.23	0.03	0.00	0.13	0.07	0.00	0.17	0.33	0.03	0.10	0.37	0.40	2.57	2.17	0.40	10.20	0.19	0.02	0.29	0.15	0.03	0.22
11	0.97	0.53	0.07	0.03	0.07	0.10	0.03	0.00	0.00	0.17	0.00	0.07	0.17	0.27	2.43	1.93	0.50	10.20	0.20	0.06	0.29	0.15	0.04	0.22
12	1.60	0.80	0.10	0.07	0.00	0.07	0.03	0.03	0.03	0.40	0.03	0.03	0.07	0.73	4.13	2.97	1.17	10.20	0.24	0.11	0.33	0.14	0.06	0.19
13	3.20	1.03	0.93	0.17	0.07	0.20	0.40	0.00	0.13	0.23	0.07	0.00	0.50	0.60	3.27	2.60	0.67	10.20	0.29	0.12	0.36	0.19	0.08	0.24
14	2.53	0.92	0.87	0.13	0.07	0.17	0.33	0.07	0.03	0.23	0.03	0.00	0.13	0.57	3.20	2.43	0.77	10.20	0.25	0.09	0.33	0.17	0.07	0.23
15	7.87	1.33	0.57	1.33	1.47	1.57	1.00	0.43	0.87	1.10	0.17	0.03	0.33	1.67	8.17	2.23	5.93	10.20	0.08	0.01	0.15	0.06	0.01	0.10
16	1.27	0.83	0.03	0.00	0.00	0.07	0.07	0.00	0.03	0.10	0.03	0.00	0.33	0.60	3.50	2.77	0.73	10.20	0.28	0.13	0.35	0.18	0.09	0.23
17	1.80	0.92	0.00	0.03	0.03	0.10	0.00	0.03	0.33	0.37	0.00	0.03	0.13	0.77	3.23	2.17	1.07	10.20	0.17	0.03	0.26	0.10	0.02	0.15
18	2.57	0.93	0.43	0.10	0.07	0.13	0.10	0.00	0.33	0.43	0.00	0.00	0.20	0.77	3.87	2.63	1.23	10.20	0.22	0.07	0.31	0.15	0.05	0.22
19	2.50	0.97	0.47	0.07	0.03	0.17	0.27	0.10	0.00	0.47	0.03	0.03	0.20	0.73	3.17	2.13	1.03	10.20	0.14	0.05	0.22	0.10	0.03	0.16
20	2.33	1.08	1.20	0.03	0.03	0.03	0.27	0.00	0.03	0.13	0.03	0.00	0.30	0.50	3.00	2.60	0.40	10.20	0.26	0.18	0.31	0.17	0.11	0.20
21	2.67	0.98	0.40	0.10	0.00	0.17	0.17	0.03	0.27	0.50	0.03	0.03	0.30	0.70	3.30	2.27	1.03	10.20	0.19	0.05	0.27	0.13	0.03	0.18
22	1.20	0.63	0.07	0.00	0.00	0.07	0.00	0.00	0.13	0.30	0.07	0.00	0.17	0.37	2.03	1.70	0.33	10.20	0.22	0.07	0.31	0.18	0.06	0.25
23	2.33	0.96	0.93	0.07	0.03	0.07	0.33	0.00	0.03	0.20	0.03	0.07	0.23	0.43	2.93	2.33	0.60	10.20	0.25	0.07	0.33	0.18	0.05	0.23
26	2.00	0.95	0.00	0.13	0.00	0.13	0.17	0.00	0.10	0.33	0.07	0.03	0.23	0.83	5.10	3.37	1.73	10.20	0.22	0.08	0.29	0.14	0.05	0.19
A	0.22	0.22	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.33	4.11	1.22	9.67	0.49	0.53	0.36	0.32	0.34	0.24
B	0.56	0.33	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.11	5.78	4.44	1.33	9.56	0.49	0.40	0.43	0.32	0.26	0.28
C	0.44	0.33	0.00	0.00	0.00	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	6.67	5.11	1.56	9.78	0.50	0.52	0.39	0.33	0.34	0.26
D	0.33	0.33	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.56	6.11	2.44	10.00	0.53	0.57	0.37	0.35	0.38	0.25
E	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.33	4.78	0.56	11.00	0.43	0.44	0.41	0.28	0.29	0.27
F	0.78	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.11	0.00	0.00	0.22	6.44	4.78	1.67	10.67	0.39	0.33	0.38	0.25	0.22	0.24
G	0.56	0.44	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00	0.22	5.11	4.22	0.89	11.44	0.43	0.39	0.44	0.28	0.26	0.29
H	1.44	0.78	0.00	0.00	0.00	0.22	0.00	0.11	0.11	0.22	0.11	0.00	0.00	0.67	9.00	6.44	2.56	10.22	0.61	0.74	0.40	0.36	0.44	0.24
I	1.44	0.67	0.11	0.11	0.00	0.11	0.11	0.00	0.22	0.33	0.00	0.00	0.00	0.44	6.11	4.89	1.22	10.33	0.46	0.48	0.39	0.29	0.31	0.25
J	0.44	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.00	0.00	0.00	0.22	5.11	3.89	1.22	9.33	0.39	0.34	0.36	0.26	0.23	0.24

Table 9: Official DUC 2003 results (Task 2).

Peer summarizer code (baseline[1-5], manual[A-J], system submission[6-26])	Count of quality questions with non-0 answers	Mean number of quality questions with non-0 answers	Q1 (0 = 0, 1 = 1-5, 2 = 6-10, 3 = 11 or more)	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Number of peer units	Number of marked peer units	Number of unmarked peer units	Number of model units	Mean coverage	Median coverage	Sample std of coverage scores	Mean length-adjusted coverage	Median length-adjusted coverage	Sample std of adjusted coverage scores
2	1.83	0.91	0.77	0.07	0.03	0.17	0.27	0.03	0.07	0.10	0.03	0.03	0.17	0.20	4.10	2.30	1.80	10.33	0.13	0.01	0.24	0.08	0.01	0.15
3	2.63	1.18	1.30	0.10	0.10	0.20	0.37	0.03	0.10	0.17	0.00	0.00	0.23	0.67	3.70	2.07	1.63	10.33	0.17	0.05	0.28	0.10	0.03	0.16
10	1.37	0.74	0.17	0.07	0.03	0.10	0.17	0.00	0.03	0.20	0.00	0.00	0.10	0.47	2.53	1.87	0.67	10.33	0.17	0.04	0.26	0.13	0.04	0.20
11	1.00	0.70	0.13	0.03	0.00	0.10	0.00	0.07	0.07	0.13	0.00	0.00	0.07	0.40	2.23	1.63	0.60	10.33	0.15	0.02	0.27	0.12	0.02	0.20
13	2.43	0.98	0.47	0.07	0.03	0.13	0.40	0.03	0.23	0.33	0.00	0.03	0.17	0.60	3.57	2.37	1.20	10.33	0.13	0.02	0.23	0.08	0.02	0.15
15	6.30	1.27	0.30	1.30	1.17	1.33	0.67	0.30	0.47	0.97	0.00	0.00	0.20	1.33	8.03	2.13	5.90	10.33	0.10	0.02	0.20	0.07	0.01	0.14
16	2.03	0.95	0.27	0.07	0.03	0.23	0.33	0.03	0.07	0.30	0.07	0.00	0.13	0.63	3.63	2.60	1.03	10.33	0.19	0.04	0.31	0.12	0.02	0.19
17	1.83	0.73	0.03	0.03	0.03	0.10	0.07	0.03	0.23	0.47	0.03	0.07	0.07	0.67	3.13	2.00	1.13	10.33	0.19	0.07	0.27	0.11	0.04	0.16
18	2.03	0.88	0.20	0.03	0.00	0.17	0.07	0.03	0.30	0.57	0.03	0.03	0.00	0.67	3.70	2.07	1.63	10.33	0.17	0.04	0.28	0.12	0.03	0.20
20	2.10	1.04	1.03	0.03	0.00	0.07	0.30	0.07	0.03	0.10	0.00	0.00	0.17	0.57	3.20	2.27	0.93	10.33	0.20	0.11	0.28	0.12	0.07	0.17
21	2.67	0.95	0.33	0.07	0.10	0.13	0.27	0.03	0.43	0.33	0.03	0.13	0.17	0.70	3.70	2.10	1.60	10.33	0.14	0.05	0.23	0.10	0.04	0.16
22	1.47	0.77	0.10	0.07	0.03	0.10	0.03	0.00	0.13	0.33	0.00	0.07	0.13	0.47	2.70	2.03	0.67	10.33	0.17	0.03	0.28	0.13	0.03	0.21
23	2.23	1.03	0.93	0.03	0.03	0.20	0.27	0.00	0.07	0.17	0.03	0.10	0.10	0.50	3.13	2.10	1.03	10.33	0.16	0.02	0.28	0.11	0.02	0.20
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.22	4.11	1.11	11.00	0.34	0.27	0.36	0.22	0.17	0.23
B	0.22	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	6.22	4.78	1.44	10.11	0.38	0.27	0.37	0.25	0.17	0.24
C	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.44	4.00	1.44	8.11	0.40	0.29	0.37	0.26	0.19	0.24
D	0.56	0.39	0.22	0.00	0.00	0.22	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.11	8.67	6.00	2.67	7.56	0.45	0.41	0.34	0.29	0.27	0.22
E	0.22	0.22	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	4.89	3.67	1.22	10.56	0.32	0.20	0.37	0.21	0.13	0.24
F	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.11	0.11	5.89	4.11	1.78	10.00	0.32	0.23	0.35	0.20	0.15	0.22
G	0.44	0.44	0.00	0.11	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.11	6.00	4.33	1.67	12.22	0.31	0.20	0.37	0.20	0.13	0.25
H	1.44	0.78	0.00	0.11	0.00	0.11	0.22	0.11	0.22	0.11	0.00	0.00	0.00	0.56	7.89	4.56	3.33	10.56	0.32	0.24	0.34	0.19	0.15	0.21
I	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	7.11	4.33	2.78	11.67	0.30	0.13	0.37	0.20	0.09	0.24
J	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.78	3.44	1.33	11.56	0.27	0.20	0.29	0.18	0.13	0.19

Table 10: Official DUC 2003 results (Task 3).

Peer summarizer code (baseline[1-5], manual[A-J], system submission[6-26])	Count of quality questions with non-0 answers	Mean number of quality questions with non-0 answers	Q1 (0 = 0, 1 = 1-5, 2 = 6-10, 3 = 11 or more)	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Number of peer units	Number of marked peer units	Number of unmarked peer units	Number of model units	Mean coverage	Median coverage	Sample std of coverage scores	Mean length-adjusted coverage	Median length-adjusted coverage	Sample std of adjusted coverage scores
4	1.17	0.57	0.17	0.00	0.00	0.07	0.13	0.03	0.10	0.40	0.00	0.00	0.00	0.27	2.67	1.67	1.00	8.77	0.16	0.01	0.26	0.11	0.01	0.18
5	2.37	0.92	0.27	0.03	0.03	0.17	0.20	0.03	0.13	0.47	0.03	0.07	0.37	0.70	3.13	2.00	1.13	8.77	0.20	0.11	0.28	0.13	0.07	0.18
10	1.07	0.73	0.10	0.00	0.00	0.10	0.03	0.00	0.10	0.33	0.00	0.03	0.13	0.27	2.30	1.60	0.70	8.77	0.14	0.02	0.24	0.10	0.02	0.19
13	2.07	0.82	0.33	0.00	0.03	0.17	0.30	0.03	0.10	0.43	0.00	0.00	0.13	0.57	3.10	1.70	1.40	8.77	0.16	0.00	0.27	0.11	0.00	0.18
14	2.30	0.90	0.13	0.07	0.07	0.30	0.17	0.00	0.17	0.50	0.03	0.10	0.13	0.63	3.90	2.47	1.43	8.77	0.19	0.07	0.29	0.13	0.05	0.20
16	1.86	0.96	0.03	0.07	0.03	0.24	0.24	0.00	0.07	0.52	0.00	0.03	0.00	0.72	4.00	2.69	1.31	8.79	0.19	0.07	0.29	0.12	0.04	0.18
17	2.20	0.89	0.27	0.13	0.00	0.20	0.23	0.03	0.20	0.57	0.00	0.00	0.07	0.70	3.13	1.70	1.43	8.77	0.14	0.05	0.23	0.08	0.03	0.13
19	3.10	0.95	0.30	0.10	0.13	0.40	0.30	0.00	0.20	0.67	0.03	0.03	0.20	0.80	4.13	1.20	2.93	8.77	0.07	0.01	0.14	0.05	0.00	0.10
20	1.67	0.87	0.50	0.00	0.00	0.10	0.17	0.03	0.07	0.13	0.00	0.00	0.13	0.70	3.43	1.67	1.77	8.77	0.14	0.02	0.24	0.09	0.01	0.15
22	2.13	0.87	0.20	0.07	0.00	0.23	0.23	0.03	0.23	0.57	0.03	0.07	0.07	0.47	3.53	2.17	1.37	8.77	0.19	0.06	0.28	0.13	0.04	0.19
23	1.27	0.78	0.17	0.03	0.03	0.13	0.13	0.00	0.03	0.37	0.00	0.00	0.03	0.43	2.20	1.83	0.37	8.77	0.19	0.05	0.30	0.14	0.03	0.21
A	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	6.22	4.00	2.22	7.78	0.31	0.33	0.33	0.20	0.21	0.21
B	0.11	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.67	4.67	1.00	7.89	0.49	0.52	0.38	0.32	0.34	0.25
C	0.67	0.44	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.00	0.00	0.33	5.33	3.67	1.67	8.89	0.44	0.39	0.37	0.29	0.25	0.24
D	0.67	0.33	0.22	0.00	0.00	0.00	0.22	0.00	0.00	0.11	0.00	0.00	0.00	0.11	6.00	3.78	2.22	9.33	0.35	0.23	0.42	0.23	0.15	0.27
E	0.56	0.33	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.22	0.11	5.78	4.33	1.44	9.56	0.38	0.29	0.37	0.25	0.19	0.23
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.75	2.88	2.88	8.63	0.28	0.24	0.31	0.18	0.15	0.20
G	0.78	0.44	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.22	0.00	0.00	0.00	0.33	5.00	4.11	0.89	9.67	0.34	0.27	0.36	0.22	0.17	0.23
H	0.56	0.33	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.11	0.11	0.22	6.44	3.78	2.67	9.22	0.46	0.50	0.35	0.28	0.31	0.21
I	0.22	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	6.11	4.56	1.56	8.11	0.44	0.36	0.37	0.29	0.24	0.24
J	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	4.78	4.00	0.78	8.67	0.35	0.27	0.39	0.23	0.18	0.25

Table 11: Official DUC 2003 results (Task 4).