CLASSY and TAC 2008 Metrics

John M. Conroy Judith D. Schlesinger
IDA/Center for Computing Sciences
{conroy, judith}@super.org

Abstract

We present results of CLASSY's submissions to TAC 2008 update and blog opinion summarization tasks. Additionally, we evaluate and analyze many of the results for all of the systems participating in each of these tasks.

1 Introduction

CLASSY (Clustering, Linguistics, and Statistics for Summarization Yield) is the summarization system developed mainly[1] by IDA/Center for Computing Sciences. We participated in both the update summary and the opinion summary tasks for TAC this year. This paper will discuss innovations we made for both of these tasks. Incorporated with this will be a metric retrospective of both our own and all submissions to these tasks.

We compute a statistical analysis of the ROUGE, pyramid, and overall responsiveness metrics. We will present results on all TAC systems testing to see if, indeed, complete sentence summary generation gives overall higher pyramid and responsiveness scores. In addition, we compute the correlations between the various human and automatic metrics.

2 CLASSY 2008

CLASSY 2008 is quite similar to CLASSY 2007 [3] and consists of the following six steps:

1. Data preparation and sentence trimming.
2. Query term selection from the topic descriptions.
3. Signature term computation for each of the document sets.
4. Sentence scoring using the approximate oracle.
5. Redundancy removal using the LSI/L1-QR algorithm.

[1] Over the years we've collaborated with several colleagues from the DoD and the University of Maryland.

6. Sentence ordering based on an approximate TSP algorithm.

Specific modifications made to any of these tasks, for either of the tasks, will be described where appropriate.

3 Update Task

[1] reported that for the DUC 2007 *main* task, systems that did not truncate the final sentence of the summary, i.e., used only complete sentences, received, as a group, higher "content responsiveness" scores than systems that truncated. At the same time, there was no significant difference in the ROUGE scores for both these groups. Conversely, an analysis of the DUC 2007 *update* task reveals that summaries ending with a fragment received higher scores for both ROUGE and content responsiveness.

Table 1 shows the average ROUGE and content responsiveness scores for the systems participating in the DUC 2007 update task. The mean system performance was computed conditioned upon ending summaries with a fragment versus ending with a complete sentence. A system was judged to use fragments if at least half of its summaries ended in a fragment. For the DUC 2007 update task, just *over* 50% of the submissions ended their summaries at least half the time with a fragment. Notice that with the exception of ROUGE-BE, the ROUGE and content responsiveness differences are statistically significant since the p-values (significance) are less than 0.05, the accepted threshold of significance. There is *no* significance between the pyramid scores for the 2 sets of systems.

Table 1: DUC 2007 Update Task: Average scores of systems ending with a complete sentence vs. ending with a fragment.

Metric	Sentence	Fragment	p-Value
ROUGE-BE	0.035	0.046	0.084
ROUGE-2	0.064	0.083	0.018
ROUGE-SU4	0.102	0.122	0.006
Pyramid	0.125	0.133	0.786
Content Responsiveness	2.146	2.482	0.019

CLASSY has always truncated summaries to the exact required length irrespective of sentence boundaries. For 2008, we decided to test the consistency of the observed behavior. To do this, we modified our summary generation to optionally produce complete sentences.

For our priority 1 run (system 6), we used a "greedy bin packing algorithm" to select the final sentence of each summary. The goal of greedy bin packing is to find a high scoring sentence of the correct length to fill out the summary to the desired length. Testing the method on DUC 2007 update data showed that, in general, summaries could consist of

complete sentences and suffer only minor reduction in ROUGE scores. Our priority two run (system 37) consisted of the corresponding truncated summaries.

For the TAC 2008 update task, just *under* 50% of the submissions ended their summaries at least half the time with a fragment. Table 2 shows the average scores conditioned on systems which end with fragments versus those that end with a complete sentence. The differences are small to nonexistent and none of the differences are statistically significant since all the p-values (significance) are greater than 0.05. Only the linguistic evaluation shows full sentences with a slightly higher (but still not significant) score and this is expected.

Responsiveness dropped dramatically from 2007 to 2008. For DUC 2007, systems were judged based on *content responsiveness* while for TAC 2008, systems were judged based on *overall responsiveness*, which includes, in part, a measure of linguistic quality. Therefore, the two scores cannot be directly compared.

Our entries, systems 6 and 37, corresponding to ending with a complete sentence and ending with a fragment, respectively, followed the overall trend in that the content scores and overall responsiveness were slightly (but not significantly) higher when the summary ended with a fragment and these summaries, on a whole, had lower linguistic scores.

Table 2: TAC 2008 Update Task: Average scores of systems ending with a complete sentence vs. those ending with a fragment.

Metric	Sentence	Fragment	p-Value
ROUGE-BE	0.044	0.044	0.976
ROUGE-2	0.072	0.073	0.761
ROUGE-SU4	0.110	0.111	0.664
Linguistic	2.380	2.298	0.457
Jacknife Pyramid	0.232	0.240	0.658
Overall Responsiveness	2.175	2.172	0.967

In summary, for the update task and the main task of DUC 2007, systems did not suffer a significant decrease in ROUGE scores by ending their summaries with a complete sentence. Unlike the main task of 2007, there was no advantage in the responsiveness score for the update systems (either DUC 2007 or TAC 2008). This is remarkable in that for the main task in 2007 the responsiveness metric was "content"-based and for TAC 2008 it was "overall", the former of which should not be greatly affected by linguistic quality while the latter one is. These results further emphasizes that the responsiveness score, both the overall and content varieties, is an inconsistent measure.

We believe that responsiveness should be a surrogate for a *task based* summary evaluation such as that done in SUMMAC [4]. We would welcome a return to a task based evaluation of summaries and, more generally, to see further research that would give more reliable human evaluation metrics.

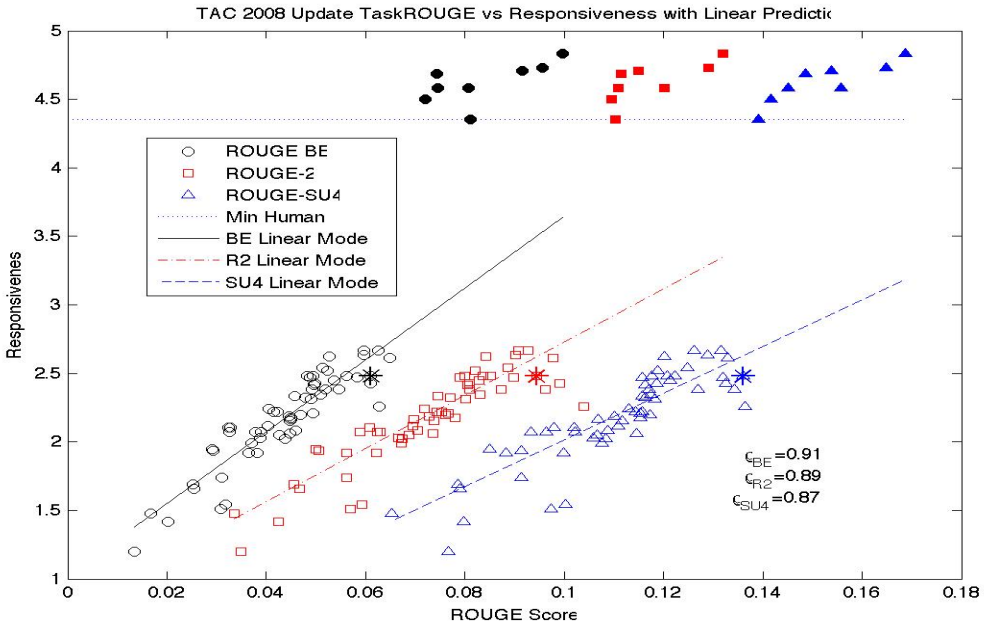

Figure 1: Scatter Plot of ROUGE and Responsiveness Scores for Update Task

We utilized the third submission to explore the robustness of the summaries based on the background used. For this submission, we used AQUAINT 1 (used for DUC 2007) for the background model instead of AQUAINT 2 (used for TAC 2008). We wanted to measure to what extent having a slightly different background model would affect the quality of the summaries. This submission also ended summaries with a fragment. System 60 is the CLASSY priority 3 submission.

The ROUGE-BE, ROUGE-2, and ROUGE-SU4 results of system 37 and system 60 (that both used truncation) were nearly identical. This would seem to indicate that if the background corpus is large enough and reasonably similar, it does not really make too much of a difference what corpus is used, which is an encouraging result.

Our last observation is that humans greatly outperform the machine systems in pyramid scores and overall responsiveness, while systems are approaching human ROUGE performance. Figure 1 gives a scatter plot of the 3 "official" ROUGE measures versus overall responsiveness. CLASSY is among the systems which produce summaries that generate near human ROUGE performance. The "*" in the plot is system 37. A Tukey honestly significant test shows that only 3 humans have, for example, a ROUGE-2 score which is significantly higher than system 37. See Figure 2.

The approximate oracle score, our sentence selection algorithm, [2] is a very strong

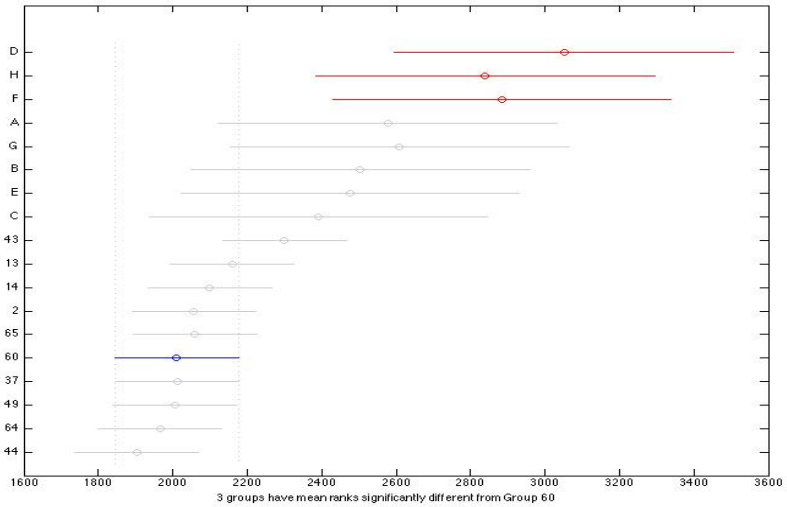

Figure 2: Tukey multi-comparison test for ROUGE-2 Scores

ROUGE-1 approximation. In fact, for the update task, all three of the CLASSY runs scored higher than human E and ranked 1, 2, and 3 among machine systems. See Table 3. ROUGE-1 has a 0.85 correlation with overall responsiveness compared with ROUGE-BE, which has a correlation of 0.91. Clearly, a new approximate oracle based on ROUGE-BE needs to be considered. It will be interesting to see if, by modeling the ROUGE score with a stronger human correlation, we can narrow the gap in the human evaluation between human summaries and system summaries.

System 37 scored 4th in the pyramid evaluation among machine systems. Furthermore, all 3 entries scored within the top 11 systems. A non-parametric ANOVA test indicates that there is no statistically significant difference in the median pyramid scores of the top 15 systems, which, unfortunately, includes baseline 1. See Figure 3. No system comes close to the human pyramid performance as illustrated in Figure 4. Note, too, that ROUGE correlates better with pyramid scoring than it does with human scoring; however, it still has a "metric" gap in that systems approaching human ROUGE performance still miss the human pyramid performance.

4 Opinion Task

The opinion task introduced many data processing challenges, independent of the summarization task. Identifying the portion of the blog text that is relevant to the topic was

Table 3: Highest ROUGE-1 scores with lower and upper 95% confidence level bounds.

System	ROUGE-1	Lower	Upper
D	0.43275	0.41974	0.44598
H	0.42253	0.40931	0.43457
F	0.41993	0.40319	0.43643
G	0.41823	0.40426	0.43286
A	0.40993	0.39426	0.42490
B	0.39408	0.38046	0.40811
C	0.39237	0.37267	0.41072
60	0.38313	0.37773	0.38856
37	0.38051	0.37483	0.38603
6	0.37448	0.36919	0.37981
E	0.37344	0.35945	0.38926
43	0.37297	0.36666	0.37975

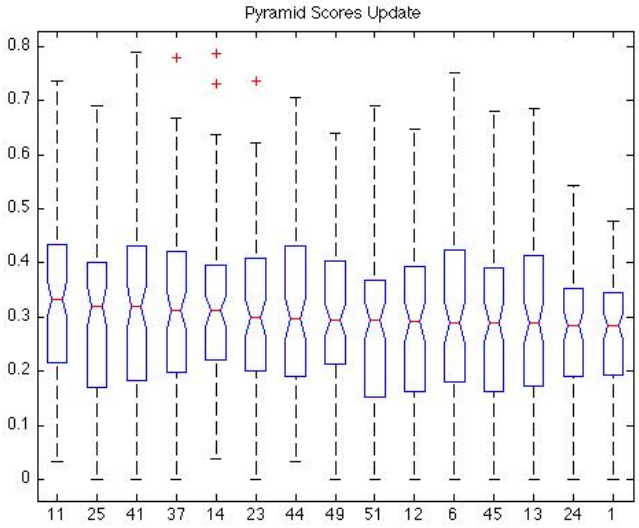

Figure 3: Pyramid Scores for Update Task

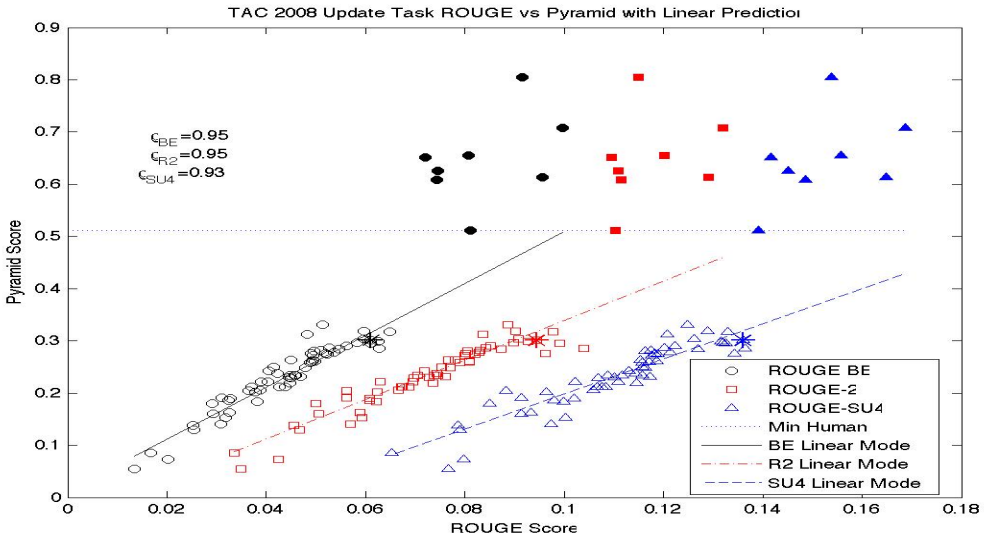

Figure 4: Scatter Plot of ROUGE and Pyramid Scores for Update Task

a major effort that is not yet sufficiently resolved. Unlike newswire, blog pages contain much, and sometimes nearly all, text that is totally unrelated (ads, personal data, etc.). Additionally, punctuation and capitalization are often nearly non-existent, rendering sentence splitting and trimming algorithms all but useless. Using the blog training data, we were partially successful in identifying which text sections should be used to generate the summaries. We were also able to identify necessary changes to handle this data although many of these modifications remain to be implemented.

We again submitted three experiments, based on variations in the use of the nugget information. We wanted to test the hypothesis that nuggets returned by the QA systems would help improve summaries. CLASSY scores sentences based on two features: query terms and signature terms. Query terms were extracted from the squishy question list. The nuggets were optionally used to help determine the signature terms:

1. Signature terms based on nuggets (system 5).
2. Signature terms based on document clusters (system 36).
3. Signature terms based on both document clusters and nuggets (not evaluated by NIST).

In each case, an approximately 250 word summary was generated, consisting of complete sentences. As the length of the requested summary was not required to be 250 words, the

greedy bin packing algorithm described in Section 2 for choosing the final sentences was not used for this task. We chose this length to remain consistent with our experience from DUC 2005–2007, which had similarly complex questions and had a target length of 250 words.

The blog data posed a surprising challenge to generate non-redundant summaries. The type of redundancy was not quite the same as seen in newswire data. The non-negative QR factorization used by CLASSY selects sentences based on their approximate oracle score and the approximate distance of the sentence to a subspace of the currently chosen sentences. It favors shorter sentences. The problem observed in the blog training data is that often, pairs of sentences exist where a longer sentence "subsumes" the content of a shorter sentence. The sentence selection algorithm was adapted to minimize this effect. With this change, the redundancy level was greatly reduced.

Table 4 gives the rank, in each of 6 metrics used to evaluate this task, for the 4 submissions with shortest summary length (which includes both CLASSY submissions) along with the top 4 scoring systems. (Note that ties in rank are handled in the usual way by replacing assigning the average of the ranks for which there was a tie.) We computed the length of each system's summaries. The lengths were converted to ranks (with shortest having rank 1) to compute a Spearman correlation between summary length and each of the human evaluation metrics.

The top scoring systems were, on average, about 4.4 times as large as the CLASSY summaries. The top-scoring systems did *not* generate the longest summaries—indeed, half the systems generated summaries longer than any of these 4. It is, therefore, reasonable to assume that at some point the summaries can get too long and adversely affect their scores. Not surprisingly, CLASSY did extremely well (often best) in grammar, non-redundancy, coherence, and fluency; indeed, shorter summaries are both easier to read and to make readable.

Table 4: TAC 2008 Opinion Blog Track: Rank of CLASSY Submissions vs. Top Performing Systems.

System	Length	Pyramid	Gram.	Non-Red.	Coher.	Fluency	Responsiveness
13	10	1	25	24	7	9	4
9	18	2	10.5	29	9	11	1
32	17	3	13.5	27.5	19	19.5	2
2	16	4	27.5	26	16	11	3
15	1	9	4	3	1.5	3	19
5 (Priority 1)	3	22	2	1	10	2	29
7	4	24	17	14	17	7.5	32.5
36 (Priority 2)	2	32	1	2	1.5	1	32.5

Table 5 shows the correlation between summary length and the 6 metrics for all systems.

Table 5: Correlation between length and responsiveness, pyramid, and grammaticality

	Pyramid	Gram.	Non-Red.	Coher.	Fluency	Responsiveness
Correlation	0.2918	0.2241	0.3939	0.6870	0.5282	-0.3298
p-Value	0.0842	0.1889	0.0174	0.0000	0.0009	0.0495

There is significant *negative* correlation between length and overall responsiveness, i.e., the shorter the length of the summary, the lower its overall responsiveness score. There is very strong and significant correlation between the length of a summary and non-redundancy, fluency and coherence. CLASSY's scores in these categories are at least in part due to the shorter summaries. Note, conversely, that for grammar, the correlation is not significant. Therefore, CLASSY's top performance in this metric is not entirely due to the summaries being short. Most surprising, perhaps, is that there is no significant correlation between summary length and pyramid score.

Lastly, Table 4 shows that for the CLASSY submissions, using the nuggets returned by the QA systems as the basis for the signature terms (System 5) most definitely improved the pyramid score although the responsiveness score was only minimally affected. We hypothesize that systems that used the full nuggets to retrieve text for the summaries not only did quite well with the pyramid evaluation (and, possibly, responsiveness) but did not have the difficulty we did in identifying the relevant text in the blog files. We await hearing about the other systems to see if this hypothesis is validated.

5 Conclusions and Future Work

Automatic summarization is not yet a solved problem although there has been great progress, especially since the advent of DUC/TAC. Much work remains to be done to ensure CLASSY is as good as it can be.

A major stumbling block in improving performance remains with the limitations in the current automatic metrics. ROUGE has been the standard bearer in automatic evaluation and has allowed systems to tremendously improve the level of relevant content in machine generated summaries. However, it is not sufficient to model more sophisticated manual content measures such as the pyramid metric, much less overall responsiveness which also reflects the linguistic quality of a summary. Our group would welcome a return to task based evaluation metrics. Ongoing research by many groups is seeking to address this challenging problem.

With this said, ROUGE can still be used to leverage further improvements. Based on the ROUGE/human correlations mentioned earlier and our success with ROUGE-1 scores, it is reasonable to assume that an approximate oracle based on basic elements (ROUGE-BE) would significantly improve our pyramid and content responsiveness scores.

Developing such an algorithm is a major priority for the future.

If we continue to work with blog data, we must greatly improve our relevant text extraction and the ability to sentence split and trim poorly formed (both lexically and syntactically) sentences. We also need to refine our redundancy removal capabilities.

6 Acknowledgments

The authors would like to thank Dianne P. O'Leary of the University of Maryland for her thoughtful suggestions and comments.

References

[1] John M. Conroy and Hoa Trang Dang. 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.

[2] John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. Topic-Focused Multi-document Summarization Using an Approximate Oracle Score. In *Proceedings of the ACL'06/COLING'06*, pages 152–159, Sydney, Australia, July 2006.

[3] John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. CLASSY 2007 at DUC 2007, http://www-nlpir.nist.gov/projects/duc/pubs.html, 2007.

[4] Inderjeet Mani, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette-Hirschman. The TIPSTER SUMMAC Text Summarization Evaluation. In *In Proceedings of EACL'99: Ninth Conference of the European Chapter of the Association for Compuational Linguistics,*, pages pages 77–85, Bergen, Norway.