The University of Michigan at DUC 2004

Güneş Erkan[1], **Dragomir R. Radev**[2,1]
[1]Department of EECS, [2]School of Information
University of Michigan
{gerkan,radev}@umich.edu

Abstract

We present the results of Michigan's participation in DUC 2004. Our system, MEAD, ranked as one of the top systems in four of the five tasks. We introduce our new feature, LexPageRank, a new measure of sentence centrality inspired by the prestige concept in social networks. LexPageRank gave promising results in multi-document summarization. Our approach for Task 5, biographical summarization, was simplistic, yet succesful. We used regular expression matching to boost up the scores of the sentences that are likely to contain biographical information patterns.

1 Introduction

The year 2004 marked the fourth 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), in all of the five tasks, submitting 3 runs for each of the first 4 tasks and only one run for Task 5. The performance of our runs (peers numbered 135-147 in the official results) was quite good - our system ranked one of the top with respect to most of the metrics used in 4 of the 5 tasks.

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 2004 evaluation

The DUC 2004 evaluation consisted of 5 tasks. Tasks 1 and 2 were essentially the same as in last year's evaluation. Tasks 3 and 4 were also similar to Tasks 1 and 2, respectively, with the exception that the documents used in summarization are either automatic or manual English translations of Arabic documents.

The goal of Task 1 was to produce a 75-byte single-document summaries (headlines) for each of the documents in the given 50 TDT clusters. In Task 2, same set of clusters were used to produce a 665-byte multi-document short summary for each cluster.

The goals of Tasks 3 and 4 were essentially the same as Tasks 1 and 2, respectively, except that they focused on cross-lingual summaries of 25 Arabic TDT clusters. There were 3 subtasks for each of the Tasks 3 and 4 according to the data sets being used. In subtasks 3a and 4a, automatic English translations of the Arabic clusters were used. In subtasks 3b and 4b, manual translations were provided instead. Finally, in subtasks 3c and 4c, automatic translations plus some original English documents that were relevant to each cluster were given as the input.

Task 5 involved a newly introduced summarization concept for DUC, which is known as "biographical summarization". Given a news cluster of approximately 10 documents, and a question of the form "Who is X", where X is the name of a person, the task was to produce a 665-byte multi-document summary of the cluster that would respond to the question. 50 TREC news clusters, each of which focused on events about a different person, were used in this task.

3 Evaluation measures

In DUC 2004, an automatic evaluation metric for summarization, ROUGE[1], was used for the first time. ROUGE is a recall-based metric for fixed-length summaries which is based on n-gram co-occurence. It reports separate scores for 1, 2, 3, and 4-gram, and also for longest common subsequence co-occurences. Tasks 1-4 were evaluated solely by means of ROUGE.

Task 5 summaries were evaluated manually for quality, coverage, and responsiveness to the question. Here is the complete list of metrics used in the evaluation of Task 5:

[1]`http://www.isi.edu/~cyl/ROUGE`

Quality metrics:

Answers to 7 quality questions: The assessors answered 7 questions (see Figure 1) for each summary. These were multiple choice questions with answers numbered from 1 to 5, where 1 was the best answer, 5 was the worst answer.

Coverage metrics:

Number of peer units: Number of rough sentences in a summary.

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

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

Fraction of unmarked peer units at least related to the model's subject: The fraction of the number of peer units which did not overlap at all in meaning with any model unit, but was at least related to the subject of the model.

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

Mean coverage: 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.

Responsiveness metrics:

Responsiveness score: An integer grade between [0,4], where 0 is the worst and 4 is the best score, indicating how responsive the summary is to the question relative to the other summaries.

Q1: Does the summary build from sentence to sentence to a coherent body of information about the topic?
Q2: If you were editing the summary to make it more concise and to the point, how much useless, confusing or repetitive text would you remove from the existing summary?
Q3: To what degree does the summary say the same thing over again?
Q4: How much trouble did you have identifying the referents of noun phrases in this summary? Are there nouns, pronouns or personal names that are not well-specified? For example, a person is mentioned and it is not clear what his role in the story is, or any other entity that is referenced but its identity and relation with the story remains unclear.
Q5: To what degree do you think the entities (person/thing/event/place/...) were re-mentioned in an overly explicit way, so that readability was impaired? For example, a pronoun could have been used instead of a lengthy description, or a shorter description would have been more appropriate?
Q6: Are there any obviously ungrammatical sentences, e.g., missing components, unrelated fragments or any other grammar-related problem that makes the text diffcult to read?
Q7: Are there any datelines, system-internal formatting or capitalization errors that can make the reading of the summary difficult?

Figure 1: The seven *quality* questions used in DUC 2004.

```
Centroid 1 Position 1 LengthCutoff 9 SimWithFirst 2
LexPageRank 1

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

Figure 2: Sample MEAD policy.

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 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 Features

We used four of the old features of MEAD in all of the tasks. These are Centroid, Position, LengthCutoff, and SimWithFirst. Additionally, we introduced two new features for this year's DUC evaluation.

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

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday, against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister, Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

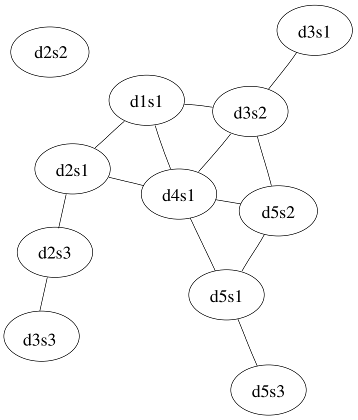
	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

Figure 3: Intra-sentence cosine similarities in a subset of cluster d1003t from DUC 2004.

4.1.1 LexPageRank feature

Extractive summarization can be viewed as choosing the most *central* sentences in a cluster that give the necessary and enough amount of information related to the main theme of the cluster. In the case of centroid-based summarization (Radev et al., 2000), *centrality* is defined in terms of the centroid of the cluster, i.e. the sentences that contain the highest ranked words in the centroid of the cluster are considered as central to the topic.

We propose a new measure of sentence centrality inspired by the concept of *prestige* in social networks and its application in the Web. PageRank (Page et al., 1998) is a method proposed for assigning a prestige score to each page in the web independent of a specific query. In PageRank, the score of a page is determined depending on the number of pages that link to that page as well as the individual scores of the linking pages. This is achieved



ID	Score
d1s1	0.0932
d2s1	0.0982
d2s2	0.0909
d2s3	0.0879
d3s1	0.0583
d3s2	0.1170
d3s3	0.0670
d4s1	0.1342
d5s1	0.0993
d5s2	0.0935
d5s3	0.0604

Figure 4: Similarity graph and corresponding LexPageRank scores for the cluster in Figure 3. Sentence d4s1 is the most central page (score = 0.1342).

by forming the normalized adjacency matrix of the network of pages and using a power method to converge to the principal eigenvector of this matrix. In the context of summarization, we recursively define the prestige of a sentence in terms of the number of sentences that it is similar to, and the corresponding prestiges of these similar sentences. We use cosine as the similarity measure and define the cosine adjacency matrix of a cluster as the binary connectivity matrix, M, where $M(u, v) = 1$ if the cosine similarity between the sentence u and the sentence v is above a certain threshold. Figure 3 shows a subset of a cluster used in Task 4b of DUC 2004, and the cosine similarity between each pair of sentences. Sentence ID dXsY indicates the Yth sentence in the Xth document. Figure 4 shows the graph that corresponds to the similarity matrix with threshold 0.2. The LexPageRank scores in the same figure are computed by an approximation of the principal eigenvector of the adjacency matrix for the graph. Unlike the original PageRank method, the graph is undirected since cosine similarity is symmetric.

We used LexPageRank feature in our priority 3 submission for Task 2, and priority 1 submission for Task 4. The cosine threshold used was 0.3. Although we did not experiment with enough number of different thresholds and feature weights, the results are better according to many of the metrics compared to our submissions without LexPageRank (Table 5). A more detailed performance analysis for LexPageRank will be presented in an extended version of this paper.

4.1.2 QueryPhraseMatch feature

Task 5 needed a very different treatment from other tasks since it involved producing biographical summaries, which are quite different from general purpose summaries. In order to do this, we searched for a way of giving higher scores to the sentences that gives information about the person in the question. There could be two ways of incorporating such a ranking of sentences into MEAD; either introducing a new feature that gives higher scores to such sentences, or a new reranker that modifies the rankings of the sentences in a general purpose summary. We chose the first one, and introduced a new feature called QueryPhraseMatch.

In order to compute the QueryPhraseMatch feature for a sentence, we look for certain regular expressions in the sentence and increase the feature value for each expression that matches the sentence. Note that QueryPhraseMatch is a general purpose feature, not designed solely for the purpose of biographical summaries, and can be used to look for any kind of regular expressions in a text. In the case of a biographical summary, these should be the expressions that often occur in sentences that describe a person (e.g. an occurance of person's name, relative clauses, etc.). We looked at some biographical texts and collected regular expressions that appear in these texts in common. A subset of these expressions are shown in Figure 5. Every expression has a weight that determines by how much we will increase the score of a sentence that matches the expression.

Due to lack of time, the regular expressions and corresponding weights were determined totally empirically and depending on subjective judgements. However, the results are promising as we discuss in Section 5.

<table>
<tr><th>Expression</th><th>Weight</th></tr>
<tr><td>X</td><td>0.25</td></tr>
<tr><td>X grew up</td><td>1</td></tr>
<tr><td>X attended</td><td>1</td></tr>
<tr><td>X (turns|turned) [1-9][0-9]?</td><td>1</td></tr>
<tr><td>X, (an?|the|who|whom|whose) [\w]*[,.]</td><td>1.5</td></tr>
<tr><td>X, [1-9][0-9]?(,|years)</td><td>1.5</td></tr>
<tr><td>X began</td><td>0.35</td></tr>
<tr><td>X (lives|lived)</td><td>0.5</td></tr>
<tr><td>X made</td><td>0.5</td></tr>
</table>

Figure 5: Some regular expressions that are searched to determine the sentences that describe a person. 'X' is unified with the person name at run time.

4.2 Training

We participated in the headline extraction tasks (Tasks 1 and 3) for the first time in DUC 2004. We used MEAD to extract only one sentence from a document as the headline of that document. No real training was done for these tasks. However, we observed that setting the weight of the Position feature high results in higher ROUGE scores for DUC 2003 data, meaning that the sentences that occur at the beginning of a document are more likely to get higher ROUGE scores with respect to the manual headlines. This issue is discussed in Section 5. Table 1 shows the feature weights we used in our submissions for Tasks 1 and 3.

Feature Weights			Task 1	Task3
Centroid	**Position**	**Length**	**Priority**	**Priority**
1.0	1.0	7	1	3
1.0	1.5	7	2	1
1.0	10.0	7	3	2

Table 1: Feature weights used in Tasks 1 and 3.

To train MEAD for Tasks 2 and 4, we used the DUC 2003 data set. We split the data into two parts as training and devtest. We fixed our reranker to be the MMR-reranker since initial experiments showed that it outperforms the other rerankers in almost any policy as far as ROUGE scores are concerned. All of the experiment results shown in this paper were derived from MEAD policies that use MMR-reranker.

We made a local search with random restarts on the feature weights space. For each set of feature weights, we ran MEAD on the test data, and tried to maximize the ROUGE-1 (unigram) and ROUGE-W (weighted longest common subsequence) scores. The search step size was decreased as we got better ROUGE scores. Table 2 shows the ROUGE scores we got on the devtest data with the best policies observed on the training data. The highlighted rows are the two policies we used in our submissions. One of them gave the highest ROUGE-W score while the other was the one which gave the highest ROUGE-1 score among the policies which were substantially different than the first one.

Centroid	Position	Length	SimWithFirst	ROUGE-1	ROUGE-W
1.5	1.0	9	2.0	0.32256	0.11712
1.5	1.0	9	3.0	0.32256	0.11712
2.0	1.0	9	2.0	0.32750	0.11875
3.0	1.0	9	2.0	0.33311	0.11847
3.0	1.0	15	2.0	0.34171	0.12132
3.0	1.0	20	2.0	0.33566	0.11906
2.0	**1.0**	**9**	**4.0**	**0.34361**	**0.12511**
3.0	1.0	9	4.0	0.33926	0.12301
4.0	1.0	9	4.0	0.34560	0.12171
3.0	1.0	9	5.0	0.34623	0.12174
2.0	1.0	15	4.0	0.34470	0.12223
2.0	1.0	9	5.0	0.34707	0.12304
3.0	**1.0**	**15**	**3.0**	**0.34684**	**0.12197**
3.0	1.0	15	4.0	0.34606	0.12170

Table 2: Best policies and corresponding ROUGE scores for DUC 2003 devtest data

Due to lack of time, we were not able to experiment with the weight of our new LexPageRank feature. Instead, we picked the policies that performed the best without LexPageRank and reran them with LexPageRank using a constant weight. Table 3 shows the ROUGE scores for these policies. The highlighted policy is the one we used in our submissions for Tasks 2 and 4.

Centroid	Position	Length	SimWithFirst	LexPageRank	ROUGE-1	ROUGE-W
2.0	1.0	9	4.0	2.0	0.34129	0.12245
3.0	1.0	9	4.0	2.0	0.34051	0.12182
1.0	1.0	10	4.0	2.0	0.33662	0.12059
1.0	1.0	11	5.0	2.0	0.33806	0.12137
4.0	1.0	9	4.0	2.0	0.34513	0.12233
3.0	1.0	9	5.0	2.0	0.34450	0.12322
2.0	1.0	15	4.0	2.0	0.32436	0.11532
2.0	**1.0**	**9**	**5.0**	**2.0**	**0.34450**	**0.12370**
3.0	1.0	15	3.0	2.0	0.33922	0.12090
3.0	1.0	15	4.0	2.0	0.34467	0.12280

Table 3: Policies with LexPageRank feature and corresponding ROUGE scores for DUC 2003 devtest data

No training was done for Task 5. MEAD was used with default feature weights. However, the weights of the regular expressions incorporated into the QueryPhraseMatch feature were high enough to override default MEAD features.

5 Results

The official results for Tasks 1-4 are summarized in Table 5 and Table 6. We show ROUGE-1, ROUGE-2, and ROUGE-W scores and omit ROUGE-3, ROUGE-4, and ROUGE-L scores due to lack of space. ROUGE-1 was shown to be the best correlating metric with the human judges among these different ROUGE metricsTable 7 shows the rankings of our submissions for Tasks 1–4 with respect to several ROUGE metrics among all of the submissions. We chose to separate the results for the subtasks of Tasks 3 and 4 since the data sets for these subtasks were different from each other.

On Task 1, our best system (peer code: 137) was the priority-3 submission for this task with the highest weight for Position feature among our tasks (Table 1), and supports our observations in the experiments that sentences that occur near the beginning of a document are more likely to get higher ROUGE scores. This also explains the surprisingly successful performance of the baseline system on Task 1, which simply extracts the first sentence of a document.

Task 2 was our most unsuccessful task, where our system showed an average performance among all of the submissions. Our submission that included the LexPageRank feature (peer code: 140) achieved the best ROUGE-1 score among our three submissions for this task, which is a promising result for this new feature.

Peer Code	Task	Priority	Rankings						Total Submissions
			ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-W	
135	1	1	7	4	4	4	5	5	39
136	1	2	5	3	3	3	4	3	39
137	1	3	2	2	1	1	2	2	39
138	2	1	18	16	16	13	27	27	34
139	2	2	21	18	17	17	29	29	34
140	2	3	15	20	19	16	26	26	34
141	3a	NA	5	2	1	1	2	2	11
142	3b	NA	5	1	1	1	4	3	10
144	4a	NA	1	2	1	1	6	6	11
145	4b	NA	3	1	1	1	7	7	11
146	4c	NA	1	2	2	2	4	4	6

Table 7: Official rankings of our submissions for Tasks 1–4 with respect to ROUGE metrics.

On Task 3, almost all systems participated in the evaluation got the best score on Task 3b since the data set was composed of manual translations, which are expected to be more similar to model human summaries compared to automatic translations. Our best score came from the best policy on Task 1, which we again used here for Task 3b. Since only two systems participated in Task 3c, we did not include the ranking of our submission for this task in Table 7. Furthermore, our submission can essentially be compared to the submissions for Task 3a since we did not use the relevant documents provided by NIST for Task 3c. The difference between the scores of our submissions for Task 3a and Task 3c is due to the difference between the weights of the Position feature in two policies. As in Task 1, we got higher ROUGE scores with the higher Position weight.

Our results on Task 4 are also very successful. The results we got using LexPageRank feature (peer code: 145) are as successful as our other submissions for this task, which is again a promising fact.

There are several metrics for evaluating Task 5. However, these metrics are not defined as an overall evaluation of system performance but rather as an assessment of each individual summary. Table 4 shows our own analysis of average performance of each system. Numbers in the parantheses show the number of clusters in which the system was one of the best performers among all systems (excluding manual summaries) considering the metric of that column only. First seven columns show the average scores of the systems for each quality question in Figure 1. A lower number shows a better system. Since our system produces extractive summaries, its performance is one of the best in questions about grammaticality and formatting (e.g. Q6 and Q7). However, it performs worse in questions about the semantic structure of the summary (e.g. Q3, Q4, and Q5). Next column shows the mean of the mean coverage scores that a system got for each

Peer code	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Mean coverage**	**Unmarked but related units (%)**	**Responsiveness**
A	1.00	1.06	1.00	1.00	1.00	1.11	1.06	0.477	100	3.60
B	1.06	1.06	1.00	1.06	1.00	1.06	1.00	0.554	100	3.92
C	1.24	1.36	1.28	1.04	1.16	1.40	1.12	0.350	86	2.84
D	1.11	1.22	1.00	1.06	1.17	1.33	1.00	0.469	96	3.60
E	1.39	1.33	1.00	1.11	1.00	1.33	1.22	0.468	90	3.32
F	1.24	1.24	1.06	1.00	1.06	1.12	1.18	0.517	84	3.64
G	1.17	1.00	1.00	1.00	1.00	1.28	1.06	0.537	94	3.48
H	1.17	1.22	1.11	1.00	1.06	1.06	1.00	0.438	96	3.52
5	1.62 (38)	2.20 (28)	1.44 (37)	1.44 (37)	1.10 (47)	1.40 (33)	1.82 (16)	0.190 (2)	62	1.52 (16)
16	3.12 (8)	2.70 (12)	1.58 (34)	2.04 (20)	1.70 (28)	1.36 (37)	1.46 (36)	0.158 (1)	69	1.30 (7)
24	3.86 (1)	3.40 (2)	1.44 (36)	2.98 (7)	1.30 (38)	1.88 (25)	1.88 (20)	0.213 (8)	55	1.16 (6)
30	2.90 (6)	2.42 (20)	1.38 (39)	2.00 (24)	1.46 (33)	1.30 (39)	1.30 (40)	0.199 (3)	72	1.42 (13)
43	3.78 (1)	2.94 (13)	1.30 (38)	2.84 (8)	1.46 (32)	1.76 (28)	1.52 (40)	0.198 (6)	70	1.30 (14)
49	3.34 (5)	2.88 (9)	1.54 (36)	2.12 (19)	1.62 (33)	1.78 (25)	1.36 (38)	0.206 (3)	78	1.40 (12)
62	3.62 (3)	3.12 (10)	1.98 (24)	1.90 (24)	2.14 (18)	1.84 (21)	1.56 (36)	0.200 (3)	49	1.64 (17)
71	3.32 (2)	2.96 (3)	1.84 (25)	1.44 (36)	2.42 (12)	1.64 (26)	1.66 (29)	0.214 (2)	64	1.50 (13)
86	3.86 (3)	3.22 (5)	1.16 (44)	2.96 (10)	1.18 (41)	2.22 (16)	1.68 (32)	0.145 (2)	51	1.06 (9)
96	3.30 (10)	3.06 (8)	1.60 (29)	1.82 (28)	1.66 (30)	1.68 (26)	3.54 (3)	0.216 (6)	60	1.44 (11)
109	3.14 (6)	2.76 (11)	1.86 (26)	1.74 (28)	1.96 (25)	1.42 (35)	1.54 (38)	0.241 (7)	76	1.76 (19)
116	4.52 (0)	4.04 (1)	1.48 (35)	3.40 (3)	1.76 (24)	2.14 (17)	2.14 (15)	0.173 (7)	65	1.00 (8)
122	2.94 (11)	2.32 (27)	1.68 (30)	1.24 (43)	2.08 (18)	1.22 (43)	1.36 (42)	0.184 (1)	74	1.26 (7)
125	3.52 (4)	3.22 (6)	1.62 (28)	2.12 (23)	1.94 (21)	2.76 (8)	2.18 (23)	0.189 (4)	71	1.40 (13)
147	**3.18 (8)**	**2.64 (14)**	**1.84 (24)**	**1.92 (24)**	**1.80 (25)**	**1.24 (39)**	**1.40 (39)**	**0.215 (4)**	**69**	**1.54 (19)**

Table 4: An overall analysis of official results for Task 5.

cluster. Our system ranked third overall, and ranked best in 4 of the 50 clusters. Another metric for Task 5 is the fraction of the unmarked peer units, i.e. the units which do not overlap at all in meaning with any model unit, that are somewhat related to the model's subject. We took the average of these fractions of a peer system for all clusters, weighting by the number of unmarked units of the system for each cluster. As seen in Table 4, about 70% of the unmarked units of our summaries were at least related to the subject. The last column shows the average responsiveness scores. Our system ranked third overall while ranking best in 19 of 50 clusters.

6 Acknowledgments

This work was partially supported by the National Science Foundation under grant 0329043 "Probabilistic and link-based Methods for Exploiting Very Large Textual Repositories" administered through the IDM program. All opinions, findings, conclusions, and recommendations in this paper are made by the authors and do not necessarily reflect the views of the National Science Foundation.

We would like to thank the members of the CLAIR group at Michigan and in particular Jahna Otterbacher and Siwei Shen for their assistance with this project.

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.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford University, Stanford, CA*.

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 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.

TASK 1						
Peer Code	**ROUGE-1 (unigram)**	**95% Confidence Interval**	**ROUGE-2 (bigram)**	**95% Confidence Interval**	**ROUGE-W (LCS)**	**95% Confidence Interval**
A	0.4676	[0.4536,0.4815]	0.3059	[0.2961,0.3157]	0.2713	[0.2653,0.2774]
B	0.4912	[0.4789,0.5035]	0.3351	[0.3265,0.3436]	0.2923	[0.2863,0.2982]
C	0.4750	[0.4620,0.4881]	0.3170	[0.3068,0.3272]	0.2769	[0.2708,0.2830]
D	0.4770	[0.4650,0.4890]	0.3172	[0.3080,0.3264]	0.2796	[0.2741,0.2852]
E	0.4893	[0.4744,0.5042]	0.3320	[0.3219,0.3422]	0.2878	[0.2811,0.2946]
F	0.4610	[0.4473,0.4746]	0.3009	[0.2931,0.3086]	0.2698	[0.2641,0.2754]
G	0.4620	[0.4501,0.4740]	0.3113	[0.3028,0.3197]	0.2761	[0.2705,0.2817]
H	0.4239	[0.4102,0.4376]	0.2830	[0.2739,0.2921]	0.2553	[0.2488,0.2617]
1	0.2214	[0.2098,0.2329]	0.0637	[0.0576,0.0698]	0.1174	[0.1115,0.1233]
6	0.2503	[0.2397,0.2609]	0.0650	[0.0592,0.0708]	0.1196	[0.1146,0.1245]
9	0.2115	[0.2040,0.2190]	0.0245	[0.0211,0.0280]	0.0964	[0.0936,0.0993]
10	0.1696	[0.1585,0.1806]	0.0409	[0.0352,0.0465]	0.0902	[0.0846,0.0958]
18	0.1532	[0.1449,0.1615]	0.0366	[0.0320,0.0412]	0.0822	[0.0780,0.0863]
25	0.1206	[0.1101,0.1311]	0.0252	[0.0210,0.0293]	0.0654	[0.0605,0.0703]
26	0.1207	[0.1102,0.1312]	0.0277	[0.0228,0.0325]	0.0654	[0.0600,0.0707]
31	0.1868	[0.1774,0.1962]	0.0429	[0.0384,0.0475]	0.0992	[0.0948,0.1036]
32	0.1919	[0.1825,0.2014]	0.0451	[0.0404,0.0499]	0.1016	[0.0971,0.1060]
33	0.1865	[0.1769,0.1961]	0.0430	[0.0384,0.0476]	0.0990	[0.0945,0.1036]
50	0.1895	[0.1790,0.1999]	0.0476	[0.0422,0.0529]	0.0993	[0.0940,0.1045]
51	0.1872	[0.1768,0.1976]	0.0469	[0.0416,0.0522]	0.0983	[0.0930,0.1035]
52	0.1960	[0.1853,0.2067]	0.0499	[0.0445,0.0553]	0.1023	[0.0970,0.1075]
53	0.2042	[0.1949,0.2136]	0.0433	[0.0386,0.0481]	0.0972	[0.0935,0.1008]
54	0.2039	[0.1946,0.2132]	0.0435	[0.0389,0.0481]	0.0970	[0.0934,0.1006]
75	0.1245	[0.1159,0.1331]	0.0256	[0.0220,0.0293]	0.0689	[0.0642,0.0736]
76	0.1743	[0.1631,0.1855]	0.0427	[0.0368,0.0487]	0.0916	[0.0861,0.0971]
77	0.2195	[0.2083,0.2306]	0.0591	[0.0530,0.0652]	0.1137	[0.1082,0.1192]
78	0.2056	[0.1979,0.2134]	0.0073	[0.0057,0.0090]	0.0876	[0.0850,0.0903]
79	0.1691	[0.1588,0.1794]	0.0415	[0.0369,0.0461]	0.0895	[0.0844,0.0946]
80	0.1893	[0.1789,0.1997]	0.0476	[0.0425,0.0527]	0.1003	[0.0950,0.1056]
87	0.1845	[0.1762,0.1928]	0.0364	[0.0319,0.0409]	0.0924	[0.0888,0.0960]
88	0.1882	[0.1792,0.1972]	0.0400	[0.0353,0.0447]	0.0939	[0.0899,0.0978]
89	0.1836	[0.1754,0.1918]	0.0367	[0.0324,0.0409]	0.0919	[0.0881,0.0956]
90	0.1746	[0.1660,0.1832]	0.0325	[0.0279,0.0371]	0.0880	[0.0836,0.0924]
91	0.1706	[0.1634,0.1779]	0.0294	[0.0257,0.0331]	0.0859	[0.0825,0.0893]
92	0.1501	[0.1399,0.1603]	0.0395	[0.0338,0.0451]	0.0827	[0.0773,0.0881]
98	0.1479	[0.1378,0.1581]	0.0311	[0.0265,0.0356]	0.0785	[0.0736,0.0834]
99	0.1816	[0.1717,0.1914]	0.0538	[0.0481,0.0594]	0.0998	[0.0947,0.1050]
100	0.1300	[0.1211,0.1389]	0.0261	[0.0214,0.0307]	0.0713	[0.0668,0.0757]
101	0.1716	[0.1612,0.1819]	0.0518	[0.0462,0.0574]	0.0950	[0.0896,0.1004]
110	0.1941	[0.1857,0.2025]	0.0430	[0.0383,0.0477]	0.0964	[0.0926,0.1002]
128	0.1989	[0.1904,0.2073]	0.0326	[0.0283,0.0369]	0.0949	[0.0910,0.0988]
129	0.1994	[0.1911,0.2077]	0.0328	[0.0286,0.0370]	0.0953	[0.0914,0.0992]
130	0.2167	[0.2087,0.2246]	0.0243	[0.0208,0.0278]	0.0974	[0.0942,0.1007]
131	0.2195	[0.2117,0.2272]	0.0255	[0.0220,0.0289]	0.1016	[0.0985,0.1048]
132	0.1537	[0.1444,0.1631]	0.0229	[0.0191,0.0266]	0.0773	[0.0728,0.0818]
135	**0.2127**	**[0.2002,0.2251]**	**0.0598**	**[0.0532,0.0664]**	**0.1129**	**[0.1069,0.1190]**
136	**0.2174**	**[0.2053,0.2294]**	**0.0618**	**[0.0556,0.0680]**	**0.1154**	**[0.1092,0.1217]**
137	**0.2210**	**[0.2095,0.2325]**	**0.0638**	**[0.0575,0.0700]**	**0.1175**	**[0.1116,0.1234]**

TASK 2						
Peer Code	**ROUGE-1 (unigram)**	**95% Confidence Interval**	**ROUGE-2 (bigram)**	**95% Confidence Interval**	**ROUGE-W (LCS)**	**95% Confidence Interval**
A	0.5437	[0.5272,0.5602]	0.3155	[0.3060,0.3250]	0.2155	[0.2107,0.2202]
B	0.5564	[0.5347,0.5781]	0.3263	[0.3127,0.3399]	0.2163	[0.2079,0.2246]
C	0.5391	[0.5242,0.5540]	0.3171	[0.3078,0.3264]	0.2133	[0.2084,0.2182]
D	0.5537	[0.5400,0.5675]	0.3291	[0.3196,0.3385]	0.2182	[0.2133,0.2230]
E	0.5585	[0.5429,0.5742]	0.3250	[0.3136,0.3364]	0.2179	[0.2124,0.2235]
F	0.5604	[0.5449,0.5759]	0.3194	[0.3081,0.3306]	0.2175	[0.2126,0.2223]
G	0.5413	[0.5250,0.5577]	0.3140	[0.3030,0.3251]	0.2143	[0.2094,0.2191]
H	0.5637	[0.5506,0.5769]	0.3288	[0.3181,0.3395]	0.2234	[0.2175,0.2292]
2	0.3242	[0.3104,0.3380]	0.0641	[0.0545,0.0737]	0.1186	[0.1130,0.1242]
11	0.3029	[0.2818,0.3240]	0.0699	[0.0605,0.0793]	0.1088	[0.1022,0.1154]
19	0.3739	[0.3602,0.3875]	0.0803	[0.0712,0.0893]	0.1315	[0.1261,0.1368]
27	0.3035	[0.2850,0.3221]	0.0475	[0.0389,0.0560]	0.1078	[0.1021,0.1136]
28	0.2952	[0.2793,0.3111]	0.0489	[0.0408,0.0570]	0.1036	[0.0986,0.1086]
29	0.2767	[0.2599,0.2935]	0.0405	[0.0328,0.0482]	0.1011	[0.0957,0.1064]
34	0.3700	[0.3565,0.3834]	0.0760	[0.0680,0.0839]	0.1325	[0.1279,0.1372]
35	0.3743	[0.3615,0.3871]	0.0837	[0.0737,0.0936]	0.1338	[0.1291,0.1384]
36	0.3235	[0.3116,0.3353]	0.0534	[0.0455,0.0612]	0.1229	[0.1184,0.1273]
44	0.3501	[0.3382,0.3620]	0.0759	[0.0677,0.0842]	0.1233	[0.1188,0.1277]
45	0.3671	[0.3548,0.3795]	0.0739	[0.0652,0.0826]	0.1297	[0.1248,0.1345]
55	0.3568	[0.3451,0.3685]	0.0708	[0.0630,0.0785]	0.1259	[0.1211,0.1308]
56	0.3500	[0.3384,0.3616]	0.0670	[0.0587,0.0753]	0.1237	[0.1191,0.1284]
57	0.3453	[0.3323,0.3583]	0.0639	[0.0557,0.0721]	0.1217	[0.1173,0.1261]
65	0.3822	[0.3708,0.3937]	0.0922	[0.0827,0.1016]	0.1333	[0.1290,0.1375]
66	0.3667	[0.3524,0.3810]	0.0887	[0.0798,0.0977]	0.1308	[0.1259,0.1358]
67	0.3697	[0.3570,0.3825]	0.0906	[0.0817,0.0995]	0.1310	[0.1268,0.1353]
81	0.3684	[0.3553,0.3816]	0.0805	[0.0722,0.0888]	0.1287	[0.1244,0.1329]
93	0.3290	[0.3113,0.3468]	0.0747	[0.0664,0.0830]	0.1198	[0.1153,0.1244]
94	0.2752	[0.2582,0.2923]	0.0568	[0.0495,0.0642]	0.1022	[0.0970,0.1074]
95	0.3317	[0.3197,0.3437]	0.0676	[0.0600,0.0751]	0.1227	[0.1187,0.1267]
102	0.3696	[0.3587,0.3806]	0.0847	[0.0764,0.0929]	0.1273	[0.1233,0.1314]
103	0.3304	[0.3154,0.3454]	0.0764	[0.0683,0.0845]	0.1172	[0.1121,0.1223]
104	0.3744	[0.3635,0.3854]	0.0855	[0.0770,0.0939]	0.1284	[0.1244,0.1324]
111	0.2419	[0.2307,0.2531]	0.0188	[0.0153,0.0222]	0.0936	[0.0900,0.0972]
117	0.3429	[0.3308,0.3550]	0.0591	[0.0521,0.0662]	0.1204	[0.1159,0.1249]
118	0.3438	[0.3301,0.3575]	0.0580	[0.0520,0.0641]	0.1184	[0.1138,0.1229]
119	0.3402	[0.3255,0.3549]	0.0580	[0.0504,0.0655]	0.1207	[0.1159,0.1255]
120	0.3042	[0.2835,0.3249]	0.0762	[0.0664,0.0860]	0.1140	[0.1070,0.1209]
121	0.3205	[0.3040,0.3370]	0.0715	[0.0623,0.0808]	0.1180	[0.1119,0.1241]
123	0.2854	[0.2715,0.2993]	0.0485	[0.0419,0.0551]	0.1046	[0.1001,0.1090]
124	0.3706	[0.3578,0.3835]	0.0829	[0.0748,0.0909]	0.1293	[0.1252,0.1334]
138	**0.3430**	**[0.3274,0.3586]**	**0.0727**	**[0.0640,0.0815]**	**0.1081**	**[0.1036,0.1127]**
139	**0.3396**	**[0.3239,0.3553]**	**0.0714**	**[0.0630,0.0798]**	**0.1074**	**[0.1029,0.1118]**
140	**0.3456**	**[0.3309,0.3604]**	**0.0706**	**[0.0619,0.0793]**	**0.1081**	**[0.1039,0.1124]**

Table 5: Summary of official ROUGE scores for Tasks 1 and 2. Peer codes: baseline[1-2], manual[A-H], and system submissions

TASK 3						
Peer Code	**ROUGE-1 (unigram)**	**95% Confidence Interval**	**ROUGE-2 (bigram)**	**95% Confidence Interval**	**ROUGE-W (LCS)**	**95% Confidence Interval**
W	0.38801	[0.36997,0.40605]	0.14438	[0.12788,0.16088]	0.19714	[0.18700,0.20728]
X	0.40318	[0.38690,0.41946]	0.15522	[0.13962,0.17082]	0.20769	[0.19836,0.21702]
Y	0.37694	[0.35936,0.39452]	0.13639	[0.12017,0.15261]	0.18957	[0.17977,0.19937]
Z	0.40277	[0.38474,0.42080]	0.14873	[0.13316,0.16430]	0.20291	[0.19352,0.21230]
TASK 3a						
3	0.13737	[0.12408,0.15066]	0.02929	[0.02186,0.03672]	0.07431	[0.06776,0.08086]
7	0.21412	[0.20026,0.22798]	0.04368	[0.03562,0.05174]	0.10427	[0.09753,0.11101]
12	0.18235	[0.16838,0.19632]	0.01836	[0.01329,0.02343]	0.08564	[0.07980,0.09148]
20	0.20166	[0.18765,0.21567]	0.04441	[0.03649,0.05233]	0.10254	[0.09606,0.10902]
37	0.16511	[0.15251,0.17771]	0.02765	[0.02136,0.03394]	0.08992	[0.08353,0.09631]
58	0.18402	[0.17096,0.19708]	0.02977	[0.02342,0.03612]	0.09034	[0.08475,0.09593]
74	0.23564	[0.22043,0.25085]	0.05161	[0.04244,0.06078]	0.12518	[0.11771,0.13265]
82	0.18794	[0.17450,0.20138]	0.03975	[0.03203,0.04747]	0.10218	[0.09508,0.10928]
105	0.19412	[0.17765,0.21059]	0.04678	[0.03810,0.05546]	0.10350	[0.09573,0.11127]
112	0.17961	[0.16614,0.19308]	0.03449	[0.02717,0.04181]	0.09318	[0.08692,0.09944]
133	0.20320	[0.19023,0.21617]	0.01713	[0.01210,0.02216]	0.09536	[0.08997,0.10075]
141	**0.19856**	**[0.18252,0.21460]**	**0.05036**	**[0.04042,0.06030]**	**0.10633**	**[0.09783,0.11483]**
TASK 3b						
8	0.25492	[0.24006,0.26978]	0.07458	[0.06391,0.08525]	0.12730	[0.11974,0.13486]
13	0.23743	[0.22475,0.25011]	0.03617	[0.03007,0.04227]	0.10814	[0.10292,0.11336]
21	0.20713	[0.19282,0.22144]	0.05771	[0.04922,0.06620]	0.11470	[0.10728,0.12212]
38	0.20266	[0.18928,0.21604]	0.04209	[0.03454,0.04964]	0.10936	[0.10252,0.11620]
59	0.25548	[0.24154,0.26942]	0.07144	[0.06135,0.08153]	0.12555	[0.11884,0.13226]
106	0.14930	[0.13295,0.16565]	0.04700	[0.03806,0.05594]	0.08733	[0.07869,0.09597]
113	0.21578	[0.19948,0.23208]	0.05154	[0.04167,0.06141]	0.11100	[0.10256,0.11944]
134	0.25866	[0.24463,0.27269]	0.04732	[0.03985,0.05479]	0.12919	[0.12273,0.13565]
142	**0.21751**	**[0.20183,0.23319]**	**0.07625**	**[0.06626,0.08624]**	**0.12644**	**[0.11729,0.13559]**
151	0.20562	[0.19071,0.22053]	0.05772	[0.04864,0.06680]	0.11459	[0.10664,0.12254]
TASK 3c						
39	0.16511	[0.15251,0.17771]	0.02765	[0.02136,0.03394]	0.08992	[0.08353,0.09631]
143	**0.19257**	**[0.17682,0.20832]**	**0.04592**	**[0.03669,0.05515]**	**0.10265**	**[0.09446,0.11084]**

TASK 4						
Peer Code	**ROUGE-1 (unigram)**	**95% Confidence Interval**	**ROUGE-2 (bigram)**	**95% Confidence Interval**	**ROUGE-W (LCS)**	**95% Confidence Interval**
W	0.41188	[0.38696,0.43680]	0.10609	[0.08905,0.12313]	0.13542	[0.12620,0.14464]
X	0.42925	[0.40680,0.45170]	0.12213	[0.10180,0.14246]	0.14147	[0.13361,0.14933]
Y	0.44450	[0.42298,0.46602]	0.12815	[0.10965,0.14665]	0.14348	[0.13456,0.15240]
Z	0.43263	[0.40875,0.45651]	0.11953	[0.10186,0.13720]	0.14019	[0.13056,0.14982]
TASK 4a						
4	0.32939	[0.30321,0.35557]	0.07294	[0.06118,0.08470]	0.10209	[0.09471,0.10947]
14	0.30072	[0.24207,0.35937]	0.08208	[0.06100,0.10316]	0.09690	[0.07785,0.11595]
22	0.38654	[0.36352,0.40956]	0.09063	[0.07794,0.10332]	0.11621	[0.10980,0.12262]
40	0.37960	[0.35809,0.40111]	0.09408	[0.08367,0.10449]	0.12240	[0.11659,0.12821]
46	0.29599	[0.26607,0.32591]	0.05773	[0.04882,0.06664]	0.09068	[0.08530,0.09606]
60	0.36055	[0.34233,0.37877]	0.07432	[0.06371,0.08493]	0.10311	[0.09830,0.10792]
68	0.38156	[0.36420,0.39892]	0.09808	[0.08686,0.10930]	0.11888	[0.11255,0.12521]
83	0.37812	[0.34957,0.40667]	0.10311	[0.08659,0.11963]	0.12023	[0.11143,0.12903]
107	0.38615	[0.35548,0.41682]	0.09851	[0.08225,0.11477]	0.11951	[0.11004,0.12898]
114	0.28589	[0.26687,0.30491]	0.04185	[0.03278,0.05092]	0.09538	[0.08937,0.10139]
144	**0.38827**	**[0.36261,0.41393]**	**0.10109**	**[0.08680,0.11538]**	**0.11140**	**[0.10471,0.11809]**
148	0.18853	[0.17300,0.20406]	0.00199	[0.00022,0.00376]	0.05058	[0.04729,0.05387]
TASK 4b						
15	0.34628	[0.31291,0.37965]	0.10452	[0.08836,0.12068]	0.11403	[0.10323,0.12483]
23	0.41577	[0.39333,0.43821]	0.12828	[0.10994,0.14662]	0.13823	[0.12995,0.14651]
41	0.39223	[0.36910,0.41536]	0.10862	[0.09293,0.12431]	0.12735	[0.11981,0.13489]
47	0.36269	[0.31643,0.40895]	0.10967	[0.09180,0.12754]	0.10828	[0.09438,0.12218]
61	0.39125	[0.37061,0.41189]	0.10943	[0.09355,0.12531]	0.12797	[0.12040,0.13554]
69	0.39844	[0.37440,0.42248]	0.11395	[0.09885,0.12905]	0.12861	[0.12000,0.13722]
84	0.41012	[0.38543,0.43481]	0.12510	[0.10506,0.14514]	0.13574	[0.12638,0.14510]
108	0.40059	[0.37002,0.43116]	0.12087	[0.10212,0.13962]	0.13011	[0.12029,0.13993]
115	0.30037	[0.28271,0.31803]	0.04916	[0.03855,0.05977]	0.10307	[0.09549,0.11065]
145	**0.40602**	**[0.36783,0.44421]**	**0.12833**	**[0.10375,0.15291]**	**0.12221**	**[0.11128,0.13314]**
149	0.23381	[0.21750,0.25012]	0.00259	[0.00093,0.00425]	0.05654	[0.05372,0.05936]
TASK 4c						
42	0.35679	[0.33655,0.37703]	0.08647	[0.07373,0.09921]	0.12272	[0.11590,0.12954]
48	0.30711	[0.27417,0.34005]	0.06253	[0.05050,0.07456]	0.09881	[0.09008,0.10754]
70	0.38576	[0.37034,0.40118]	0.10146	[0.08866,0.11426]	0.12129	[0.11597,0.12661]
85	0.37841	[0.35642,0.40040]	0.09275	[0.08054,0.10496]	0.11948	[0.11197,0.12699]
146	**0.38818**	**[0.36373,0.41263]**	**0.09708**	**[0.07936,0.11480]**	**0.11060**	**[0.10268,0.11852]**
150	0.31581	[0.28416,0.34746]	0.06458	[0.05216,0.07700]	0.09466	[0.08831,0.10101]

Table 6: Summary of official ROUGE scores for Tasks 3 and 4. Peer codes: baseline[3-4], manual[W-Z], and system submissions