Overview of the TAC 2008 Update Summarization Task

Hoa Trang Dang and **Karolina Owczarzak**
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899
hoa.dang@nist.gov, karolina.owczarzak@nist.gov

Abstract

The summarization track at the Text Analysis Conference (TAC) is a direct continuation of the Document Understanding Conference (DUC) series of workshops, focused on providing common data and evaluation framework for research in automatic summarization. In the TAC 2008 summarization track, the main task was to produce two 100-word summaries from two related sets of 10 documents, where the second summary was an update summary. While all of the 71 submitted runs were automatically scored with the ROUGE and BE metrics, NIST assessors manually evaluated only 57 of the submitted runs for readability, content, and overall responsiveness.

1 Introduction

The TAC summarization track is a continuation of the Document Understanding Conference (DUC) series of workshops, which focused on evaluation of automatic text summarization systems. The main task of the TAC 2008 summarization track was a refinement of the update summarization pilot task of DUC 2007 and consisted of two types of summaries[1]:

1. Initial (Summary A): a 100-word summary of a set of 10 newswire articles about a particular topic.

2. Update (Summary B): a 100-word summary of a subsequent set of 10 newswire articles for the same topic, under the assumption that the reader has already read the first 10 documents. The purpose of the update summary is to inform the reader of new information about the topic.

The task is based on a scenario in which a user has a standing question that gets asked of an IR/Summarization system at two different times. The first time, the system retrieves a number of relevant newswire articles, which the user reads completely. Later (perhaps the next day, or even weeks later), the user has time to return to the system to see if there are any updates concerning his question of interest. New articles have arrived, and the system must generate an update summary of the new articles, under the assumption that the user has already read the initial articles.

NIST assessors acted as surrogate users in the task and manually assessed the summaries in terms of their content, readability (or linguistic quality), and overall responsiveness. The summary content was assessed with the Pyramid annotation method developed at Columbia University (Passonneau et al., 2005). Because ROUGE (Lin, 2004) and BE (Hovy et al., 2005) are widely used by the summarization community to automatically score summary content during system development, NIST also computed ROUGE/BE scores for all summaries, in order

[1] An additional pilot task, summarizing opinions from blog documents, was run jointly with the TAC 2008 Question Answering track, and results of this pilot are reported with the TAC 2008 QA results.

to track how well the automatic measures correlate with manual ones.

2 Task and Data

The Update Summarization task at TAC 2008 consisted of two components: an initial summary and an update summary, following the pilot update task at DUC 2007.

Participants were required to summarize information from multiple documents, guided by a topic statement describing the reader's need for information. An example topic statement, including a title and a narrative, is shown below:

> **num**: D0842G
> **title**: Natural Gas Pipeline
> **narr**: Follow the progress of pipelines being built to move natural gas from Asia to Europe. Include any problems encountered and implications resulting from the pipeline construction.

The documents for summarization came from the AQUAINT-2 collection of newswire articles. The AQUAINT-2 collection is a subset of the LDC English Gigaword Third Edition (LDC catalog number LDC2007T07) and comprises approximately 2.5 GB of text (about 907K documents) spanning the time period of October 2004 - March 2006. Articles are in English and come from a variety of sources including Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, and the Associated Press.

48 topics were developed by 8 NIST assessors, who also selected 20 AQUAINT-2 documents relevant to each topic. The retrieved documents were ordered chronologically and divided into two sets of 10 documents each, such that Set B followed Set A in the temporal order. The assessors constructed a topic statement, which was a request for information that could be answered using the selected documents. The topic statement could be in the form of a question or statement and could include background information that the assessor thought would help clarify his/her information need.

The summarization task was the same for each peer (human or automatic) summarizer: Given a topic and a set of documents relevant to the topic, the summarization task was to create from the documents two brief, well-organized, fluent summaries, A and B, that answer the need for information expressed in the topic. Summary A was the summary of the first 10 documents (Set A), while Summary B was the summary of the second 10 documents (Set B), on the assumption that the reader is already familiar with documents from Set A.

The summaries could be no longer than 100 words (whitespace-delimited tokens). Automatic summaries over the size limit were truncated, and no bonus was given for creating a shorter summary. No specific formatting other than linear was allowed.

For each document set, NIST assessors wrote 4 human summaries; one of these summaries was always written by the topic developer.

Each team participating in the update summarization task was allowed to submit up to three prioritized runs, where a run consist of exactly one summary per document set. All submitted summaries were required to be fully automatic.

The TAC 2008 Summarization track had 33 participating teams from around the world. 14 teams submitted 3 system runs, 10 teams submitted 2 runs; the rest submitted single runs only. The teams submitted a total of 71 runs, and each of these runs was assigned a numeric peer ID. The participating organizations, their submitted runs, and the peer IDs are presented in Tables 1 and 2. For comparison purposes, NIST also created a baseline automatic summarizer (peer ID=0), which selected the first few sentences of the most recent document in the relevant document set, such that their combined length did not exceed 100 words. In addition to automatic peers, the 8 human peers were assigned alphabetic IDs, A-H.

3 System Approaches

Without exception, all the submitted systems produced extractive summaries, ranking sentences in documents according to their value for a prospective summary, then extracting those sentences, sometimes with partial compression, up to a length limit (100 words in this year's task). To produce an adequate update summary, the most popular approach was to use the same anti-redundancy techniques as for the main summary, but this time comparing each

First-priority runs

ID	Run name	Organization
0		NIST (baseline)
1	abawakid1	University of Birmingham
2	AUEBNLP1	Athens University of Economics and Business
3	CCNU1	Huazhong Normal University
4	ceaList1	French Atomic Energy Commission
5	ClaC1	Concordia University
6	CLASSY1	IDA Center for Computing Sciences
7	crchowdary1	AIDB Lab
8	csiro1	CSIRO
9	DemokritosGR1	National Center of Scientific Research "Demokritos"
10	EMLR1	EML Research gGmbH
11	HITIRTMDS1	Information Retrieval Lab, Harbin Institute of Technology
12	ICL081	Peking University
13	ICSI1	International Computer Science Institute
14	ICTCAS1	Institute of Computing Technology
15	IIITSum081	Language Technologies Research Centre
16	kkireyev1	University of Colorado - Boulder
17	LIA1	Université d'Avignon
18	LIPN1	Universite Paris 13
19	Miracl1	MIRACL Laboratory
20	NUS1	National University of Singapore
21	OGI1	Oregon Health and Science University
22	PolyU1	The Hong Kong Polytechnic University
23	THUSUM1	Tsinghua University
24	RaliLatl1	Université de Montreal
25	Sutler1	University of West Bohemia
26	TOC1	Thomson Corp
27	txsumm1	University of Houston
28	uavua1	University of Antwerp and Vrije Universiteit Amsterdam
29	UBC1	University of British Columbia
30	UMD1	University of Maryland
31	UofL1	University of Lethbridge
32	UofO1	University of Ottawa
33	VensesTeam1	Università Ca' Foscari

Second-priority runs

ID	Run name	Organization
34	abawakid2	University of Birmingham
35	CCNU2	Huazhong Normal University
36	ceaList2	French Atomic Energy Commission
37	CLASSY2	IDA Center for Computing Sciences
38	csiro2	CSIRO
39	DemokritosGR2	National Center of Scientific Research "Demokritos"
40	EMLR2	EML Research gGmbH
41	HITIRTMDS2	Information Retrieval Lab, Harbin Institute of Technology
42	ICL082	Peking University
43	ICSI2	International Computer Science Institute
44	ICTCAS2	Institute of Computing Technology
45	IIITSum082	Language Technologies Research Centre
46	LIA2	Université d'Avignon
47	LIPN2	Universite Paris 13
48	Miracl2	MIRACL Laboratory
49	THUSUM2	Tsinghua University
50	RaliLatl2	Université de Montreal
51	Sutler2	University of West Bohemia
52	TOC2	Thomson Corp
53	txsumm2	University of Houston
54	uavua2	University of Antwerp and Vrije Universiteit Amsterdam
55	UBC2	University of British Columbia
56	UofL2	University of Lethbridge
57	UofO2	University of Ottawa

Table 1: Participants and first- and second-priority runs in the TAC 2008 update summarization task.

Third-priority runs

ID	Run name	Organization
58	CCNU3	Huazhong Normal University
59	ceaList3	French Atomic Energy Commission
60	CLASSY3	IDA Center for Computing Sciences
61	EMLR3	EML Research gGmbH
62	HITIRTMDS3	Information Retrieval Lab, Harbin Institute of Technology
63	ICL083	Peking University
64	ICSI3	International Computer Science Institute
65	ICTCAS3	Institute of Computing Technology
66	IIITSum083	Language Technologies Research Centre
67	Miracl3	MIRACL Laboratory
68	RaliLatl3	Université de Montreal
69	TOC3	Thomson Corp
70	uavua3	University of Antwerp and Vrije Universiteit Amsterdam
71	UofL3	University of Lethbridge

Table 2: Participants and third-priority runs in the TAC 2008 update summarization task.

candidate sentence against the first (main) set of documents as well as against the sentences previously included in the update summary. Systems often used post-processing to improve readability.

The first step for many teams was query expansion, where related words or phrases would be added to the topic and/or narrative in order to increase the possibility of finding matching sentences in the documents. This query expansion was achieved through a variety of means: WordNet (as in the submissions from NCSR Demokritos and University of Montreal), Wikipedia (EML Research), or word co-occurrences harvested from a corpus (Oregon Health and Science University). In a similar vein, these external semantic resources were used in the process of calculating the degree of overlap between the summary topic and document sentences, helping to determine each sentence's similarity or relevance to the topic. University of Birmingham and University of Paris 13 both employed WordNet for this purpose, University of Ottawa used Roget's Thesaurus, whereas University of Houston employed a slightly more complex method of calculating the distance between terms based on their place in the WordNet hierarchy. The submission of French Atomic Energy Commission, on the other hand, used senses generated from word co-occurrences in a news corpus (a common method in the field of word sense disambiguation) to assign senses to words in documents, and it calculated term frequency on the word senses instead of word forms.

Other means of improving the search for relevant sentences, whether in topic overlap or in the search for central concepts in the document sets, were lemmatization (used, among others, by University of Avignon and University of Montreal), part-of-speech tagging (Peking University, University of Paris 13), and Named Entity recognition (University of Lethbridge, University of Maryland, Athens University of Economics and Business, and many others).

The most popular features used to determine sentence relevance were sentence position in the document and sentence length. Early sentences were considered more likely to contain focused, important information, and very short and very long sentences were considered unlikely to be useful. Some participants experimented with excluding sentences that contain quotations (University of Ottawa), those that start with anaphora (Oregon Health and Science University), or pre-selecting only those sentences that have the same features as document opening sentences, on the assumption that such sentences are most likely to be focused on the topic and contain no problematic anaphoric expressions (Thomson Corp). Others ignored sentences that did not have at least some term overlap with the query (ICSI, University of Colorado-Boulder).

A number of systems employed a degree of deeper linguistic processing for the documents in question. For instance, Concordia University based their similarity measure on clustered NPs, CSIRO compared clauses instead of sentences, and EML Research looked at graphs representing Named Entities connected by dependency relations. In University of Lethbridge's submission one of the features

to determine similarity was Basic Elements overlap, and University of Montreal took into consideration word position in a parse tree to find more important words, on the assumption that they would be in higher positions. Syntactic parsing was also very often used in the process of sentence compression, allowing the systems to remove unnecessary parts of sentences, either pre- or post-selection (University of Antwerp, ICSI, University of Maryland, Huazhong Normal University). Some submissions compressed sentences without parsing, eliminating parenthetical expressions delimited by paired punctuation (OHSU), removing certain phrases like *As a matter of fact* (Thomson Corp), or *he/she said* (University of Montreal).

There were two main approaches to sentence selection: ranking and clustering. In the first case, sentences in documents were ranked according to a number of features, like *n*-gram or content word overlap with the summary topic (ICSI, University of Avignon, and others), probability of sentence given the query (Language Technologies Research Centre), or Levenstein distance between the sentence and the query (Athens University of Economics and Business). The ranking could also be a result of a ranking algorithm applied to a sentence graph (National University of Singapore, University of Antwerp). In the second approach, sentences were clustered according to similarity, and a central sentence from each cluster would be chosen for the summary (University of Colorado-Boulder, University of Paris 13). Similarity in general was evaluated either on the basis of the well-known *tf*idf* formula (University of Montreal, University of Ottawa) or Latent Semantic Analysis (University of Colorado-Boulder, University of West Bohemia).

Many participants attempted to improve their performance by combining multiple similarity features in a machine learning approach. Thomson Corp and University of Lethbridge, for example, used a Support Vector Machine based on select features, Oregon Health and Science University implemented a perceptron ranker, while University of British Columbia experimented with both supervised and unsupervised learning methods.

Producing the update summary for each topic required additional strategies to avoid redundant information which has already been covered by the main summary. In most cases, participants extended the techniques they used in the production of main summaries, but this time checking the candidate update sentences against the main documents that served to produce the main summary. One of the most frequently used methods was Maximal Marginal Relevance (CSIRO, National University of Singapore, AIDB Lab); alternatively, systems employed an upper bound on permissible similarity between sentences to reduce redundancy (University of Ottawa, Peking University).

Finally, there were a number of techniques involved in post-processing the summaries, with the goal of improving readability of the summary. Language Technologies Research Institute replaced temporal expressions in the text with dates, and eliminated sentences with too many non-content words; University of Avignon employed rewriting of acronyms and numbers in addition to temporal expressions, and deleted discourse particles such as *but*, *he/she said*, etc., and parenthetical expressions.

4 Evaluation Results

All peer summaries (manual and automatic) were evaluated with the automatic metrics ROUGE and BE. First- and second-priority runs and the model summaries were also evaluated manually in terms of content (according to the Pyramid method), readability, and overall responsiveness. The manual evaluation was performed by 8 NIST assessors. All summaries for a given topic were evaluated by a single assessor, who was also one of the summarizers, and was usually the topic developer.

In order to determine whether there were any statistically significant differences between the peers, we performed an analysis of variance (ANOVA) on all scores, followed by a multiple comparison test between individual scores according to Tukey's honestly significant difference criterion, to determine which pairs of peers are significantly different at the 95 % confidence level.

Additionally, two-way ANOVA was performed to see if there was a significant difference in scores between initial summaries (A) and update summaries (B).

4.1 Manual Evaluation

4.1.1 Overall Responsiveness and Readability

Overall responsiveness evaluated the degree to which a summary is responding to the information need contained in the topic statement, considering the summary's content as well as its linguistic quality. The readability score reflected the fluency and structure of the summary, independently of content, and was based on such aspects as grammaticality, non-redundancy, referential clarity, focus, structure, and coherence. Both overall responsiveness and readability were evaluated according to a five-point scale:

1. Very Poor
2. Poor
3. Barely Acceptable
4. Good
5. Very Good

Table 3 presents the overall responsiveness scores obtained by the models and the participants' first- and second-priority runs. Scores marked with the same letter were determined not to be significantly different in the multiple comparison test. Table 4 contains similarly marked results of the readability evaluation.

In terms of readability and overall responsiveness, it is clear that all human peers were significantly better than all automatic peers. The gap in overall responsiveness scores is larger than the gap in readability, although this difference decreases if we ignore the NIST baseline summarizer, which obtained the highest readability score of all automatic peers. The high readability score for the baseline should not surprise; as a continuous sequence of complete sentences extracted from the beginning of a human-written document, it was always well-formed linguistically. The fact that it was extracted from the *most recent* document in the set also contributed to its overall responsiveness score, as it could be expected that a most recent document would attempt to summarize previous developments on the subject. A sample baseline summary, rated "5" for readability and "3" for overall responsiveness, is shown below:

The superjumbo Airbus A380, the world's largest commercial airliner, took off Wednesday into cloudy skies over southwestern France for its second test flight. The European aircraft maker, based in the French city of Toulouse, said the second flight – which came exactly a week after the A380's highly anticipated maiden voyage – would last about four hours. As opposed to the international media hype that surrounded last week's flight, with hundreds of journalists on site to capture the historic moment, Airbus chose to conduct Wednesday's test more discreetly.

4.1.2 Pyramid Evaluation

In addition to overall responsiveness and readability, NIST assessors evaluated the content of each summary within the Pyramid evaluation framework developed at Columbia University (Passonneau et al., 2005). In the Pyramid evaluation, assessors first extract all possible "information nuggets", or Summary Content Units (SCUs) from the four model summaries on a given topic. Each SCU is assigned a weight equal to the number of model summaries in which it appears. An example SCU with its four contributors from the four model summaries is shown below:

num: D0820D-A
SCU: Mini-submarine trapped underwater
contr1: mini-submarine... became trapped... on the sea floor
contr2: a small... submarine... snagged... at a depth of 625 feet
contr3: mini-submarine was trapped... below the surface
contr4: A small... submarine... was trapped on the seabed

The number of contributors is equal to the weight of the SCU, i.e. an SCU with four contributors has a weight of 4, an SCU with 3 contributors has the weight of 3, etc. Once all SCUs are harvested from the model summaries, assessors determine how many of these SCUs can be found in each of the automatic summaries. Repetitive information is not rewarded, as only one contributor per SCU is counted for the total peer SCU count. The final Pyramid score for an automatic summary is its total SCU weight divided by the maximum SCU weight available to a summary of average length (where the av-

erage length is determined by the mean SCU count of the model summaries for this topic).

A Pyramid score was also calculated for the model summaries, evaluating each of them against the remaining three models for a given topic. In order to provide a fair comparison with the scores obtained by the automatic submissions (evaluated against all four models), a mean of four scores, computed with three model summaries each, was calculated for the automatic submissions as well.

Table 5 gives the results of the multiple comparison of the Pyramid scores. All the scores in the table were calculated with 3 models. As in previous tables, scores which share the same letter are not significantly different at the 95% level. The Pyramid evaluation, similarly to readability and overall responsiveness, makes a clear distinction between all human peers and all automatic peers, although it also differentiates between the human peers themselves.

4.1.3 Analysis

Since overall responsiveness measures both the linguistic quality and the content quality of a summary, and the Pyramid score can be thought of as a pure content measure, it is interesting to examine the relations between the manual evaluation metrics. Table 6 contains correlations between overall responsiveness on the one hand, and readability and the Pyramid score on the other. Correlations for automatic peers are all statistically significant with p-values close to zero; however, the situation for human peers looks different, as the correlations between overall responsiveness and the Pyramid score are not significant at the 95% confidence level.

Because content is a large component of overall responsiveness, it is not surprising that the Pyramid score and overall responsiveness are highly correlated for automatic peers. What is surprising is their low and insignificant correlation for human peers. Note, however, that for human peers a higher correlation is obtained by readability, which is especially visible in Spearman's rank comparison. This could be explained by the fact that parallel human summaries can display certain variability in content, and yet still be thought of as relevant and responsive. In that case, readability would be a greater influence on the overall responsiveness than the strict content identity measured by the Pyramid method.

The situation is reversed for automatic peers, where we see high correlations of overall responsiveness with the Pyramid score but lower with readability. This might be for two reasons: first, it is easy to produce a perfectly fluent summary with little or no relevant content (*vide* NIST baseline), which would explain the relatively low correlation with readability. Second, content cannot ever be fully divorced from linguistic quality; rather, some well-formedness (at least of short sequences) is necessary for the meaning to be conveyed. Therefore, it is likely that a summary assessed as high in content cannot be completely unreadable, and conversely, a completely unreadable summary cannot be high in content. A detailed comparison of manual scores on a summary level could shed more light on these relations.

In order to see whether the performance of peers was different for initial versus update summaries, we separately calculated average per-topic scores over all automatic peers and human peers. Table 7 shows macroaverages of per-topic scores. The overall responsiveness of the automatic peers was significantly different between Summaries A and B; this difference was confirmed by a t-test on per-topic scores, and by a two-way ANOVA on mean submission scores. The Pyramid scores of automatic peers was also significantly lower for Summary B than for Summary A. This suggests that, for most participants, it was more difficult to produce a responsive update summary than an initial summary, while the lingustic quality and well-formedness were not dependent on the summary type (initial vs. update).

4.2 Automatic ROUGE/BE Evaluation

While the manual evaluation could only be applied to the first- and second-priority runs submitted by the participating teams, automatic scores were produced for all submitted runs. We calculated two versions of ROUGE: ROUGE-2 and ROUGE-SU4, as well as BE-HM recall. For the BE evaluation, summaries were parsed with Minipar, and BE-F were extracted and matched using the Head-Modifier criterion. Each automatic score was computed using stemming and implementing jackknifing for each [peer, topic] pair so that human and automatic peers could be compared.

Similarly to manual evaluation, NIST conducted an analysis of variance (ANOVA) and a multiple comparison of the scores in order to determine which differences were statistically significant. Tables 11-13 show the results of the multiple comparison.

The profile emerging from the automatic evaluation is rather different than the one based on manual assessment. Not only is there no significant gap between models and systems, but in many cases certain automatic peers are scored higher than some human models. Moreover, as can be seen in the tables, ROUGE and BE scoring leads to larger confidence intervals in the case of models than in the case of similarly scored systems. This effect is directly tied to the greater variance of model scores in the ROUGE/BE evaluation. Automatic metrics, based on string matching, are unable to appreciate a summary that uses different phrases than the reference text, even if such a summary is perfectly fine by human standards. This leads to low scores for some models, and, consequently, wider confidence intervals.

However, both ROUGE-2 and ROUGE-SU4 show the same distinction between the quality of initial and update summaries that overall responsiveness provided: the significant gap in scores is present for the automatic peers, but not for the human peers. Table 8 presents macro-averaged per-topic scores; values in bold mark those pairs of averages for Summary A and Summary B which are significantly different. BE-HM, according to a two-tailed t-test, makes no distinction between the two types of summaries at the 95% confidence level.

4.2.1 Correlation

To check how well the automatic evaluation metrics correlate with manual scores, we computed Pearson's and Spearman's correlation coefficients for recall of ROUGE-2, ROUGE-SU4, and BE-HM vs the Pyramid score (Table 9) and overall responsiveness (Table 10). The tables show the correlations separately for human peers and automatic peers. Using Fisher's z' transformation, we obtained a normal distribution for the correlations and calculated confidence intervals. We concluded that none of the differences in Pearson's or Spearman's correlations between metrics are statistically significant at the 95% level, i.e. ROUGE-2, ROUGE-SU4, and BE-HM are equally good at predicting manual scores.

While ROUGE/BE correlations with manual assessment are universally high for automatic peers, when it comes to human peers, only overall responsiveness is relatively well reflected in their scores. This might be because there were only 8 human peers, in contrast to 58 automatic ones, so the data is too sparse to notice a trend.

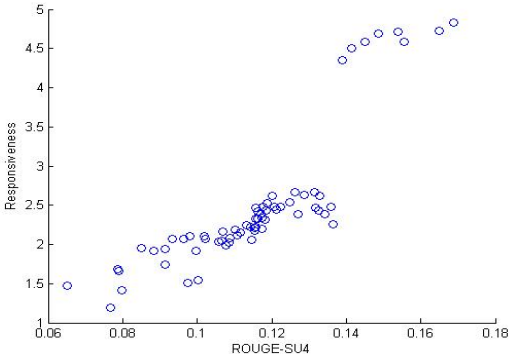

Figure 1: Average overall responsiveness vs. average ROUGE-SU4 recall with stemming.

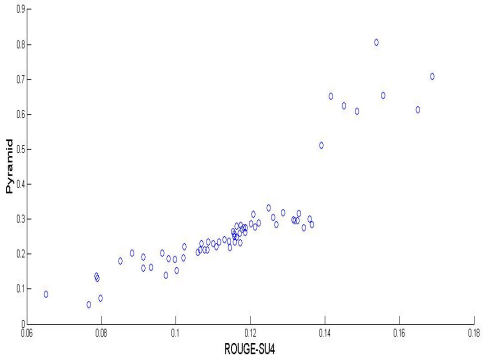

Figure 2: Average Pyramid score vs. average ROUGE-SU4 recall with stemming.

Figures 1 and 2 show the relation between ROUGE-SU4 and overall responsiveness and Pyramid scores, respectively. Such a comparison shows where human peers and automatic peers are placed on the evaluation scales of different metrics. The figures are similar for ROUGE-2 and BE, where the 8 human peers are clustered into a group with the highest manual scores. The gap between humans

and systems, clearly visible in the overall responsiveness and in the Pyramid scores, is not present in ROUGE and BE scores.

5 Conclusions

The TAC 2008 update summarization task showed that there still exists a significant gap between automatic summarizers and human summarizers based on manual evaluations of summary quality: readability, content (Pyramid), and overall responsiveness. Additionally, while humans are equally adept at writing update summaries versus initial summaries, automatic summarizers have greater difficulty with selecting content and producing responsive summaries for the update summaries.

A comparison of the automatic evaluation metrics developed for DUC showed that ROUGE-2, ROUGE-SU4, and BE-HM all correlate highly (indeed, equally highly) with the manual metrics. However, BE-HM fails to detect that automatic summarizers (as a group) perform more poorly on update summaries than on initial summaries. While automatic evaluation metrics have been evaluated based only on their Pearson/Spearman correlations with manual metrics, it is reasonable to want automatic metrics to be able to mimic manual metrics in other aspects, including discriminative power (between human and automatic peers, or between different tasks). To formalize these goals, TAC 2009 will include a new AESOP task (Automatically Evaluating Summaries of Peers), which will more systematically evaluate automatic evaluation metrics.

References

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating duc 2005 using basic elements. In *Proceedings of the Fifth Document Understanding Conference (DUC)*, Vancouver, Canada.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in duc 2005. In *Proceedings of the Fifth Document Understanding Conference (DUC)*, Vancouver, Canada.

```
-------Summaries A---------------------------------Summaries B--------------------------
ID      Score    Significance                    ID      Score    Significance

F       4.7917   A                               D       4.875    A
D       4.7917   A                               G       4.75     A
A       4.75     A                               H       4.6667   A
G       4.6667   A                               F       4.6667   A
B       4.5833   A                               A       4.625    A
H       4.5      A                               B       4.5833   A
C       4.4583   A                               C       4.5417   A
E       4.4167   A                               E       4.2917   A
50      2.7917     B                             14      2.6042     B
26      2.7917     B                             49      2.5833     B
12      2.7708     B C                           23      2.5833     B
49      2.75       B C                           11      2.5625     B
44      2.75       B C                           44      2.5208     B
42      2.75       B C                           24      2.5        B C
23      2.75       B C                           50      2.4583     B C
52      2.7083     B C D                         41      2.4375     B C
13      2.6875     B C D                         37      2.4167     B C
25      2.6667     B C D                         6       2.3958     B C
51      2.6458     B C D                         19      2.3333     B C D
14      2.625      B C D                         17      2.3333     B C D
45      2.6042     B C D                         1       2.3333     B C D
2       2.6042     B C D                         34      2.3125     B C D E
37      2.5417     B C D E                       25      2.2917     B C D E
24      2.5417     B C D E                       52      2.25       B C D E F
6       2.5417     B C D E                       51      2.25       B C D E F
41      2.5208     B C D E F                     46      2.25       B C D E F
11      2.5208     B C D E F                     4       2.2083     B C D E F G
1       2.5208     B C D E F                     45      2.1667     B C D E F G
35      2.5        B C D E F                     13      2.1667     B C D E F G
30      2.5        B C D E F                     2       2.1667     B C D E F G
3       2.4792     B C D E F                     43      2.1458     B C D E F G
34      2.4583     B C D E F                     26      2.1458     B C D E F G
15      2.4375     B C D E F                     48      2.125      B C D E F G H
46      2.4167     B C D E F                     5       2.125      B C D E F G H
22      2.3958     B C D E F                     42      2.0833     B C D E F G H I
43      2.375      B C D E F                     29      2.0625     B C D E F G H I J
10      2.375      B C D E F                     16      2.0625     B C D E F G H I J
54      2.3542     B C D E F G                   10      2.0625     B C D E F G H I J
36      2.3542     B C D E F G                   54      2.0208     B C D E F G H I J K
56      2.3333     B C D E F G H                 32      2          B C D E F G H I J K
33      2.3125     B C D E F G H                 22      2          B C D E F G H I J K
17      2.3125     B C D E F G H                 3       2          B C D E F G H I J K
0       2.2917     B C D E F G H                 55      1.9792     B C D E F G H I J K
19      2.2917     B C D E F G H                 15      1.9792     B C D E F G H I J K
57      2.2292     B C D E F G H I               36      1.9583     B C D E F G H I J K
48      2.2292     B C D E F G H I               35      1.9375     B C D E F G H I J K
27      2.2292     B C D E F G H I               57      1.9167     B C D E F G H I J K
20      2.2292     B C D E F G H I               21      1.9167     B C D E F G H I J K
55      2.1667     B C D E F G H I               12      1.9167     B C D E F G H I J K
16      2.1667     B C D E F G H I               33      1.8958     B C D E F G H I J K
40      2.125      B C D E F G H I               20      1.8958     B C D E F G H I J K
21      2.125      B C D E F G H I               31      1.875      B C D E F G H I J K
4       2.125      B C D E F G H I               27      1.875      B C D E F G H I J K
29      2.1042     B C D E F G H I               0       1.8542     B C D E F G H I J K L
7       2.1042     B C D E F G H I               40      1.8542     B C D E F G H I J K L
53      2.0833     B C D E F G H I               53      1.75         C D E F G H I J K L
5       2.0833     B C D E F G H I               28      1.6042         D E F G H I J K L
32      2.0625       C D E F G H I               56      1.5625           E F G H I J K L
31      2              D E F G H I J             47      1.5                F G H I J K L
28      1.875            E F G H I J             8       1.4583               G H I J K L
47      1.8125             F G H I J             38      1.375                  H I J K L
38      1.6458               G H I J             30      1.3333                   I J K L
18      1.6458               G H I J             18      1.3125                     J K L
8       1.625                  H I J             39      1.2917                       K L
39      1.5417                   I J             7       1.2708                       K L
9       1.2917                     J             9       1.1042                         L
```

Table 3: Overall responsiveness results for the TAC 2008 update summarization task for summaries A and B. Peers not sharing a common letter are significantly different at the 95% confidence level.

-------Summaries A----------			-------Summaries B----------		
ID	Score	Significance	ID	Score	Significance
F	4.9167	A	D	4.9583	A
G	4.875	A	F	4.875	A
D	4.875	A	A	4.875	A
B	4.8333	A	G	4.8333	A
A	4.7917	A	B	4.7917	A
E	4.75	A	H	4.75	A
H	4.625	A	E	4.7083	A
C	4.625	A	C	4.5833	A
0	3.25	B	0	3.4167	B
50	3	B C	49	3.2083	B C
49	2.9375	B C D	23	3.1042	B C D
24	2.9375	B C D	52	2.9792	B C D E
26	2.875	B C D E	26	2.8958	B C D E F
51	2.8333	B C D E F	25	2.8958	B C D E F
52	2.8125	B C D E F G	44	2.8542	B C D E F
23	2.8125	B C D E F G	34	2.8542	B C D E F
1	2.75	B C D E F G	46	2.8333	B C D E F G
44	2.7292	B C D E F G H	24	2.8333	B C D E F G
34	2.6667	B C D E F G H I	14	2.8333	B C D E F G
33	2.6458	B C D E F G H I	51	2.7917	B C D E F G H
25	2.6458	B C D E F G H I	50	2.7917	B C D E F G H
14	2.5833	B C D E F G H I	1	2.6875	B C D E F G H I
47	2.5625	B C D E F G H I J	45	2.6667	B C D E F G H I J
17	2.5417	B C D E F G H I J K	6	2.6667	B C D E F G H I J
6	2.5208	B C D E F G H I J K	37	2.6458	B C D E F G H I J K
12	2.5	B C D E F G H I J K L	17	2.5833	C D E F G H I J K L
56	2.4792	B C D E F G H I J K L M	11	2.5417	C D E F G H I J K L
20	2.4792	B C D E F G H I J K L M	10	2.5	C D E F G H I J K L
13	2.4792	B C D E F G H I J K L M	5	2.5	C D E F G H I J K L
46	2.4583	B C D E F G H I J K L M	31	2.4792	C D E F G H I J K L
37	2.4583	B C D E F G H I J K L M	16	2.4792	C D E F G H I J K L
35	2.4583	B C D E F G H I J K L M	13	2.4792	C D E F G H I J K L
16	2.4375	B C D E F G H I J K L M	4	2.4583	C D E F G H I J K L
54	2.4167	C D E F G H I J K L M	22	2.4375	C D E F G H I J K L
57	2.3958	C D E F G H I J K L M	35	2.3958	C D E F G H I J K L M
31	2.3958	C D E F G H I J K L M	27	2.3958	C D E F G H I J K L M
15	2.3958	C D E F G H I J K L M	2	2.3958	C D E F G H I J K L M
10	2.3958	C D E F G H I J K L M	48	2.375	D E F G H I J K L M N
3	2.3958	C D E F G H I J K L M	53	2.3542	D E F G H I J K L M N
45	2.375	C D E F G H I J K L M	15	2.3333	D E F G H I J K L M N
41	2.375	C D E F G H I J K L M	36	2.2917	D E F G H I J K L M N
22	2.375	C D E F G H I J K L M	19	2.2917	D E F G H I J K L M N
4	2.375	C D E F G H I J K L M	41	2.2708	E F G H I J K L M N
27	2.3542	C D E F G H I J K L M N	3	2.2708	E F G H I J K L M N
5	2.3542	C D E F G H I J K L M N	33	2.2292	E F G H I J K L M N
2	2.3125	C D E F G H I J K L M N	20	2.2292	E F G H I J K L M N
11	2.2708	C D E F G H I J K L M N O	57	2.1667	E F G H I J K L M N O
53	2.25	C D E F G H I J K L M N O	54	2.1667	E F G H I J K L M N O
21	2.2083	C D E F G H I J K L M N O	21	2.1458	F G H I J K L M N O
36	2.1875	C D E F G H I J K L M N O P	47	2.125	F G H I J K L M N O
7	2.1667	D E F G H I J K L M N O P	32	2.0208	G H I J K L M N O P
42	2.0833	E F G H I J K L M N O P	40	2	H I J K L M N O P Q
19	2.0833	E F G H I J K L M N O P	43	1.9792	H I J K L M N O P Q
48	2.0417	F G H I J K L M N O P	42	1.9792	H I J K L M N O P Q
43	2.0208	F G H I J K L M N O P	30	1.875	I J K L M N O P Q
32	2	G H I J K L M N O P	29	1.8542	J K L M N O P Q
30	2	G H I J K L M N O P	56	1.8333	K L M N O P Q
40	1.9167	H I J K L M N O P	18	1.8333	K L M N O P Q
55	1.8958	I J K L M N O P	39	1.8125	L M N O P Q
29	1.75	J K L M N O P	9	1.8125	L M N O P Q
39	1.7292	K L M N O P	55	1.7708	L M N O P Q
18	1.6875	L M N O P	28	1.5833	M N O P Q
28	1.6667	M N O P	12	1.5625	N O P Q
38	1.5417	N O P	38	1.3542	O P Q
9	1.4583	O P	8	1.25	P Q
8	1.375	P	7	1.1875	Q

Table 4: Readability results for the TAC 2008 update summarization task for summaries A and B. Peers not sharing a common letter are significantly different at the 95% confidence level.

```
-------Summaries A-----------------------------------------------
ID   Score    Significance
G    0.84808  A
D    0.71892  A B
F    0.67217    B C
H    0.64383    B C
C    0.64375    B C
A    0.629      B C
B    0.62442    B C
E    0.52783      C
30   0.35929        D
13   0.34204        D E
41   0.34006        D E F
44   0.3334         D E F G
42   0.33004        D E F G H
11   0.328          D E F G H
45   0.32744        D E F G H
49   0.32448        D E F G H I
6    0.3229         D E F G H I
43   0.32133        D E F G H I
23   0.31467        D E F G H I
37   0.31413        D E F G H I
52   0.31065        D E F G H I
12   0.30475        D E F G H I
50   0.30438        D E F G H I J
2    0.30265        D E F G H I J K
14   0.29854        D E F G H I J K L
25   0.29796        D E F G H I J K L M
26   0.29792        D E F G H I J K L M
3    0.29675        D E F G H I J K L M
48   0.29554        D E F G H I J K L M N
35   0.29531        D E F G H I J K L M N
19   0.28973        D E F G H I J K L M N
51   0.28969        D E F G H I J K L M N
15   0.28356        D E F G H I J K L M N
22   0.2769         D E F G H I J K L M N
24   0.27404        D E F G H I J K L M N
54   0.2735         D E F G H I J K L M N
34   0.27244        D E F G H I J K L M N
1    0.27081        D E F G H I J K L M N
10   0.26565        D E F G H I J K L M N
17   0.26271        D E F G H I J K L M N
36   0.25969        D E F G H I J K L M N
46   0.25763        D E F G H I J K L M N
27   0.25504        D E F G H I J K L M N
20   0.24496        D E F G H I J K L M N
29   0.24227        D E F G H I J K L M N
16   0.24217        D E F G H I J K L M N
4    0.2391         D E F G H I J K L M N
56   0.2381           E F G H I J K L M N
21   0.23785          E F G H I J K L M N
55   0.2296           E F G H I J K L M N O
40   0.22667          E F G H I J K L M N O
57   0.22333          E F G H I J K L M N O
28   0.22206          E F G H I J K L M N O
32   0.22079            F G H I J K L M N O P
53   0.21788              G H I J K L M N O P
7    0.21727              G H I J K L M N O P
5    0.20967                H I J K L M N O P
33   0.20473                  I J K L M N O P
0    0.18354                    J K L M N O P Q
31   0.18227                      K L M N O P Q
38   0.18052                        L M N O P Q
8    0.17698                          M N O P Q
47   0.1749                             N O P Q
39   0.11444                              O P Q
18   0.10002                                P Q
9    0.080583                                 Q

-------Summaries B-------------------------
ID   Score    Significance
G    0.76104  A
D    0.69617  A B
H    0.66525  A B
C    0.65875  A B C
B    0.62617  A B C
A    0.588      B C
F    0.55317    B C
E    0.495        C
14   0.33581        D
11   0.33323        D E
44   0.30535        D E F
23   0.29298        D E F G
37   0.28835        D E F G H
41   0.28652        D E F G H I
25   0.28156        D E F G H I J
24   0.27692        D E F G H I J K
49   0.27415        D E F G H I J K
50   0.26952        D E F G H I J K
6    0.26894        D E F G H I J K
51   0.266          D E F G H I J K L
19   0.26179        D E F G H I J K L M
12   0.25931        D E F G H I J K L M
1    0.25029        D E F G H I J K L M N
2    0.24962        D E F G H I J K L M N
43   0.24873        D E F G H I J K L M N
34   0.24815        D E F G H I J K L M N
13   0.24777        D E F G H I J K L M N
15   0.24167        D E F G H I J K L M N O
45   0.24042        D E F G H I J K L M N O P
17   0.23525        D E F G H I J K L M N O P
52   0.23415        D E F G H I J K L M N O P
48   0.23131        D E F G H I J K L M N O P
42   0.23008        D E F G H I J K L M N O P
29   0.22633        D E F G H I J K L M N O P
26   0.2186         D E F G H I J K L M N O P
4    0.218          D E F G H I J K L M N O P
55   0.21452          E F G H I J K L M N O P Q
10   0.21               F G H I J K L M N O P Q
46   0.20988            F G H I J K L M N O P Q
36   0.2081             F G H I J K L M N O P Q
35   0.20392            F G H I J K L M N O P Q
16   0.20217            F G H I J K L M N O P Q R
40   0.19802            F G H I J K L M N O P Q R
20   0.19225            F G H I J K L M N O P Q R
32   0.19081            F G H I J K L M N O P Q R
3    0.18688            F G H I J K L M N O P Q R S
22   0.1864               G H I J K L M N O P Q R S
54   0.18627              G H I J K L M N O P Q R S
21   0.18517              G H I J K L M N O P Q R S
57   0.18087              G H I J K L M N O P Q R S
5    0.1704                 H I J K L M N O P Q R S T
27   0.16815                  I J K L M N O P Q R S T
33   0.16725                  J K L M N O P Q R S T
28   0.15981                    K L M N O P Q R S T U
53   0.14942                      L M N O P Q R S T U V
0    0.14321                        M N O P Q R S T U V W
31   0.13846                          N O P Q R S T U V W
8    0.12819                            O P Q R S T U V W
56   0.12219                              P Q R S T U V W
38   0.098604                               Q R S T U V W
47   0.084458                                 R S T U V W
18   0.069896                                   S T U V W
7    0.059187                                     T U V W
30   0.048042                                       U V W
39   0.031958                                         V W
9    0.030312                                           W
```

Table 5: Pyramid evaluation results for the TAC 2008 update summarization task for summaries A and B. Peers not sharing a common letter are significantly different at the 95% confidence level.

Metric	Pearson		Spearman	
	humans	systems	humans	systems
Readability	**0.778** (p=0.02)	**0.763** (0.628-0.853)	**0.910** (p=0.003)	**0.750**
Pyramid	0.637 (p=0.09)	**0.950** (0.916-0.970)	0.455 (p=0.26)	**0.941**

Table 6: Correlation between average overall responsiveness and the remaining manual evaluation metrics for all summaries. Correlations in bold are statistically significant ($p <= 0.05$); 95% confidence intervals are included for significant Pearson correlations.

	Responsiveness		Readability		Pyramid	
	humans	systems	humans	systems	humans	systems
Summary A	4.620	**2.324**	4.786	2.347	0.663	**0.260**
Summary B	4.625	**2.024**	4.800	2.337	0.630	**0.204**

Table 7: Macro-average per-topic manual scores calculated over all automatic peers and all human summarizers. Pairs of values in bold for Summary A vs. B are significantly different from each other at the 95% confidence level.

	ROUGE-2		ROUGE-SU4		BE-HM	
	humans	systems	humans	systems	humans	systems
Summary A	0.117	**0.079**	0.154	**0.116**	0.078	0.038
Summary B	0.117	**0.068**	0.150	**0.107**	0.089	0.039

Table 8: Macro-average per-topic automatic scores calculated over all automatic peers and all human peers. Pairs of values in bold for Summary A vs. Summary B are significantly different from each other at the 95% confidence level.

Metric	Pearson's		Spearman's	
	humans	systems	humans	systems
ROUGE-2	0.276 (p=0.5)	**0.946** (0.910-0.968)	0.429 (p=0.3)	**0.967**
ROUGE-SU4	0.457 (p=0.25)	**0.928** (0.880-0.957)	0.595 (p=0.2)	**0.951**
BE-HM	0.423 (p=0.3)	**0.949** (0.915-0.969)	0.309 (p=0.46)	**0.950**

Table 9: Correlation between average Pyramid score and average ROUGE-2/ROUGE-SU4/BE-HM recall for all summaries. Correlations in bold are significant with p-values close to zero.

Metric	Pearson's		Spearman's	
	humans	systems	humans	systems
ROUGE-2	0.725 (p=0.04)	**0.894** (0.827-0.936)	0.874 (p=0.007)	**0.920**
ROUGE-SU4	0.866 (p=0.005)	**0.874** (0.796-0.924)	0.898 (p=0.005)	**0.909**
BE-HM	0.656 (p=0.08)	**0.9106** (0.853-0.946)	0.683 (p=0.07)	**0.910**

Table 10: Correlation between average overall responsiveness and average ROUGE-2/ROUGE-SU4/BE-HM recall for all summaries. Correlations in bold are significant with p-values close to zero.

```
-------Summaries A-----------------------------------------------------
ID   Score      Significance
D    0.13133    A
F    0.12988    A B
G    0.12058    A B C
H    0.11846    A B C D
C    0.11221    A B C D E
43   0.1114     A B C D E F
13   0.11044    A B C D E F G
E    0.10992    A B C D E F G H
A    0.10983    A B C D E F G H
B    0.10842    A B C D E F G H I
60   0.10379    A B C D E F G H I J
37   0.10338    A B C D E F G H I J K
6    0.10133    A B C D E F G H I J K L
2    0.10012    A B C D E F G H I J K L M
64   0.099271   A B C D E F G H I J K L M N
45   0.095188   A B C D E F G H I J K L M N O
12   0.094979   A B C D E F G H I J K L M N O
65   0.094229   A B C D E F G H I J K L M N O P
14   0.094229   A B C D E F G H I J K L M N O P
49   0.093438     B C D E F G H I J K L M N O P
63   0.092729     B C D E F G H I J K L M N O P Q
42   0.092729     B C D E F G H I J K L M N O P Q
23   0.091812       C D E F G H I J K L M N O P Q
50   0.090583       C D E F G H I J K L M N O P Q
44   0.090292       C D E F G H I J K L M N O P Q
25   0.088563       C D E F G H I J K L M N O P Q R
11   0.088188       C D E F G H I J K L M N O P Q R
69   0.088063       C D E F G H I J K L M N O P Q R S
51   0.088          C D E F G H I J K L M N O P Q R S
41   0.087813       C D E F G H I J K L M N O P Q R S
26   0.085792       C D E F G H I J K L M N O P Q R S T
70   0.085583       C D E F G H I J K L M N O P Q R S T
54   0.085583       C D E F G H I J K L M N O P Q R S T
52   0.084958       C D E F G H I J K L M N O P Q R S T
58   0.084646       C D E F G H I J K L M N O P Q R S T
62   0.084625       C D E F G H I J K L M N O P Q R S T
22   0.084188       C D E F G H I J K L M N O P Q R S T
17   0.083917       C D E F G H I J K L M N O P Q R S T
35   0.083104         D E F G H I J K L M N O P Q R S T U
48   0.082667         D E F G H I J K L M N O P Q R S T U
68   0.082479         D E F G H I J K L M N O P Q R S T U
24   0.082479         D E F G H I J K L M N O P Q R S T U
19   0.081812         D E F G H I J K L M N O P Q R S T U
46   0.081229         D E F G H I J K L M N O P Q R S T U
15   0.081229         D E F G H I J K L M N O P Q R S T U
3    0.081187           E F G H I J K L M N O P Q R S T U
61   0.080021           E   G H I J K L M N O P Q R S T U V
10   0.080021           E   G H I J K L M N O P Q R S T U V
34   0.079667           E     H I J K L M N O P Q R S T U V
1    0.079604           E     H I J K L M N O P Q R S T U V
36   0.078375           E     H I J K L M N O P Q R S T U V
30   0.078229           E     H I J K L M N O P Q R S T U V
67   0.076833           E     H I J K L M N O P Q R S T U V
27   0.074521                 H I J K L M N O P Q R S T U V
20   0.073458                 H I J K L M N O P Q R S T U V
7    0.073146                 H I   K L M N O P Q R S T U V
66   0.071604                   I     L M N O P Q R S T U V W
16   0.07075                          M N O P Q R S T U V W
40   0.069896                         M N O P Q R S T U V W
4    0.069875                         M N O P Q R S T U V W
21   0.069083                           N O P Q R S T U V W X
53   0.069042                           N O P Q R S T U V W X
57   0.068375                             O P Q R S T U V W X
33   0.067604                             O P Q R S T U V W X
32   0.067604                             O P Q R S T U V W X
71   0.067542                             O P Q R S T U V W X
29   0.067542                             O P Q R S T U V W X
8    0.067167                             O P Q R S T U V W X
28   0.06575                              O P Q R S T U V W X
56   0.065333                             O P Q R S T U V W X
38   0.064875                             O P Q R S T U V W X
5    0.064542                               P Q R S T U V W X
55   0.062354                                 Q R S T U V W X
0    0.058229                                   R S T U V W X
47   0.05775                                      S T U V W X
59   0.056                                          T U V W X
31   0.052833                                         U V W X
39   0.050583                                           V W X
9    0.042354                                             W X
18   0.039188                                               X

-------Summaries B-------------------------
ID   Score      Significance
D    0.13171    A
F    0.12779    A B
H    0.12137    A B C
A    0.11325    A B C D
B    0.11288    A B C D E
E    0.11129    A B C D E F
G    0.10933    A B C D E F G
C    0.10617    A B C D E F G H
14   0.10108    A B C D E F G H I
65   0.096729   A B C D E F G H I J
43   0.096542   A B C D E F G H I J
2    0.092375   A B C D E F G H I J K
49   0.091667   A B C D E F G H I J K
62   0.089813     B C D E F G H I J K
44   0.089813     B C D E F G H I J K
23   0.089479     B C D E F G H I J K L
11   0.088958     B C D E F G H I J K L
69   0.088167     B C D E F G H I J K L M
13   0.087521     B C D E F G H I J K L M N
64   0.085604       C D E F G H I J K L M N O
37   0.085333       C D E F G H I J K L M N O
60   0.085208       C D E F G H I J K L M N O
1    0.082187       C D E F G H I J K L M N O P
34   0.081875       C D E F G H I J K L M N O P Q
24   0.081604       C D E F G H I J K L M N O P Q
25   0.081375       C D E F G H I J K L M N O P Q
68   0.080354         D E F G H I J K L M N O P Q R
41   0.079208         D E F G H I J K L M N O P Q R S
45   0.079083         D E F G H I J K L M N O P Q R S
19   0.078333         D E F G H I J K L M N O P Q R S
6    0.078167         D E F G H I J K L M N O P Q R S
50   0.078042         D E F G H I J K L M N O P Q R S
51   0.0775           D E F G H I J K L M N O P Q R S
52   0.074896         D E F G H I J K L M N O P Q R S
20   0.073562         D E F G H I J K L M N O P Q R S
29   0.073417         D E F G H I J K L M N O P Q R S
48   0.073292         D E F G H I J K L M N O P Q R S
63   0.072896           E F G H I J K L M N O P Q R S
15   0.072167             F G H I J K L M N O P Q R S
26   0.072104             F G H I J K L M N O P Q R S
12   0.071042             F G H I J K L M N O P Q R S
61   0.070104               G H I J K L M N O P Q R S
17   0.070021               G H I J K L M N O P Q R S
67   0.069271               G H I J K L M N O P Q R S
36   0.069208               G H I J K L M N O P Q R S
4    0.069208               G H I J K L M N O P Q R S
10   0.068667                 H I J K L M N O P Q R S T
42   0.068312                 H I J K L M N O P Q R S T
16   0.068021                 H   J K L M N O P Q R S T
22   0.067875                 H   J K L M N O P Q R S T
46   0.067771                 H   J K L M N O P Q R S T
58   0.067375                 H   J K L M N O P Q R S T U
35   0.067187                 H   J K L M N O P Q R S T U V
21   0.065979                 H   J K L M N O P Q R S T U V W
32   0.065625                     J K L M N O P Q R S T U V W
40   0.064333                     J K L M N O P Q R S T U V W X
55   0.063208                       K L M N O P Q R S T U V W X
27   0.06275                        K L M N O P Q R S T U V W X
3    0.062125                       K L M N O P Q R S T U V W X
66   0.060083                       K L M N O P Q R S T U V W X Y
70   0.059958                       K L M N O P Q R S T U V W X Y
0    0.059875                       K L M N O P Q R S T U V W X Y
54   0.059708                       K L M N O P Q R S T U V W X Y
5    0.056875                         L M N O P Q R S T U V W X Y
57   0.056042                           M N O P Q R S T U V W X Y
59   0.0555                             M N O P Q R S T U V W X Y
53   0.055                                N O P Q R S T U V W X Y
33   0.054021                               O P Q R S T U V W X Y
8    0.0515                                   P Q R S T U V W X Y
38   0.049167                                   Q R S T U V W X Y Z
31   0.048479                                     R S T U V W X Y Z
28   0.046354                                       S T U V W X Y Z
47   0.035854                                         T U V W X Y Z
56   0.034729                                           U V W X Y Z
39   0.034396                                             V W X Y Z
30   0.034208                                               W X Y Z
71   0.032417                                                 X Y Z
18   0.028042                                                   Y Z
9    0.027333                                                   Y Z
7    0.017729                                                     Z
```

Table 11: ROUGE-2 results for the TAC 2008 update summarization task for summaries A and B. Peers not sharing a common letter are significantly different at the 95% confidence level.

-------Summaries A----------

ID	Score	Significance
D	0.17	A
F	0.16733	A B
G	0.15533	A B C
H	0.15375	A B C D
A	0.15254	A B C D E
B	0.14667	A B C D E F
C	0.14612	A B C D E F G
43	0.14298	A B C D E F G H
37	0.14277	A B C D E F G H
60	0.142	A B C D E F G H I
E	0.14046	A B C D E F G H I J
13	0.13985	A B C D E F G H I J K
6	0.13977	A B C D E F G H I J K
2	0.13694	B C D E F G H I J K L
45	0.13394	C D E F G H I J K L M
49	0.13104	C D E F G H I J K L M N
64	0.12917	C D E F G H I J K L M N O
65	0.12896	C D E F G H I J K L M N O
14	0.12896	C D E F G H I J K L M N O
44	0.12727	C D E F G H I J K L M N O P
12	0.12725	C D E F G H I J K L M N O P
23	0.12665	C D E F G H I J K L M N O P Q
63	0.126	C D E F G H I J K L M N O P Q
42	0.126	C D E F G H I J K L M N O P Q
69	0.1244	C D E F G H I J K L M N O P Q R
58	0.12417	C D E F G H I J K L M N O P Q R
25	0.12371	C D E F G H I J K L M N O P Q R
51	0.12363	C D E F G H I J K L M N O P Q R
35	0.12325	C D E F G H I J K L M N O P Q R
41	0.12321	C D E F G H I J K L M N O P Q R
22	0.12315	C D E F G H I J K L M N O P Q R
11	0.12296	C D E F G H I J K L M N O P Q R
50	0.12254	D E F G H I J K L M N O P Q R
3	0.12248	D E F G H I J K L M N O P Q R
54	0.1215	D E F G H I J K L M N O P Q R
26	0.12079	E F G H I J K L M N O P Q R S
70	0.12058	E F G H I J K L M N O P Q R S
52	0.12033	E F G H I J K L M N O P Q R S
46	0.11958	F G H I J K L M N O P Q R S T
17	0.11881	F G H I J K L M N O P Q R S T U
62	0.11877	F G H I J K L M N O P Q R S T U
61	0.11852	F G H I J K L M N O P Q R S T U
10	0.11852	F G H I J K L M N O P Q R S T U
19	0.11794	F G H I J K L M N O P Q R S T U
48	0.11787	F G H I J K L M N O P Q R S T U
15	0.11783	F G H I J K L M N O P Q R S T U
68	0.11748	F G H I J K L M N O P Q R S T U
24	0.11748	F G H I J K L M N O P Q R S T U
34	0.11596	F G I J K L M N O P Q R S T U
1	0.11583	F G I J K L M N O P Q R S T U
20	0.11512	F G J K L M N O P Q R S T U
36	0.11379	G J K L M N O P Q R S T U
67	0.11325	J L M N O P Q R S T U
66	0.11298	J L M N O P Q R S T U
16	0.11187	J L M N O P Q R S T U V
21	0.11067	J L M N O P Q R S T U V
27	0.11019	J M N O P Q R S T U V
7	0.10863	J M N O P Q R S T U V W
40	0.10837	J M N O P Q R S T U V W
30	0.10733	N O P Q R S T U V W
4	0.10583	N O P Q R S T U V W
8	0.10558	N O P Q R S T U V W
32	0.10527	N O P Q R S T U V W
29	0.10477	N O P Q R S T U V W
53	0.10433	O P Q R S T U V W
38	0.10398	O P Q R S T U V W
71	0.10327	O P Q R S T U V W
5	0.10273	O P Q R S T U V W
33	0.1024	P Q R S T U V W
56	0.10169	P Q R S T U V W
55	0.10125	P Q R S T U V W
57	0.10044	Q R S T U V W
28	0.099458	R S T U V W X
59	0.094542	S T U V W X
47	0.0935	T U V W X
31	0.092833	U V W X
0	0.092687	U V W X
39	0.086188	V W X
9	0.082625	W X
18	0.073625	X

-------Summaries B----------

ID	Score	Significance
D	0.16692	A
F	0.16225	A B
H	0.15758	A B C
G	0.15242	A B C D
A	0.14454	A B C D E
B	0.14342	A B C D E F
E	0.13833	A B C D E F G
14	0.13669	A B C D E F G H
C	0.13658	A B C D E F G H I
65	0.13379	A B C D E F G H I J
49	0.13183	A B C D E F G H I J K
2	0.1316	A B C D E F G H I J K
44	0.13025	B C D E F G H I J K L
43	0.13004	B C D E F G H I J K L
60	0.12962	B C D E F G H I J K L
37	0.12904	B C D E F G H I J K L
11	0.12677	B C D E F G H I J K L M
69	0.12635	C D E F G H I J K L M
62	0.126	C D E F G H I J K L M
23	0.12544	C D E F G H I J K L M N
13	0.12531	C D E F G H I J K L M N
6	0.12388	C D E F G H I J K L M N O
1	0.12121	D E F G H I J K L M N O P
64	0.12069	D E F G H I J K L M N O P
25	0.12063	D E F G H I J K L M N O P
45	0.12004	D E F G H I J K L M N O P
24	0.12004	D E F G H I J K L M N O P
68	0.11898	D E F G H I J K L M N O P Q
51	0.11892	D E F G H I J K L M N O P Q
19	0.1184	D E F G H I J K L M N O P Q R
41	0.11833	D E F G H I J K L M N O P Q R
34	0.1181	D E F G H I J K L M N O P Q R
50	0.11792	D E F G H I J K L M N O P Q R
52	0.11506	E F G H I J K L M N O P Q R
17	0.1139	E F G H I J K L M N O P Q R
20	0.11377	E F G H I J K L M N O P Q R
15	0.11304	E F G H I J K L M N O P Q R
48	0.11281	E F G H I J K L M N O P Q R
29	0.11244	E F G H I J K L M N O P Q R
46	0.1121	E F G H I J K L M N O P Q R S
63	0.11198	E F G H I J K L M N O P Q R S
22	0.11146	E F G H I J K L M N O P Q R S
26	0.11071	E F G H I J K L M N O P Q R S
61	0.11056	E F G H I J K L M N O P Q R S
67	0.11046	E F G H I J K L M N O P Q R S
10	0.11004	E F G H I J K L M N O P Q R S
16	0.1099	E F G H I J K L M N O P Q R S
36	0.10923	E F G H I J K L M N O P Q R S
35	0.10796	F G H I J K L M N O P Q R S
4	0.10771	G H I J K L M N O P Q R S
12	0.10748	G I J K L M N O P Q R S
40	0.10706	G I J K L M N O P Q R S
58	0.1069	G I J K L M N O P Q R S
42	0.10656	G I J K L M N O P Q R S
32	0.10638	G I J K L M N O P Q R S
21	0.10625	G I J K L M N O P Q R S
3	0.10352	G I K L M N O P Q R S
66	0.10308	G I K L M N O P Q R S
55	0.10306	G I K L M N O P Q R S
27	0.10277	G I K L M N O P Q R S
5	0.10119	I L M N O P Q R S T
70	0.099625	M N O P Q R S T U
54	0.098417	M N O P Q R S T U
59	0.096813	N O P Q R S T U V
53	0.095125	O P Q R S T U V W
8	0.094896	O P Q R S T U V W
0	0.093896	P Q R S T U V W
33	0.0935	P Q R S T U V W
57	0.092229	P Q R S T U V W X
38	0.090896	Q R S T U V W X
31	0.089792	R S T U V W X
28	0.083042	S T U V W X Y
39	0.073375	T U V W X Y Z
9	0.070604	U V W X Y Z
30	0.069083	V W X Y Z
56	0.068396	V W X Y Z
71	0.066542	W X Y Z
47	0.064229	X Y Z
18	0.056854	Y Z
7	0.048542	Z

Table 12: ROUGE-SU4 results for the TAC 2008 update summarization task for summaries A and B. Peers not sharing a common letter are significantly different at the 95% confidence level.

```
-------Summaries A-----------------------------------------Summaries B-------------------------
ID  Score       Significance                                ID  Score      Significance

D   0.094708    A                                           F   0.10342    A
G   0.087875    A B                                         D   0.10342    A
F   0.087875    A B                                         G   0.095667   A B
H   0.076083    A B C                                       E   0.09025    A B C
E   0.0725      A B C D                                     H   0.085083   A B C D
C   0.069333    A B C D E                                   B   0.080625   A B C D E
B   0.067833    A B C D E                                   A   0.080583   A B C D E F
A   0.067333    A B C D E F                                 14  0.075604   A B C D E F G
43  0.063896    A B C D E F G                               C   0.073917   A B C D E F G H
13  0.063021      B C D E F G                               65  0.071958   A B C D E F G H I
37  0.061229      B C D E F G H                             64  0.065583   A B C D E F G H I J
6   0.060979      B C D E F G H                             69  0.065417   A B C D E F G H I J
60  0.060375      B C D E F G H I                           49  0.065229   A B C D E F G H I J
49  0.059458      B C D E F G H I J                         44  0.064542     B C D E F G H I J K
64  0.058333      B C D E F G H I J K                       23  0.063937     B C D E F G H I J K
23  0.055187        C D E F G H I J K L                     60  0.063667     B C D E F G H I J K
51  0.054688        C D E F G H I J K L                     43  0.061437     B C D E F G H I J K L
44  0.054625        C D E F G H I J K L                     62  0.060938     B C D E F G H I J K L
45  0.054354        C D E F G H I J K L                     37  0.060896     B C D E F G H I J K L
65  0.05375         C D E F G H I J K L                     24  0.059646     B C D E F G H I J K L M
14  0.05375         C D E F G H I J K L                     25  0.059396     B C D E F G H I J K L M
25  0.052771        C D E F G H I J K L                     13  0.058813     B C D E F G H I J K L M
70  0.052333        C D E F G H I J K L                     68  0.056896     B C D E F G H I J K L M N
54  0.052333        C D E F G H I J K L                     1   0.05675        C D E F G H I J K L M N O
12  0.051417        C D E F G H I J K L                     50  0.056562       C D E F G H I J K L M N O
69  0.051083        C D E F G H I J K L                     34  0.055813       C D E F G H I J K L M N O
30  0.051021        C D E F G H I J K L                     6   0.055625       C D E F G H I J K L M N O
2   0.050979        C D E F G H I J K L                     11  0.055167       C D E F G H I J K L M N O
63  0.050854        C D E F G H I J K L                     45  0.054958       C D E F G H I J K L M N O
42  0.050854        C D E F G H I J K L                     2   0.053083       C D E F G H I J K L M N O P
15  0.050813        C D E F G H I J K L                     29  0.052979       C D E F G H I J K L M N O P
22  0.049708        C D E F G H I J K L                     63  0.05275        C D E F G H I J K L M N O P
50  0.048854        C D E F G H I J K L M                   51  0.052729       C D E F G H I J K L M N O P
17  0.048354        C D E F G H I J K L M                   52  0.052062       C D E F G H I J K L M N O P
19  0.047687        C D E F G H I J K L M                   26  0.051354         D E F G H I J K L M N O P Q
11  0.047604        C D E F G H I J K L M                   12  0.050333         D E F G H I J K L M N O P Q
48  0.047083        C D E F G H I J K L M                   19  0.050188         D E F G H I J K L M N O P Q
41  0.046708        C D E F G H I J K L M                   41  0.04975          D E F G H I J K L M N O P Q
52  0.046437        C D E F G H I J K L M                   20  0.049271         D E F G H I J K L M N O P Q
46  0.046396        C D E F G H I J K L M                   15  0.048292         D E F G H I J K L M N O P Q R
26  0.046313        C D E F G H I J K L M                   42  0.048208         D E F G H I J K L M N O P Q R
67  0.045437        C D E F G H I J K L M                   17  0.047563         D E F G H I J K L M N O P Q R
68  0.044979        C D E F G H I J K L M                   4   0.046896         D E F G H I J K L M N O P Q R
24  0.044979        C D E F G H I J K L M                   36  0.046354         D E F G H I J K L M N O P Q R S
3   0.044583        C D E F G H I J K L M                   21  0.045229           E F G H I J K L M N O P Q R S
35  0.044417          D E F G H I J K L M                   46  0.045125           E F G H I J K L M N O P Q R S
61  0.043958          D E F G H I J K L M                   55  0.044771           E F G H I J K L M N O P Q R S
10  0.043958          D E F G H I J K L M                   67  0.044083           E F G H I J K L M N O P Q R S
62  0.0435            D E F G H I J K L M N                 22  0.044063           E F G H I J K L M N O P Q R S
36  0.043188          D E F G H I J K L M N                 27  0.043708           E F   H I J K L M N O P Q R S
58  0.042771          D E F G H I J K L M N                 48  0.042792           E F   H I J K L M N O P Q R S T
4   0.042729          D E F G H I J K L M N                 32  0.041792             F   H I J K L M N O P Q R S T U
1   0.042646          D E F G H I J K L M N                 40  0.040375                 H I J K L M N O P Q R S T U
34  0.042396          D E F G H I J K L M N                 61  0.040208                 H   J K L M N O P Q R S T U
21  0.042375          D E F G H I J K L M N                 10  0.040208                 H   J K L M N O P Q R S T U
27  0.041521          D E F G H I J K L M N                 16  0.039917                 H   J K L M N O P Q R S T U
16  0.040854            E F G H I J K L M N O               35  0.038437                 H   J K L M N O P Q R S T U
66  0.040687            E F G H I J K L M N O               70  0.037396                 H   J K L M N O P Q R S T U
53  0.040458            E F G H I J K L M N O               58  0.03675                  H   J K L M N O P Q R S T U
20  0.040271            E F G H I J K L M N O               54  0.036729                 H   J K L M N O P Q R S T U
7   0.039417            E F G H I J K L M N O               59  0.036646                 H   J K L M N O P Q R S T U
57  0.038979            E F G H I J K L M N O               3   0.036417                 H   J K L M N O P Q R S T U
29  0.038583            E F G H I J K L M N O               57  0.036354                 H   J K L M N O P Q R S T U
32  0.035854              F   H I J K L M N O               53  0.035688                 H   J K L M N O P Q R S T U
28  0.035583                  H I J K L M N O               0   0.035083                     J K L M N O P Q R S T U
71  0.035563                  H I J K L M N O               66  0.034021                     J K L M N O P Q R S T U
8   0.035063                    I J K L M N O               5   0.033271                       K L M N O P Q R S T U
40  0.033771                      J K L M N O               33  0.032896                       K L M N O P Q R S T U
55  0.033021                        K L M N O               38  0.029833                         L M N O P Q R S T U
56  0.032937                        K L M N O               8   0.028458                           M N O P Q R S T U
59  0.03225                           L M N O               31  0.026854                             N O P Q R S T U
5   0.032187                          L M N O               28  0.0265                               N O P Q R S T U
33  0.031896                          L M N O               56  0.025083                               O P Q R S T U
38  0.031854                          L M N O               30  0.021896                                 P Q R S T U
31  0.031854                          L M N O               71  0.021688                                 P Q R S T U
47  0.031167                          L M N O               47  0.019792                                   Q R S T U
0   0.030333                          L M N O               39  0.017                                        R S T U
39  0.023479                            M N O               18  0.015188                                       S T U
18  0.018229                              N O               9   0.011333                                       T U
9   0.015417                                O               7   0.011063                                         U
```

Table 13: BE-HM results for the TAC 2008 update summarization task for summaries A and B. Peers not sharing a common letter are significantly different at the 95% confidence level.