Looking for a Few Good Metrics: Automatic Summarization Evaluation — How Many Samples Are Enough?

Chin-Yew Lin
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
cyl@isi.edu

Abstract

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. This paper discusses the validity of the evaluation method used in the Document Understanding Conference (DUC) and evaluates five different ROUGE metrics: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU included in the ROUGE summarization evaluation package using data provided by DUC. A comprehensive study of the effects of using single or multiple references and various sample sizes on the stability of the results is also presented.

Keywords: *Summarization, automatic evaluation, Document Understanding Conference, DUC, ROUGE.*

1 Introduction

Large scale evaluations of automatic text summarization such as the Document Understanding Conference (DUC) [8] sponsored by NIST in the United States and the Text Summarization Challenge (TSC) [2] sponsored by the NTCIR Workshop in Japan usually involve expensive human efforts and can only be conducted on a less frequent basis. For example, evaluation is held once per year in DUC and once per one and a half years in TSC. Simple manual evaluation of summaries over a few linguistic quality questions and content coverage as in the Document Understanding Conference (DUC) [8] would require over 3,000 hours of human effort. This human evaluation bottleneck has hindered the advance of the field and researchers have been actively looking for methods to evaluate summaries automatically. For example, Saggion et al. [10] proposed three content-based evaluation methods that measure similarity between summaries. These methods are: *cosine similarity*, *unit overlap* (i.e. unigram or bigram), and *longest common subsequence*. However, they did not show how the results of these automatic evaluation methods correlate to human judgments. Following the successful application of automatic evaluation methods, such as BLEU [9], in machine translation evaluation, Lin and Hovy [5] showed that methods similar to BLEU, i.e. n-gram co-occurrence statistics, could be applied to evaluate summaries. Hori et al. [6] concluded that an automatic metric WSumACCY that rewarded consensus matches performed better and was more stable than two other metrics (SumACCY and BLEU) that did not take advantage of the consensus matches. Their experiments were conducted over speech summaries of 50 utterances in Japanese TV broadcast news with 25 manual summaries for each utterance. Van Halteren and Teufel [11] collected 50 manual summaries of one text and showed that per-summary evaluation based on single reference summary was insufficient because any two randomly chosen summaries from the summary pool were very different. However, stable consensus summary could be obtained if a large number of summaries were considered. Following their work, Nenkova and Passonneau [7] also provided evidence that using multiple reference summaries in multiple document summarization evaluation could reach more stable and robust results by manually evaluating three DUC 2003 topic sets using 10 manual summaries per topic. We can summarize these recent results as follows:

- They used small data sets on single collections;
- They did not provide estimation of the statistical significance of their results; and
- They did not investigate the effect of sample size but focused on the effect of multiple references.

In this paper, we briefly review the DUC evaluation procedure in Section 2, introduce ROUGE, an automatic summary evaluation package, in Section 3, and then summarize its evaluations on single and multiple document summarization tasks over DUC 2001, 2002, and 2003 data in Section 4. In Section 5, we present an in-depth analysis of the effect of sample

size and number of references on the correlation of two ROUGE metrics, ROURGE-1 and ROURGE-SU4, and human assigned mean coverage score using data from DUC 2001 single and multiple document summarization tasks. Section 6 concludes this paper and discusses future directions.

2 Document Understanding Conference

The Document Understanding Conference included the follow tasks:

- Fully automatic single-document summarization: participants were required to create a generic 100-word summary for each document in a set of 30 topics in DUC 2001 and 2002. Single document summarization task was dropped in DUC 2003 and 2004.
- Fully automatic single-document very short summary (headline-like) summarization: participants were required to create a generic 10-word summary for each document in a set of 60 topics in DUC 2003; in DUC 2004, participants were required to create a generic 75-byte summary for each document in a set of 50 topics and each translated document (from Arabic to English, manual or machine translated) in a set of 25 topics.
- Fully automatic multi-document summarization: participants were required to create summaries of 10 (DUC 2002), 50 (DUC 2001 and 2002), 100 (DUC 2001, 2002, and 2003[1]), 200 (DUC 2001 and 2002), and 400 (DUC 2001) words for a set of documents related to a topic, for example, *Hurricane Andrew* or *Mad Cow Disease*. In DUC 2004, the requirement was changed to 665 bytes. There were 30 topics in DUC 2001 and 2003, 60 topics in DUC 2002, and 50 topics in DUC 2004.

For each document or document set (per topic), several human summaries (or reference summaries) were created as the 'ideal' model summaries at each specified length. We will refer to each document in the single document summarization task and each document set in the multi-document summarization task as a "*sample point*" from now on. Three references were created by NIST assessors for each sample point in DUC 2001, two in DUC 2002, four in DUC 2003, and 4 in DUC 2004.

To evaluate system performance, NIST assessors who created the 'ideal' written summaries did pairwise comparisons of their summaries to system-generated summaries, other assessors' summaries, and baseline summaries. They used the Summary Evaluation Environment[2] (SEE) to support the process. Using SEE, the assessors compared the system's text (the *peer* text) to the ideal (the *model* text). Each text was decomposed into a list of units (sentences or elementary discourse unit (EDU) and displayed in separate windows. SEE provides interfaces for assessors to judge both the content and the quality of summaries. We are only concerned with the content selection part in this study. To measure content, assessors step through each model unit, mark all system units sharing content with the current model unit, and specify how much of the content of the current model unit[3] expresses the marked system units.

Instead of pure sentence recall score, DUC used mean coverage score *C*. We define it as follows:

$$C = \frac{(\text{Number of MUs marked}) \bullet E}{\text{Total number of MUs in the model summary}} \quad (1)$$

E, the ratio of completeness, ranges from 1 to 0. If we ignore *E* (set it to 1), we obtain simple sentence recall score. We use mean coverage scores derived from human judgments as the references to evaluate various automatic scoring methods in the following sections.

2.1 Is the DUC Evaluation Methodology Sound?

Lin and Hovy [4] investigated the DUC 2001 human assessment data and found that humans agreed with themselves about 82% in 5,921 total judgments on the single document summarization evaluation task when they assigned different ratings to the same peer and model pair coming from different systems, and about 92.4% in 6,963 total judgments on the multi-document summarization task. They cautioned that future evaluation of summarization should take into account this instability of human judgment.

Based on Lin and Hovy's observation, Nenkova and Passonneau [7] criticized the DUC evaluation methodology by showing a scatterplot (Figure 1 in their paper) of human vs. human mean coverage scores using the task 2 multi-document summarization evaluation results of DUC 2003. We recreated the scatterplot with additional markings of individual human summarizer identification (letters A-J) at each data point in Figure 1 and three lines connecting data points belonging to three assessors, H, I, and J respectively. According to Figure 1 (without the addi-

[1] There were three different multi-document summarization tasks in DUC 2003 and 2004. Please see DUC website at http://duc.nist.gov for details. Only the English TDT (Topic Detection and Tracking) event cluster summarization task, i.e. task 2 in DUC 2003 and 2004 was used in this study.

[2] SEE is free for research purposes and can be downloaded from: http://www.isi.edu/~cyl/SEE.

[3] Categorical ratings: *all*, *most*, *some*, *hardly any*, and *none* were used in DUC 2001. These were converted to 5 points scale from 4 to 0 and normalized to numbers between 1 and 0 in this study. Direct numerical ratings: 0%, 20%, 40%, 60%, 80%, and 100% were used in subsequent DUCs.

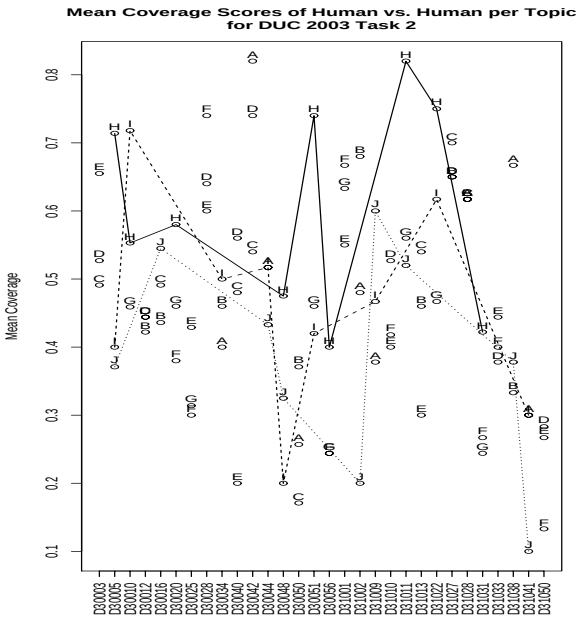

Figure 1. Scatterplot for DUC 2003 mean coverage score of human summaries for different topics. The lines connect human summarizers H, I, and J where they contributed summaries.

tional assessor identifications and the three lines), Nenkova and Passonneau made the following observation: "*an apparently random relation of summarizers to each other, and to document sets*." They further made two conclusions based on this observation: (1) "*DUC scores cannot be used to distinguish a good human summarizer from a bad one*" and (2) "*The DUC method is not powerful enough to distinguish between systems*". However, we found that we could not make the same observation and conclusions as they did according to Figure 1 for the following reasons:

(1) DUC reference summaries were not created for assessing summarization performance of a particular human in the reference pool but for evaluating systems or other human summarizers who in principle should create as many summaries as other humans or systems participating an evaluation. For example, there should be 30 multi-document summaries from each human summarizer whom we want to evaluate if the DUC 2003 data set is used. Figure 1 clearly shows that this is not the case. Usually different combinations of three human summarizers contributed summaries for each topic and no single human summarizer wrote summaries for all topics. Therefore, simple averages of human summarizer scores across topics were not comparable.

(2) The seemingly random relation of summarizers to each other and to document sets actually demonstrates the sound foundation of the DUC evaluation method because mixing different summarizers over different sets of topics prevent DUC evaluation results being biased to a particular human summarizer. Therefore, the seemingly randomness is due to well considered evaluation design not negligence.

(3) With further investigation of Figure 1 following the three lines that connect three different summarizers originating from topic D30005 (the second topic from the left), we found that summarizer H was better than J in 3 out of 3 topics, i.e. D30005, D30048, and D31013, when they co-contributed to a topic. H was also the best summarizer in 8 out of the 9 topics that H contributed. The relative performance of I and J was not clear. I was better than J in 3 (D30005, D30044, and D31041) out of 5 topics (+ D30048 and D31009) where I and J co-contributed. Therefore, the DUC method was able to identify a good summarizer, H, from a bad one, J, showing that H was not only a good one but a very good one. This also indicates that the DUC method is a reasonable approach and we should have confidence in the evaluations based on this method. However, we also need to pay attention to estimation errors to claim significance of the results. For example, do the facts that H beat J three out of three times or H won eight out of nine times mean anything significant? We can only answer this question after rigorous statistical analysis.

(4) We could not make any conclusion about the effectiveness of the DUC method in distinguishing systems based on Figure 1. We could only make the conclusion that a set of documents could have multiple, very different, equally valid summaries. And in this we agree with Nenkova and Passonneau [7]. Based on this observation, we ask the following questions: (a) Can we obtain stable evaluation results despite using only a single reference summary per sample point as we did in DUC? (b) If the answer to (a) is yes, then how is the stability of evaluation affected by sample size? (c) Will inclusion of multiple summaries make the evaluation results more or less stable? (d) How can multiple references be used in improving the stability of evaluations?

For questions (c) and (d), Hori et al. [6] demonstrated that using many references could be counterproductive if the evaluation metric adopted did not take advantage of the consensus among multiple summaries; while a metric utilizing the consensus would stabilize eventually and perform better than using just a few references. This was independently confirmed by Nenkova and Passonneau [7] on a small scale human-vs.-human experiment using their *pyramid* method. Results reported by Lin [3] also indicated that using multiple references tend to increase evaluation stability although human judgments only referred to single reference summary.

For question (a), Lin and Hovy [4, 5] showed that stable evaluation results could be obtained but the variability of human judgments and evaluation metrics must be factored in. This was followed extensively in [3] where bootstrap resampling method [1] was used to estimate confidence intervals (i.e. reliability) of the evaluation results in all experiments in

that paper and was implemented in the publicly available summary evaluation package ROUGE. In the remaining of this paper, we focused on the remaining question (b). Before we detail our experiments in answering this question, we provide a brief overview of metrics included in ROUGE in the next section and summarize their evaluations as described in [3] in Section 4.

3 ROUGE: a Package for Automatic Evaluation of Summaries

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. We summarize each metric in the following sections.

3.1 ROUGE-N: N-gram Co-Occurrence Statistics

Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

$$\frac{\sum_{S \in \{Referemce\,Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\,Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. Note that the number of n-grams in the denominator of the ROUGE-N formula increases as we add more references. Therefore multiple references can be easily integrated into this metric. Also note that the numerator sums over all reference summaries. This effectively gives more weight to matching n-grams occurring in multiple references. Therefore a candidate summary that contains words shared by more references is favored by the ROUGE-N measure.

3.2 ROUGE-L: Longest Common Subsequence

Given two sequences X and Y, the longest common subsequence (LCS) of X and Y is a common subsequence with maximum length. Saggion et al. [10] used normalized pairwise LCS to compare similarity between two texts in automatic summarization evaluation. To apply LCS in summarization evaluation, we view a summary sentence as a sequence of words and the LCS-based metric, ROUGE-L, computes the ratio between the length of the two summaries' LCS and the length of the reference summary. One advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order as n-grams. The other advantage is that it automatically includes longest in-sequence common n-grams, therefore no predefined n-gram length is necessary.

3.3 ROUGE-W: Weighted Longest Common Subsequence

The basic LCS also has a problem that it does not differentiate LCSes of different spatial relations within their embedding sequences. To improve the basic LCS method, we introduce another metric called ROUGE-W or weighted longest common subsequence that favors LCS with consecutive matches. ROUGE-W can be computed efficiently using dynamic programming [3].

3.4 ROUGE-S: Skip-Bigram Co-Occurrence Statistics

Skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics, ROUGE-S, measure the overlap ratio of skip-bigrams between a candidate summary and a set of reference summaries. For example, sentence "*police killed the gunman*" has $C(4,2)$[4] = 6 skip-bigrams:

("*police killed*", "*police the*", "*police gunman*", "*killed the*", "*killed gunman*", "*the gunman*")

Comparing skip-bigram with LCS, skip-bigram counts all in-order matching word pairs while LCS only counts one longest common subsequence. To reduce spurious matches such as "*the the*" or "*of in*", we can limit the maximum skip distance, d_{skip}, between two in-order words that is allowed to form a skip-bigram. ROUGE-S with maximum skip distance of N is called ROUGE-SN.

3.5 ROUGE-SU: Extension of ROUGE-S

One potential problem for ROUGE-S is that it does not give any credit to a candidate sentence if the sentence does not have any word pair co-occurring with its references. To accommodate this, we extend ROUGE-S with the addition of unigram as counting unit. The extended version is called ROUGE-SU. We presented the evaluations of variants of these ROUGE metrics in the next section using three years' DUC data.

4 Evaluation of ROUGE

To assess the effectiveness of ROUGE measures, we compute the correlation between ROUGE assigned summary scores and human assigned mean coverage scores. The intuition is that a good evaluation measure should assign a good score to a good summary and a bad score to a bad summary. The ground truth is based on human assigned scores. Acquiring human judgments are usually very expensive; fortunately,

[4] Combination: $C(4,2) = 4!/(2!*2!) = 6$.

	DUC 2001 100 WORDS SINGLE DOC						DUC 2002 100 WORDS SINGLE DOC					
	1 REF			3 REFS			1 REF			2 REFS		
Method	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP
R-1	0.76	0.76	0.84	0.80	0.78	0.84	0.98	0.98	**0.99**	0.98	0.98	**0.99**
R-2	0.84	0.84	0.83	**0.87**	0.87	0.86	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-3	0.82	0.83	0.80	0.86	0.86	0.85	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-4	0.81	0.81	0.77	0.84	0.84	0.83	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**
R-5	0.79	0.79	0.75	0.83	0.83	0.81	**0.99**	**0.99**	0.98	**0.99**	**0.99**	0.98
R-6	0.76	0.77	0.71	0.81	0.81	0.79	0.98	**0.99**	0.97	**0.99**	**0.99**	0.98
R-7	0.73	0.74	0.65	0.79	0.80	0.76	0.98	0.98	0.97	**0.99**	**0.99**	0.97
R-8	0.69	0.71	0.61	0.78	0.78	0.72	0.98	0.98	0.96	**0.99**	**0.99**	0.97
R-9	0.65	0.67	0.59	0.76	0.76	0.69	0.97	0.97	0.95	0.98	0.98	0.96
R-L	0.83	0.83	0.83	0.86	0.86	0.86	**0.99**	**0.99**	0.99	**0.99**	**0.99**	**0.99**
R-S*	0.74	0.74	0.80	0.78	0.77	0.82	0.98	0.98	0.98	0.98	0.97	0.98
R-S4	0.84	**0.85**	0.84	**0.87**	**0.88**	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-S9	0.84	**0.85**	0.84	**0.87**	**0.88**	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-SU*	0.74	0.74	0.81	0.78	0.77	0.83	0.98	0.98	0.98	0.98	0.98	0.98
R-SU4	0.84	0.84	**0.85**	**0.87**	0.87	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-SU9	0.84	0.84	**0.85**	**0.87**	0.87	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
R-W-1.2	**0.85**	**0.85**	**0.85**	**0.87**	0.87	**0.87**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**

Table 1. Pearson's correlations of 17 ROUGE measure scores vs. human judgments for the DUC 2001 and 2002 100 words single document summarization tasks.

	DUC 2003 10 WORDS SINGLE DOC					
	1 REF	4REFS	1 REF	4 REFS	1 REF	4 REFS
Method	CASE		STEM		STOP	
R-1	0.96	0.95	0.95	0.95	0.90	0.90
R-2	0.75	0.76	0.75	0.75	0.76	0.77
R-3	0.71	0.70	0.70	0.68	0.73	0.70
R-4	0.64	0.65	0.62	0.63	0.69	0.66
R-5	0.62	0.64	0.60	0.63	0.63	0.60
R-6	0.57	0.62	0.55	0.61	0.46	0.54
R-7	0.56	0.56	0.58	0.60	0.46	0.44
R-8	0.55	0.53	0.54	0.55	0.00	0.24
R-9	0.51	0.47	0.51	0.49	0.00	0.14
R-L	**0.97**	**0.96**	**0.97**	**0.96**	0.97	0.96
R-S*	0.89	0.87	0.88	0.85	0.95	0.92
R-S4	0.88	0.89	0.88	0.88	0.95	0.96
R-S9	0.92	0.92	0.92	0.91	0.97	0.95
R-SU*	0.93	0.90	0.91	0.89	0.96	0.94
R-SU4	**0.97**	**0.96**	0.96	0.95	**0.98**	**0.97**
R-SU9	**0.97**	0.95	0.96	0.94	0.97	0.95
R-W-1.2	0.96	**0.96**	0.96	**0.96**	0.96	0.96

Table 2. Pearson's correlations of 17 ROUGE measure scores vs. human judgments for the DUC 2003 very short summary task.

	(A1) DUC 2001 100 WORDS MULTI						(A2) DUC 2002 100 WORDS MULTI						(A3) DUC 2003 100 WORDS MULTI					
	1 RFF			3 REFS			1 REF			2 REFS			1 REF			4 REFS		
Method	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP
R-1	0.48	0.56	**0.86**	0.53	0.57	**0.87**	0.66	0.66	0.77	0.71	0.71	0.78	0.58	0.57	0.71	0.58	0.57	0.71
R-2	0.55	**0.57**	0.64	0.59	**0.61**	0.71	0.83	0.83	**0.80**	0.88	0.87	**0.85**	**0.69**	**0.67**	0.71	**0.79**	**0.79**	**0.81**
R-3	0.46	0.45	0.47	0.53	0.53	0.55	**0.85**	**0.84**	0.76	**0.89**	**0.88**	0.83	0.54	0.51	0.48	0.76	0.75	0.74
R-4	0.39	0.39	0.43	0.48	0.49	0.47	0.80	0.80	0.63	0.83	0.82	0.75	0.37	0.36	0.36	0.62	0.61	0.52
R-5	0.38	0.39	0.33	0.47	0.48	0.43	0.73	0.73	0.45	0.73	0.73	0.62	0.25	0.25	0.27	0.45	0.44	0.38
R-6	0.39	0.39	0.20	0.45	0.46	0.39	0.71	0.72	0.38	0.66	0.64	0.46	0.21	0.21	0.26	0.34	0.31	0.29
R-7	0.31	0.31	0.17	0.44	0.44	0.36	0.63	0.65	0.33	0.56	0.53	0.44	0.20	0.20	0.23	0.29	0.27	0.25
R-8	0.18	0.19	0.09	0.40	0.40	0.31	0.55	0.55	0.52	0.50	0.46	0.52	0.18	0.18	0.21	0.23	0.22	0.23
R-9	0.11	0.12	0.06	0.38	0.38	0.28	0.54	0.54	0.52	0.45	0.42	0.52	0.16	0.16	0.19	0.21	0.21	0.21
R-L	**0.49**	0.49	0.49	0.56	0.56	0.56	0.62	0.62	0.62	0.65	0.65	0.65	0.50	0.50	0.50	0.53	0.53	0.53
R-S*	0.45	0.52	0.84	0.51	0.54	0.86	0.69	0.69	0.77	0.73	0.73	0.79	0.60	0.60	0.67	0.61	0.60	0.70
R-S4	0.46	0.50	0.71	0.54	0.57	0.78	0.79	0.80	0.79	0.84	0.85	0.82	0.63	0.64	0.70	0.73	0.73	0.78
R-S9	0.42	0.49	0.77	0.53	0.56	0.81	0.79	0.80	0.78	0.83	0.84	0.81	0.65	0.65	0.70	0.70	0.70	0.76
R-SU*	0.45	0.52	0.84	0.51	0.54	**0.87**	0.69	0.69	0.77	0.73	0.73	0.79	0.60	0.59	0.67	0.60	0.60	0.70
R-SU4	0.47	0.53	0.80	0.55	0.58	0.83	0.76	0.76	0.79	0.80	0.81	0.81	0.64	0.64	**0.74**	0.68	0.68	0.76
R-SU9	0.44	0.50	0.80	0.53	0.57	0.84	0.77	0.78	0.78	0.81	0.82	0.81	0.65	0.65	0.72	0.68	0.68	0.75
R-W-1.2	0.52	0.52	0.52	**0.60**	0.60	0.60	0.67	0.67	0.67	0.69	0.69	0.69	0.53	0.53	0.53	0.58	0.58	0.58

	(C) DUC02 10			(D1) DUC01 50			(D2) DUC02 50			(E1) DUC01 200			(E2) DUC02 200			(F) DUC01 400		
Method	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP
R-1	0.71	0.68	0.49	0.49	0.49	0.73	0.44	0.48	**0.80**	0.81	0.81	**0.90**	0.84	0.84	0.91	0.74	0.73	**0.90**
R-2	0.82	0.85	**0.80**	0.43	0.45	0.59	0.47	0.49	0.62	**0.84**	**0.85**	0.86	**0.93**	**0.93**	**0.94**	**0.88**	**0.88**	0.87
R-3	0.59	0.74	0.75	0.32	0.33	0.39	0.36	0.36	0.45	0.80	0.80	0.81	0.90	0.91	0.91	0.84	0.84	0.82
R-4	0.25	0.36	0.16	0.28	0.26	0.36	0.28	0.28	0.39	0.77	0.78	0.78	0.87	0.88	0.88	0.80	0.80	0.75
R-5	-0.25	-0.25	-0.24	0.30	0.29	0.31	0.28	0.30	0.49	0.77	0.76	0.72	0.82	0.83	0.84	0.77	0.77	0.70
R-6	0.00	0.00	0.00	0.22	0.23	0.41	0.18	0.21	-0.17	0.75	0.75	0.67	0.78	0.79	0.77	0.74	0.74	0.63
R-7	0.00	0.00	0.00	0.26	0.23	0.50	0.11	0.16	0.00	0.72	0.72	0.62	0.72	0.73	0.74	0.70	0.70	0.58
R-8	0.00	0.00	0.00	0.32	0.32	0.34	-0.11	-0.11	0.00	0.68	0.68	0.54	0.71	0.71	0.70	0.66	0.66	0.52
R-9	0.00	0.00	0.00	0.30	0.30	0.34	-0.14	-0.14	0.00	0.64	0.64	0.48	0.70	0.69	0.59	0.63	0.62	0.46
R-L	0.78	0.78	0.78	**0.56**	**0.56**	0.56	0.50	0.50	0.50	0.81	0.81	0.81	0.88	0.88	0.88	0.82	0.82	0.82
R-S*	0.83	0.82	0.69	0.46	0.45	**0.74**	0.46	0.49	**0.80**	0.80	0.80	**0.90**	0.84	0.85	0.93	0.75	0.74	0.89
R-S4	**0.85**	**0.86**	0.76	0.40	0.41	0.69	0.42	0.44	0.73	0.82	0.82	0.87	0.91	0.91	0.93	0.85	0.85	0.85
R-S9	0.82	0.81	0.69	0.42	0.41	0.72	0.40	0.43	0.78	0.81	0.82	0.86	0.90	0.90	0.92	0.83	0.83	0.84
R-SU*	0.75	0.74	0.56	0.46	0.46	**0.74**	0.46	0.49	**0.80**	0.80	0.80	**0.90**	0.84	0.85	0.93	0.75	0.74	0.89
R-SU4	0.76	0.75	0.58	0.45	0.45	0.72	0.44	0.46	0.78	0.82	0.83	0.89	0.90	0.90	0.93	0.84	0.84	0.88
R-SU9	0.74	0.73	0.56	0.44	0.44	0.73	0.41	0.45	0.79	0.82	0.82	0.88	0.89	0.89	0.92	0.83	0.82	0.87
R-W-1.2	0.78	0.78	0.78	**0.56**	**0.56**	0.56	**0.51**	**0.51**	0.51	**0.84**	0.84	0.84	0.90	0.90	0.90	0.86	0.86	0.86

Table 3. Pearson's correlations of 17 ROUGE measure scores vs. human judgments for the DUC 2001, 2002, and 2003 multi-document summarization tasks.

we have DUC 2001, 2002, and 2003 evaluation data that include human judgments as described earlier in Section 2. With the DUC data, we computed Pearson's product moment correlation coefficients between systems' average ROUGE scores and their human assigned mean coverage scores using single reference and multiple references. To investigate the effect of stemming and inclusion or exclusion of stopwords, we also ran experiments over original automatic and manual summaries (CASE set), stemmed[5] version of the summaries (STEM set), and stopped version of the summaries (STOP set). To assess the significance of the results, we applied bootstrap resampling technique [1] to estimate 95% confidence intervals for every correlation computation.

17 ROUGE measures were tested for each run: ROUGE-N with N = 1 to 9, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU with maximum skip distance d_{skip} = 1, 4, and 9. Table 1 shows the Pearson's correlation coefficients of the 17 ROUGE measures vs. human judgments on DUC 2001 and 2002 100 words single document summarization data. The best values in each column are marked with dark (green) color and statistically equivalent values to the best values are marked with gray. We found that correlations were not affected by stemming or removal of stopwords in this data set, ROUGE-2 performed better among the ROUGE-N variants, ROUGE-L, ROUGE-W, and ROUGE-S were all performing well, and using multiple references improved performance though not much. All ROUGE measures achieved very good correlation with human judgments in the DUC 2002 data. This might be due to the double sample size in

[5] Porter's stemmer was used.

Sample Size			1		6		11		16		21		26		31		36		41		46		51		56		61		66	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Single Document	Pearson	1 REF	0.02	1.20	0.19	1.13	0.33	1.04	0.43	0.92	0.49	0.80	0.55	0.78	0.60	0.69	0.63	0.62	0.66	0.56	0.68	0.55	0.70	0.55	0.71	0.53	0.72	0.50	0.74	0.47
		2 REF	0.05	1.19	0.25	1.09	0.39	0.99	0.48	0.83	0.54	0.76	0.60	0.66	0.64	0.61	0.67	0.55	0.70	0.53	0.72	0.48	0.74	0.45	0.75	0.46	0.76	0.42	0.77	0.39
		3 REF	0.04	1.14	0.24	1.10	0.38	0.98	0.48	0.82	0.54	0.73	0.59	0.66	0.63	0.60	0.66	0.55	0.69	0.52	0.71	0.50	0.72	0.47	0.74	0.45	0.75	0.42	0.76	0.40
	Spearman	1 REF	0.01	1.26	0.19	1.15	0.32	1.10	0.42	0.96	0.49	0.85	0.55	0.79	0.59	0.71	0.63	0.65	0.66	0.61	0.68	0.58	0.70	0.56	0.72	0.54	0.73	0.52	0.74	0.49
		2 REF	0.04	1.20	0.24	1.09	0.38	1.00	0.47	0.85	0.54	0.82	0.59	0.72	0.64	0.64	0.67	0.60	0.70	0.56	0.72	0.53	0.74	0.53	0.75	0.53	0.76	0.49	0.77	0.46
		3 REF	0.03	1.25	0.24	1.11	0.37	0.99	0.47	0.83	0.53	0.77	0.58	0.69	0.63	0.61	0.66	0.60	0.69	0.55	0.71	0.52	0.72	0.51	0.74	0.53	0.75	0.44	0.76	0.44
	C (R)	X REF	0.345	0.121	0.342	0.054	0.342	0.041	0.342	0.033	0.344	0.028	0.344	0.025	0.344	0.023	0.344	0.022	0.344	0.020	0.344	0.019	0.344	0.018	0.344	0.018	0.344	0.017	0.344	0.016
	R-1 (R)	1 REF	0.401	0.557	0.404	0.234	0.404	0.175	0.402	0.141	0.403	0.120	0.403	0.107	0.403	0.101	0.403	0.094	0.403	0.087	0.403	0.083	0.403	0.078	0.403	0.074	0.403	0.074	0.403	0.070
		2 REF	0.403	0.546	0.404	0.219	0.404	0.164	0.402	0.134	0.404	0.116	0.403	0.099	0.404	0.092	0.404	0.092	0.404	0.085	0.404	0.081	0.404	0.076	0.404	0.074	0.404	0.070	0.404	0.066
		3 REF	0.391	0.476	0.391	0.201	0.392	0.151	0.391	0.122	0.392	0.104	0.392	0.093	0.392	0.085	0.392	0.082	0.392	0.077	0.392	0.073	0.392	0.071	0.392	0.066	0.392	0.060	0.392	0.059

Sample Size			71		76		81		86		91		96		101		106		111		116		121		126		131		136	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Single Document	Pearson	1 REF	0.75	0.44	0.76	0.40	0.77	0.39	0.78	0.39	0.79	0.36	0.80	0.36	0.80	0.36	0.81	0.35	0.81	0.33	0.82	0.32	0.82	0.32	0.83	0.30	0.83	0.30	0.83	0.28
		2 REF	0.78	0.39	0.79	0.35	0.80	0.35	0.81	0.34	0.82	0.33	0.82	0.31	0.83	0.31	0.83	0.30	0.84	0.29	0.84	0.28	0.85	0.26	0.85	0.25	0.85	0.26	0.86	0.25
		3 REF	0.77	0.38	0.78	0.36	0.79	0.34	0.79	0.33	0.80	0.32	0.81	0.31	0.81	0.30	0.82	0.31	0.82	0.30	0.83	0.28	0.83	0.27	0.83	0.26	0.84	0.26	0.84	0.25
	Spearman	1 REF	0.75	0.47	0.77	0.44	0.77	0.43	0.79	0.42	0.79	0.40	0.80	0.39	0.80	0.38	0.81	0.38	0.82	0.35	0.82	0.36	0.83	0.34	0.83	0.33	0.84	0.32	0.84	0.31
		2 REF	0.79	0.42	0.80	0.38	0.80	0.38	0.81	0.37	0.82	0.35	0.83	0.36	0.83	0.33	0.84	0.32	0.85	0.33	0.85	0.31	0.85	0.32	0.86	0.30	0.86	0.29	0.87	0.27
		3 REF	0.77	0.41	0.78	0.39	0.79	0.41	0.80	0.38	0.80	0.36	0.81	0.35	0.82	0.35	0.82	0.35	0.83	0.34	0.83	0.34	0.84	0.32	0.84	0.32	0.84	0.31	0.85	0.30
	C (R)	X REF	0.344	0.016	0.344	0.016	0.344	0.016	0.344	0.015	0.344	0.015	0.344	0.014	0.344	0.013	0.344	0.013	0.344	0.013	0.344	0.013	0.344	0.012	0.344	0.012	0.344	0.012	0.344	0.011
	R-1 (R)	1 REF	0.403	0.068	0.403	0.065	0.403	0.065	0.403	0.062	0.403	0.061	0.403	0.060	0.403	0.059	0.403	0.055	0.403	0.055	0.403	0.054	0.403	0.053	0.403	0.050	0.403	0.049	0.403	0.048
		2 REF	0.404	0.063	0.404	0.065	0.404	0.063	0.404	0.061	0.404	0.058	0.404	0.057	0.404	0.056	0.404	0.054	0.404	0.054	0.404	0.054	0.404	0.053	0.404	0.049	0.404	0.048	0.404	0.047
		3 REF	0.392	0.057	0.392	0.058	0.392	0.057	0.392	0.053	0.392	0.052	0.392	0.052	0.392	0.050	0.392	0.049	0.392	0.048	0.392	0.047	0.392	0.045	0.392	0.044	0.392	0.043	0.392	0.042

Table 3. Pearson's and Spearman's correlation coefficients of ROUGE-1 (R-1) vs. mean coverage (C) and system R's scores of these two metrics over different sample sizes on DUC 2001 single document summarization task.

Sample Size			1		2		3		4		5		6		7		8		9		10		11		12		13	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Multi Document	Pearson	1 REF	0.24	1.07	0.39	0.86	0.49	0.76	0.56	0.67	0.61	0.60	0.65	0.58	0.68	0.52	0.71	0.49	0.73	0.46	0.75	0.42	0.76	0.40	0.78	0.38	0.78	0.36
		2 REF	0.23	1.01	0.38	0.83	0.48	0.74	0.55	0.67	0.60	0.60	0.64	0.54	0.68	0.49	0.70	0.46	0.72	0.46	0.74	0.40	0.76	0.39	0.77	0.38	0.78	0.37
		3 REF	0.26	1.01	0.41	0.80	0.52	0.75	0.59	0.66	0.64	0.59	0.67	0.52	0.71	0.47	0.73	0.45	0.75	0.43	0.77	0.40	0.78	0.37	0.79	0.35	0.80	0.33
	Spearman	1 REF	0.24	1.11	0.40	0.90	0.50	0.79	0.57	0.74	0.62	0.63	0.65	0.60	0.69	0.56	0.72	0.51	0.74	0.49	0.75	0.45	0.77	0.43	0.78	0.43	0.79	0.41
		2 REF	0.25	1.06	0.39	0.95	0.51	0.79	0.57	0.72	0.63	0.62	0.66	0.58	0.70	0.53	0.72	0.48	0.74	0.48	0.76	0.42	0.77	0.40	0.79	0.38	0.79	0.37
		3 REF	0.29	1.23	0.42	0.94	0.54	0.79	0.60	0.71	0.66	0.60	0.69	0.53	0.72	0.50	0.75	0.46	0.77	0.42	0.78	0.40	0.80	0.38	0.81	0.35	0.81	0.34
	C (T)	X REF	0.187	0.113	0.187	0.092	0.187	0.077	0.186	0.066	0.187	0.060	0.187	0.056	0.187	0.055	0.187	0.048	0.187	0.046	0.187	0.045	0.187	0.041	0.187	0.039	0.179	0.038
	R-1 (T)	1 REF	0.249	0.203	0.249	0.153	0.249	0.122	0.250	0.105	0.249	0.092	0.249	0.088	0.249	0.083	0.249	0.077	0.249	0.073	0.249	0.069	0.249	0.067	0.249	0.063	0.249	0.060
		2 REF	0.239	0.276	0.239	0.173	0.239	0.131	0.239	0.114	0.239	0.103	0.239	0.096	0.239	0.086	0.239	0.082	0.239	0.077	0.239	0.071	0.239	0.068	0.239	0.065	0.238	0.065
		3 REF	0.238	0.274	0.238	0.166	0.238	0.132	0.238	0.116	0.238	0.104	0.238	0.097	0.237	0.092	0.237	0.085	0.237	0.078	0.237	0.076	0.237	0.072	0.237	0.070	0.237	0.067

Sample Size			14		15		16		17		18		19		20		21		22		23		24		25		26	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Multi Document	Pearson	1 REF	0.79	0.37	0.80	0.34	0.81	0.32	0.82	0.32	0.82	0.29	0.83	0.29	0.83	0.28	0.84	0.26	0.84	0.26	0.85	0.25	0.85	0.24	0.85	0.24	0.86	0.22
		2 REF	0.79	0.35	0.80	0.35	0.80	0.34	0.81	0.31	0.82	0.29	0.82	0.28	0.83	0.28	0.83	0.27	0.84	0.26	0.84	0.25	0.85	0.23	0.85	0.22	0.86	0.22
		3 REF	0.81	0.33	0.82	0.32	0.83	0.31	0.83	0.28	0.84	0.26	0.84	0.26	0.85	0.26	0.85	0.25	0.86	0.23	0.86	0.22	0.87	0.21	0.87	0.20	0.87	0.20
	Spearman	1 REF	0.80	0.38	0.81	0.36	0.81	0.37	0.82	0.35	0.83	0.32	0.83	0.33	0.84	0.32	0.84	0.29	0.85	0.30	0.85	0.28	0.85	0.27	0.86	0.26	0.86	0.25
		2 REF	0.80	0.35	0.81	0.34	0.82	0.32	0.82	0.31	0.83	0.29	0.84	0.28	0.84	0.29	0.85	0.27	0.85	0.26	0.86	0.26	0.86	0.24	0.87	0.23	0.87	0.22
		3 REF	0.82	0.36	0.83	0.33	0.84	0.30	0.84	0.31	0.85	0.29	0.85	0.28	0.86	0.26	0.87	0.25	0.87	0.25	0.88	0.23	0.88	0.22	0.88	0.21	0.89	0.20
	C (T)	X REF	0.180	0.037	0.180	0.035	0.180	0.033	0.181	0.032	0.181	0.033	0.181	0.031	0.182	0.031	0.183	0.030	0.183	0.030	0.183	0.029	0.183	0.028	0.183	0.028	0.183	0.028
	R-1 (T)	1 REF	0.249	0.058	0.249	0.056	0.249	0.056	0.249	0.053	0.249	0.051	0.249	0.048	0.249	0.048	0.249	0.048	0.249	0.047	0.249	0.045	0.249	0.044	0.249	0.043	0.249	0.044
		2 REF	0.238	0.061	0.238	0.057	0.238	0.056	0.238	0.054	0.238	0.054	0.238	0.052	0.238	0.051	0.238	0.052	0.238	0.050	0.238	0.048	0.238	0.045	0.238	0.045	0.238	0.045
		3 REF	0.237	0.064	0.237	0.059	0.237	0.059	0.237	0.057	0.237	0.056	0.238	0.053	0.237	0.054	0.238	0.053	0.237	0.052	0.238	0.050	0.237	0.047	0.237	0.047	0.237	0.045

Table 4. Pearson's and Spearman's correlation coefficients of ROUGE-1 (R-1) vs. mean coverage (C) and system T's scores of these two metrics over different sample sizes on DUC 2001 multiple document summarization task.

DUC 2002 (295 vs. 149 in DUC 2001) for each system.

Table 2 shows the correlation analysis results on the DUC 2003 single document very short summary data. We found that ROUGE-1, ROUGE-L, ROUGE-SU4 and 9, and ROUGE-W were very good measures in this category, ROUGE-N with N > 1 performed significantly worse than all other measures, and exclusion of stopwords improved performance in general except for ROUGE-1. Due to the large number of samples (624) in this data set, using multiple references did not improve correlations.

In Table 3 columns A1, A2, and A3, we show correlation analysis results on DUC 2001, 2002, and 2003 100 words multi-document summarization data. The results indicated that using multiple references improved correlation and exclusion of stopwords usually improved performance. ROUGE-1, 2, and 3 performed fine but were not consistent. ROUGE-1, ROUGE-S4, ROUGE-SU4, ROUGE-S9, and ROUGE-SU9 with stopword removal had correlation above 0.70. ROUGE-L and ROUGE-W did not work well in this set of data.

Table 3 columns C, D1, D2, E1, E2, and F show the correlation analyses using multiple references on the rest of DUC data. These results again suggested that exclusion of stopwords achieved better performance especially in multi-document summaries of 50 words. Better correlations (> 0.70) were observed on long summary tasks, i.e. 200 and 400 words summaries. The relative performance of ROUGE measures followed the pattern of the 100 words multi-document summarization task.

Comparing the results in Table 3 with Tables 1 and 2, we found that correlation values in the multi-document tasks rarely reached high 90% except in long summary tasks. One possible explanation of this outcome is that we did not have large amount of samples for the multi-document tasks. In the single document summarization tasks we had over 100 samples; while we only had about 30 samples in the multi-document tasks. The only tasks that had over 30 samples was from DUC 2002; and here as expected the correlations of ROUGE measures with human judgments on the 100 words summary task were much better and more stable than similar tasks in DUC 2001 and 2003. Statistically stable human judgments of system performance might not be obtained due to lack of samples and this in turn caused instability of correlation analyses.

5 The Effect of Multiple References and Different Sample Sizes

Sample Size			1		6		11		16		21		26		31		36		41		46		51		56		61		66	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Single Document	Pearson	1 REF	0.01	1.20	0.17	1.20	0.30	1.08	0.39	0.96	0.46	0.84	0.53	0.81	0.58	0.74	0.61	0.65	0.64	0.62	0.67	0.58	0.69	0.56	0.70	0.53	0.72	0.51	0.73	0.48
		2 REF	0.04	1.17	0.24	1.07	0.39	1.03	0.48	0.84	0.54	0.78	0.59	0.69	0.64	0.64	0.67	0.57	0.70	0.51	0.72	0.50	0.74	0.47	0.75	0.45	0.76	0.43	0.78	0.40
		3 REF	0.04	1.17	0.24	1.09	0.39	1.00	0.48	0.84	0.54	0.75	0.60	0.66	0.65	0.62	0.68	0.54	0.71	0.50	0.73	0.48	0.74	0.45	0.76	0.43	0.77	0.41	0.78	0.38
	Spearman	1 REF	0.01	1.24	0.17	1.20	0.29	1.09	0.39	0.98	0.46	0.90	0.52	0.80	0.57	0.77	0.61	0.69	0.64	0.66	0.67	0.57	0.68	0.58	0.70	0.57	0.71	0.55	0.73	0.52
		2 REF	0.04	1.20	0.23	1.08	0.38	1.04	0.47	0.91	0.53	0.82	0.59	0.75	0.64	0.67	0.67	0.60	0.69	0.56	0.71	0.54	0.73	0.53	0.75	0.50	0.76	0.48	0.77	0.45
		3 REF	0.04	1.19	0.23	1.09	0.38	1.00	0.47	0.88	0.54	0.82	0.59	0.73	0.64	0.63	0.68	0.57	0.70	0.56	0.72	0.52	0.74	0.50	0.75	0.50	0.77	0.45	0.78	0.44
	C (R)	X REF	0.345	0.121	0.342	0.054	0.342	0.041	0.342	0.033	0.344	0.028	0.344	0.025	0.344	0.023	0.344	0.022	0.344	0.020	0.344	0.019	0.344	0.018	0.344	0.018	0.344	0.017	0.344	0.016
	R-SU4 (R)	1 REF	0.200	0.536	0.203	0.214	0.203	0.155	0.201	0.129	0.202	0.111	0.202	0.098	0.202	0.089	0.202	0.084	0.202	0.078	0.202	0.076	0.202	0.072	0.202	0.066	0.202	0.065	0.202	0.062
		2 REF	0.210	0.429	0.209	0.194	0.210	0.144	0.208	0.111	0.209	0.098	0.209	0.083	0.210	0.079	0.209	0.077	0.210	0.073	0.210	0.070	0.210	0.065	0.210	0.062	0.210	0.061	0.210	0.058
		3 REF	0.194	0.400	0.193	0.163	0.194	0.121	0.193	0.092	0.194	0.080	0.194	0.070	0.194	0.067	0.194	0.064	0.194	0.061	0.194	0.058	0.194	0.056	0.194	0.052	0.194	0.050	0.194	0.049

Sample Size			71		76		81		86		91		96		101		106		111		116		121		126		131		136	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Single Document	Pearson	1 REF	0.75	0.44	0.76	0.42	0.77	0.42	0.78	0.39	0.79	0.37	0.80	0.34	0.81	0.33	0.82	0.33	0.82	0.32	0.83	0.30	0.83	0.30	0.84	0.28	0.84	0.28	0.85	0.26
		2 REF	0.79	0.38	0.80	0.36	0.81	0.37	0.81	0.36	0.82	0.33	0.83	0.32	0.83	0.31	0.84	0.30	0.84	0.30	0.85	0.30	0.85	0.29	0.86	0.28	0.86	0.27	0.86	0.25
		3 REF	0.80	0.37	0.81	0.34	0.81	0.34	0.82	0.32	0.83	0.31	0.84	0.30	0.84	0.28	0.85	0.29	0.85	0.28	0.86	0.27	0.86	0.27	0.86	0.26	0.87	0.25	0.87	0.24
	Spearman	1 REF	0.74	0.47	0.76	0.44	0.77	0.43	0.78	0.42	0.79	0.40	0.79	0.39	0.80	0.36	0.81	0.38	0.82	0.35	0.82	0.34	0.83	0.33	0.83	0.34	0.84	0.32	0.84	0.30
		2 REF	0.78	0.43	0.79	0.41	0.80	0.38	0.81	0.38	0.81	0.37	0.82	0.38	0.83	0.33	0.83	0.34	0.84	0.32	0.84	0.31	0.85	0.32	0.85	0.30	0.85	0.31	0.86	0.29
		3 REF	0.79	0.40	0.80	0.39	0.81	0.37	0.82	0.37	0.83	0.34	0.83	0.36	0.84	0.33	0.85	0.31	0.85	0.31	0.86	0.31	0.86	0.30	0.86	0.30	0.87	0.29	0.87	0.27
	C (R)	X REF	0.344	0.016	0.344	0.016	0.344	0.016	0.344	0.015	0.344	0.015	0.344	0.014	0.344	0.013	0.344	0.013	0.344	0.013	0.344	0.013	0.344	0.012	0.344	0.012	0.344	0.012	0.344	0.011
	R-SU4 (R)	1 REF	0.202	0.060	0.202	0.059	0.202	0.057	0.202	0.056	0.202	0.056	0.202	0.054	0.202	0.053	0.202	0.051	0.202	0.050	0.202	0.049	0.202	0.048	0.202	0.047	0.202	0.045	0.202	0.042
		2 REF	0.210	0.056	0.210	0.057	0.210	0.054	0.210	0.052	0.210	0.051	0.210	0.049	0.210	0.049	0.210	0.048	0.210	0.047	0.210	0.046	0.210	0.044	0.210	0.043	0.210	0.041	0.210	0.041
		3 REF	0.194	0.048	0.194	0.047	0.194	0.046	0.194	0.044	0.194	0.042	0.194	0.041	0.194	0.040	0.194	0.039	0.194	0.039	0.194	0.039	0.194	0.038	0.194	0.037	0.194	0.036	0.194	0.035

Table 5. Pearson's and Spearman's correlation coefficients of ROUGE-SU4 (R-SU4) vs. mean coverage (C) and system R's scores of these two metrics over different sample sizes on DUC 2001 single document summarization task.

Sample Size			1		2		3		4		5		6		7		8		9		10		11		12		13	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Multi Document	Pearson	1 REF	0.16	1.27	0.32	0.99	0.42	0.90	0.49	0.80	0.54	0.75	0.58	0.71	0.61	0.68	0.64	0.65	0.66	0.62	0.68	0.57	0.70	0.53	0.71	0.54	0.71	0.54
		2 REF	0.17	1.14	0.30	0.88	0.40	0.85	0.46	0.73	0.51	0.70	0.55	0.63	0.59	0.59	0.62	0.56	0.64	0.54	0.66	0.50	0.67	0.48	0.69	0.47	0.70	0.44
		3 REF	0.21	1.19	0.35	0.90	0.45	0.85	0.52	0.75	0.57	0.67	0.61	0.61	0.65	0.57	0.67	0.54	0.69	0.51	0.71	0.46	0.73	0.45	0.74	0.44	0.75	0.41
	Spearman	1 REF	0.20	1.35	0.35	1.01	0.46	0.88	0.52	0.77	0.57	0.72	0.61	0.65	0.64	0.62	0.66	0.57	0.68	0.56	0.70	0.51	0.71	0.52	0.72	0.50	0.72	0.49
		2 REF	0.20	1.19	0.33	0.93	0.43	0.87	0.49	0.74	0.54	0.66	0.57	0.62	0.61	0.59	0.63	0.54	0.65	0.52	0.67	0.48	0.68	0.50	0.69	0.47	0.70	0.46
		3 REF	0.24	1.26	0.37	0.99	0.47	0.87	0.54	0.76	0.59	0.64	0.62	0.60	0.66	0.57	0.68	0.52	0.70	0.52	0.72	0.47	0.73	0.47	0.74	0.44	0.75	0.44
	C (T)	X REF	0.187	0.113	0.187	0.092	0.187	0.077	0.186	0.066	0.187	0.060	0.187	0.056	0.187	0.055	0.187	0.048	0.187	0.046	0.187	0.045	0.187	0.041	0.187	0.039	0.179	0.038
	R-SU4 (T)	1 REF	0.075	0.130	0.075	0.084	0.075	0.073	0.075	0.062	0.075	0.056	0.075	0.050	0.075	0.045	0.075	0.043	0.075	0.043	0.075	0.039	0.075	0.039	0.075	0.037	0.075	0.036
		2 REF	0.073	0.142	0.073	0.090	0.073	0.068	0.073	0.060	0.073	0.054	0.073	0.051	0.073	0.047	0.073	0.044	0.073	0.040	0.073	0.038	0.073	0.037	0.073	0.036	0.073	0.035
		3 REF	0.074	0.129	0.074	0.089	0.074	0.069	0.074	0.060	0.074	0.052	0.074	0.050	0.074	0.046	0.074	0.043	0.074	0.040	0.074	0.039	0.074	0.038	0.074	0.036	0.074	0.035

Sample Size			14		15		16		17		18		19		20		21		22		23		24		25		26	
			M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI	M	dCI
Multi Document	Pearson	1 REF	0.72	0.50	0.73	0.47	0.74	0.48	0.75	0.47	0.76	0.45	0.76	0.43	0.77	0.43	0.77	0.41	0.78	0.42	0.78	0.41	0.79	0.38	0.79	0.37	0.80	0.36
		2 REF	0.71	0.42	0.72	0.42	0.73	0.41	0.74	0.40	0.75	0.38	0.75	0.36	0.76	0.36	0.76	0.35	0.77	0.34	0.77	0.32	0.78	0.31	0.78	0.31	0.79	0.31
		3 REF	0.76	0.39	0.77	0.36	0.78	0.36	0.79	0.36	0.79	0.34	0.80	0.34	0.80	0.32	0.81	0.30	0.81	0.30	0.82	0.30	0.82	0.28	0.83	0.27	0.83	0.26
	Spearman	1 REF	0.73	0.48	0.74	0.46	0.75	0.45	0.75	0.44	0.76	0.44	0.76	0.44	0.77	0.41	0.77	0.42	0.78	0.40	0.78	0.39	0.78	0.38	0.79	0.38	0.79	0.37
		2 REF	0.71	0.44	0.72	0.41	0.72	0.40	0.73	0.41	0.73	0.39	0.74	0.39	0.75	0.39	0.75	0.37	0.76	0.35	0.76	0.35	0.76	0.32	0.77	0.33	0.77	0.32
		3 REF	0.76	0.41	0.77	0.39	0.77	0.39	0.78	0.38	0.79	0.36	0.79	0.36	0.80	0.35	0.80	0.34	0.81	0.33	0.81	0.32	0.82	0.30	0.82	0.29	0.82	0.30
	C (T)	X REF	0.180	0.037	0.180	0.035	0.180	0.033	0.181	0.032	0.181	0.033	0.181	0.031	0.182	0.031	0.183	0.030	0.183	0.030	0.183	0.029	0.183	0.028	0.183	0.028	0.183	0.028
	R-SU4 (T)	1 REF	0.075	0.035	0.075	0.033	0.075	0.033	0.075	0.032	0.075	0.031	0.075	0.029	0.075	0.028	0.075	0.028	0.075	0.028	0.075	0.027	0.075	0.027	0.075	0.026	0.075	0.026
		2 REF	0.073	0.033	0.073	0.032	0.073	0.031	0.073	0.029	0.073	0.029	0.073	0.028	0.073	0.028	0.073	0.027	0.073	0.027	0.073	0.026	0.073	0.024	0.073	0.024	0.073	0.024
		3 REF	0.074	0.034	0.074	0.032	0.074	0.031	0.074	0.030	0.074	0.029	0.074	0.029	0.074	0.028	0.074	0.028	0.074	0.027	0.074	0.026	0.074	0.025	0.074	0.025	0.074	0.024

Table 6. Pearson's and Spearman's correlation coefficients of ROUGE-SU4 (R-SU4) vs. mean coverage (C) and system T's scores of these two metrics over different sample sizes on DUC 2001 multiple document summarization task.

We showed that ROUGE metrics can be used to evaluate summaries fairly reliably providing that there are enough samples. Although the results indicated that multiple references helped, the dominating factor that affected the stability and reliability of evaluations seems to be the number of samples. To further quantify the sample size effect on evaluation of summarization, we conducted the following experiments: (1) examine the effect of sample size on human assigned mean coverage score by computing mean coverage score at different sample sizes and different number of references for each participating system; (2) examine the effect of sample size on automatic evaluation metrics by computing them at different sample sizes and different number of references; (3) examine the effect of sample size on correlation between mean coverage score and automatic evaluation metrics by using the results from (1) and (2) and computing the Pearson's and Spearman's correlation coefficients between them.

The Pearson's correlation coefficient[6] measures the strength and direction of a *linear* relationship between any two variables, i.e. automatic metric score and human assigned mean coverage score in our case. It ranges from +1 to -1. A correlation of 1 means that there is a perfect positive linear relationship between the two variables, a correlation of -1 means that there is a perfect negative linear relationship between them, and a correlation of 0 means that there is no linear relationship between them. Since we would like to use automatic evaluation metric not only in comparing systems but also in in-house system development, a good linear correlation with human judgment would enable us to use automatic scores to predict corresponding human judgment scores. Therefore, Pearson's correlation coefficient is a good measure to look at.

Spearman's correlation coefficient[7] is also a measure of correlation between two variables. It is a non-parametric measure and is a special case of the Pearson's correlation coefficient when the values of data are converted into ranks before computing the coefficient. Spearman's correlation coefficient does not assume the correlation between the variables is linear. Therefore it is a useful correlation indicator even when good linear correlation, for example, according

[6] For a quick overview of the Pearson's coefficient, see: http://davidmlane.com/hyperstat/A34739.html.

[7] For a quick overview of the Spearman's coefficient, see: http://davidmlane.com/hyperstat/A62436.html.

to Pearson's correlation coefficient between two variables could not be found. It also suits the DUC evaluation scenario where multiple systems are ranked according to some performance metrics.

We used the data from DUC 2001 100 words single and multiple document summarization tasks for these experiments. To ensure reliability of our results, we only used 142 documents from the single document summarization task and 26 topics from the multiple document summarization task because these documents and topics included submissions from all participants and all submissions were judged by NIST assessors. Based on the evaluation results presented in Section 4, we ran our experiments on a stopped version of the DUC 2001 data. The stopped version was obtained by applying Porter's stemmer and excluding stopwords on the DUC 2001 data. However, the evaluation methodology described in this paper can be applied to other versions of the DUC data.

At sample size N, we applied standard bootstrap resampling procedure [1]. We randomly sampled N summaries from each system or human summarizer, calculated the mean of the sample, put these N summaries back to the summary pool, and then repeated this procedure 1,000 times. We then cut the lower and upper 2.5% of the 1,000 samples to get the 95% confidence interval of the mean. The Pearson's correlation coefficient was calculated in a similar way. Instead of computing the mean of a sample, we computed the Pearson's correlation of a set of system means vs. their mean coverage scores 1,000 times. The same procedure was used in calculation of the Spearman's correlation coefficient. The width of this 95% confidence interval is shown in column dCI in Tables 3, 4, 5, and 6.

Table 3 shows Pearson's and Spearman's correlation coefficients between mean coverage score (C) and ROUGE-1 (R1) with sample size, i.e. number of documents used in evaluation, ranging from 1 to 136 with increment of 5 on the DUC 2001 single document summarization task. It also provides the mean coverage score and ROUGE-1 score of a DUC 2001 participant system (R) at different sample sizes. Table 5 presents similar information when ROUGE-SU4 was used.

The mean scores are listed in the M column. The dCI column gives the size of 95% confidence interval around the mean scores. A smaller value indicates more reliable estimation of means. ROUGE-1 and ROUGE-SU4 scores were calculated using single reference (1 REF), two references (2 REF), and three references (3 REF). ROUGE-1 and ROUGE-SU4 scores are comparable across different sample sizes but not across different number of references because raw ROUGE-1 and ROUGE-SU4 scores do not normalize according to number of references used in computation. However, Pearson's and Spearman's correlation coefficients can be compared across different number of references since they were computed according to one set of mean coverage scores at each sample size. Mean coverage score was assigned by NIST assessors during the DUC evaluation; therefore it was not affected by the number of references used (X REF).

Tables 4 and 6 show similar information as Tables 3 and 5 for the DUC 2001 multiple document summarization task but use system T as example. Results of 26 sample sizes are presented. Figure 2 displays the Pearson's correlation over different sample sizes and different number of references (RH: 1 reference, R12: two references, and R123: three references) based on the DUC 2001 single (S100) and multiple (M100) document summarization data in 6 boxplots. Each sample size includes 1,000 resampling data points. The gray box contains the middle 50% data points, i.e. points between the first and the third quartiles, and the line in the gray box marks the median. The length of whiskers of each box is 1.5 times of the length of the box. Data points outside of the whiskers are marked as circles indicating potential outliers. Figure 3 shows similar information as Figure 2 when Spearman's correlation coefficient was used.

Based on Tables 3, 4, 5, and 6 and Figures 2 and 3, we make the following observations:

(1) Correlation of automatic metrics (ROUGE-1 and ROUGE-SU4) and mean coverage (C) improve and become more accurate (smaller gray boxes) as the size of sample increases. The critical values for Pearson's correlation at 95% confidence with 10 (single document task) and 12 (multiple document task) degrees of freedom are 0.576 and 0.532 respectively. Therefore, using ROUGE-1 as the automatic metric, we reached significant level at sample size of 31 documents for the single document summarization task and at sample size of 4 topics for the multiple document summarization task when single reference was used. However, these numbers do not include estimation errors. To obtain a more accurate estimation, we need to find the sample size where the Pearson's correlation is at least half dCI larger than its critical value. After we factored in the estimation errors, the critical number of documents for single document task was 86 (0.78 - 0.39/2 = 0.585 > 0.576) and 10 (0.75 – 0.42/2 = 0.54 > 0.532) for the multi-document task. These numbers reduced to 71 (0.77 – 0.38/2 = 0.58 > 0.576) and 9 (0.75 – 0.43/2 = 0.535 > 0.532) respectively with three references. In case of using ROUGE-SU4 with single reference, the critical number of documents for single document task was 86 (0.78 - 0.39/2 = 0.585 > 0.576) and 18 (0.76 – 0.45/2 = 0.535 > 0.532) for the multi-document task. These values reduced to 66 (0.78 – 0.38/2 = 0.59 > 0.576) and 13 (0.75 – 0.41/2 = 0.54.5 > 0.532) with three references. These results suggest that the number of documents and topics provided in DUC were large enough for making significant correlation analysis and using multiple references reduced the sample size of obtaining statistically significant results.

(2) Using multiple references could help automatic metrics achieve better correlation with human judgments and improve the reliability of evaluations as reflected in the shorter confidence intervals. There were a few abnormalities to this general trend when 2 references were used. These conditions might be the same phenomena observed by Hori et al. [6] that there existed a critical number of references that should be used to achieve better and stable evaluation results. However, the number of reference summaries in our experiments was not large enough for a comprehensive investigation of this effect.

(3) Although multiple references are useful, the results in this and previous sections suggest that calibrating automatic metrics against single reference human judgment data such as the DUC data is a valid and sound approach. The only caveat is that we need to pay attention to estimation errors and use enough samples.

(4) Similar results and trends were observed when Spearman's correlation coefficient was used. The critical numbers of documents required to achieve statistical significant results can be calculated following the procedure described above.

6 Conclusions

In this paper, we reviewed the DUC evaluation method and showed that it was a sound and valid approach. We introduced ROUGE, an automatic evaluation package for summarization, and conducted comprehensive evaluations of the automatic measures included in the ROUGE package using three years of DUC data. To check the significance of the results, we estimated confidence intervals of correlations using bootstrap resampling. We found that (1) ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S worked well in single document summarization tasks, (2) ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4, and ROUGE-SU9 performed great in evaluating very short summaries (or headline-like summaries), (3) correlation of high 90% was hard to achieve for multi-document summarization tasks but ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-S9, ROUGE-SU4, and ROUGE-SU9 worked reasonably well when stopwords were excluded from matching, (4) exclusion of stopwords usually improved correlation, (5) correlations to human judgments were increased by using multiple references but using single reference summary with enough number of samples was a valid alternative, (6) sample size did affect the stability and reliability of evaluations, (7) to reach any statistical significant result, a critical number of samples had to be used, and (8) our study confirmed that the number of documents and topics used in DUC evaluations provided enough samples to claim significant results and human judgments over these data can be used to calibrate automatic evaluation metrics.

In summary, we showed that DUC evaluation data was a very valuable resource to the research community and the ROUGE package could be used effectively in automatic evaluation of summaries. Achieving statistically significant results in automatic summarization evaluation using single reference was feasible if sample size was large enough. However, in order to perform summary level diagnostic error analysis instead of overall system level performance analysis, multiple references would be necessary. Nenkova and Passonneau's [7] pyramid method provides a very good starting point in this topic, but how to fully automate their method is still an open research topic.

References

[1] Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press.

[2] Fukusima, T., M. Okumura, and H. Nanba. 2003. Text Summarization Challenge 2 – Text Summarization Evaluation at NTCIR Workshop 3. In *Proceedings of the 3rd NTCIR Workshop*, Tokyo, Japan.

[3] Lin, C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain.

[4] Lin, C.-Y. and E.H. Hovy 2002. Manual and Automatic Evaluations of Summaries. In *Proceedings of the Workshop on Automatic Summarization* post-conference workshop of ACL-02, Philadelphia, U.S.A.

[5] Lin, C.-Y. and E.H. Hovy 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada.

[6] Hori, C., T. Hori, and S. Furui. Evaluation Methods for Automatic Speech Summarization. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland.

[7] Nenkova, A. and R. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

[8] Over, P. and J. Yen. 2003. An Introduction to DUC 2003 – Intrinsic Evaluation of Generic News Text Summarization Systems. http://duc.nist.gov.

[9] Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, USA.

[10] Saggion H., D. Radev, S. Teufel, and W. Lam. 2002. Meta-Evaluation of Summaries in a Cross-Lingual Environment Using Content-Based Metrics. In *Proceedings of COLING-2002*, Taipei, Taiwan.

[11] Van Halteren and S. Teufel. 2003. Examining the Consensus between Human Summaries: Initial Experiments with Factoid Analysis. *In Proceedings of the Document Understanding Workshop 2003* (DUC 2003), Edmonton, Canada.

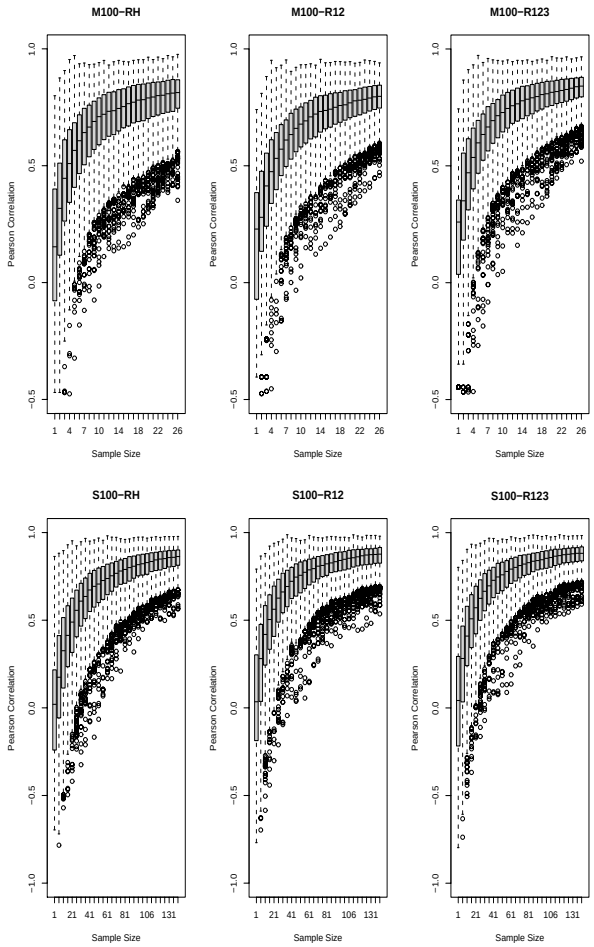

Figure 2. Pearson's correlation of ROUGE-SU4 (R-SU4) vs. mean coverage (C) for DUC 2001 single (S100) and multiple (M100) document summarization tasks with different number of references. RH: 1 reference, R12: 2 references, and R123: 3 references.

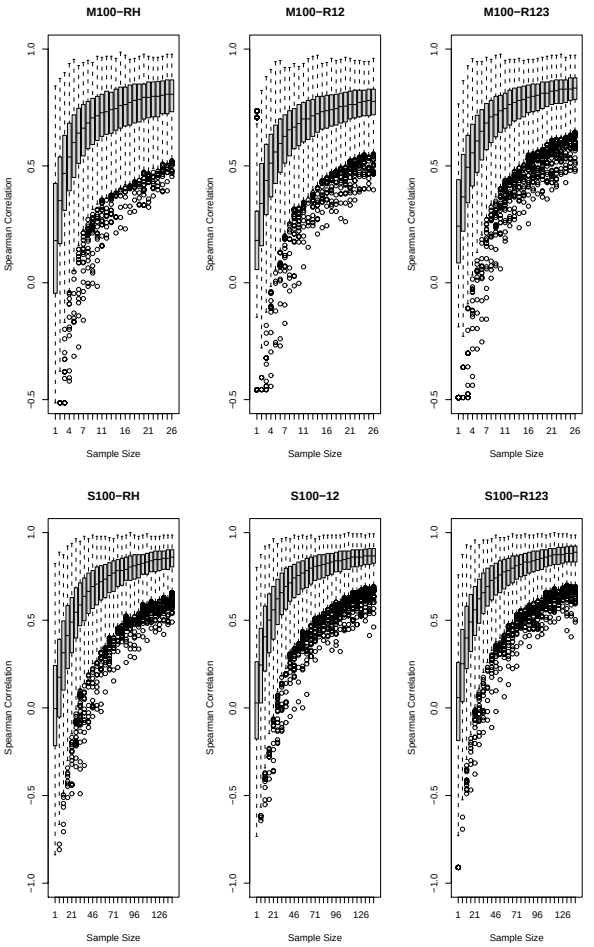

Figure 3. Spearman's correlation coefficient of ROUGE-SU4 (R-SU4) vs. mean coverage (C) for DUC 2001 single (S100) and multiple (M100) document summarization tasks with different number of references. RH: 1 reference, R12: 2 references, and R123: 3 references.