DUC 2005: Evaluation of Question-Focused Summarization Systems

Hoa Trang Dang
Information Access Division
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD, 20899
hoa.dang@nist.gov

Abstract

The Document Understanding Conference (DUC) 2005 evaluation had a single user-oriented, question-focused summarization task, which was to synthesize from a set of 25-50 documents a well-organized, fluent answer to a complex question. The evaluation shows that the best summarization systems have difficulty extracting relevant sentences in response to complex questions (as opposed to representative sentences that might be appropriate to a generic summary). The relatively generous allowance of 250 words for each answer also reveals how difficult it is for current summarization systems to produce fluent text from multiple documents.

1 Introduction

The Document Understanding Conference (DUC) is a series of evaluations of automatic text summarization systems. It is organized by the National Institute of Standards of Technology with the goals of furthering progress in automatic summarization and enabling researchers to participate in large-scale experiments.

In DUC 2001-2004 a growing number of research groups participated in the evaluation of generic and focused summaries of English newspaper and newswire data. Various target sizes were used (10-400 words) and both single-document summaries and summaries of multiple documents were evaluated (around 10 documents per set). Summaries were manually judged for both content and readability. To evaluate content, each peer (human or automatic) summary was compared against a single model (human) summary using SEE (http://www.isi.edu/ cyl/SEE/) to estimate the percentage of information in the model that was covered in the peer. Additionally, automatic evaluation of content coverage using ROUGE (Lin, 2004) was explored in 2004.

Human summaries vary in both writing style and content. For example, (Harman and Over, 2004) noted that a human summary can vary in its level of *granularity*, whether the summary has a very high-level analysis or primarily contains details. They analyzed the effects of human variaion in the DUC evaluations and concluded that despite large variation in model summaries, the rankings of the systems when compared against a single model for each document set remained stable *when averaged over a large number of document sets* and human assessors. The use of a large test set to smooth over natural human variation is not a new technique; it is the approach that has been taken in TREC (Text Retrieval Conference) for many years (Voorhees and Buckley, 2002).

While evaluators can achieve stable overall system rankings by averaging scores over a large number of document sets, system builders are still faced with the challenge of producing a summary for a given document set that is *most likely* to satisfy any human user (since they cannot know ahead of time which human will be using or judging the summary). Thus, system developers desire an evaluation methodology that takes into account human variation in summaries *for any given document set*.

DUC 2005 marked a major change in direction from previous years. The road mapping committee had strongly recommended that new tasks be undertaken that were strongly tied to a clear user application. At the same time, the program committee wanted to work on new evaluation methodologies and metrics that would take into

Proceedings of the Workshop on Task-Focused Summarization and Question Answering, pages 48–55,
Sydney, July 2006. ©2006 Association for Computational Linguistics

account variation of content in human-authored summaries.

Therefore, DUC 2005 had a single user-oriented system task that allowed the community to put some time and effort into helping with a new evaluation framework. The system task modeled real-world complex question answering (Amigo et al., 2004). Systems were to synthesize from a set of 25-50 documents a brief, well-organized, fluent answer to a need for information that could not be met by just stating a name, date, quantity, etc. Summaries were evaluated for both content and readability.

The task design attempted to constrain two parameters that could produce summaries with widely different content: focus and granularity. Having a question to focus the summary was intended to improve agreement in content between the model summaries. Additionally, the assessor who developed each topic specified the desired granularity (level of generalization) of the summary. Granularity was a way to express one type of user preference; one user might want a general background or overview summary, while another user might want specific details that would allow him to answer questions about specific events or situations.

Because it is both impossible and unnatural to eliminate all human variation, our assessors created as many manual summaries as feasible for each topic, to provide examples of the range of normal human variability in the summarization task. These multiple models would provide more representative training data to system developers, while enabling additional experiments to investigate the effect of human variability on the evaluation of summarization systems.

As in past DUCs, assessors manually evaluated each summary for readability using a set of linguistic quality questions. Summary content was manually evaluated using the pseudo-extrinsic measure of responsiveness, which does not attempt pairwise comparison of peers against a model summary but gives a coarse ranking of all the summaries based on responsiveness of the summary to the topic. In parallel, ISI and Columbia University led the summarization research community in two exploratory efforts at intrinsic evaluation of summary content; these evaluations compared peer summaries against multiple reference summaries, using Basic Elements at ISI and Pyramids at Columbia University.

This paper describes the DUC 2005 task and the results of our evaluations of summary content and readability. (Hovy et al., 2005) and (Passonneau et al., 2005) provide additional details and results of the evaluations of summary content using Basic Elements and Pyramids.

2 Task Description

The DUC 2005 task was a complex question-focused summarization task that required summarizers to piece together information from multiple documents to answer a question or set of questions as posed in a topic.

Assessors developed a total of 50 topics to be used as test data. For each topic, the assessor selected 25-50 related documents from the *Los Angeles Times* and *Financial Times of London* and formulated 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 set of related questions and could include background information that the assessor thought would help clarify his/her information need.

The assessor also indicated the “granularity” of the desired response for each topic. That is, they indicated whether they wanted the answer to their question(s) to name *specific* events, people, places, etc., or whether they wanted a *general*, high-level answer. Only one value of granularity was given for each topic, since the goal was not to measure the effect of different granularities on system performance for a given topic, but to provide additional information about the user’s preferences to both human and automatic summarizers.

An example DUC topic follows:

> **num**: D345
> **title**: American Tobacco Companies Overseas
> **narr**: In the early 1990’s, American tobacco companies tried to expand their business overseas. What did these companies do or try to do and where? How did their parent companies fare?
> **granularity**: specific

The summarization task was the same for both human and automatic summarizers: Given a DUC topic with granularity specification and a set of documents relevant to the topic, the summarization task was to create from the documents a brief, well-organized, fluent summary that answers the need for information expressed in the topic, at the specified level of granularity. The summary could be no longer than 250 words (whitespace-delimited tokens). 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. The summary should include (in some form or other) all the information in the documents that contributed to meeting the information need.

Ten assessors produced a total of 9 human summaries for each of 20 topics, and 4 human summaries for each of the remaining 30 topics. The summarization task was a relatively difficult task, requiring about 5 hours to manually create each summary. Thus, there would be a real benefit to users if the task could be performed automatically.

3 Participants

There was much interest in the longer, question-focused summaries required in the DUC 2005 task. 31 participants submitted runs to the evaluation; they are identified by numeric Run IDs (2-32) in the remainder of this paper. We also developed a simple baseline system that returned the first 250 words of the most recent document for each topic (Run ID = 1). In addition to the automatic peers, there were 10 human peers, assigned alphabetic Run IDs, A-J.

Most system developers treated the summarization task as a passage retrieval task. Sentences were ranked according to relevance to the topic. The most relevant sentences were then selected for inclusion in the summary while minimizing redundancy within the summary, up to the maximum 250-word allowance. A significant minority of systems first decomposed the topic narrative into a set of simpler questions, and then extracted sentences to answer each subquestion. Systems differed in the approach taken to compute relevance and redundancy, using similarity metrics ranging from simple term frequency to semantic graph matching. In order to include more relevant information in the summary, systems attempted within-sentence compression by removing phrases such as parentheticals and relative clauses.

Many systems simply ignored the granularity specification. The systems that addressed granularity did so by preferring to extract sentences that contained proper names for topics with a “specific” granularity but not for topics with “general” granularity.

Cross-sentence dependencies had to be handled, including anaphora. Strategies for dealing with pronouns that occurred in relevant sentences included co-reference resolution, including the previous sentence for additional context, or simply excluding all sentences containing any pronouns.

Most systems made no attempt to reword the extracted sentences to improve the readability of the final summary. Although some systems grouped related sentences together to improve cohesion, the most common heuristic to improve readability was simply to order the extracted sentences by document date and position in the document. System 12 achieved high readability scores by choosing a single representative document and extracting sentences in the order of appearance in that document. This approach is similar to the baseline summarizer and produces summaries that are more fluent than those constructed from multiple documents.

4 Evaluation Results

Summaries were manually evaluated by 10 assessors. All summaries for a given topic were judged by a single assessor (who was usually the same as the topic developer). In all cases, the assessor was one of the summarizers for the topic. All summaries for the topic (including the one written by the assessor) were anonymously presented to the assessor, in a random order, and the ssessor judged each summary for readability and responsiveness to the topic, giving separate scores for responsiveness and each of 5 linguistic qualities. This allowed participants who could not work on optimizing all 6 manual scores, to focus on only the elements that they were interested in or had the resources to address.

No single score was reported that reflected a combination of readability and content. In previous years, responsiveness considered both the content and readability of the summary. While it tracked SEE coverage, responsiveness could not be seen as a direct measure of content due to possible effects of readability on the score. Because we needed an inexpensive manual measure of coverage, we revised the definition of responsiveness in 2005 so that it considered only the information content and not the readability of the summary, to the extent possible.

4.1 Evaluation of Readability

The readability of the summaries was assessed using five linguistic quality questions which measured qualities of the summary that *do not* involve comparison with a reference summary or DUC topic. The linguistic qualities measured were *Grammaticality, Non-redundancy, Referential clarity, Focus, and Structure and coherence*.

Q1: Grammaticality The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

Q2: Non-redundancy There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., “Bill Clinton”) when a pronoun (“he”) would suffice.

Q3: Referential clarity It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

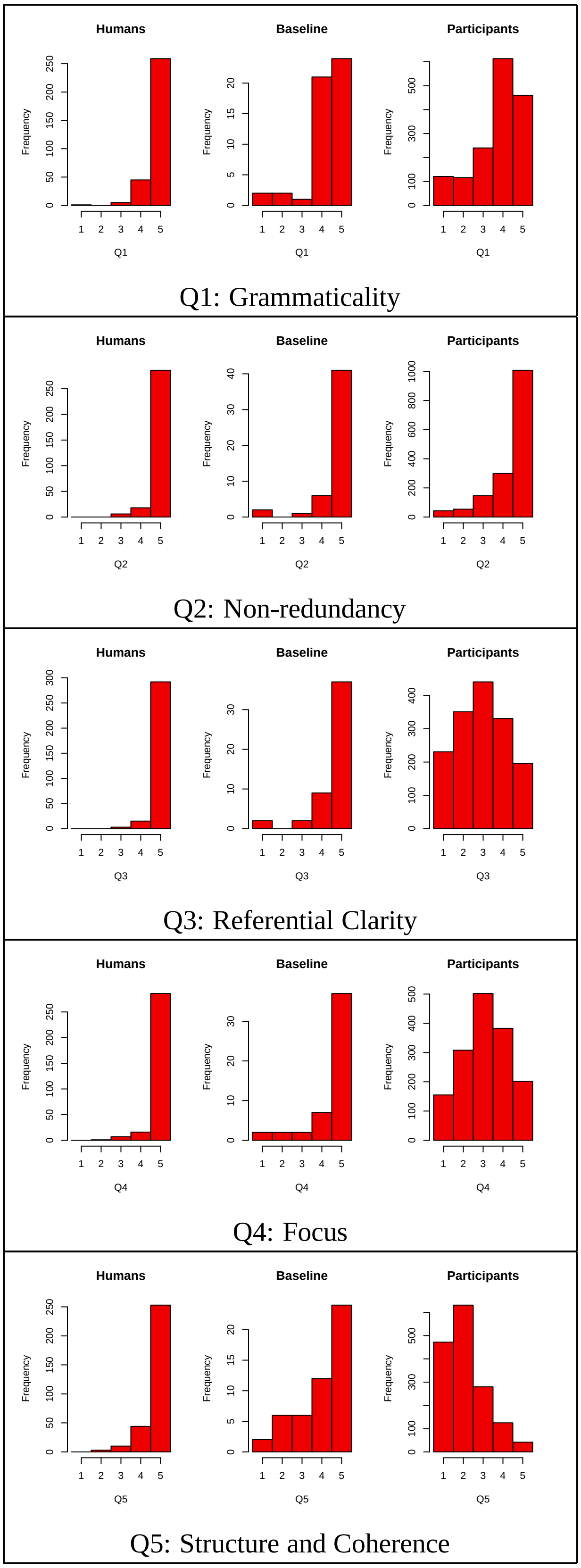
Q4: Focus The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

Q5: Structure and Coherence The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Each linguistic quality question was assessed on a five-point scale:

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

Table 1 shows the distribution of the scores across all the summaries, broken down by the type of summarizer (Human, Baseline, or Participants). All summarizers generally performed well on the first two linguistic qualities. The high scores on non-redundancy show that most participants have



Q1: Grammaticality

Q2: Non-redundancy

Q3: Referential Clarity

Q4: Focus

Q5: Structure and Coherence

Table 1: Frequency of scores for each linguistic quality, broken down by source of summary (Humans, Baseline, Participants).

successfully achieved this capability. Humans and the baseline system also scored well on the last 3 linguistic qualities. The multi-document summarization systems submitted by participants, on the other hand, still struggle with referential clarity and focus, and perform very poorly on structure and coherence.

4.1.1 Comparison by system

For each linguistic quality question, we performed a multiple comparison test between the scores of all peers using Tukey's honestly significant difference criterion. A multiple comparison test between all human and automatic peers was performed using the Kruskall-Wallis test, to see how the individual automatic peers performed relative to human peers. For grammaticality, the best human summarizer is significantly better than 28 of the 32 systems; the worst human summarizer is better than 8 systems. For non-redundancy, the two best humans are significantly better than 6 systems, and the two worst humans are not significantly different from any system. For referential clarity, all humans are significantly better than all but 2 automatic peers (baseline and System 12). For focus, the best human is significantly better than all automatic peers except the baseline; all other humans are significantly better than all automatic peers except the baseline and System 12. For structure and coherence, the two best humans are significantly better than 31 systems (all automatic peers except the baseline); all humans are better than 30 of the automatic peers (all automatic peers except baseline and System 12).

4.2 Evaluation of Content

We performed manual pseudo-extrinsic evaluation of peer summaries in the form of assessment of responsiveness. Responsiveness is different from SEE coverage in that it does not compare a peer summary against a single reference; however, responsiveness tracked SEE coverage in DUC 2003 and 2004, and was used to provide a coarse-grained measure of content in 2005. We also computed ROUGE scores as was done in DUC 2004.

4.2.1 Responsiveness

Assessors assigned a raw responsiveness score to each summary. The score provides a coarse ranking of the summaries for each topic, according to the amount of information in the summary that helps to satisfy the information need expressed in the topic statement, at the level of granularity requested in the user profile. The score was an integer between 1 and 5, with 1 being least responsive and 5 being most responsive. For a given topic, some summary was required to receive each of the five possible scores, but no distribution was specified for how many summaries had to receive each score. The number of human summaries scored per topic also varied. Therefore, raw responsiveness scores should not be directly added and compared across topics. Assigning responsiveness scores can be seen as a clustering task in which peers are partitioned into exactly 5 clusters, where members of a cluster are more similar to each other in quality.

RunID							
10	A						
5	A						
4	A	B					
15	A	B	C				
29	A	B	C	D			
11	A	B	C	D			
17	A	B	C	D			
8	A	B	C	D			
7	A	B	C	D	E		
14	A	B	C	D	E		
6	A	B	C	D	E		
28	A	B	C	D	E	F	
21	A	B	C	D	E	F	
19	A	B	C	D	E	F	
24	A	B	C	D	E	F	
9	A	B	C	D	E	F	
16	A	B	C	D	E	F	
32	A	B	C	D	E	F	
12	A	B	C	D	E	F	
25	A	B	C	D	E	F	
18	A	B	C	D	E	F	
27	A	B	C	D	E	F	
20	A	B	C	D	E	F	
3	A	B	C	D	E	F	
2		B	C	D	E	F	
13			C	D	E	F	
30				D	E	F	
22					E	F	
1					E	F	
26						F	
31						F	G
23							G

Table 2: Multiple comparison of systems based on Friedman's test on responsiveness

For each topic, we computed the scaled responsiveness score for each summary, such that the sum of the scaled responsiveness score is proportional to the number of summaries for the topic. The scaled responsiveness is the rank of the summary based on the raw responsiveness score. We computed the average scaled responsiveness score of each summarizer across all topics. Since the number of human summaries varied across topics, we also computed the average scaled responsiveness score of only the automatic summaries (ignoring the human summaries in scaling responsiveness).

Table 2 shows the results of a multiple comparison of scaled responsiveness of the automatic peers using Tukey's honestly significant criterion and Friedman's test, with the best peers on top; peers not sharing a common letter are significantly different at the 95.5% confidence level. None of the automatic peers performed significantly better than the majority of the remaining peers, and only eight of the automatic peers performed significantly better than the simple baseline. In multiple comparison of all peers using the Kruskal-Wallis test, all human peers were significantly better than all the automatic peers.

4.2.2 ROUGE

We computed two ROUGE scores: ROUGE-2 and ROUGE-SU4 recall, both with stemming and implementing jackknifing for each $[peer, topic]$ pair so that human and automatic peers could be compared. Since the number of ROUGE evaluations per topic varied depending on the number of reference summaries, we computed a macro-average of each score for each peer, where the macro-average score is the mean over all topics of the mean per-topic score for the peer.

Unlike responsiveness and linguistic quality scores, which are ordinal data and are best suited for non-parametric analyses, ROUGE scores, can be measured on an interval scale and are suitable for parametric analysis. Analysis of variance showed significant effects from peer and topic ($p = 0$ for each factor) for both ROUGE-2 and ROUGE-SU4 recall. To see which peers were different, a multiple comparison of population marginal means (PMM) was performed for each type of ROUGE score. The population marginal means remove any effect of an unbalanced design (since not all human peers created summaries for all topics) by fixing the values of the "peer" factor, and averaging out the effects of the "topic" factor as if each factor combination occurred the same number of times.

Table 3 shows multiple comparison of all peers based on ANOVA of ROUGE-2 recall (ROUGE-SU4 shows similar results). ROUGE-2 and ROUGE-SU4 both distinguish human peers from automatic ones. The difference in the ROUGE-2 score of the best system and worst human is not considered significant (possibly due to the very conservative nature of the multiple comparison test) but is still relatively large. On the other hand, ANOVA of ROUGE-2 found more significant differences between the automatic peers than did Friedman's test of responsiveness.

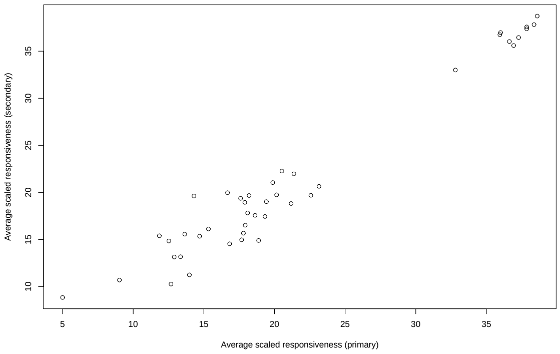


Figure 1: Primary vs. secondary average scaled responsiveness

4.3 Correlation

A metric must produce stable rankings of systems in the face of human variation. Intrinsic measures like ROUGE rely on multiple model summaries to take into account human variation (although Pyramids add another level of human variation in the manual pyramid and peer annotation). For a metric like responsiveness, which does not depend on comparison of peer summaries against a model or set of model summaries, it is appropriate to consider the stability of the measure across different assessors.

A secondary assessment was done on responsiveness for the 20 topics that had 9 summaries each. The secondary assessor had written a summary for the topic but was generally not the same person who developed the topic. As seen in Figure 1, average scaled responsiveness scores from the two sets of assessments (averaged over the 20 topics) track each other very well. The human summaries are clustered on the upper right side of the graph, while the automatic summaries form a second cluster on the lower left side.

The actual responsiveness scores for each system and each topic do vary between assessors, but this variation in human judgment is smoothed out by averaging over multiple topics. Table 4 shows that the correlation between the primary and sec-

RunID	PMM of R2													
C	0.1172	A												
A	0.1156	A	B											
I	0.1023	A	B	C										
B	0.1014	A	B	C										
J	0.1012	A	B	C										
E	0.1009	A	B	C										
D	0.0986	A	B	C										
G	0.0970		B	C										
F	0.0947			C										
H	0.0897			C	D									
15	0.0725				D	E								
17	0.0717					E								
10	0.0698					E	F							
8	0.0696					E	F							
4	0.0686					E	F	G						
5	0.0675					E	F	G						
11	0.0643					E	F	G	H					
14	0.0635					E	F	G	H	I				
16	0.0633					E	F	G	H	I				
19	0.0632					E	F	G	H	I				
7	0.0628					E	F	G	H	I	J			
9	0.0625					E	F	G	H	I	J			
29	0.0609					E	F	G	H	I	J	K		
25	0.0609					E	F	G	H	I	J	K		
6	0.0609					E	F	G	H	I	J	K		
24	0.0597					E	F	G	H	I	J	K		
28	0.0594					E	F	G	H	I	J	K		
3	0.0594					E	F	G	H	I	J	K		
21	0.0573					E	F	G	H	I	J	K		
12	0.0563						F	G	H	I	J	K		
18	0.0553						F	G	H	I	J	K	L	
26	0.0547						F	G	H	I	J	K	L	
27	0.0546						F	G	H	I	J	K	L	
32	0.0534							G	H	I	J	K	L	
20	0.0515								H	I	J	K	L	
13	0.0497								H	I	J	K	L	
30	0.0496								H	I	J	K	L	
31	0.0487									I	J	K	L	
2	0.0478										J	K	L	
22	0.0462											K	L	
1	0.0403												L	M
23	0.0256													M

Table 3: Multiple comparison of all peers based on ANOVA of ROUGE-2 recall

	Spearman	Pearson
All peers	0.900	0.976 [0.960, 1.000]
Auto peers	0.775	0.822 [0.695, 1.000]

Table 4: Correlation between primary and secondary average scaled responsiveness (20 topics), with 95% confidence intervals for Pearson's *r*.

ondary average scaled responsiveness scores is respectable despite the low number of topics. The correlation suggests that responsiveness would give a stable ranking of the systems when averaged over the entire set of 50 topics.

Table 5 shows that there is high correlation between macro-average ROUGE scores (intrinsic measures) and average scaled responsiveness (a pseudo-extrinisic measure). The correlation is high even when the human summaries are ignored.

Metric	Spearman	Pearson
ROUGE-2 (all)	0.951	0.972 [0.953, 1.000]
ROUGE-SU4 (all)	0.942	0.958 [0.930, 1.000]
ROUGE-2 (auto)	0.901	0.928 [0.872, 1.000]
ROUGE-SU4 (auto)	0.872	0.919 [0.855, 1.000]

Table 5: Correlation between average scaled responsiveness and macro-average ROUGE recall over all topics and either all peers or only automatic peers.

5 Conclusion

The DUC 2005 task was to summarize the answer to a complex question, as found in a set of documents. The evaluation showed that only the top systems are able to extract sentences whose information content is more responsive to the question than a simple baseline. Additionally, systems require much additional work to produce coherent, well-structured text, which is apparent in the longer summary sizes of DUC 2005. On the other hand, systems do well on non-redundancy, since text summarization has historically been formulated as a text compression task. Since DUC 2005 is the first time question-focused summarization has been evaluated on a large-scale, we have repeated the task in 2006, with some modifications.

We eliminated the "granularity" specification in DUC 2006. Assessors had appreciated the theory behind the granularity specification, but found that the size limit for the summaries was a much bigger factor in determining what information to include; some "specific" summaries ended up being very general given the large amount of information and limited space allowed. From a human perspective, the actual granularity of the resulting summary mostly fell out naturally from the topic question and the content that was available in the source documents.

The definition of responsiveness scores was meant to yield a coarse ranking of the peer summaries into 5 ordered clusters. However, assessors found it difficult to form these 5 clusters because of the large number (36+) of summaries that needed to be compared with one another, and the impression that many sets of human and automatic summaries could not be separated into as many as 5 groups. We therefore changed the scoring of responsiveness in 2006 so that it is based on the same scale as the linguistic quality questions; this may reduce the discriminative power of the responsiveness measure but should produce scores that more accurately reflect the true differences between summaries.

References

Enrique Amigo, Julio Gonzalo, Victor Peinado, Anselmo Penas, and Felisa Verdejo. 2004. An empirical study of information synthesis tasks. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 207–214, Barcelona, Spain.

Donna Harman and Paul Over. 2004. The effects of human variation in duc summarization evaluation. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 10–17, Barcelona, Spain.

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.

Ellen M. Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, Tampere, Finland, August.