Overview of DUC 2006

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

Abstract

The DUC 2006 summarization task was to synthesize from a set of 25 documents a well-organized, fluent answer to a complex question. The task and evaluation measures were basically the same as in DUC 2005, except that an additional "overall" responsiveness measure was added which took into account both content and readability of the summary. The average performance of systems in 2006 was noticeably better than in 2005; systems achieved better focus on average, and many attempted to provide greater coherence to their summaries. The overall responsiveness metric showed that readability plays an important role in the perceived quality of the summaries.

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 and Technology (NIST) with the goal of furthering progress in automatic summarization and enabling researchers to participate in large-scale experiments.

In DUC 2001-2005 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. Additionally, DUC began exploring automatic evaluation of content coverage using ROUGE (Lin, 2004) in 2004 and Basic Elements (BE) (Hovy et al., 2005) in 2005.

DUC 2005 (Dang, 2005) marked a major change in direction from previous years. DUC 2005 had a single user-oriented, question-focused summarization task that allowed researchers to devote some resources to helping with the evaluation. Prior to 2005, 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. SEE provided detailed feedback about which sentences contained overlapping information in the peer and model. However, since model summaries vary in content, the research community wanted an evaluation measure that would not depend on a single model summary.

NIST has since moved to a pseudo-extrinsic evaluation of content, called responsiveness, which does not attempt pairwise comparison of peers against a model summary but assigns a value from a 5-point scale to each summary based on its responsiveness to a specified topic. Responsiveness is only a coarse-grained measure of content, so in DUC 2005, researchers also participated in an optional manual Pyramid evaluation led by Columbia University (Passonneau et al., 2005). The Pyramid evaluation gives researchers detailed feedback about which information is contained in each of several model summaries, assigns different importance to each piece of information based on the number of model summaries it appears in, and says which information is also contained in the peer summaries.

DUC 2006 repeated the DUC 2005 task and evaluation. The system task modeled real-world complex question answering. Systems were to synthesize from a set of 25 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.

As in DUC 2005, 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. In 2006, two variants of responsiveness were measured: content responsiveness (based only on the amount of information in the summary that responded to the topic) and overall responsiveness (based on both content and readability). NIST also computed automatic ROUGE and BE scores as in 2005, and Columbia University again led the summarization research community in a voluntary Pyramid evaluation of summary content.

This paper describes the DUC 2006 task and the results of NIST's evaluations of summary content, readability, and overall quality. (Passonneau et al., 2006) provides additional details and results of the evaluation of summary content using the Pyramid method.

2 Task and Data

The DUC 2006 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 related documents from the *Associated Press, New York Times*, and *Xinhua* newswire 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.

An example topic from DUC 2006 follows:

> **num**: D0641E
> **title**: global warming
> **narr**: Describe theories concerning the causes and effects of global warming and arguments against these theories.

The summarization task was the same for both human and automatic summarizers: Given a topic 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. 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.[1] No specific formatting other than linear was allowed.

[1] A number of summaries were erroneously truncated to fewer than 250 words before being evaluated in the official manual (Responsiveness, Linguistic Quality, Pyramid) and automatic (ROUGE, BE) evaluations. In particular, the performance of Systems 2, 8, 9, 15, and 19 may be higher than indicated by the evaluation scores reported in this paper. A subsequent automatic evaluation of all summaries, using correctly truncated summaries, yielded sigificantly higher ROUGE-2 and ROUGE-SU4 scores for Systems 8 and 15, according to the 95% confidence intervals computed by ROUGE-1.5.5; none of the changes in any automatic scores for any other runs were significant.

Ten NIST assessors produced a total of 4 human summaries for each of 50 topics, and 34 participants submitted runs to be evaluated. NIST also developed a simple baseline system that returned all the leading sentences of the "TEXT" field of the most recent document for each topic, up to 250 words. 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.

3 Evaluation Results

Summaries were manually evaluated by 10 NIST 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. Assessors first judged each summary for a topic for *readability*, assigning a separate score for each of 5 linguistic qualities; each summary for the topic was then judged for *content responsiveness*. After all summaries for all topics had been judged for readability and content responsiveness, assessors then judged each summary for *overall responsiveness*. The content responsiveness score provides a coarse manual measure of information coverage; overall responsiveness reflects a combination of readability and content.

Each of these manual evaluations was based on a five-point scale:

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

Responsiveness and readability scores are ordinal data and should technically be analyzed with non-parametric statistical tests. However, parametric and non-parametric analyses yield similar results for these metrics, with the parametric tests finding slightly more statistically significant differences between peers. Since ROUGE and BE scores *are* suitable for parametric analysis, we uniformly perform an analysis of variance (ANOVA) for all metrics to determine if there is a statistically significant difference between peers according to the metric. We then performed a multiple comparison test between the scores of the peers using Tukey's honestly significant difference criterion, to determine which pairs of peers are significantly different at the 95% confidence level.

Run ID	System ID	Organization
1	Baseline	(NIST)
2	OGI.OHSU06	Oregon Health & Science University
3	IS_SUM	Chinese Academy of Sciences
4	CLResearch.duc06	CL Research
5	Columbia06	Columbia University
6	FDUSUM	Fudan University
7	ISI-Webcl	Information Sciences Institute (Zhou)
8	JIKD	IDA CCS *and* University of Maryland Joint Institute for Knowledge Discovery
9	MQ06	Macquarie University
10	MSR	Microsoft Research
11	NKTrust	NK Trust, Inc.
12	OnModer	National University of Singapore
13	SFU_v36	Simon Fraser University
14	TUTNII	Toyohashi University of Technology
15	CCS06	IDA Center for Computing Sciences
16	UConnDG	University of Connecticut
17	BCBB-DUC	National Central University
18	UTwente06	University of Twente
19	envQASUM	Universitat Politcnica de Catalunya
20	ERSS06	University of Karlsruhe *and* Concordia University
21	FSC-wm-pairs=.3	Fitchburg State College
22	HKPolyU	Hong Kong Polytechnic University
23	ICL_SUM	Peking University
24	IIITH-Sum	International Institute of Information Technology
25	IIRG-UCD-2006	University College Dublin
26	ISI-BQFS	Information Sciences Institute (Daume)
27	lcc.duc06	Language Computer Corporation
28	LIA_THALES	University of Avignon
29	MIRACL06	Larim Unit (MIRACL Laboratory)
30	titech-uam	Tokyo Institute of Technology *and* Universidad Autonoma de Madrid
31	TLR	Thomson Legal & Regulatory
32	UMD_BBN	University of Maryland *and* BBN Technologies
33	UMich	University of Michigan
34	LAKE06	University of Salerno
35	UofO	University of Ottawa

Table 1: Participants and runs in DUC 2006.

3.1 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.

Table 2 shows the distribution of the scores across all the summaries, broken down by the type of summarizer (Human, Baseline, or Participants). As in DUC 2005, all summarizers generally performed well on the first two linguistic qualities. Participants scored higher on Focus in 2006 than in 2005, with the best systems achieving scores comparable to humans. As a group, participants' performance remained unchanged on referential clarity and structure and coherence, though the best individual participants do come close to human performance on these qualities.

Tables 3-7 show the results of multiple comparison of the automatic peers for each linguistic quality, with best peers on top; peers not sharing a common letter are significantly different at the 95.5% confidence level. An analysis using the non-parametric Friedman's test instead of ANOVA yields similar results.

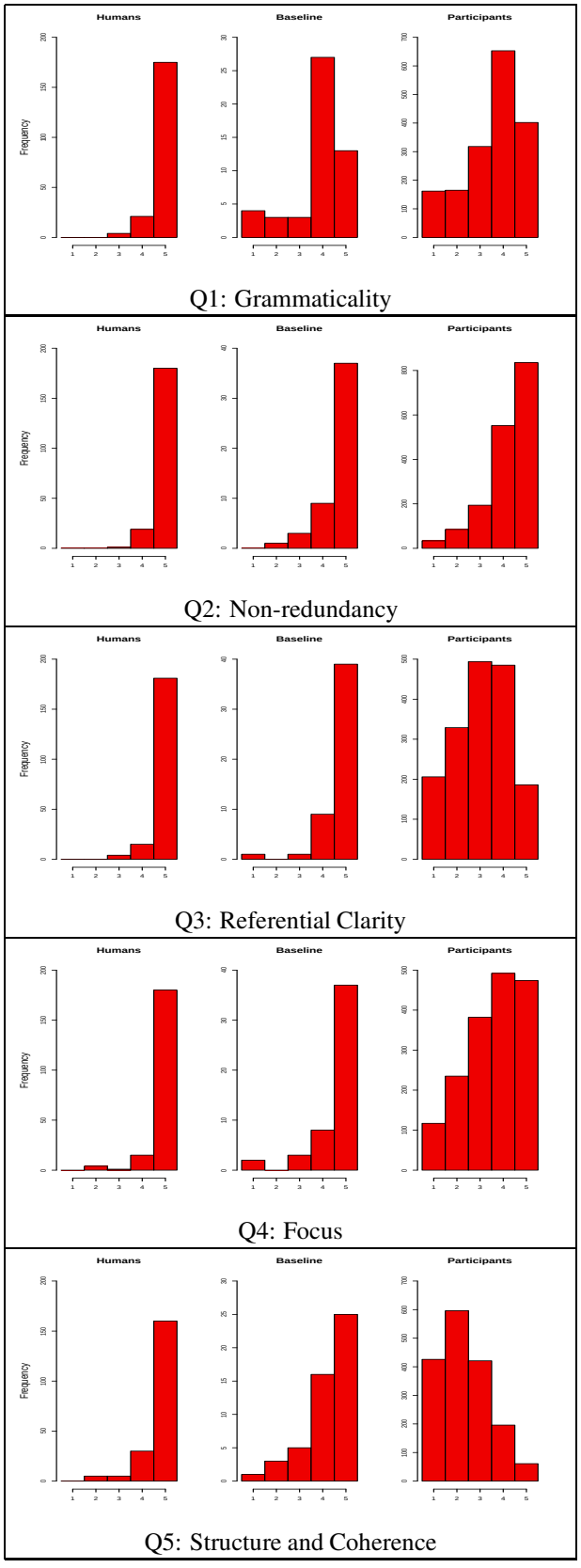

Table 2: Frequency of scores for each linguistic quality, broken down by source of summary (Humans, Baseline, Participants).

RunID	score											
27	4.6200	A										
35	4.5200	A	B									
22	4.4200	A	B	C								
18	4.4200	A	B	C								
29	4.2200	A	B	C	D							
23	4.1600	A	B	C	D	E						
28	4.0800	A	B	C	D	E						
13	4.0000	A	B	C	D	E	F					
20	3.9600	A	B	C	D	E	F					
26	3.8600		B	C	D	E	F	G				
1	3.8400		B	C	D	E	F	G				
3	3.8200		B	C	D	E	F	G	H			
2	3.8000			C	D	E	F	G	H	I		
5	3.7400			C	D	E	F	G	H	I		
21	3.7200			C	D	E	F	G	H	I	J	
24	3.6400				D	E	F	G	H	I	J	
16	3.6400				D	E	F	G	H	I	J	
4	3.6000				D	E	F	G	H	I	J	
14	3.5800				D	E	F	G	H	I	J	
17	3.5600				D	E	F	G	H	I	J	
7	3.5200				D	E	F	G	H	I	J	K
30	3.5200				D	E	F	G	H	I	J	K
31	3.5000					E	F	G	H	I	J	K
15	3.3400						F	G	H	I	J	K L
9	3.3000						F	G	H	I	J	K L
25	3.2200							G	H	I	J	K L
19	3.2200							G	H	I	J	K L
33	3.1600							G	H	I	J	K L
10	3.1200								H	I	J	K L
8	3.1000									I	J	K L
6	3.0200										J	K L
34	3.0200										J	K L
12	2.8400											K L
32	2.7400											L
11	1.3800											M

Table 3: Multiple comparison of systems based on ANOVA of Q1: Grammaticality

RunID	score											
35	4.6600	A										
1	4.6400	A	B									
26	4.5800	A	B	C								
30	4.5600	A	B	C	D							
27	4.5000	A	B	C	D							
18	4.5000	A	B	C	D							
11	4.5000	A	B	C	D							
7	4.4800	A	B	C	D							
22	4.4600	A	B	C	D							
10	4.4200	A	B	C	D	E						
34	4.4000	A	B	C	D	E						
5	4.3600	A	B	C	D	E	F					
29	4.3600	A	B	C	D	E	F					
17	4.3600	A	B	C	D	E	F					
4	4.3400	A	B	C	D	E	F					
3	4.3200	A	B	C	D	E	F					
2	4.3000	A	B	C	D	E	F					
14	4.2600	A	B	C	D	E	F					
13	4.2400	A	B	C	D	E	F					
9	4.2000	A	B	C	D	E	F					
33	4.1800	A	B	C	D	E	F					
16	4.1200	A	B	C	D	E	F					
25	4.1000	A	B	C	D	E	F					
20	4.0800	A	B	C	D	E	F					
23	4.0600	A	B	C	D	E	F					
21	4.0400		B	C	D	E	F					
12	4.0200			C	D	E	F					
24	4.0000			C	D	E	F					
6	3.9800			C	D	E	F					
19	3.9600				D	E	F					
8	3.8400					E	F					
28	3.8400					E	F					
15	3.8200					E	F					
31	3.7800						F					
32	3.7600						F					

Table 4: Multiple comparison of systems based on ANOVA of Q2: Non-Redundancy

RunID	score											
1	4.7000	A										
34	4.0000	A	B									
23	3.8600		B	C								
27	3.7200		B	C	D							
21	3.4600		B	C	D	E						
28	3.4200		B	C	D	E	F					
24	3.4200		B	C	D	E	F					
5	3.4000		B	C	D	E	F					
2	3.4000		B	C	D	E	F					
13	3.3800		B	C	D	E	F					
18	3.3200		B	C	D	E	F	G				
31	3.2600			C	D	E	F	G				
30	3.2200			C	D	E	F	G				
14	3.2200			C	D	E	F	G				
12	3.2200			C	D	E	F	G				
33	3.2000			C	D	E	F	G				
8	3.1600			C	D	E	F	G	H			
4	3.1600			C	D	E	F	G	H			
35	3.1600			C	D	E	F	G	H			
6	3.0800				D	E	F	G	H	I		
9	3.0600				D	E	F	G	H	I		
3	3.0200				D	E	F	G	H	I		
17	3.0000					E	F	G	H	I		
15	2.9800					E	F	G	H	I		
16	2.8800					E	F	G	H	I		
32	2.8400					E	F	G	H	I		
19	2.8000					E	F	G	H	I		
25	2.7600					E	F	G	H	I	J	
22	2.7600					E	F	G	H	I	J	
29	2.7200						F	G	H	I	J	
10	2.6400							G	H	I	J	
20	2.4600								H	I	J	K
7	2.3800									I	J	K
11	2.0600										J	K
26	1.9000											K

Table 5: Multiple comparison of systems based on ANOVA of Q3: Referential Clarity

RunID	score											
1	4.5600	A										
27	4.2800	A	B									
34	4.1200	A	B	C								
24	3.9400		B	C	D							
31	3.8600		B	C	D	E						
5	3.8400		B	C	D	E						
21	3.8200		B	C	D	E						
4	3.8000		B	C	D	E						
33	3.8000		B	C	D	E						
23	3.8000		B	C	D	E						
2	3.7800		B	C	D	E						
13	3.7800		B	C	D	E						
28	3.7400		B	C	D	E	F					
15	3.7400		B	C	D	E	F					
18	3.7200		B	C	D	E	F					
22	3.6800		B	C	D	E	F					
12	3.6600			C	D	E	F					
30	3.6200			C	D	E	F					
8	3.6000			C	D	E	F					
3	3.5800			C	D	E	F					
6	3.5200			C	D	E	F					
17	3.5200			C	D	E	F					
14	3.5200			C	D	E	F					
35	3.5000				D	E	F					
16	3.4600				D	E	F					
25	3.4400				D	E	F					
32	3.4200				D	E	F					
9	3.3600				D	E	F					
19	3.3600				D	E	F					
29	3.3400				D	E	F					
20	3.3200					E	F					
10	3.3200					E	F					
7	3.1600						F					
26	2.5200							G				
11	2.5000							G				

Table 6: Multiple comparison of systems based on ANOVA of Q4: Focus

RunID	score										
1	4.2200	A									
27	3.2800		B								
34	3.0800		B	C							
18	2.8200		B	C	D						
24	2.8000		B	C	D	E					
30	2.7800		B	C	D	E					
13	2.7200		B	C	D	E	F				
23	2.6400			C	D	E	F	G			
22	2.6400			C	D	E	F	G			
21	2.5800			C	D	E	F	G	H		
33	2.5600			C	D	E	F	G	H		
5	2.5200			C	D	E	F	G	H		
35	2.5000			C	D	E	F	G	H		
31	2.5000			C	D	E	F	G	H		
4	2.4800				D	E	F	G	H		
2	2.4800				D	E	F	G	H		
14	2.4200				D	E	F	G	H	I	
3	2.3000				D	E	F	G	H	I	J
20	2.2800				D	E	F	G	H	I	J
28	2.2600				D	E	F	G	H	I	J
17	2.2600				D	E	F	G	H	I	J
29	2.2200					E	F	G	H	I	J
25	2.2200					E	F	G	H	I	J
15	2.1600						F	G	H	I	J
6	2.1400						F	G	H	I	J
16	2.1200							G	H	I	J
9	2.1000							G	H	I	J
7	2.0800							G	H	I	J K
8	2.0600							G	H	I	J K
19	2.0600							G	H	I	J K
12	2.0400								H	I	J K
32	1.8400									I	J K
10	1.8000										J K
26	1.5000										K L
11	1.1600										L

Table 7: Multiple comparison of systems based on ANOVA of Q5: Structure and Coherence

3.2 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 and BE scores as was done in DUC 2005.

3.2.1 Manual Responsiveness

NIST assessors assigned two types of responsiveness scores to each summary. The *content* responsiveness score indicated the amount of information in the summary that helped to satisfy the information need expressed in the topic statement. For content responsiveness, 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 *overall* responsiveness score was based on both information content and readability. Assessors judged overall responsiveness only *after* judging all their topics for readability and content responsiveness; however, they were not given direct access to these previously assigned scores, but were told to give their "gut" reaction to the overall responsiveness of each summary. Many assessors found it helpful to recast overall responsiveness as asking "How much money would I pay for this summary?" and judged accordingly. Poor readability in the automatic peers generally resulted in the average overall responsiveness for each peer being much lower than its average content responsiveness.

Table 8 shows the results of a multiple comparison of content responsiveness of the automatic peers, and Table 9 shows the same analysis on overall responsiveness. An analysis using Friedman's test yields similar results.

RunID	score								
27	3.0800	A							
23	3.0000	A	B						
10	2.9400	A	B	C					
12	2.9200	A	B	C	D				
24	2.8800	A	B	C	D	E			
31	2.8600	A	B	C	D	E			
14	2.8200	A	B	C	D	E	F		
28	2.7800	A	B	C	D	E	F		
5	2.7600	A	B	C	D	E	F		
13	2.7000	A	B	C	D	E	F		
6	2.6200	A	B	C	D	E	F	G	
3	2.6000	A	B	C	D	E	F	G	
32	2.6000	A	B	C	D	E	F	G	
19	2.6000	A	B	C	D	E	F	G	
8	2.5800	A	B	C	D	E	F	G	
33	2.5800	A	B	C	D	E	F	G	
30	2.5800	A	B	C	D	E	F	G	
22	2.5600	A	B	C	D	E	F	G	
4	2.5400	A	B	C	D	E	F	G	
2	2.5400	A	B	C	D	E	F	G	
20	2.5200	A	B	C	D	E	F	G	
7	2.5000	A	B	C	D	E	F	G	
15	2.4800	A	B	C	D	E	F	G	
29	2.4400		B	C	D	E	F	G	
35	2.4200		B	C	D	E	F	G	
17	2.3800			C	D	E	F	G	
9	2.3600			C	D	E	F	G	
21	2.3600			C	D	E	F	G	
25	2.3400			C	D	E	F	G	
18	2.3200				D	E	F	G	
16	2.3000					E	F	G	
34	2.2400						F	G	H
26	2.0600							G	H
1	2.0400							G	H
11	1.6800								H

Table 8: Multiple comparison of systems based on ANOVA of content responsiveness

In a multiple comparison of all peers, all human peers were significantly better than all the automatic peers, in both content and overall responsiveness, and the humans were indistinguishable from one another. While the system with the highest average content and overall responsiveness scores, System 27, is still not performing at human level, there are certain topics where its overall responsiveness is as high as the human scores. For topic D0641E, for example, the system is given an overall responsiveness score of 5 (very good) and a content score of 3 (barely acceptable) for the following summary:

"*The dominant view is that the surface warming is at least partly attributable to emissions of heat-trapping waste industrial gases like carbon dioxide, a product of the burning of fossil fuels like coal, oil and natural gas. On that issue, and on the remaining big question of how the climate might change in the future, skeptics continue to differ sharply with the dominant view among climate*

RunID	score						
27	2.8400	A					
23	2.7600	A	B				
31	2.6000	A	B	C			
2	2.4600	A	B	C	D		
24	2.4400	A	B	C	D		
5	2.4200	A	B	C	D	E	
28	2.4200	A	B	C	D	E	
14	2.4200	A	B	C	D	E	
6	2.3600	A	B	C	D	E	
13	2.3600	A	B	C	D	E	
33	2.2800		B	C	D	E	
20	2.2800		B	C	D	E	
34	2.2400		B	C	D	E	
3	2.2200		B	C	D	E	F
30	2.2200		B	C	D	E	F
12	2.2200		B	C	D	E	F
35	2.2000			C	D	E	F
4	2.1800			C	D	E	F
10	2.1600			C	D	E	F
9	2.1200			C	D	E	F
22	2.1200			C	D	E	F
7	2.0800			C	D	E	F
32	2.0800			C	D	E	F
29	2.0800			C	D	E	F
21	2.0800			C	D	E	F
25	2.0600			C	D	E	F
15	2.0600			C	D	E	F
1	2.0000				D	E	F
19	1.9800				D	E	F
18	1.9800				D	E	F
16	1.9800				D	E	F
8	1.9600				D	E	F
17	1.8800					E	F G
26	1.6800						F G
11	1.3400						G

Table 9: Multiple comparison of systems based on ANOVA of overall responsiveness

experts. To them, the observed surface warming of about 1 degree over the last century_with an especially sharp rise in the last quarter century_is mostly or wholly natural, and there is no significant human influence on global climate. The West Antarctic Ice Sheet first attracted widespread attention 30 years ago when scientists suggested that global warming caused by greenhouse gases might cause its disintegration. Last year, scientists declared 1997 the warmest year on record, and the fact that nine of the past 11 years set new records for warm temperatures bolstered the view that greenhouse emissions were raising the average temperature. Over time, these increases could cause changes in climate, including the increased frequency and intensity of storms, floods, heat waves, and droughts, the scientists said. A U.N. scientific panel has predicted that unless these greenhouse gas emissions are reduced, the earth's average surface temperature will rise by some 2 to 6 degrees Fahrenheit over the next century, with a best estimate of about 3.5 degrees, compared with a rise of 5 to 9 degrees since the depths of the last ice age 18,000"

While poor readability can certainly downgrade the overall responsiveness of a summary that has very good content responsiveness, the example shows that very good readability can sometimes bolster the overall responsiveness score of a less information-laden summary.

We used a linear regression to model the effect of the 5 linguistic qualities and content responsiveness on overall responsiveness. The purpose of multiple linear regression is to establish a quantitative relationship between a group of predictor variables X and a response, y: $y = \beta X + \epsilon$. This relationship is useful for:

- Understanding which predictors have the greatest effect.
- Knowing the direction of the effect (i.e., increasing x $\in$ X increases or decreases y).
- Using the model to predict future values of the response when only the predictors are currently known.

Table 10 shows the weight β of each factor on overall responsiveness, for each of the 10 assessors. Figure 1 shows the weights when considering all 10 assessors.

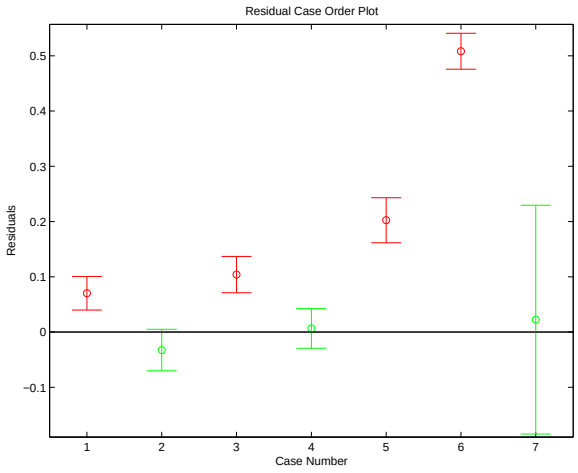

Figure 1: Multiple linear regression with all 10 assessors, $R^2 = 0.5877$. Case Numbers 1-5 correspond to the five linguistic qualities; Case Number 6 corresponds to content responsiveness.

3.2.2 Automatic ROUGE/BE

NIST computed three "official" automatic scores: ROUGE-2, ROUGE-SU4, and 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. The per-topic recall was computed for each peer, and this per-topic recall was used as the dependent variable in an analysis of variance. Tables 11-13 show the results of multiple comparison of systems based on the automatic scores.

Assessor	β: Q1	β: Q2	β: Q3	β: Q4	β: Q5	β: Content	R^2
B	0.0623	**-0.1068**	0.0604	-0.0738	**0.2996**	**0.5955**	0.7543
J	0.0419	0.0106	0.0355	-0.0902	**0.4183**	**0.5366**	0.7439
A	-0.0016	-0.0374	0.0560	0.0618	0.1033	**0.6973**	0.7316
E	0.0677	0.0153	**0.2513**	0.0463	0.0803	**0.5028**	0.6911
I	0.0789	0.0165	0.0969	-0.0135	**0.1736**	**0.5765**	0.6221
D	0.0207	-0.0289	0.0073	0.0129	**0.3415**	**0.4936**	0.5512
C	-0.0003	-0.0822	**0.1695**	-0.0223	**0.1977**	**0.5474**	0.5096
F	-0.0277	**0.2280**	**0.1635**	-0.0302	-0.0510	**0.7250**	0.4759
H	**0.1018**	-0.0169	**0.1395**	**-0.1494**	**0.2569**	**0.3909**	0.4530
G	0.0389	-0.1576	**0.2211**	-0.0286	**0.5293**	**0.1967**	0.3945

Table 10: Linear regression model for each assessor. Values of β in bold are signficiantly different from 0. R^2 measures the amount of variability in the observations accounted for by the model.

RunID	score	
24	0.0951	A
12	0.0899	A B
23	0.0879	A B C
8	0.0871	A B C
28	0.0870	A B C
15	0.0868	A B C
31	0.0858	A B C D
33	0.0845	A B C D
2	0.0841	A B C D
5	0.0827	A B C D E
6	0.0809	B C D E F
27	0.0809	B C D E F
32	0.0805	B C D E F
13	0.0799	B C D E F G
3	0.0792	B C D E F G H
10	0.0780	B C D E F G H I
4	0.0774	B C D E F G H I J
19	0.0764	C D E F G H I J
9	0.0754	C D E F G H I J K
22	0.0748	C D E F G H I J K
14	0.0729	D E F G H I J K
29	0.0703	E F G H I J K
25	0.0695	E F G H I J K L
16	0.0678	F G H I J K L
30	0.0667	G H I J K L
21	0.0663	H I J K L
18	0.0659	I J K L
34	0.0650	I J K L
20	0.0648	I J K L
17	0.0645	J K L M
7	0.0626	K L M N
35	0.0567	L M N
26	0.0516	M N
1	0.0495	N
11	0.0284	O

Table 11: Multiple comparison of systems based on ANOVA of ROUGE-2 score

RunID	score	
24	0.1547	A
12	0.1475	A B
28	0.1452	A B C
33	0.1449	A B C
23	0.1449	A B C
31	0.1438	A B C D
15	0.1417	B C D E
8	0.1413	B C D E
5	0.1402	B C D E
2	0.1391	B C D E
6	0.1374	B C D E F
10	0.1372	B C D E F
32	0.1360	B C D E F G
3	0.1359	B C D E F G
27	0.1359	B C D E F G
13	0.1353	B C D E F G H
4	0.1332	C D E F G H I
22	0.1316	D E F G H I J
19	0.1312	D E F G H I J
14	0.1291	E F G H I J
9	0.1290	E F G H I J
29	0.1252	F G H I J K
25	0.1248	F G H I J K
20	0.1239	G H I J K
16	0.1238	G H I J K
18	0.1229	H I J K
7	0.1226	H I J K
30	0.1223	I J K
21	0.1199	J K L
34	0.1189	J K L
17	0.1150	K L M
35	0.1081	L M N
26	0.1045	M N
1	0.0979	N
11	0.0640	O

Table 12: Multiple comparison of systems based on ANOVA of ROUGE-SU4

RunID	score										
24	0.0508	A									
23	0.0505	A									
28	0.0476	A	B								
2	0.0471	A	B								
12	0.0471	A	B								
8	0.0464	A	B	C							
15	0.0458	A	B	C	D						
31	0.0456	A	B	C	D						
10	0.0437	A	B	C	D	E					
6	0.0436	A	B	C	D	E	F				
27	0.0419	A	B	C	D	E	F				
13	0.0415	A	B	C	D	E	F				
32	0.0413	A	B	C	D	E	F				
5	0.0410	A	B	C	D	E	F				
4	0.0410	A	B	C	D	E	F				
3	0.0407	A	B	C	D	E	F				
33	0.0389	A	B	C	D	E	F	G			
9	0.0385		B	C	D	E	F	G			
22	0.0373		B	C	D	E	F	G			
14	0.0363		B	C	D	E	F	G	H		
19	0.0350			C	D	E	F	G	H		
25	0.0348			C	D	E	F	G	H	I	
30	0.0344				D	E	F	G	H	I	
21	0.0341				D	E	F	G	H	I	
20	0.0341				D	E	F	G	H	I	
29	0.0328					E	F	G	H	I	
16	0.0318						F	G	H	I	
18	0.0288							G	H	I	J
7	0.0285							G	H	I	J
34	0.0284							G	H	I	J
35	0.0253								H	I	J
26	0.0230									I	J
1	0.0194										J
17	0.0046											K
11	0.0046											K

Table 13: Multiple comparison of systems based on ANOVA of BE-HM

3.3 Correlation

Figure 2 plots the average content responsiveness score with average ROUGE-SU4 score for all peers; as seen in the graph, the peers form two clusters, with the humans clumped on the upper right side of the graph, and the automatic peers spread out on the lower left side. The manual content responsiveness metric (x-axis) clearly separates the humans from the automatic peers, while the difference between the humans and automatic peers is quite small based on the automatic metric (y-axis). ROUGE-2 and BE-HM yield similar graphs.

Table 14 shows the correlation between average content responsiveness and the other measures involving content, computed over only the automatic peers. Both Spearman's rank correlation rho and Pearson's product-moment correlation (with 95% confidence intervals) are shown.[2] As expected, content responsiveness and overall responsiveness were only moderately correlated, since many peers that scored well on content largely ignored readability, resulting in lower overall responsiveness.

[2]The Pearson correlation between BE and content responsiveness is low because the BEs depend on linguistic preprocessing, and any brittleness in a pre-processor can prevent BEs from being extracted from a summary. System 17, for example, has extremely low BE scores even though it has relatively better scores under the other measures of content, because most of its summaries are slightly ungrammatical (based on human standards): Punctuation marks are presented as a separate token instead of being attached to the previous word; this prevents the sentence segmenter from segmenting the summary into sentences, and the parser from generating a bracketing from which BEs can be extracted.

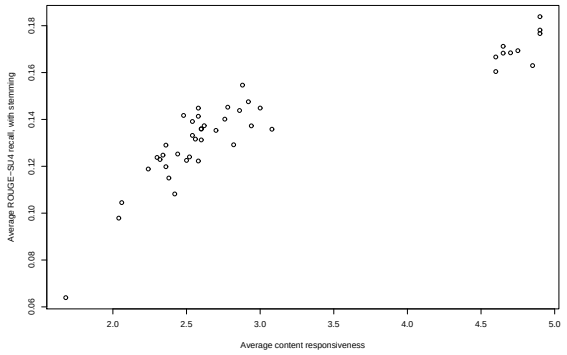

Figure 2: Average content responsiveness vs. average ROUGE-SU4 recall with stemming

Metric	Spearman	Pearson
overall responsiveness	0.718	0.833 [0.720, 1.000]
R2	0.767	0.836 [0.725, 1.000]
RSU4	0.790	0.850 [0.746, 1.000]
BE-HM	0.797	0.782 [0.641, 1.000]

Table 14: Correlation between average content responsiveness and overall responsiveness, average ROUGE-2/ROUGE-SU4 recall, and average BE-HM recall over all automatic peers.

What is surprising is that the correlation between ROUGE scores and content responsiveness is lower in DUC 2006 than in DUC 2005. There are a number of differences between the DUC 2005 and 2006 evaluation conditions that could have led to the lower correlation in 2006, including differences in number of models per topic, whether content responsiveness was first converted to a rank before being averaged, and whether or not a desired "granularity" was specified for the summary. In 2005 some topics had 9 models instead of 4; the responsiveness score for each topic was also scaled by the number of peers for the topic to compute a rank for each peer, and these *ranks* were then averaged across topics to compute "average scaled responsiveness" (Dang, 2005), which was then compared with average ROUGE scores. We mitigated the effect of these differences by using only 4 randomly selected models for each summary from 2005, and by converting the 2006 content responsiveness score to average scaled responsiveness.

Table 15 shows the correlations of average scaled content responsiveness (average of rank of system for each topic) vs. macroaveraged ROUGE recall with stemming, keeping stopwards, using jackknifing; we compared only

Metric	Spearman	Pearson
R2 (2005)	0.889	0.926 [0.868, 1.000]
RSU4 (2005)	0.867	0.917 [0.852, 1.000]
R2 (2006)	0.759	0.835 [0.722, 1.000]
RSU4 (2006)	0.780	0.849 [0.745, 1.000]

Table 15: Correlation between average scaled responsiveness and ROUGE-2/ROUGE-SU4 recall, over all automatic peers, computed using exactly 4 models per topic

2005 Metric	Spearman	Pearson
R2 (general)	0.804	0.827 [0.702, 1.000]
RSU4 (general)	0.841	0.868 [0.770, 1.000]
R2 (specific)	0.912	0.928 [0.871, 1.000]
RSU4 (specific)	0.884	0.921 [0.858, 1.000]

Table 16: Correlation between average scaled responsiveness and ROUGE-2/ROUGE-SU4 recall, over all automatic peers for DUC 2005, by granularity

the automatic peers, using only 4 models. The correlations for DUC 2006 are still lower than for DUC 2005 when considering all 50 topics. However, as shown in Table 16, when the DUC 2005 topics are broken down into topics with specific vs. general summaries, we see that correlations for general summaries are much lower than for specific summaries. This should not be surprising, since "specific" summaries in DUC 2005 had a large number of named entities (specific people, places, dates, etc.), which can be matched using simple string matching. On the other hand, general [model] summaries abstract over concepts in the documents, so string matching us less effective at detecting overlapping content between summaries.

4 Conclusion

The automatic summaries in DUC 2006 showed an improvement over those in DUC 2005, with many peers achieving better focus. Attempts at greater readability also paid off among the peers with the best overall responsiveness scores. While content responsiveness was largely responsible for determining how assessors perceived the overall quality of a summary, readability (especially structure and coherence) also played an important role, sometimes boosting the overall responsiveness of a less information-laden summary.

ROUGE was widely used by participants to develop their DUC 2006 systems. While it is less effective at predicting the quality of general/abstractive summaries, ROUGE appears to still be effective as a development tool; many participants attributed their improved performance in DUC 2006 to the ability to train their systems to optimize ROUGE scores on past data.

Acknowledgments

We are grateful to Becky Passoneau at Columbia University for organizing the Pyramid evaluation, and to all the DUC participants for contributing their system results and analyses.

References

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the Fifth Document Understanding Conference (DUC)*, Vancouver, Canada.

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.

Rebecca J. Passonneau, Kathleen McKeown, Sergey Sigelman, and Adam Goodkind. 2006. Applying the pyramid method in the 2006 document understanding conference. In *Proceedings of the Sixth Document Understanding Conference (DUC)*, New York City, NY.