A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art

Peter A. Rankel
University of Maryland
rankel@math.umd.edu

John M. Conroy
IDA / Center for Computing Sciences
conroy@super.org

Hoa Trang Dang
National Institute of Standards and Technology
hoa.dang@nist.gov

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract

How good are automatic content metrics for news summary evaluation? Here we provide a detailed answer to this question, with a particular focus on assessing the ability of automatic evaluations to identify statistically significant differences present in manual evaluation of content. Using four years of data from the Text Analysis Conference, we analyze the performance of eight ROUGE variants in terms of accuracy, precision and recall in finding significantly different systems. Our experiments show that some of the neglected variants of ROUGE, based on higher order n-grams and syntactic dependencies, are most accurate across the years; the commonly used ROUGE-1 scores find too many significant differences between systems which manual evaluation would deem comparable. We also test combinations of ROUGE variants and find that they considerably improve the accuracy of automatic prediction.

1 Introduction

ROUGE (Lin, 2004) is a suite of automatic evaluations for summarization and was introduced a decade ago as a reasonable substitute for costly and slow human evaluation. The scores it produces are based on n-gram or syntactic overlap between an automatic summary and a set of human reference summaries. However, the field does not have a good grasp of which of the many evaluation scores is most accurate in replicating human judgements. This state of uncertainty has led to problems in comparing published work, as different researchers choose to publish different variants of scores.

In this paper we reassess the strengths of ROUGE variants using the data from four years of Text Analysis Conference (TAC) evaluations, 2008 to 2011. To assess the performance of the automatic evaluations, we focus on determining statistical significance[1] between systems, where the gold-standard comes from comparing the systems using manual pyramid and responsiveness evaluations. In this setting, computing correlation coefficients between manual and automatic scores is not applicable as it does not take into account the statistical significance of the differences nor does it allow the use of more powerful statistical tests which use pairwise comparisons of performance on individual document sets. Instead, we report on the accuracy of decisions on pairs of systems, as well as the precision and recall of identifying pairs of systems which exhibit statistically significant differences in content selection performance.

2 Background

During 2008–2011, automatic summarization systems at TAC were required to create 100-word summaries. Each year there were two multi-document summarization sub-tasks, the initial summary and the update summary, usually referred to as task A and task B, respectively. The test inputs in each consisted of about 10 documents and the type of summary varied between query-focused and guided. There are between 44 and 48 test inputs on which systems are compared for each task.

In 2008 and 2009, task A was to produce a

[1] For the purpose of this study, we define a difference as significant when the test statistic attains a value corresponding to a p-value less than 0.05.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 131–136,
Sofia, Bulgaria, August 4-9 2013.

query-focused summary in response to a user information need stated both as a brief statement and a paragraph-long description of the information the user seeks to find. In 2010 and 2011 task A was "guided summarization", where the test inputs came from a small set of predefined domains. These domains included accidents and natural disasters, attacks, health and safety, endangered resources, investigations and trials. Systems were provided with a list of important aspects of information for each domain and were asked to cover as many of these aspects as possible. The writers of the reference summaries for evaluation were given similar instructions. In all four years, task B was to produce an update summary for each of the inputs given in task A (query-focused or guided). In each case, a new, subsequent set of documents related to the topic of the respective test set for task A was provided to the system. The task was to generate an update summary aimed at a user who has already read all documents in the inputs for task A.

The two manual evaluation approaches used in TAC 2008–2011 are modified pyramid (Nenkova et al., 2007) and overall responsiveness. The pyramid method requires several reference summaries for each input. These are manually analyzed to discover content units based on meaning rather than specific wording. Each content unit is assigned a weight equal to the number of reference summaries that included that content unit. The modified pyramid score is defined as the sum of weights of the content units in the summary normalized by the weight of an ideally informative summary which expresses n content units, where n is equal to the average of content units in the reference summaries. Responsiveness, on the other hand, is based on direct human judgements, without the need for reference summaries. Assessors are presented with a statement of the user's information need and the summary they need to evaluate. Then they rate how well they think the summary responds to the information need contained in the topic statement. Responsiveness was rated on a ten-point scale in 2009, and on a five-point scale in all other years.

For each sub-task during 2008–2011, we analyze the performance of only the top 30 systems, which roughly corresponds to the systems that performed better than or around the median according to each manual metric. Table 1 gives the number of significant differences among the top 30 participating systems. We keep only the best performing systems for the analysis because we are interested in studying how well automatic evaluation metrics can correctly compare very good systems.

Year	Pyr A	Pyr B	Resp A	Resp B
2008	82	109	68	105
2009	146	190	106	92
2010	165	139	150	128
2011	39	83	5	11

Table 1: Number of pairs of significantly different systems among the top 30 across the years. There is a total of 435 pairs in each year.

3 Which ROUGE is best?

In this section, we study the performance of several ROUGE variants, including ROUGE-n, for $n = 1, 2, 3, 4$, ROUGE-L, ROUGE-W-1.2, ROUGE-SU4, and ROUGE-BE-HM (Hovy et al., 2006). ROUGE-n measures the n-gram recall of the evaluated summary compared to the available reference summaries. ROUGE-L is the ratio of the number of words in the longest common subsequence between the reference and the evaluated summary and the number of words in the reference. ROUGE-W-1.2 is a weighted version of ROUGE-L. ROUGE-SU4 is a combination of skip bigrams and unigrams, where the skip bigrams are formed for all words that appear in the text with no more than four intervening words in between. ROUGE-BE-HM computes recall of dependency syntactic relations between the summary and the reference.

To evaluate how well an automatic evaluation metric reproduces human judgments, we use prediction *accuracy* similar to Owczarzak et al. (2012). For each pair of systems in each subtask, we compare the results of two Wilcoxon signed-rank tests, one using the manual evaluation scores for each system and one using the automatic evaluation scores for each system (Rankel et al., 2011).[2] The accuracy then is simply the percent agreement between the results of these two tests.

[2]We use the Wilcoxon test as it was demonstrated by Rankel et al. (2011) to give more statistical power than unpaired tests. As reported by Yeh (2000), other tests such as randomized testing, may also be appropriate. There is considerable variation in system performance for different inputs (Nenkova and Louis, 2008) and paired tests remove the effect of the input.

	Responsiveness				Pyramid			
Metric	Acc	P	R	BA	Acc	P	R	BA
R1	0.58 (0.61)	0.24	0.64	0.57	0.62 (0.66)	0.37	0.67	0.61
R2	0.64 (0.63)	0.28	0.60	0.59	0.68 (0.69)	0.43	0.63	0.64
R3	0.70 (0.63)	0.31	0.48	0.60	0.73 (0.68)	0.49	0.53	0.66
R4	0.73 (0.64)	0.33	0.40	0.60	0.74 (0.65)	0.50	0.45	0.65
RL	0.50 (0.59)	0.20	0.56	0.54	0.54 (0.63)	0.29	0.60	0.55
R-SU4	0.61(0.62)	0.26	0.61	0.58	0.65 (0.68)	0.40	0.65	0.63
R-W-1.2	0.52(0.62)	0.21	0.54	0.55	0.57(0.64)	0.32	0.62	0.57
R-BE-HM	0.70 (0.63)	0.30	0.49	0.59	0.74(0.68)	0.49	0.56	0.66

Table 2: Accuracy, Precision, Recall, and Balanced Accuracy of each ROUGE variant, averaged across all eight tasks in 2008-2011, with and (without) significance.

As can be seen in Table 1, the manual evaluation metrics often did not show many significant differences between systems.[3] Thus, it is clear that the percent agreement will be high for an approach for automatic evaluation that always predicts zero significant differences. As traditionally done when dealing which such skewed distributions of classes, we also examine the *precision* and *recall* with respect to finding significant differences of several ROUGE variants, to better assess the quality of their prediction. To identify a measure that is strong at both predicting significant and non-significant differences we compute balanced accuracy, the mean of the accuracy of predicting significant differences and the accuracy of predicting no significant difference.[4]

Each of these four measures for judging the performance of ROUGE variants has direct intuitive interpretation, unlike other opaque measures such as correlation coefficients and F-measure which have formal definitions which do not readily yield to intuitive understanding.

[3]This is a somewhat surprising finding which may warrant further investigation. One possible explanation is that different systems generate similar summaries. Recent work has shown that this is unlikely to be the case because the collection of summaries from several systems indicates better what content is important than the single best summary (Louis and Nenkova, 2013). The short summary length for which the summarizers are compared may also contribute to the fact that there are few significant difference. In early NIST evaluations manual evaluations could not distinguish automatic and human summaries based on summaries of length 50 and 100 words and there were more significant differences between systems for 200-word summaries than for 100-word summaries (Nenkova, 2005).

[4]More generally, one could define a utility function which gives costs associated with errors and benefits to correct prediction. Balanced accuracy weighs all errors as equally bad and all correct prediction as equally good (von Neumann and Morgenstern, 1953).

Few prior studies have taken statistical significance into account during the assessment of automatic metrics for evaluation. For this reason we first briefly discuss ROUGE accuracy without taking significance into account. In this special case, agreement simply means that the automatic and manual evaluations agree on which of two systems is better, based on each system's average score for all test inputs for a given task. It is very rare that the average scores of two systems are equal, so there is always a better system in each pair, and random prediction would have 50% accuracy.

Many papers do not report the significance of differences in ROUGE scores (for the ROUGE variant of their choice), but simply claim that their system X with higher average ROUGE score than system Y is better than system Y. Table 2 lists the average accuracy with significance taken into account and then in parentheses, accuracy without taking significance into account. The data demonstrate that the best accuracy of the eight ROUGE metrics is a meager 64% for responsiveness when significance is not taken into account. So the conclusion about the relative merit of systems would be different from that based on manual evaluation in one out of three comparisons. However, the best accuracy rises to 73% when significance is taken into account; an incorrect conclusion will be drawn in one out of four comparisons. The reduction in error is considerable.

Furthermore, ROUGE-3 and ROUGE-4, which are rarely reported, are among the most accurate. Note also, these results differ considerably from those reported by Owczarzak et al. (2012), where ROUGE-2 was shown to have accuracy of 81% for responsiveness and 89% for pyramid. The wide differences are due to the fact we are only consid-

ering systems which scored in the top 30. This illustrates that our automatic metrics are not as good at discriminating systems near the top. These findings give strong support for the idea of requiring authors to report the significance of the difference between their summarization system and the chosen baseline; the conclusions about relative merits of the system would be more similar to those one would draw from manual evaluation.

In addition to accuracy, Table 2 gives precision, recall and balanced accuracy for each of the eight ROUGE measures when significance is taken into account. ROUGE-1 is arguably the most widely used score in the literature and Table 2 reveals an interesting property: ROUGE-1 has high recall but low precision. This means that it reports many significant differences, most of which do not exist according to the manual evaluations.

Balanced accuracy helps us identify which ROUGE variants are most accurate in finding statistical significance and correctly predicting that two systems are not significantly different. For the pyramid evaluation, the variants with best balanced accuracy (66%) are ROUGE-3 and ROUGE-BE, with ROUGE-4 just a percent lower at 65%. For responsiveness the configuration is similar, with ROUGE-3 and ROUGE-4 tied for best (60%), and ROUGE-BE just a percent lower.

The good performance of higher-order n-grams is quite surprising because these are practically never used for reporting results in the literature. Based on our results however, they are much more likely to accurately reproduce conclusions that would have been drawn from manual evaluation of top-performing systems.

4 Multiple hypothesis tests to combine ROUGE variants

We now consider a method to combine multiple evaluation scores in order to obtain a stronger ensemble metric. The idea of combining ROUGE variants has been explored in the prior literature. Conroy and Dang (2008), for example, proposed taking linear combinations of ROUGE metrics. This approach was extended by Rankel et al. (2012) by including measures of linguistic quality. Recently, Amigó et al. (2012) applied the "heterogeneity principle" and combined ROUGE scores to improve the *precision* relative to a human evaluation metric. Their results demonstrate that a consensus among ROUGE scores can predict more accurately if an improvement in a human evaluation metric will be achieved.

Along the lines of these investigations, we examine the performance of a simple combination of variants: Call the difference between two systems significant only when *all* the variants in the combination indicate significance. As in the section above, a paired Wilcoxon signed-rank test is used to determine the level of significance.

ROUGE Combination	Acc	Prec	Rec	BA
R1_R2_R4_RBE	0.76	0.77	0.36	0.76
R1_R4_RBE	0.76	0.76	0.36	0.76
R2_R4_RBE	0.76	0.74	0.40	0.75
R4_RBE	0.76	0.73	0.41	0.75
R1_R2_R4	0.76	0.71	0.40	0.74
R1_R4	0.75	0.70	0.40	0.73
R2_R4	0.75	0.68	0.44	0.73
R1_R2_RBE	0.75	0.66	0.48	0.72
R2_RBE	0.75	0.64	0.52	0.72
R4	0.74	0.62	0.47	0.70
R1_RBE	0.74	0.62	0.49	0.70
R1_R2	0.73	0.57	0.62	0.70
RBE	0.73	0.57	0.58	0.68
R2	0.71	0.53	0.69	0.68
R1	0.62	0.43	0.69	0.63

Table 3: Accuracy, Precision, Recall, and Balanced Accuracy of each ROUGE combination on TAC 2008-2010 pyramid.

We considered all possible combinations of four ROUGE metrics that exhibited good properties in the analyses presented so far: ROUGE-1 (because of its high recall), ROUGE-2 (because of high accuracy when significance is not taken into account) and ROUGE-4 and ROUGE-BE, which showed good balanced accuracy.

The performance of these combinations for reproducing the decisions in TAC 2008-2010 based on the pyramid[5] evaluation are given in Table 3. The best balanced accuracy (76%) is for the combination of all four variants. As more variants are combined, precision increases but recalls drops.

5 Comparison with automatic evaluations from AESOP 2011

In 2009-2011, TAC ran the task of Automatically Evaluating Summaries of Peers (AESOP), to com-

[5]The ordering of the metric combinations relative to responsiveness was almost identical to the ordering relative to the pyramid evaluation, and precision and recall exhibited the same trend as more metrics were added to the combination.

	Pyramid A				Pyramid B				Responsiveness A				Responsiveness B			
Evaluation Metric	Acc	P	R	BA	Acc	P	R	BA	Acc	P	R	BA	Acc	P	R	BA
CLASSY1	0.60	0.02	**0.60**	0.50	**0.84**	0.03	0.18	0.50	0.61	0.14	0.64	0.54	0.70	0.21	0.22	0.52
DemokritosGR1	0.59	0.01	0.20	0.50	0.79	0.07	0.55	0.53	0.66	0.18	**0.79**	0.58	0.64	0.17	0.24	0.49
uOttawa3	0.44	0.01	**0.60**	0.50	0.48	0.02	0.36	0.50	0.52	0.13	0.77	0.55	0.43	0.13	0.36	0.46
DemokritosGR2	0.78	0.01	0.20	0.50	0.76	0.06	0.55	0.52	0.76	0.23	0.69	0.60	0.67	0.22	0.29	0.52
C-S-IIITH4	0.69	0.01	0.20	0.50	0.77	0.07	0.64	0.53	0.82	**0.29**	0.74	**0.63**	0.60	0.15	0.24	0.47
C-S-IIITH1	0.60	0.01	0.40	0.50	0.70	0.06	**0.82**	0.53	0.69	0.20	**0.79**	0.59	0.60	0.22	**0.42**	0.52
BEwT-E	0.73	0.01	0.20	0.50	0.80	0.01	0.09	0.49	0.79	0.25	0.72	**0.61**	0.72	**0.31**	0.39	**0.58**
R1-R2-R4-RBE	**0.89**	**0.40**	0.44	**0.67**	0.76	0.27	0.17	0.55	**0.88**	0.00	0.00	0.49	**0.91**	0.03	0.09	0.50
R1-R4-RBE	**0.89**	**0.40**	0.44	**0.67**	0.77	**0.35**	0.24	**0.59**	**0.88**	0.00	0.00	0.49	0.90	0.03	0.09	0.50
All ROUGEs	**0.89**	**0.40**	0.44	**0.67**	0.75	0.26	0.16	0.54	**0.88**	0.00	0.00	0.49	**0.91**	0.04	0.09	0.51

Table 4: Best performing AESOP systems from TAC 2011; Scores within the 95% confidence interval of the best are in bold face.

pare automatic evaluation methods for automatic summarization. Here we show how the submitted AESOP metrics compare to the best ROUGE variants that we have established so far. We report the results on 2011 only, because even when the same team participated in more than one year, the metrics submitted were different and the 2011 results represent the best effort of these teams. However, as we saw in Table 1, in 2011 there were very few significant differences between the top summarization systems. In this sense the tasks that year represent a challenging dataset for testing automatic evaluations.

The results for the best AESOP systems (according to one or more measures), and the corresponding results for the ROUGE combinations are shown in Table 4. These AESOP systems are: CLASSY1 (Conroy et al., 2011; Rankel et al., 2012), DemokritosGR1 and 2 (Giannakopoulos et al., 2008; Giannakopoulos et al., 2010), uOttawa3 (Kennedy et al., 2011), C-S-IITH1 and 4 (Kumar et al., 2011; Kumar et al., 2012), and BEwT-E (Tratz and Hovy, 2008).[6] The combination metrics achieve the highest accuracy by generally predicting correctly when there are no significant differences between the systems. In addition, for 2008-2010, where far more differences between systems occur, the results of Table 3 show the combination metrics outperformed use of a single metric and are competitive with the best metrics of AESOP 2011. Thus, the combination metrics have the ability to discriminate under both conditions giving good prediction of human evaluation.

[6]To perform the comparison in the table the scores for each system and document set were needed. Some systems have changed after TAC 2011, but the data needed for these comparisons were not available. BEwT-E did not participate in AESOP 2011 and these data were provided by Stephen Tratz. Special thanks to Stephen for providing these data.

6 Conclusion

We have tested the best-known automatic evaluation metrics (ROUGE) on several years of TAC data and compared their performance with recently developed AESOP metrics. We discovered that some of the rarely used variants of ROUGE perform surprisingly well, and that by combining different ROUGEs together, one can create an evaluation metric that is extremely competitive with metrics submitted to the latest AESOP task. Our results were reported in terms of several different measures, and in each case, compared how well the automatic metric predicted significant differences found in manual evaluation. We believe strongly that developers should include statistical significance when reporting differences in ROUGE scores of theirs and other systems, as this improves the accuracy and credibility of their results. Significant improvement in multiple ROUGE scores is a significantly stronger indicator that the developers have made a noteworthy improvement in text summarization. Systems that report significant improvement using a combination of ROUGE-BE (or its improved version BEwT-E) in conjunction with ROUGE-1, 2, and 4, are more likely to give rise to summaries that humans would judge as significantly better.

Acknowledgments

The authors would like to thank Ed Hovy who raised the question "How well do automatic metrics perform when comparing top systems?" Ed's comments helped motivate this work. In addition, we would like to thank our anonymous referees for their insightful comments, which contributed *significantly* to this paper.

References

Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2012. The heterogeneity principle in evaluation measures for automatic summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 36–43, Montréal, Canada, June. Association for Computational Linguistics.

John M. Conroy and Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK, August. Coling 2008 Organizing Committee.

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. 2011. Nouveau-ROUGE: A Novelty Metric for Update Summarization. *Computational Linguistics*, 37(1):1–8.

George Giannakopoulos, Vangelis Karkaletsis, George A. Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *TSLP*, 5(3).

George Giannakopoulos, George A. Vouros, and Vangelis Karkaletsis. 2010. Mudos-ng: Multi-document summaries using n-gram graphs (tech report). *CoRR*, abs/1012.2042.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 899–902.

Alistair Kennedy, Anna Kazantseva Saif Mohammad, Terry Copeck, Diana Inkpen, and Stan Szpakowicz. 2011. Getting emotional about news. In *Fourth Text Analysis Conference (TAC 2011)*.

Niraj Kumar, Kannan Srinathan, and Vasudeva Varma. 2011. Using unsupervised system with least linguistic features for tac-aesop task. In *Fourth Text Analysis Conference (TAC 2011)*.

N. Kumar, K. Srinathan, and V. Varma. 2012. Using graph based mapping of co-occurring words and closeness centrality score for summarization evaluation. *Computational Linguistics and Intelligent Text Processing*, pages 353–365.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39:267–300.

Ani Nenkova and Annie Louis. 2008. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *ACL*, pages 825–833.

Ani Nenkova, Rebecca J. Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4(2).

Ani Nenkova. 2005. Discourse factors in multi-document summarization. In *AAAI*, pages 1654–1655.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada, June. Association for Computational Linguistics.

Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. 2011. Ranking human and machine summarization systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Peter A. Rankel, John M. Conroy, and Judith D. Schlesinger. 2012. Better metrics to automatically predict the quality of a text summary. *Algorithms*, 5(4):398–420.

Stephen Tratz and Eduard Hovy. 2008. Summarisation evaluation using transformed basic elements. In *Proceedings TAC 2008*. NIST.

John von Neumann and Oskar Morgenstern. 1953. *Theory of games and economic behavior*. Princeton Univ. Press, Princeton, NJ, 3. ed. edition.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA. Association for Computational Linguistics.