ROUGE: A Package for Automatic Evaluation of Summaries

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 introduces four different ROUGE measures: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S included in the ROUGE summarization evaluation package and their evaluations. Three of them have been used in the Document Understanding Conference (DUC) 2004, a large-scale summarization evaluation sponsored by NIST.

1 Introduction

Traditionally evaluation of summarization involves human judgments of different quality metrics, for example, coherence, conciseness, grammaticality, readability, and content (Mani, 2001). However, even simple manual evaluation of summaries on a large scale over a few linguistic quality questions and content coverage as in the Document Understanding Conference (DUC) (Over and Yen, 2003) would require over 3,000 hours of human efforts. This is very expensive and difficult to conduct in a frequent basis. Therefore, how to evaluate summaries automatically has drawn a lot of attention in the summarization research community in recent years. For example, Saggion et al. (2002) 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 (Papineni et al., 2001), in machine translation evaluation, Lin and Hovy (2003) showed that methods similar to BLEU, i.e. n-gram co-occurrence statistics, could be applied to evaluate summaries. In this paper, we introduce a package, ROUGE, for automatic evaluation of summaries and its evaluations. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes several automatic evaluation methods that measure the similarity between summaries. We describe ROUGE-N in Section 2, ROUGE-L in Section 3, ROUGE-W in Section 4, and ROUGE-S in Section 5. Section 6 shows how these measures correlate with human judgments using DUC 2001, 2002, and 2003 data. Section 7 concludes this paper and discusses future directions.

2 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:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

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.

It is clear that ROUGE-N is a recall-related measure because the denominator of the equation is the total sum of the number of n-grams occurring at the reference summary side. A closely related measure, BLEU, used in automatic evaluation of machine translation, is a precision-based measure. BLEU measures how well a candidate translation matches a set of reference translations by counting the percentage of n-grams in the candidate translation overlapping with the references. Please see Papineni et al. (2001) for details about BLEU.

Note that the number of n-grams in the denominator of the ROUGE-N formula increases as we add more references. This is intuitive and reasonable because there might exist multiple good summaries.

Every time we add a reference into the pool, we expand the space of alternative summaries. By controlling what types of references we add to the reference pool, we can design evaluations that focus on different aspects of summarization. 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. This is again very intuitive and reasonable because we normally prefer a candidate summary that is more similar to consensus among reference summaries.

2.1 Multiple References

So far, we only demonstrated how to compute ROUGE-N using a single reference. When multiple references are used, we compute pairwise summary-level ROUGE-N between a candidate summary *s* and every reference, r_i, in the reference set. We then take the maximum of pairwise summary-level ROUGE-N scores as the final multiple reference ROUGE-N score. This can be written as follows:

$$ROUGE\text{-}N_{multi} = \text{argmax}_i\, ROUGE\text{-}N(r_i, s)$$

This procedure is also applied to computation of ROUGE-L (Section 3), ROUGE-W (Section 4), and ROUGE-S (Section 5). In the implementation, we use a Jackknifing procedure. Given M references, we compute the best score over M sets of M-1 references. The final ROUGE-N score is the average of the M ROUGE-N scores using different M-1 references. The Jackknifing procedure is adopted since we often need to compare system and human performance and the reference summaries are usually the only human summaries available. Using this procedure, we are able to estimate average human performance by averaging M ROUGE-N scores of one reference vs. the rest M-1 references. Although the Jackknifing procedure is not necessary when we just want to compute ROUGE scores using multiple references, it is applied in all ROUGE score computations in the ROUGE evaluation package.

In the next section, we describe a ROUGE measure based on longest common subsequences between two summaries.

3 ROUGE-L: Longest Common Subsequence

A sequence $Z = [z_1, z_2, ..., z_n]$ is a subsequence of another sequence $X = [x_1, x_2, ..., x_m]$, if there exists a strict increasing sequence $[i_1, i_2, ..., i_k]$ of indices of X such that for all $j = 1, 2, ..., k$, we have $x_{ij} = z_j$ (Cormen et al., 1989). Given two sequences X and Y, the longest common subsequence (LCS) of X and Y is a common subsequence with maximum length. LCS has been used in identifying cognate candidates during construction of N-best translation lexicon from parallel text. Melamed (1995) used the ratio (LCSR) between the length of the LCS of two words and the length of the longer word of the two words to measure the cognateness between them. He used LCS as an approximate string matching algorithm. Saggion et al. (2002) used normalized pairwise LCS to compare similarity between two texts in automatic summarization evaluation.

3.1 Sentence-Level LCS

To apply LCS in summarization evaluation, we view a summary sentence as a sequence of words. The intuition is that the longer the LCS of two summary sentences is, the more similar the two summaries are. We propose using LCS-based F-measure to estimate the similarity between two summaries X of length m and Y of length n, assuming X is a reference summary sentence and Y is a candidate summary sentence, as follows:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (2)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (3)$$

$$F_{lcs} = \frac{(1+b^2)R_{lcs}P_{lcs}}{R_{lcs} + b^2 P_{lcs}} \quad (4)$$

Where $LCS(X,Y)$ is the length of a longest common subsequence of X and Y, and $ß = P_{lcs}/R_{lcs}$ when $?F_{lcs}/?R_{lcs} = ?F_{lcs}/?P_{lcs}$. In DUC, $ß$ is set to a very big number (? 8). Therefore, only R_{lcs} is considered. We call the LCS-based F-measure, i.e. Equation 4, ROUGE-L. Notice that ROUGE-L is 1 when $X = Y$; while ROUGE-L is zero when $LCS(X,Y) = 0$, i.e. there is nothing in common between X and Y. F-measure or its equivalents has been shown to have met several theoretical criteria in measuring accuracy involving more than one factor (Van Rijsbergen, 1979). The composite factors are LCS-based recall and precision in this case. Melamed et al. (2003) used unigram F-measure to estimate machine translation quality and showed that unigram F-measure was as good as BLEU.

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.

ROUGE-L as defined in Equation 4 has the property that its value is less than or equal to the minimum of unigram F-measure of X and Y. Unigram

recall reflects the proportion of words in *X* (reference summary sentence) that are also present in *Y* (candidate summary sentence); while unigram precision is the proportion of words in *Y* that are also in *X*. Unigram recall and precision count all co-occurring words regardless their orders; while ROUGE-L counts only in-sequence co-occurrences.

By only awarding credit to in-sequence unigram matches, ROUGE-L also captures sentence level structure in a natural way. Consider the following example:

S1. police killed the gunman
S2. police kill the gunman
S3. the gunman kill police

We only consider ROUGE-2, i.e. *N*=2, for the purpose of explanation. Using S1 as the reference and S2 and S3 as the candidate summary sentences, S2 and S3 would have the same ROUGE-2 score, since they both have one bigram, i.e. "the gunman". However, S2 and S3 have very different meanings. In the case of ROUGE-L, S2 has a score of 3/4 = 0.75 and S3 has a score of 2/4 = 0.5, with ß = 1. Therefore S2 is better than S3 according to ROUGE-L. This example also illustrated that ROUGE-L can work reliably at sentence level.

However, LCS suffers one disadvantage that it only counts the main in-sequence words; therefore, other alternative LCSes and shorter sequences are not reflected in the final score. For example, given the following candidate sentence:

S4. the gunman police killed

Using S1 as its reference, LCS counts either "the gunman" or "police killed", but not both; therefore, S4 has the same ROUGE-L score as S3. ROUGE-2 would prefer S4 than S3.

3.2 Summary-Level LCS

Previous section described how to compute sentence-level LCS-based F-measure score. When applying to summary-level, we take the union LCS matches between a reference summary sentence, r_i, and every candidate summary sentence, c_j. Given a reference summary of *u* sentences containing a total of *m* words and a candidate summary of *v* sentences containing a total of *n* words, the summary-level LCS-based F-measure can be computed as follows:

$$R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m} \quad (5)$$

$$P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{n} \quad (6)$$

$$F_{lcs} = \frac{(1+b^2) R_{lcs} P_{lcs}}{R_{lcs} + b^2 P_{lcs}} \quad (7)$$

Again ß is set to a very big number (? 8) in DUC, i.e. only R_{lcs} is considered. $LCS_{\cup}(r_i, C)$ is the LCS score of the *union* longest common subsequence between reference sentence r_i and candidate summary *C*. For example, if $r_i = w_1\ w_2\ w_3\ w_4\ w_5$, and *C* contains two sentences: $c_1 = w_1\ w_2\ w_6\ w_7\ w_8$ and $c_2 = w_1\ w_3\ w_8\ w_9\ w_5$, then the longest common subsequence of r_i and c_1 is "$w_1\ w_2$" and the longest common subsequence of r_i and c_2 is "$w_1\ w_3\ w_5$". The union longest common subsequence of r_i, c_1, and c_2 is "$w_1\ w_2\ w_3\ w_5$" and $LCS_{\cup}(r_i, C) = 4/5$.

3.3 ROUGE-L vs. Normalized Pairwise LCS

The normalized pairwise LCS proposed by Radev et al. (page 51, 2002) between two summaries S1 and S2, $LCS(S_1, S_2)_{MEAD}$, is written as follows:

$$\frac{\sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j) + \sum_{s_j \in S_2} \max_{s_i \in S_1} LCS(s_i, s_j)}{\sum_{s_i \in S_1} length(s_i) + \sum_{s_j \in S_2} length(s_j)} \quad (8)$$

Assuming S1 has *m* words and S2 has *n* words, Equation 8 can be rewritten as Equation 9 due to symmetry:

$$\frac{2 * \sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j)}{m+n} \quad (9)$$

We then define MEAD LCS recall ($R_{lcs\text{-}MEAD}$) and MEAD LCS precision ($P_{lcs\text{-}MEAD}$) as follows:

$$R_{lcs\text{-}MEAD} = \frac{\sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j)}{m} \quad (10)$$

$$P_{lcs\text{-}MEAD} = \frac{\sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j)}{n} \quad (11)$$

We can rewrite Equation (9) in terms of $R_{lcs\text{-}MEAD}$ and $P_{lcs\text{-}MEAD}$ with a constant parameter ß = 1 as follows:

$$LCS(S_1, S_2)_{MEAD} = \frac{(1+b^2) R_{lcs-MEAD} P_{lcs-MEAD}}{R_{lcs-MEAD} + b^2 P_{lcs-MEAD}} \quad (12)$$

Equation 12 shows that normalized pairwise LCS as defined in Radev et al. (2002) and implemented in MEAD is also a F-measure with ß = 1. Sentence-level normalized pairwise LCS is the same as ROUGE-L with ß = 1. Besides setting ß = 1, summary-level normalized pairwise LCS is different from ROUGE-L in how a sentence gets its LCS score from its references. Normalized pairwise LCS takes

the best LCS score while ROUGE-L takes the union LCS score.

4 ROUGE-W: Weighted Longest Common Subsequence

LCS has many nice properties as we have described in the previous sections. Unfortunately, the basic LCS also has a problem that it does not differentiate LCSes of different spatial relations within their embedding sequences. For example, given a reference sequence X and two candidate sequences Y_1 and Y_2 as follows:

X: [A B C D E F G]
Y_1: [A B C D H I K]
Y_2: [A H B K C I D]

Y_1 and Y_2 have the same ROUGE-L score. However, in this case, Y_1 should be the better choice than Y_2 because Y_1 has consecutive matches. To improve the basic LCS method, we can simply remember the length of consecutive matches encountered so far to a regular two dimensional dynamic program table computing LCS. We call this weighted LCS (WLCS) and use k to indicate the length of the current consecutive matches ending at words x_i and y_j. Given two sentences X and Y, the WLCS score of X and Y can be computed using the following dynamic programming procedure:

(1) For (i = 0; i <=m; i++)
  $c(i,j)$ = 0 // *initialize c-table*
  $w(i,j)$ = 0 // *initialize w-table*
(2) For (i = 1; i <= m; i++)
  For (j = 1; j <= n; j++)
    If x_i = y_j Then
      // *the length of consecutive matches at*
      // *position i-1 and j-1*
      k = $w(i\text{-}1,j\text{-}1)$
      $c(i,j) = c(i\text{-}1,j\text{-}1) + f(k+1) – f(k)$
      // *remember the length of consecutive*
      // *matches at position i, j*
      $w(i,j) = k+1$
    Otherwise
      If $c(i\text{-}1,j) > c(i,j\text{-}1)$ Then
        $c(i,j) = c(i\text{-}1,j)$
        $w(i,j)$ = 0 // *no match at i, j*
      Else $c(i,j) = c(i,j\text{-}1)$
        $w(i,j)$ = 0 // *no match at i, j*
(3) $WLCS(X,Y) = c(m,n)$

Where c is the dynamic programming table, $c(i,j)$ stores the WLCS score ending at word x_i of X and y_j of Y, w is the table storing the length of consecutive matches ended at c table position i and j, and f is a function of consecutive matches at the table position, $c(i,j)$. Notice that by providing different weighting function f, we can parameterize the WLCS algorithm to assign different credit to consecutive in-sequence matches.

The weighting function f must have the property that $f(x+y) > f(x) + f(y)$ for any positive integers x and y. In other words, consecutive matches are awarded more scores than non-consecutive matches. For example, $f(k) = ak – b$ when $k >= 0$, and $a, b > 0$. This function charges a gap penalty of $–b$ for each non-consecutive n-gram sequences. Another possible function family is the polynomial family of the form k^a where $a > 1$. However, in order to normalize the final ROUGE-W score, we also prefer to have a function that has a close form inverse function. For example, $f(k) = k^2$ has a close form inverse function $f^{-1}(k) = k^{1/2}$. F-measure based on WLCS can be computed as follows, given two sequences X of length m and Y of length n:

$$R_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(m)}\right) \quad (13)$$

$$P_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(n)}\right) \quad (14)$$

$$F_{wlcs} = \frac{(1+b^2)R_{wlcs}P_{wlcs}}{R_{wlcs} + b^2 P_{wlcs}} \quad (15)$$

Where f^{-1} is the inverse function of f. In DUC, ß is set to a very big number (? 8). Therefore, only R_{wlcs} is considered. We call the WLCS-based F-measure, i.e. Equation 15, ROUGE-W. Using Equation 15 and $f(k) = k^2$ as the weighting function, the ROUGE-W scores for sequences Y_1 and Y_2 are 0.571 and 0.286 respectively. Therefore, Y_1 would be ranked higher than Y_2 using WLCS. We use the polynomial function of the form k^a in the ROUGE evaluation package. In the next section, we introduce the skip-bigram co-occurrence statistics.

5 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 measure the overlap of skip-bigrams between a candidate translation and a set of reference translations. Using the example given in Section 3.1:

S1. police killed the gunman
S2. police kill the gunman
S3. the gunman kill police
S4. the gunman police killed

each sentence has $C(4,2)$[1] = 6 skip-bigrams. For example, S1 has the following skip-bigrams:

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

S2 has three skip-bigram matches with S1 ("*police the*", "*police gunman*", "*the gunman*"), S3 has one skip-bigram match with S1 ("*the gunman*"), and S4 has two skip-bigram matches with S1 ("*police killed*", "*the gunman*"). Given translations X of length m and Y of length n, assuming X is a reference translation and Y is a candidate translation, we compute skip-bigram-based F-measure as follows:

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)} \quad (16)$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)} \quad (17)$$

$$F_{skip2} = \frac{(1+b^2)R_{skip2}P_{skip2}}{R_{skip2} + b^2 P_{skip2}} \quad (18)$$

Where $SKIP2(X,Y)$ is the number of skip-bigram matches between X and Y, ß controlling the relative importance of P_{skip2} and R_{skip2}, and C is the combination function. We call the skip-bigram-based F-measure, i.e. Equation 18, ROUGE-S.

Using Equation 18 with ß = 1 and S1 as the reference, S2's ROUGE-S score is 0.5, S3 is 0.167, and S4 is 0.333. Therefore, S2 is better than S3 and S4, and S4 is better than S3. This result is more intuitive than using BLEU-2 and ROUGE-L. One advantage of skip-bigram vs. BLEU is that it does not require consecutive matches but is still sensitive to word order. Comparing skip-bigram with LCS, skip-bigram counts all in-order matching word pairs while LCS only counts one longest common subsequence.

Applying skip-bigram without any constraint on the distance between the words, spurious matches such as "*the the*" or "*of in*" might be counted as valid matches. To reduce these spurious matches, we can limit the maximum skip distance, d_{skip}, between two in-order words that is allowed to form a skip-bigram. For example, if we set d_{skip} to 0 then ROUGE-S is equivalent to bigram overlap F-measure. If we set d_{skip} to 4 then only word pairs of at most 4 words apart can form skip-bigrams.

Adjusting Equations 16, 17, and 18 to use maximum skip distance limit is straightforward: we only count the skip-bigram matches, $SKIP2(X,Y)$, within the maximum skip distance and replace denominators of Equations 16, $C(m,2)$, and 17, $C(n,2)$, with the actual numbers of within distance skip-bigrams from the reference and the candidate respectively.

[1] C(4,2) = 4!/(2!*2!) = 6.

5.1 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. For example, the following sentence has a ROUGE-S score of zero:

S5. gunman the killed police

S5 is the exact reverse of S1 and there is no skip bigram match between them. However, we would like to differentiate sentences similar to S5 from sentences that do not have single word co-occurrence with S1. To achieve this, we extend ROUGE-S with the addition of unigram as counting unit. The extended version is called ROUGE-SU. We can also obtain ROUGE-SU from ROUGE-S by adding a begin-of-sentence marker at the beginning of candidate and reference sentences.

6 Evaluations of ROUGE

To assess the effectiveness of ROUGE measures, we compute the correlation between ROUGE assigned summary scores and human assigned summary 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, we have DUC 2001, 2002, and 2003 evaluation data that include human judgments for the following:

- Single document summaries of about 100 words: 12 systems[2] for DUC 2001 and 14 systems for 2002. 149 single document summaries were judged per system in DUC 2001 and 295 were judged in DUC 2002.
- Single document very short summaries of about 10 words (headline-like, keywords, or phrases): 14 systems for DUC 2003. 624 very short summaries were judged per system in DUC 2003.
- Multi-document summaries of about 10 words: 6 systems for DUC 2002; 50 words: 14 systems for DUC 2001 and 10 systems for DUC 2002; 100 words: 14 systems for DUC 2001, 10 systems for DUC 2002, and 18 systems for DUC 2003; 200 words: 14 systems for DUC 2001 and 10 systems for DUC 2002; 400 words: 14 systems for DUC 2001. 29 summaries were judged per system per summary size in DUC 2001, 59 were judged in DUC 2002, and 30 were judged in DUC 2003.

[2] All systems include 1 or 2 baselines. Please see DUC website for details.

	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

Besides these human judgments, we also have 3 sets of manual summaries for DUC 2001, 2 sets for DUC 2002, and 4 sets for DUC 2003. Human judges assigned content coverage scores to a candidate summary by examining the percentage of content overlap between a manual summary unit, i.e. elementary discourse unit or sentence, and the candidate summary using Summary Evaluation Environment[3] (SEE) developed by the University of Southern California's Information Sciences Institute (ISI). The overall candidate summary score is the average of the content coverage scores of all the units in the manual summary. Note that human judges used only one manual summary in all the evaluations although multiple alternative summaries were available.

With the DUC data, we computed Pearson's product moment correlation coefficients, Spearman's rank order correlation coefficients, and Kendall's correlation coefficients between systems' average ROUGE scores and their human assigned average 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[4] version of the summaries (STEM set), and stopped version of the summaries (STOP set). For example, we computed ROUGE scores for the 12 systems participated in the DUC 2001 single document summarization evaluation using the CASE set with single reference and then calculated the three correlation scores for these 12 systems' ROUGE scores vs. human assigned average coverage scores. After that we repeated the process using multiple references and then using STEM and STOP sets. Therefore, 2 (multi or single) x 3 (CASE, STEM, or STOP) x 3 (Pearson, Spearman, or Kendall) = 18 data points were collected for each ROUGE measure and each DUC task. To assess the significance of the results, we applied bootstrap resampling technique (Davison and Hinkley, 1997) to estimate 95% confidence intervals for every correlation computation.

17 ROUGE measures were tested for each run using ROUGE evaluation package v1.2.1: ROUGE-N with N = 1 to 9, ROUGE-L, ROUGE-W with weighting factor α = 1.2, ROUGE-S and ROUGE-SU with maximum skip distance d_{skip} = 1, 4, and 9. Due to limitation of space, we only report correlation analysis results based on Pearson's correlation coefficient. Correlation analyses based on Spearman's and Kendall's correlation coefficients are tracking Pearson's very closely and will be posted later at the ROUGE website[5] for reference. The critical value[6] for Pearson's correlation is 0.632 at 95% confidence with 8 degrees of freedom.

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 due to the double sample size in 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-

[3] SEE is available online at http://www.isi.edu/~cyl.

[4] Porter's stemmer was used.

[5] ROUGE website: http://www.isi.edu/~cyl/ROUGE.

[6] The critical values for Pearson's correlation at 95% confidence with 10, 12, 14, and 16 degrees of freedom are 0.576, 0.532, 0.497, and 0.468 respectively.

	(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

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 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 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 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.

7 Conclusions

In this paper, 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, and (5) correlations to human judgments were increased by using multiple references.

In summary, we showed that the ROUGE package could be used effectively in automatic evaluation of summaries. In a separate study (Lin and Och, 2004),

ROUGE-L, W, and S were also shown to be very effective in automatic evaluation of machine translation. The stability and reliability of ROUGE at different sample sizes was reported by the author in (Lin, 2004). However, how to achieve high correlation with human judgments in multi-document summarization tasks as ROUGE already did in single document summarization tasks is still an open research topic.

8 Acknowledgements

The author would like to thank the anonymous reviewers for their constructive comments, Paul Over at NIST, U.S.A, and ROUGE users around the world for testing and providing useful feedback on earlier versions of the ROUGE evaluation package, and the DARPA TIDES project for supporting this research.

References

Cormen, T. R., C. E. Leiserson, and R. L. Rivest. 1989. *Introduction to Algorithms*. The MIT Press.

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

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.

Lin, C.-Y. 2004. Looking for a few good metrics: ROUGE and its evaluation. In *Proceedings of NTCIR Workshop 2004*, Tokyo, Japan.

Lin, C.-Y. and F. J. Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of 42nd Annual Meeting of ACL* (ACL 2004), Barcelona, Spain.

Mani, I. 2001. *Automatic Summarization*. John Benjamins Publishing Co.

Melamed, I. D. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora (WVLC3)*. Boston, U.S.A.

Melamed, I. D., R. Green and J. P. Turian (2003). Precision and recall of machine translation. In *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada.

Over, P. and J. Yen. 2003. An introduction to DUC 2003 – Intrinsic evaluation of generic news text summarization systems. http://www-nlpir.nist.gov/projects/duc/pubs/2003slides/duc2003intro.pdf

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. *IBM Research Report RC22176 (W0109-022)*.

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.

Radev, D. S. Teufel, H. Saggion, W. Lam, J. Blitzer, A. Gelebi, H. Qi, E. Drabek, and D. Liu. 2002. *Evaluation of Text Summarization in a Cross-Lingual Information Retrieval Framework*. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.

Van Rijsbergen, C. J. 1979. *Information Retrieval*. Butterworths. London.