Squibs

Nouveau-ROUGE: A Novelty Metric for Update Summarization

John M. Conroy*
IDA/Center for Computing Sciences

Judith D. Schlesinger*
IDA/Center for Computing Sciences

Dianne P. O'Leary**
University of Maryland

An update summary should provide a fluent summarization of new information on a time-evolving topic, assuming that the reader has already reviewed older documents or summaries. In 2007 and 2008, an annual summarization evaluation included an update summarization task. Several participating systems produced update summaries indistinguishable from human-generated summaries when measured using ROUGE. However, no machine system performed near human-level performance in manual evaluations such as pyramid and overall responsiveness scoring.

We present a metric called Nouveau-ROUGE that improves correlation with manual evaluation metrics and can be used to predict both the pyramid score and overall responsiveness for update summaries. Nouveau-ROUGE can serve as a less expensive surrogate for manual evaluations when comparing existing systems and when developing new ones.

1. Introduction

Update summaries focus on what is new relative to a previous body of information. They pose new challenges both to algorithm developers and to evaluation of summaries. In 2007, DUC (Document Understanding Conference) introduced an update summarization task, repeated in 2008 for TAC (Text Analysis Conference).[1] This task consisted of producing a multi-document summary for a set of articles on a single topic, followed by one (2008) or two (2007) multi-document summaries for sets of articles on

* Institute for Defense Analyses, Center for Computing Sciences, 17100 Science Drive, Bowie, MD 20715 USA. E-mail: {judith,conroy}@super.org.
** Computer Science Department, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA. E-mail: oleary@cs.umd.edu.

1 DUC (http://duc.nist.gov), the summarization evaluation event, was replaced in 2008 by TAC (http://www.nist.gov/tac). Both were sponsored by NIST, the U.S. National Institute of Standards and Technology.

Submission received: 14 January 2010; revised submission received: 15 May 2010; accepted for publication: 27 September 2010.

the same topic published at later dates. The goal was to generate a good first summary, along with update(s) that contained new content and minimized redundancy.

The modifier **manual** is used to identify evaluations, and the corresponding scores, produced by humans. The modifier **automatic** is used to identify evaluations, and the corresponding scores, produced by machines. Similarly, **human-generated** and **machine-generated** will be used to distinguish between summaries created by humans and those generated by machine systems, respectively.[2]

Because we are working with update summarization, there is a minimum of two summaries for a set of documents. The first summary for the document set is called the **original** (Task A) summary and a later summary is called an **update** (Task B) summary.

Several machine summarizing systems produced update summaries that were statistically indistinguishable from human-generated summaries, as measured by the ROUGE metrics, the standard metrics for *automatic evaluation* of summaries. However, none of these machine systems performed near human levels in overall responsiveness or pyramid evaluation, the currently used *manual evaluation* metrics.

We define the **metric gap** (or gap) as the distance between a prediction of a manual score, based on automatic scores, and the observed manual score.

The purpose of our work is to investigate and mitigate this metric gap by introducing an automatic evaluation that is a better predictor of manual evaluation. Reducing the metric gap is important for two reasons. First, the gap severely limits the usefulness of automatic evaluation and forces the use of much more expensive manual evaluation for comparing existing systems. More importantly, however, this gap is a severe handicap to research on new update summarization methods because it makes it difficult to evaluate new ideas and compare them with existing methods.

In Section 2, we analyze the results of the TAC 2008 summarization task, demonstrating the large gap between ROUGE automatic metrics and manual evaluation of update summaries. In Section 3, we modify ROUGE to produce scores that correlate significantly better with manual evaluation. We evaluate our new metric on TAC 2008 data in Section 4, demonstrating its superiority as a predictor of manual evaluations.

2. State-of-the-Art Evaluation of Update Summaries

TAC 2008 presented 48 20-document sets, with 10 documents in each of two subsets, A and B. Subset B documents were more recent. *Original* summaries were generated for the A subsets and *update* summaries were then produced for the B subsets.

In TAC 2008, ROUGE was used for automatic evaluation. ROUGE (Lin and Hovy 2000) compares any summary to any other (typically human-generated) summary using a recall-oriented approach. ROUGE-1 and -2 are based on unigrams and bigrams, respectively; ROUGE-SU4 uses bigrams with a maximum skip distance of 4 between bigrams; ROUGE-BE (Hovy, Lin, and Zhou 2005) is an *n*-gram approach based on *basic elements*, computed via parsing or automatic entity recognition. ROUGE-2, ROUGE-SU4, and ROUGE-BE were the official automatic metrics for TAC 2008, used to compare machine-generated summaries to human-generated summaries, and to compare human-generated summaries to each other using a jackknife approach. In addition to the three official metrics, we include ROUGE-1 in our study as it is often competitive with the official metrics.

2 NIST uses "model" for human-generated summaries and "peer" for machine-generated summaries.

Three manual evaluation metrics were used in TAC 2008: pyramid, overall responsiveness, and linguistic quality (not considered in our work). The pyramid method (Nenkova and Passonneau 2004) is a content-based metric for which human annotators mark *content units* in the human-generated summaries. The content units are collected across a set of human-generated summaries for a topic, and a weight is computed based on how many human-generated summaries include this content unit. TAC 2008 also used a manual overall responsiveness score. After evaluating data from 2005–2007 (Dang 2007; Conroy and Dang 2008), NIST decided that this score, which evaluates summary usefulness including linguistic quality, is a reliable and stable manual evaluation.

We analyzed the three official TAC 2008 *automatic* evaluation scores to see how well they predict the *manual* evaluation metrics of overall responsiveness and pyramid score. Figure 1 shows scatter plots of responsiveness and pyramid scores vs. the three official ROUGE measures for the TAC 2008 update task. Each solid data point represents the average score for a human summarizer over 24 document sets, and a dashed line marks the minimum; each open data point represents the average score for a machine system. We also show robust linear least squares fits to the data as well as the Pearson correlation coefficients between ROUGE-BE, -2, and -SU4 and the manual-evaluation scores. Surprisingly, these correlations are higher for the update task than for the original summarization task (data not shown); nevertheless, the *gap* between the lines' predictions and the scores for the human-generated summaries is larger.

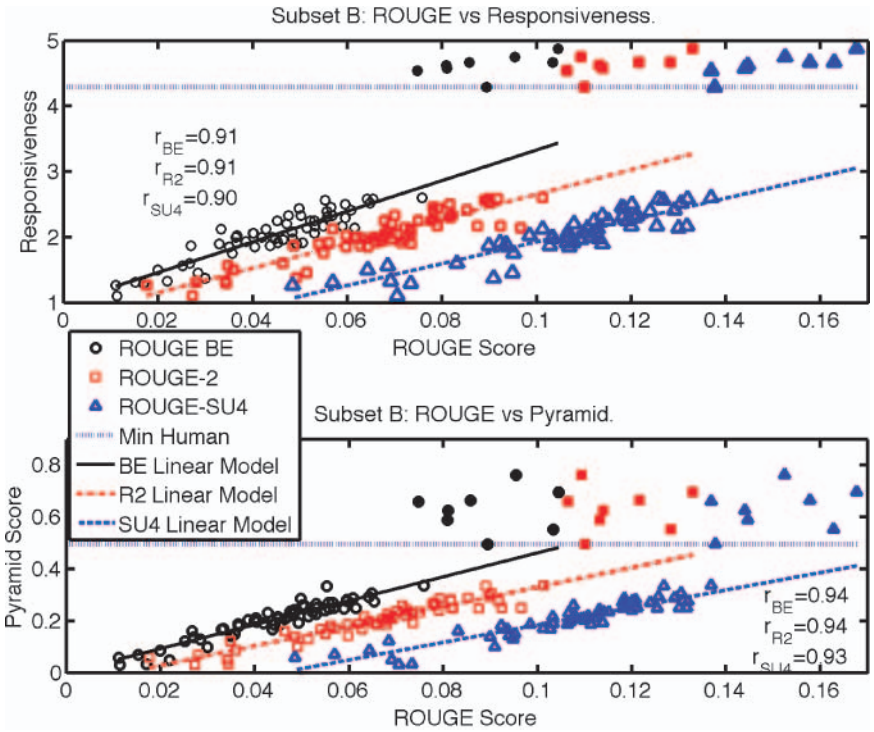

Figure 1
TAC 2008: The update task (Task B) responsiveness and pyramid scores vs. ROUGE scores; human-generated summaries (solid points) and machine-generated summaries (open points).

We report correlation coefficients only for the machine-generated summaries. Scores for the human-generated summaries are distributed differently, and correlation for the set of human-generated summaries is often not significant due to the small number of human summarizers.

The correlation coefficients in Figure 1 show that the automatic metrics do well in predicting responsiveness and pyramid scoring for *machine-generated* summaries. In contrast, the scores for human-generated summaries far exceed the predictions, with a large gap between predicted and actual scores. As may be expected, ROUGE is more highly correlated with the pyramid evaluation, which is a pure content evaluation score, whereas the responsiveness score also reflects linguistic quality.

3. Improving Automatic Evaluation—Nouveau-ROUGE

We more formally define the *metric gap* to be the absolute value of the difference between a manual evaluation score and our prediction of it based only on automatic evaluation scores. A number of TAC 2008 machine systems performed within statistical confidence of human performance in the automatic evaluation metrics, but no system performed near human performance in the manual evaluations. This has also been observed in previous summarization evaluations (Conroy and Dang 2008). Progress has been made in closing this metric gap but it persists, especially for update summaries.

A good update summary must contain essential information but focus on new information. When a machine-generated update summary is good, it is *similar to the human-generated update summaries*. This is assessed quite well by a ROUGE score. But the machine-generated update summary should also be *different from the human-generated original summaries*, and we need an automatic metric to assess this difference, or lack of redundancy. We suggest using a ROUGE score to measure similarity, and thus redundancy, between a given original summary and an update summary: A high ROUGE score indicates high redundancy.

To illustrate this, we used the CLASSY algorithm (Conroy, Schlesinger, and O'Leary 2006; Schlesinger, O'Leary, and Conroy 2008) to produce original summaries and update summaries for the TAC 2008 data. We also produced update summaries using a variant, projected-CLASSY, that reduces overlap by using a linear algebra projection of the term-sentence matrix (Conroy, Schlesinger, and O'Leary 2007) of candidate sentences against the matrix for the original (Task A) summary in order to favor new information. Table 1 gives average ROUGE-2 scores and 95% confidence intervals, computed via bootstrapping (Efron and Tibshirani 1993), over the 48 document sets. Two scores are given: $R_2^{(BB)}$ compares each CLASSY update (Task B) summary to the human-generated summaries, and $R_2^{(AB)}$ compares each to the original (Task A) model summaries. Whereas the two variants score comparably using $R_2^{(BB)}$, there is a significant difference in the $R_2^{(AB)}$ metric, as desired.

Table 1
TAC 2008: Average ROUGE-2 scores and 95% confidence intervals for update summaries produced by two variants of CLASSY.

Variation	$R_2^{(BB)}$	$R_2^{(AB)}$
projected-CLASSY	0.087 (0.080, 0.094)	0.075 (0.070, 0.079)
CLASSY	0.089 (0.082, 0.096)	0.083 (0.078, 0.088)

Table 2
TAC 2008: Nouveau-ROUGE α-parameters.

	Predicting Responsiveness			Predicting Pyramid Scores		
	$\alpha_{i,0}$	$\alpha_{i,1}$	$\alpha_{i,2}$	$\alpha_{i,0}$	$\alpha_{i,1}$	$\alpha_{i,2}$
R_1	−0.0271	−7.3550	13.4227	−0.2143	−1.9011	3.1118
R_2	0.9126	−5.4536	21.1556	−0.0143	−1.3499	4.3778
R_{SU4}	1.1381	−2.6931	35.8555	0.0346	−1.1680	7.2589
R_{BE}	1.0602	−5.0811	24.8365	0.0145	−1.3156	5.0446

Given this evidence, we propose predicting manual scores for update summaries by using two ROUGE scores, $R_i^{(AB)}$ and $R_i^{(BB)}$ ($i = 1, 2, SU4, ...$), in a three-parameter model called *Nouveau-ROUGE*:

$$N_i = \alpha_{i,0} + \alpha_{i,1} R_i^{(AB)} + \alpha_{i,2} R_i^{(BB)}$$

We determine the α parameters (Table 2) using robust linear regression on the TAC 2008 evaluation data so that the Nouveau-ROUGE score N_i best predicts the manual scores of responsiveness and pyramid performance.

Nouveau-ROUGE could be used by researchers to predict how a new system would compare with the TAC 2008 systems in overall responsiveness and pyramid scoring, a comparison that up to now has been impossible.

4. Evaluating Nouveau-ROUGE

We evaluate Nouveau-ROUGE using cross validation studies to demonstrate that if manual scores are available for a subset of summaries (in this case, those from 29 machine systems, half of those that participated in TAC 2008), then Nouveau-ROUGE can predict manual scores for the remaining (held-back) summaries.

4.1 Improved Correlation with Manual Evaluation

Correlation scores between automatic and manual scores have traditionally been used as a measure of the effectiveness of automatic evaluation as a surrogate for manual evaluation. Pearson correlation coefficients, shown in Table 3, were computed for the scores for the held-back subset of summaries. Correlation is indeed higher for the Nouveau-ROUGE scores than for any of the ROUGE scores.

Table 3
Correlation scores for TAC 2008 human evaluations.

	Average Responsiveness Score				Average Pyramid Score			
Metric	$i = 1$	$i = 2$	$i = \text{SU4}$	$i = \text{BE}$	$i = 1$	$i = 2$	$i = \text{SU4}$	$i = \text{BE}$
$R_i^{(AB)}$	0.676	0.576	0.619	0.490	0.698	0.592	0.634	0.483
$R_i^{(BB)}$	0.870	0.921	0.902	0.933	0.910	0.952	0.933	0.964
N_i	0.888	0.929	0.912	0.935	0.946	0.961	0.951	0.969

Figure 2
TAC 2008: ROUGE and Nouveau-ROUGE responsiveness and pyramid predictions for subtask B.

Table 4
TAC 2008: Median Pearson correlation coefficients for automatic vs. manual evaluations.

	Average Responsiveness Score				Jackknife Pyramid Score			
Metric	$R_i^{(AB)}$	$R_i^{(BB)}$	N_i	p-value	$R_i^{(AB)}$	$R_i^{(BB)}$	N_i	p-value
R_1	0.378	0.804	0.920	5.4e-284	0.406	0.837	0.943	3.5e-307
R_2	0.149	0.889	0.925	1.6e-104	0.177	0.909	0.941	1.8e-99
R_{SU4}	0.267	0.846	0.913	1.3e-176	0.291	0.875	0.933	8.7e-214
R_{BE}	0.222	0.913	0.919	6.2e-09	0.243	0.924	0.933	1.7e-17

Figure 2 shows that ROUGE-BE and ROUGE-1 predictions of both responsiveness and pyramid scores are inferior to the Nouveau-ROUGE-BE predictions. Plots for N_2 and N_{SU4} are omitted due to space restrictions, but performance improvement relative to ROUGE is greater than that for N_{BE} and less than that for N_1.

4.2 Validation Using Bootstrapping Experiments

To show that our results are not due to a lucky partitioning of the data, we used bootstrapping (Efron and Tibshirani 1993), a resampling method, which allows us to compute our statistical confidence in the results. This model assumes that observed data (scores for the 58 systems) characterize all data. Given this model, the proper sampling method is to choose subsets with replacement. We chose 58 systems (with replacement) and used half to determine the Nouveau-ROUGE parameters and half to test the model. We repeated this process 1,000 times. Table 4 gives the correlation coefficients (for the tested-half of the data) for all four ROUGE metrics with each of the manual evaluations when comparing the machine-generated summaries with the human-generated summaries. Data in the columns labeled "p-value" result from a Mann-Whitney *U*-test for equal medians of $R_i^{(BB)}$ and N_i of the distributions of correlations returned by the bootstrapping procedure. Because all p-values are small, we can conclude that the differences between the $R_i^{(BB)}$ and N_i correlation scores are statistically significant for all variants of ROUGE.

4.3 Narrowing the Gap for Update Summaries

Table 5 gives the median gap on the TAC 2008 data for predicting responsiveness and pyramid scores. Recall that the gap is the absolute value of the difference between the

Table 5
Narrowing the TAC 2008 metric gap.

	Responsiveness Metric Gap			Pyramid Metric Gap		
Metric	*R* Gap	*N* Gap	p-value	*R* Gap	*N* Gap	p-value
R_1	2.025	1.277	7.8e-03	0.285	0.187	7.8e-03
R_2	1.655	1.518	7.8e-03	0.241	0.197	7.8e-03
R_{SU4}	1.887	1.344	7.8e-03	0.273	0.206	7.8e-03
R_{BE}	1.591	1.547	7.8e-03	0.229	0.206	7.8e-03

manual score and the prediction of it. The median gap is always smaller for Nouveau-ROUGE than for ROUGE; in fact, the gap is smaller on every trial.

We used the Wilcox sign test to test the significance of this observation. The null hypothesis is that the differences in the gaps between ROUGE and Nouveau-ROUGE has median 0. The p-values from the Wilcox test indicate that the null hypothesis is true with probability $\frac{1}{128} \approx 7.812 \times 10^{-3}$, so it can be safely rejected.

5. Conclusion

Our new metric, Nouveau-ROUGE, includes a measure of novelty for an update summary. We demonstrated that it has higher correlation to manual evaluation of overall responsiveness and to pyramid scores than does ROUGE. The most obvious deficiency in ROUGE is the lack of a *linguistic quality* measurement, which we take to encompass all language-related issues: lexical, syntactic, and semantic. Therefore, we conjecture that most remaining prediction error in Nouveau-ROUGE is a result of omitting linguistic quality and caution that better prediction would be achieved only for systems of comparable linguistic quality.

We believe that responsiveness is an imperfect surrogate for *task-based* summary evaluation such as that done in SUMMAC (Mani et al. 1999). We would welcome a return to task-based evaluation, as well as research increasing the reliability and consistency of manual evaluation metrics. Investigation could also quantify the impact of low responsiveness and pyramid scores on the ability to perform a specific task.

References

Conroy, John M. 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.

Conroy, John M., Judith D. Schlesinger, and Dianne P. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 152–159, Sydney.

Conroy, John M., Judith D. Schlesinger, and Dianne P. O'Leary. 2007. CLASSY 2007 at DUC 2007. In *Proceedings of the Seventh Document Understanding Conference (DUC)*, Rochester, NY. Available at `http://duc.nist.gov/pubs.html#2007`.

Dang, Hoa Trang. 2007. Overview of DUC 2007. In *Proceedings of the Seventh Document Understanding Conference (DUC)*, Rochester, NY. Available at `http://duc.nist.gov/pubs.html#2005`.

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Hovy, Eduard, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating DUC 2005 using basic elements. In *Proceedings of the Fifth Document Understanding Conference (DUC)*, Vancouver. Available at `www-nlpir/nist.gov/projects/duc/pubs.html`.

Lin, Chin-Yew and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 495–501, Morristown, NJ.

Mani, Inderjeet, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette Hirschman. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of EACL'99: Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–85, Bergen.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–152, Boston, MA.

Schlesinger, Judith D., Dianne P. O'Leary, and John M. Conroy. 2008. Arabic/English multi-document summarization with CLASSY—The past and the future. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*. Springer, Haifa, pages 568–581.