Overview of DUC 2005

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

Abstract

The focus of DUC 2005 was on developing new evaluation methods that take into account variation in content in human-authored summaries. Therefore, DUC 2005 had a single user-oriented, question-focused summarization task that allowed the community to put some time and effort into helping with the new evaluation framework. The summarization task was to synthesize from a set of 25-50 documents a well-organized, fluent answer to a complex question. The relatively generous allowance of 250 words for each answer reveals how difficult it is for current summarization systems to produce fluent multi-document summaries.

1 Introduction

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 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. Consequently, a report-writing task based on a "natural disaster" scenario was proposed at the DUC 2004 workshop, but this was met with little enthusiasm in the community. At the same time, the program committee wanted to work on new evaluation methodologies and metrics that would take into account variation of content in human-authored summaries.

Therefore, DUC 2005 had a single simpler (but still 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 NIST 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, NIST 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, NIST manually evaluated each summary for readability using a set of linguistic quality questions. Summary content was manually evaluated at NIST 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 NIST's 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 DUC topic.

NIST Assessors developed a total of 50 DUC 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 DUC 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 DUC 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 NIST 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. NIST also developed a simple baseline system that returned the first 250 words of the most recent document for each topic. The systems and their Run IDs are listed in table 1. In addition to the automatic peers, the 10 human peers were assigned alphabetic Run IDs, A-J.

Organization	System ID	Run ID
(NIST)	Baseline	1
Chinese Academy of Sciences	IOS_SUMMZ	2
CL Research	CLResearch.duc05	3
Columbia University	Columbia	4
FreeText Software Technologies, Inc.	FTextST-05	5
Fudan University	FDUSUM	6
IDA Center for Computing Sciences	CCS-NSA-05	7
International Institute of Information Technology	IIITH-Sum	8
Institute for Infocomm Research	I2RNLS	9
Information Sciences Institute (Daume)	isi-bqfs	10
Information Sciences Institute (Lin)	ISI-Webcl	11
ITC-irst	LAKE05	12
Laris/Larim Laboratory	LARIS2005	13
Language Computer Corporation	lcc.duc05	14
National University of Singapore	NUS3	15
Oregon Health & Science University	OHSU-DUC05	16
The Hong Kong Polytechnic University	PolyU	17
Royal Institute of Technology KTH KOD	KTH-holsum	18
Simon Fraser University	SFU_v2.4	19
Thomson Legal & Regulatory	TLR	20
Toyohashi University of Technology	TUT/NII	21
Universidad Autonoma de Madrid	UAM2005	22
University College Dublin	UCD-IIRG	23
University of Edinburgh	EMBRA	24
University of Karlsruhe / Concordia University	ERSS2005	25
University of Lethbridge	ULETH2005	26
University of Maryland and BBN	UMDBBN	27
University of Michigan	CLAIR	28
Universite de Montreal	NLP-RALI05	29
University of Ottawa	UofO	30
Technical University of Catalonia (UPC)	QASUM-UPC	31
University of Sheffield	SHEF-BSL	32

Table 1: Participants and runs in DUC 2005.

3.1 System Approaches

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 (lcc.duc05, TLR, QASUM-UPC, UCD-IIRG, IIITH-Sum) 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 like Columbia 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. The LAKE05 system 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 document.

4 Evaluation Results

Summaries were manually evaluated by 10 NIST assessors. The primary assessment was done for all 50 topics. 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 topics. Assessors 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, NIST 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

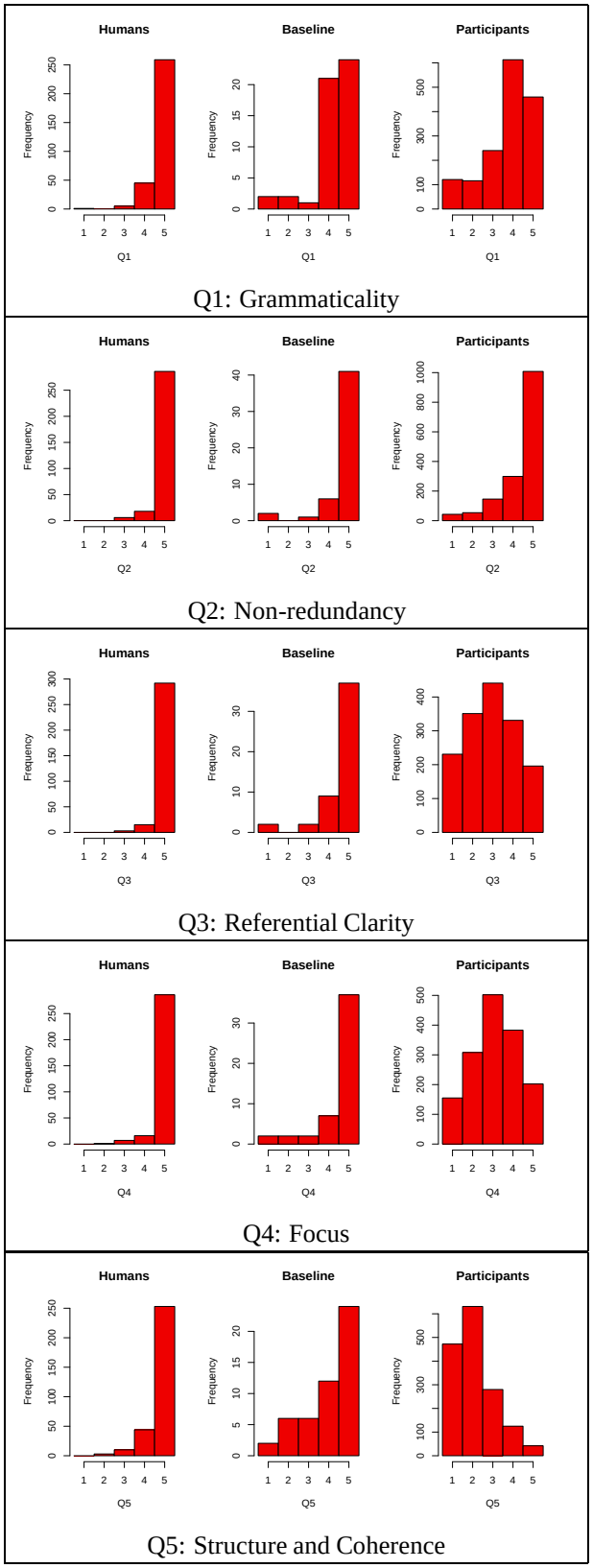

Table 2: Frequency of scores for each linguistic quality, broken down by source of summary (Humans, Baseline, Participants).

Table 2 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 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, NIST performed a multiple comparison test between the scores of all peers using Tukey's honestly significant difference criterion. Tables 3-7 compare the automatic peers using Friedman's test, with best peers on top; peers not sharing a common letter are significantly different at the 95.5% confidence level.

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

Table 3: Multiple comparison of systems based on Friedman's test on Q1: Grammaticality

For each quality question, a multiple comparison test between all human and automatic peers was also per-

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

Table 4: Multiple comparison of systems based on Friedman's test on Q2: Non-Redundancy

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

Table 5: Multiple comparison of systems based on Friedman's test on Q3: Referential Clarity

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

Table 6: Multiple comparison of systems based on Friedman's test on Q4: Focus

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

Table 7: Multiple comparison of systems based on Friedman's test on Q5: Structure and Coherence

formed using the Kruskall-Wallis test instead of Friedman's test, to see how the individual automatic peers performed relative to human peers. For grammaticality, the best human summarizer (B) is significantly better than 28 of the 32 systems (all systems except 1,5,14,16); the worst human summarizer (H) is better than 8 systems (7,10,11,15,25,26,27,31). For non-redundancy, the best humans (B,D) are significantly better than 6 systems (10,15,17,26,27,31). Five humans (I,C,G,F,E) are better than just 2 systems (15,31). One human (H) is better than 1 system (15). The worst humans (A,J) 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 (G) 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 best humans (B,G) 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

NIST 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. NIST also computed ROUGE scores as was done in DUC 2004.

4.2.1 Responsiveness

NIST 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 linguistic quality of the summary was to play a role in the assessment only insofar as it interfered with the expression of information and reduced the amount of information that was conveyed.) 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.

For each topic, NIST 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. NIST computed the average scaled responsiveness score of each summarizer across all topics. Since the number of human summaries varied across topics, NIST also computed the average scaled responsiveness score of only the automatic summaries (ignoring the human summaries in scaling responsiveness).

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

Table 8 shows the results of a multiple comparison of scaled responsiveness of the automatic peers using Tukey's honestly significant criterion and Friedman's test ($\alpha = 0.05$), with the best peers on top; none of the automatic peers performed significantly better than the majority of the remaining peers, though a few were much worse. 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

NIST computed two official ROUGE scores: ROUGE-2 and ROUGE-SU4 recall, both with stemming and im-

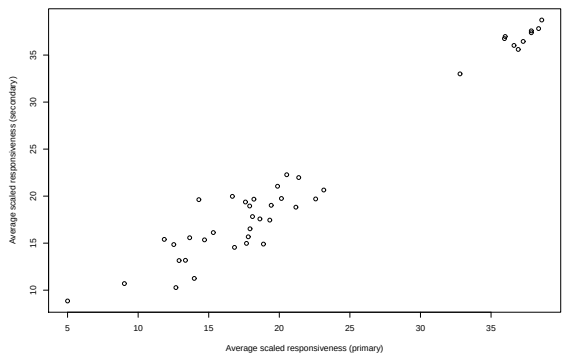

Figure 1: Primary vs. secondary average scaled responsiveness

plementing 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, NIST 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.

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. As can be seen in Tables 9-10, 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.

4.3 Correlation

A metric must produce stable rankings of systems in the face of human variation. Intrinsic measures like ROUGE/BE and Pyramids 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.

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

Table 11: Correlation between primary and secondary average scaled responsiveness (20 topics), including 95% confidence intervals for Pearson's *r*.

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 11 shows that the correlation between the primary and secondary average scaled responsiveness scores is high despite the low number of topics. The correlation indicates that responsiveness would give a stable ranking of the systems when averaged over the entire set of 50 topics.

Furthermore, Table 12 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 12: Correlation between average scaled responsiveness and macro-average ROUGE recall over all topics and either all peers or only automatic peers.

5 Future of DUC

Since DUC 2006 has a very short development cycle, the same question-focused summarization task will be repeated in 2006, with some modifications based on lessons learned from DUC 2005:

1. Eliminate "granularity" specification
2. Modify responsiveness scoring procedure

NIST assessors 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. Almost all the assessors tried

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

RunID	PMM of R-SU4																
C	0.1775	A															
A	0.1744	A	B														
I	0.1650	A	B	C													
J	0.1624	A	B	C													
B	0.1613	A	B	C													
G	0.1593	A	B	C													
D	0.1587	A	B	C													
E	0.1533		B	C													
F	0.1518			C													
H	0.1510			C													
15	0.1316				D												
17	0.1297				D	E											
8	0.1279				D	E											
4	0.1277				D	E	F										
10	0.1253				D	E	F	G									
5	0.1232				D	E	F	G	H								
11	0.1225				D	E	F	G	H								
19	0.1218				D	E	F	G	H								
16	0.1190				D	E	F	G	H	I							
7	0.1190				D	E	F	G	H	I							
6	0.1188				D	E	F	G	H	I	J						
25	0.1187				D	E	F	G	H	I	J						
14	0.1176				D	E	F	G	H	I	J						
9	0.1174				D	E	F	G	H	I	J						
24	0.1168				D	E	F	G	H	I	J						
3	0.1167				D	E	F	G	H	I	J						
28	0.1146					E	F	G	H	I	J	K					
29	0.1139					E	F	G	H	I	J	K					
21	0.1112						F	G	H	I	J	K	L				
12	0.1107							G	H	I	J	K	L				
18	0.1095							G	H	I	J	K	L	M			
27	0.1085								H	I	J	K	L	M			
32	0.1041									I	J	K	L	M			
13	0.1041									I	J	K	L	M			
26	0.1023										J	K	L	M	N		
30	0.0995											K	L	M	N		
2	0.0981											K	L	M	N		
22	0.0970												L	M	N		
31	0.0967												L	M	N		
20	0.0940													M	N		
1	0.0872														N		
23	0.0557																O

Table 10: Multiple comparison of all peers based on ANOVA of ROUGE-SU4 recall

to write their summaries according to the granularity requested, but some "specific" summaries ended up being very general given the large amount of information and small space allowance. Two assessors (A,H) simply ignored granularity. It speaks of the difficulty of controlling too many parameters in the task, even with a relatively large space allowance. 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 ranks; this can be seen as the result of 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. However, assessors found it difficult to form clusters with so many summaries, and preferred a more absolute scale by which to judge responsiveness. NIST will change the scoring of responsiveness so that it is based on the same Likert scale as the linguistic quality questions.

Acknowledgments

The DUC 2005 evaluation was organized by NIST and funded by the Advanced Research and Development Agency (ARDA). Thanks to Columbia University for organizing the Pyramid evaluation, and to ISI for making available their ROUGE/BE toolkit.

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.