Overview of DUC 2005 (DRAFT)

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 the variation in content of human-authored summaries. Therefore, DUC 2005 had a single user-oriented task that allowed the community to put some of their time and effort into helping with the new evaluation framework during 2005. The question-focused summarization task was 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.

1 Introduction

Over the past several years, we have witnessed a tremendous increase in interest in summarization research from both academia and industry. A DARPA program, Translingual Information Detection, Extraction, and Summarization (TIDES), specifically called for major advances in summarization technology, both in English and from other languages to English (cross-language summarization). In response, the National Institute of Standards and Technology (NIST) initiated the Document Understanding Conference (DUC) series to evaluate automatic text summarization. Its goal is to further progress in summarization and enable 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 (10-400 words) were used and both single-document summaries and summaries of multiple documents (around 10 documents per set) were evaluated. Summaries were manually judged for their readability, and both manual and automatic (Rouge) evaluation of content coverage have been explored. In 2004 the output of Arabic-to-English MT systems was also summarized.

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. 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, there was serious discussion in the program committee about working on new evaluation methodologies and metrics that take into account variation of content in human-authored summaries (Harman and Over, 2004). In previous DUCs, NIST had done fine-grained manual evaluation of content against a single reference (human) summary using SEE. However, given the variation in human-authored summaries, the community feared that evaluating peer summaries against a single reference might lead to unstable evaluation results.

Therefore it was decided that the main thrust of DUC 2005 would be to have a single simpler (but still user-oriented) task that would allow the community to put some of their time and effort into helping with a new evaluation framework during 2005. The system task in 2005 was 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. This task was to model real-world complex question answering and was suggested by (Amigo et al., 2004).

The main goals in DUC 2005 and their associated actions are listed below.

1. Inclusion of user/task context information for systems and human summarizers
    - create DUC topics which explicitly reflect the specific interests of a potential user in a task context

- capture some general user preferences in a simple user profile, where the user indicates the "granularity," or level of generalization desired in the summary.

2. Evaluation of content in terms of more basic units of meaning
   - develop automatic tools and manual procedures to identify basic units of meaning
   - develop automatic tools and manual procedures to estimate the importance of such units based on agreement among humans
   - use the above in evaluating systems
   - evaluate the new evaluation scheme(s)
3. Better understanding of normal human variability in a summarization task and how it may affect evaluation of summarization systems
   - create as many manual reference summaries as feasible
   - examine the relationship between the number of reference summaries, the ways in which they vary, and the effect of the number and variability on system evaluation

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.

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

Given a user profile, a DUC topic, and a cluster of documents relevant to the DUC 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 level of granularity specified in the user profile. 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.

3 Participants

There was much interest in the longer, question-focused summaries required in the DUC2005 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.

4 Evaluation of Linguistic Quality

NIST assessors judged each summary for linguistic quality and Responsiveness. All summaries for a given topic were judged by a single assessor. In most cases, the assessor judging the summaries for a topic was the same assessor who developed the topic.

The linguistic quality questions assessed how readable and fluent the summaries are, and they measure qualities of the summary that *do not* involve comparison with a reference summary or DUC topic. The five linguistic qualities that were measured were *Grammaticality, Non-redundancy, Referential clarity, Focus, and Structure and coherence*. All linguistic quality questions required a certain readability property to be assessed on a five-point scale from "A" to "E", where "A" indicated that the summary was good with the respect to the quality under question, "E" indicated that the summary is bad with respect to the quality stated in the question, and "B" to "D" showed the gradation in between.

In this section we show the results of multiple comparisons of linquistic quality scores between peers using Tukey's honestly significant difference criterion for each quality question. Tables 2-6 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). For each quality ques-

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	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.

tion, we also did multiple comparison between all human and automatic peers using the Kruskall-Wallis test instead of Friedman's test, to see how the automatic peers performed relative to human peers.

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.

Multiple comparison of all peers on Q1 shows 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).

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

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.

Multiple comparison of all peers on Q2 shows 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.

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

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.

Multiple comparison of all peers on Q3 shows that all humans are significantly better than all but 2 automatic peers (1, 12).

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

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

Multiple comparison of all peers on Q4 shows that 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.

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	I
19			C	D	E	F	G	H	I
10			C	D	E	F	G	H	I
25			C	D	E	F	G	H	I
9			C	D	E	F	G	H	I
6			C	D	E	F	G	H	I
11			C	D	E	F	G	H	I
7			C	D	E	F	G	H	I
18				D	E	F	G	H	I
13					E	F	G	H	I
27					E	F	G	H	I
31						F	G	H	I
22							G	H	I
30								H	I
23									I

Table 5: Multiple comparison of systems based on Friedman's test on Q4: Focus

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.

Multiple comparison of all peers on Q5 shows that 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).

5 Evaluation of Content

NIST did manual pseudo-extrinsic evaluation of peer summaries in the form of assessment of Responsiveness. Responsiveness is different from SEE coverage in that it

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

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 this year. NIST also computed Rouge scores as was done last year.

In addition, ISI and Columbia spearheaded two exploratory efforts at intrinsic evaluation of automatic summaries compared against multiple reference summaries, using automatic methods at ISI and manual methods at Columbia:

1. [Columbia] Manual evaluation using coincidence of meaning units (Summary Content Units, or SCUs) between a peer summary and a pyramid of SCUs constructed from multiple reference summaries. DUC participants helped with this manual evaluation, using tools and guidelines distributed by Columbia. *(See Pyramid overview paper, in this volume.)*

2. [ISI] Automatic evaluation using coincidence of meaning units (Basic Elements, or BEs) between a peer summary and one or more reference summaries. The BEs are extracted from parse structures and matched across summaries using various heuristics. *(See BE overview paper, in this volume.)*

5.1 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 cannot 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.

Figure 7 shows multiple comparisons 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 comparisons of all peers using the Kruskal-Wallis test, all human peers were significantly better than all the automatic peers.

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 7: Multiple comparison of systems based on Friedman's test on Responsiveness

5.2 Rouge

NIST computed two official Rouge scores: Rouge-2 and Rouge-SU4 recall, both with stemming and implementing jackknifing 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, we did multiple comparisons of population marginal means (PMM) 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 factor "RunID", and averaging out the effects of the other factor ("topic") as if each factor combination occurred the same number of times. As can be seen in Tables 8-9, Rouge-2 and Rouge-SU4 both clearly differentiate between human vs. automatic peers.

6 Correlation

It is very important that intrinsic measures of content (such as Rouge, Pyramids, and BEs) correlate with extrinsic ones. Since Responsiveness was the only pseudo-extrinsic evaluation of content this year, we need to measure the correlation of the intrinsic metrics with Responsiveness.

NIST 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 10 shows that there is high correlation between macro-average Rouge scores and average scaled Responsiveness, despite concerns about the low Rouge scores over all summaries. The correlation is high even when the human summaries are ignored (Table 11).

Metric	Spearman	Pearson
Rouge-2	0.95	0.97
Rouge-SU4	0.94	0.96

Table 10: Correlation between average scaled Responsiveness and macro-average Rouge recall (with jackknifing and stemming), all peers

Metric	Spearman	Pearson
Rouge-2	0.90	0.93
Rouge-SU4	0.87	0.92

Table 11: Correlation between average scaled Responsiveness and macro-average Rouge recall (with jackknifing and stemming), only automatic peers

7 Discussion

To be undertaken at the DUC 2005 workshop.

References

Enrique Amigo, Julio Gonzalo, Victor Peinado, Anselmo Peñas, and Felisa Verdejo. 2004. An empirical study of information synthesis task. In *Proceedings of the 42nd Meeting of the Association for Computational*

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 8: 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 9: Multiple comparison of all peers based on ANOVA of Rouge-SU4 recall

Linguistics (ACL'04), Main Volume, pages 207–214, Barcelona, Spain, July.

Donna Harman and Paul Over. 2004. The effects of human variation in duc summarization evaluation. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 10–17, Barcelona, Spain, July. Association for Computational Linguistics.