CLASSY 2007 at DUC 2007

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

Judith D. Schlesinger
IDA/Center for Computing Sciences
judith@super.org

Dianne P. O'Leary
University of Maryland
oleary@cs.umd.edu

Abstract

The IDA/CCS summarization system, CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield), was enhanced in several areas for this year's DUC. Our sentence splitting and trimming algorithms continue to be improved. Signature terms were improved by using the AQUAINT data as the background. Redundancy removal was also considerably improved employing LSI and a new variant of QR. We proposed a new way to determine paragraph breaks. In addition, a sub-cluster redundancy removal method was developed to tackle the update summary task. We summarize our results and analyze the relationship between ROUGE scores and responsiveness.

1 Introduction

The IDA/CCS summarization system, CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield), was enhanced in several areas for this year's DUC. In this report, we focus on the changes in the system, the results, and the relationship between ROUGE scores and responsiveness.

2 CLASSY 2007

CLASSY 2007 is quite similar to CLASSY 2006 [4] and consists of the following six steps:

1. Data preparation/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.
6. Sentence ordering based on an approximate TSP algorithm.

Document Preparation

The document preparation step begins with sentence splitting. We continue to use the sentence splitter introduced for DUC 2006, with additional error corrections made to compensate for erroneous splits (either splitting where inappropriate or not splitting where needed). Sentences are then tagged as one of 1) candidates for inclusion in a summary; 2) non-candidates that might provide useful terms; and 3) non-candidates to be ignored.

Our sentence trimming routines continue to evolve. We continue to perform all trimming without using any POS-tagging or parsing. Patterns, based on punctuation and specific word use, e.g., "who(m)", "which", "when", and "where" for relative clause appositives, and extensive usage checking, are used to ensure that only good clauses and phrases are removed. Improvements made since DUC 2006 include finding trims that were previously missed and avoiding trims that turned out to be erroneous. In addition, for attribution removal, in many cases we can now replace a pronoun with the appropriate noun that is being trimmed. See [4] and prior DUC papers for more detail on our sentence splitting and trimming algorithms.

In anticipation of the update task for DUC 2007, we investigated word usage in the DUC 2006 documents to see if there was any significant use of phrases that would indicate new or updated information from previous documents. Specifically, we evaluated the documents for phrases such as "(in) further—later—new—... developments", "updated—new—recent—... figures—data—...", "previous(ly)", and "recent(ly)". We were a bit surprised to find *almost no* evidence of any linguistic clues to changes in information in the DUC 2006 document set. This finding helped guide us to the kinds of modifications we needed to make to CLASSY for the update task.

Query Term Selection

Query terms were selected using the same procedure used in 2006 [4] except that the query terms (as well as all the terms in the document) were stemmed using the Porter stemmer [9]. Experimentation on old data determined that better ROUGE scores were achieved with stemming.

Signature Term Computation

A more inclusive set of signature terms were considered for 2007. Loosely, signature terms are those terms which occur significantly more than expected "at large" ([5]). When signature terms were first proposed for summarization by Lin and Hovy ([7]), "at large" was defined as the other documents in the small corpus, i.e., the document clusters for the DUC task. In contrast, for the Multi-Lingual Summarization Evaluation 2005 ([2]), as well as MSE 2006 and DUC 2006, we used the TREC Novelty data and found that using a larger background corpus gave superior signature terms and better summaries.

For DUC 2007, we expanded the set from which signature terms were chosen in two ways. First, we used all of the AQUAINT data for the background. Second, we did not use a stop list. The AQUAINT collection consists of about 500 billion terms, with over 361 times as many terms as in the TREC Novelty data. The stop list was removed based on testing which indicated that ROUGE scores increased when stop words were included in the signature term computation, once the larger background was used.

Sentence Scoring

Sentence scoring in CLASSY is an approximate oracle score ([3], [4]), where sentences are scored based on an approximation of the fraction of terms a sentence has in common with unknown human model summaries. A minor change in the formula in the oracle score was shown to give significant improvement. The score is built upon an estimate of the probability that a term t will be included in a human summary given a topic, τ. This probability is denoted $P(t|\tau)$. It is approximated using the query terms, signature terms, and the distribution of the terms in the relevant document cluster. Specifically, we estimate our target probability by a weighted average of three distributions: the impulse distributions for the query and signature terms and the probability that a term occurs in the sentences to be considered for extraction. Our estimate was:

$$P_{qs\rho}(t|\tau) = \frac{1}{4}q_t(\tau) + \frac{1}{4}s_t(\tau) + \frac{1}{2}\rho_t(\tau)$$

where $s_t(\tau)$=1 if t is a signature term for topic τ and 0 otherwise, $q_t(\tau) = 1$ if t is a query term for topic τ and 0 otherwise, and $\rho_t(\tau)$ is the maximum likelihood estimate of the probability that term t occurs in a sentence in the topic τ. Note that a maximum likelihood procedure could be used to give sharper

estimates for the weighting of these distributions. However, we observed no significant improvement when such an estimate was used, so we opted for the above fixed weighting.

Based on the above estimate, the approximate oracle score for sentence x is computed as

$$\omega_{qs\rho}(x) = \frac{1}{|x|} \sum_{t \in T} x(t) P_{qs\rho}(t|\tau),$$

where $x(t) = 1$ if the sentence x contains the term t and 0 otherwise. We pass to the redundancy removal step sufficient top-scoring sentences to make a summary 6 times the target length.

Redundancy Removal

In CLASSY 2006 for DUC, pivoted-QR was used for redundancy removal. CLASSY 2007 for DUC used latent semantic indexing(LSI, [8]) followed by an L1-norm variant of a pivoted-QR, which is more fully described in [1]. We truncate the summary when the last sentence added brings the length greater than or equal to the target length.

Sentence Reordering

CLASSY models the sentence ordering for output as a Traveling Salesman Problem, finding a shortest path among the sentences where term overlap is used to measure the similarity of sentences. A Monte-Carlo method is used to approximate the solution of this NP-hard problem to produce an ordering of the sentences which minimizes discontinuities in flow. This method was described in [4].

Two changes were made. First, we fix the last sentence in the summary instead of the first sentence. This change was made to insure that the lowest scoring sentence would be truncated when the summary was too long. Second, we experimented with using the similarity metric to define summary layout by inserting a paragraph break when a significant change in term overlap was found.

3 CLASSY Submissions for DUC 2007

3.1 Main Task Submission

For the main DUC task, a continuation of the DUC 2006 task, we used CLASSY as described in Section 2 with no variation from that description.

In preparation for DUC 2007, CLASSY 2007 was benchmarked against CLASSY 2006, human performance, and the oracle score (in contrast to the approximate-oracle score that is actually used in CLASSY) on the DUC 2006 data. The oracle score uses $P(t|\tau)$, the empirical probabilities based on the human abstracts, and represents a theoretical maximum performance of the sentence scoring approach. We see from Table 1 that CLASSY 2007 is an improvement over CLASSY 2006 in three ROUGE measures and achieves performance that is statistically indistinguishable from the low-end of human performance. As a measure of how much the linguistic pre-preprocessing improves ROUGE scores we benchmarked *CASSY07* and *CASSY06*, which are CLASSY07 and CLASSY06 *without* linguistic pre-processing. With but one exception, these systems are significantly worse in each of the ROUGE scores (as measured by the 95% confidence intervals) than the corresponding system which includes the linguistic pre-processing. Note, too, that significant "boilerplate", especially datelines, was included in the selected sentences. This leads us to believe that the linguistic scores (Q1–Q5) would also have suffered.

3.2 Update Task Submission

Our submission for update task A used the same procedure as for the main task. For update tasks B and C, we conjectured that an update summary should be written based on extracted information from only the documents in the sub-clusters of B and C respectively. We used all the documents available for

Table 1: Main Task Preparation Using DUC 2006 Data

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Max Human	0.479	0.133	0.184
Oracle	0.473	0.119	0.182
Min Human	0.436	0.104	0.160
CLASSY07	0.433	0.101	0.160
CASSY07	0.422	0.095	0.153
CLASSY06	0.404	0.091	0.148
CASSY06	0.400	0.081	0.141

a sub-cluster to compute signature terms. For example, for task B of the update summary, signature terms were based on the 10 documents in the A sub-cluster as well as the 8 in the B sub-cluster while sentences for the task B summary were extracted only from the 8 documents in the B sub-cluster.

To score the sentences within the documents in sub-clusters B and C, an orthogonal projection was use to project out the term subspace covered by a previous summary. The scores were then computed based on the length of the projected term vector corresponding to the sentence.

Testing the update algorithm was a challenge as there was no sample dataset. We opted to take the DUC 2006 data and carve it into sub-clusters of size 10, 8, and 7, for sub-clusters A, B, and C, respectively, based on date. Unfortunately, the human summaries were 250 words and were based on the entire document set. As a result, our evaluation method compared three 100 word summaries to the same 250 word human summaries. We tested the above approach with and without the orthogonal projection and found no significant difference in ROUGE scores. We opted to include the orthogonal projection as it perhaps would focus on the new information.

4 Results

4.1 Main Task

For the main task, CLASSY 2007, system 24, scored significantly higher than any other system in ROUGE-1. Also, this score, as well as all of the ROUGE scores for CLASSY, were within the 95% confidence intervals of one or more human summarizers. On human evaluations for the main task, CLASSY scored an average content responsiveness of 3. Figure 1 gives a scatter plot of the three main ROUGE scores for all systems as well as humans. The plot shows that ROUGE-2, SU4, and BE correlate well with the responsiveness score; however, there is a wide gap between human and machine systems in responsiveness. The top scoring system scored a 3.4 in responsiveness. All systems scored significantly below human summarizers and no system was significantly different in responsiveness (as measured by Tukey's honestly significantly different test) than CLASSY 2004, baseline 2.

Furthermore, the three linear least squares fits of the ROUGE scores and responsiveness, displayed in Figure 1, can be used to extrapolate what system performance would be if ROUGE scores were to increase to that of the highest scoring human summarizer. Sadly, the linear fits predict that even for a ROUGE score matching that of the top human summarizer, the corresponding responsiveness score would still not meet that of the lowest scoring human summarizer! (That is, no linear extrapolation for the three ROUGE scores crosses the dotted line.)

Figure 2 gives a box plot of the subset of the systems evaluated by using the pyramid evaluation method. CLASSY 2007 scored third and the ANOVA test indicates that there is a significant difference in the means. This difference is due to baseline 1 and system 6. Baseline 2 is grouped with the other systems as measured by a Tukey honestly significantly different test.

We now turn to the linguistic questions, with an eye towards understanding areas where improvement is needed. Table 2[1] displays the results of CLASSY 07, CLASSY 04 (Baseline 2), and those systems

[1]Thanks to Guy Lapalme and Fabrizio Gotti for the Excel sheet which we used to generate this table.

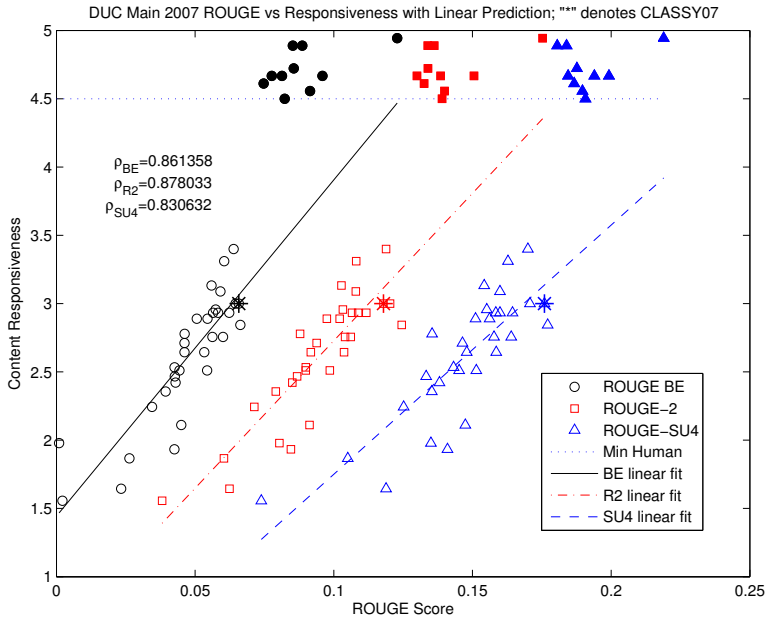

Figure 1: Scatter Plot of ROUGE Scores and Responsiveness for Main Task

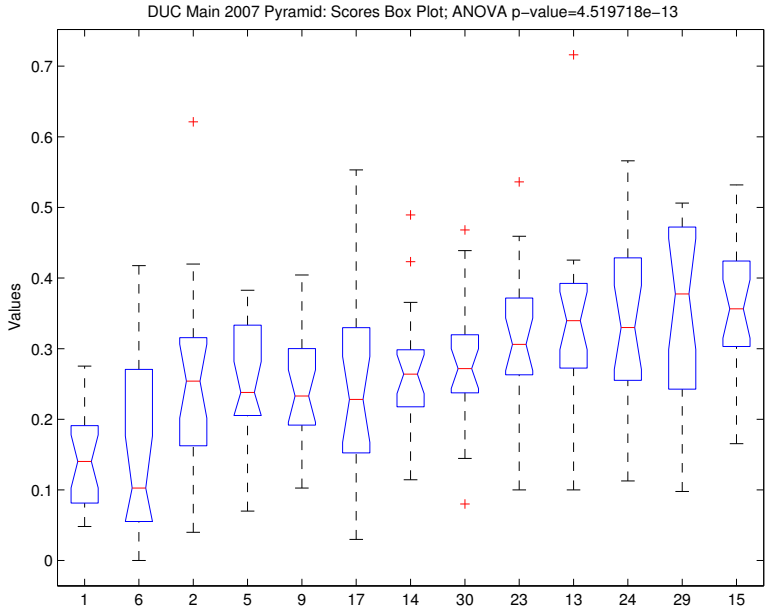

Figure 2: Box Plot for Pyramid Scores for Main Task; Systems Sorted by Mean; Notch in Plot Indicates the Median

which scored higher in responsiveness than CLASSY 07, along with their linguistic scores. We observe that, in general, systems with better average responsiveness scores also did better in linguistic quality while most did not do as well in ROUGE-2 scores.

To explore if this effect was held more broadly on the the data, we computed the correlation scores between the linguistic question scores and responsiveness. Q1, Q3, and Q4 seem to account for why ROUGE-2 is lacking at predicting the order of systems for average responsiveness: systems that had a high ROUGE-2 score but a relatively low score in these questions were penalized in responsiveness. Indeed, the correlation coefficient between Q4 (focus) and average responsiveness is 0.7. This suggests that systems could improve responsiveness by improving focus. Similarly, Q1 (grammaticality) and Q3 (referential clarity) have correlation coefficients of about 0.6. All these coefficients were statistically significant. (Not surprisingly, no linguistic questions correlated significantly with ROUGE-2).

Table 2: Linguistic Questions, ROUGE 2, and Responsiveness

System	Q1	Q2	Q3	Q4	Q5	ROUGE-2	Avg Resp.
4	4.58	3.73	3.60	3.98	3.22	0.119	3.400
23	4.64	3.89	4.09	4.24	3.69	0.108	3.311
14	4.40	3.51	3.98	3.64	2.82	0.103	3.133
7	3.22	3.33	3.98	3.42	2.27	0.108	3.089
CLASSY07	3.93	3.67	3.16	3.64	2.67	0.118	3.000
29	3.24	3.64	2.71	3.02	2.07	0.120	3.000
15	3.22	3.31	3.76	3.51	2.64	0.124	2.844
CLASSY04	3.84	4.11	3.58	3.44	2.82	0.0938	2.711

4.2 Update Task

In the update task, CLASSY 2007 was system 44, and the baselines were systems 57 and 58 corresponding to lead and CLASSY 2004, respectively. As this was a pilot task, there were only 10 document sets, so Figures 3, 4, and 5 are understandably much noisier than Figure 1. CLASSY 2007's performance was consistently good in ROUGE on the three sub-tasks, but very sporadic in responsiveness. In particular, CLASSY 2007 scored the highest in responsiveness for task B. Both humans and machine summarizers have a noted drop off in performance, although this, too, may be a reflection of the size of the data set. In the pyramid evaluation, CLASSY 2007 ranked third. Repeating this task next year might reveal a clearer trend.

5 Conclusions

As a result of DUC 2007, we conclude that CLASSY 2007 is a very strong system for extracting content for multi-document summarization. For the main task, CLASSY, as well as several of its peers, achieved ROUGE performance which was statistically indistinguishable from humans. CLASSY also achieved pyramid scores which were among the best systems. On the other hand, CLASSY and the other systems still dramatically lag behind humans in responsiveness. A preliminary analysis of the linguistic questions leads us to the hypothesis that if more attention were paid to focus, grammaticality, and referential clarity then responsiveness may improve. We aim to investigate automatic methods for scoring these linguistic properties (e.g., [6]), to improve the overall quality of a summary.

The update-summary task is a valuable addition to DUC. The adaptation of CLASSY 2007 to this task shows much promise and with the data provided by DUC 2007 we can now more seriously evaluate a number of variations on the methods we employed. Here, too, we will aim to measure linguistic quality automatically and perhaps begin to close the gap between the ROUGE indication of performance of machine systems and their responsiveness for this task.

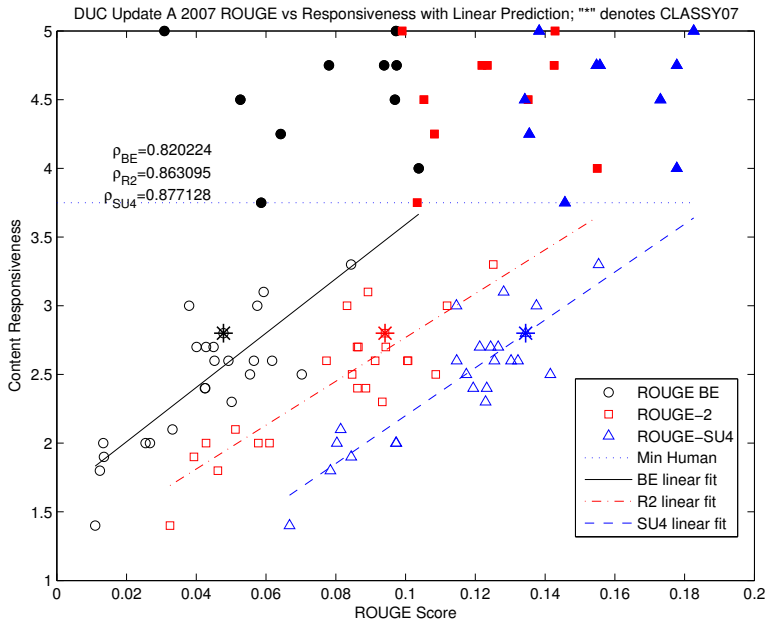

Figure 3: Scatter Plot of ROUGE Scores and Responsiveness for Update A Task

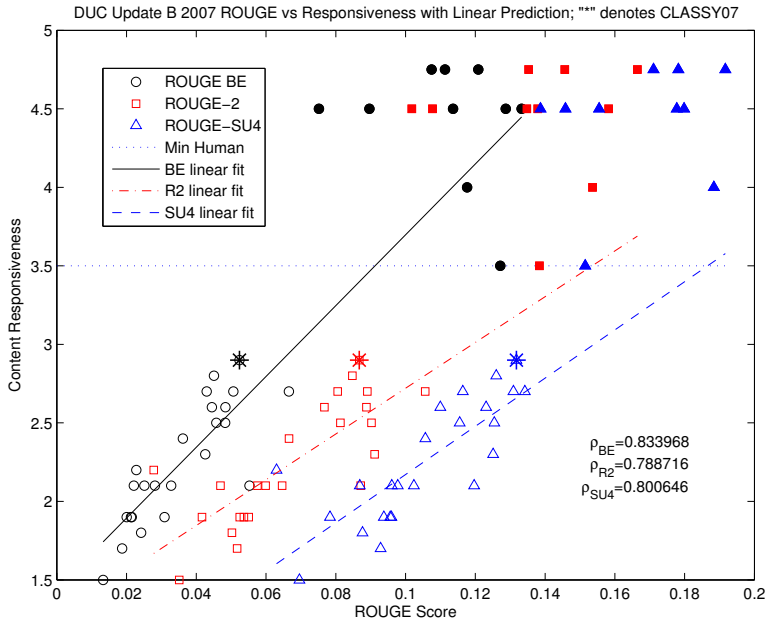

Figure 4: Scatter Plot of ROUGE Scores and Responsiveness for Update B Task

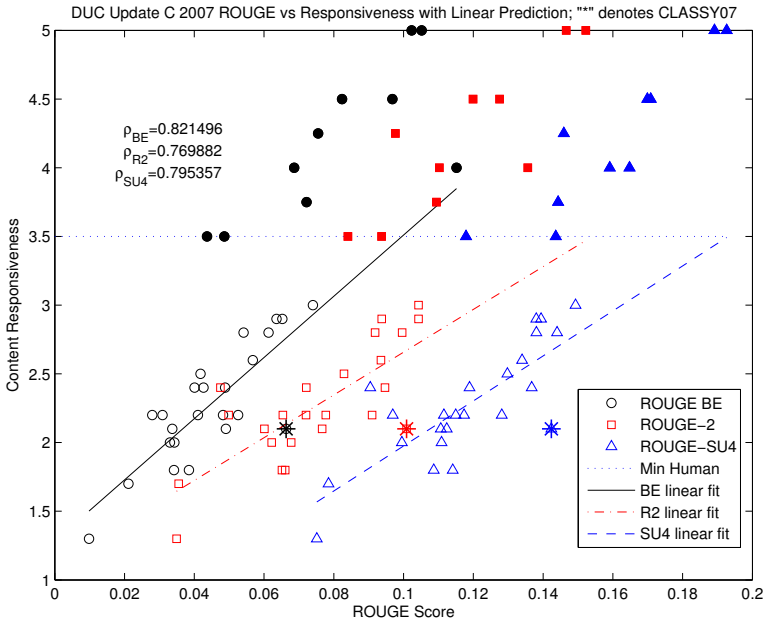

Figure 5: Scatter Plot of ROUGE Scores and Responsiveness for Update C Task

It is encouraging that at the Multi-lingual Summarization Evaluation of 2006, where systems were evaluated on the task of generating 100 word summaries given English and translated Arabic, that there was no gap between ROUGE and responsiveness. Furthermore, CLASSY achieved a responsiveness score which was indistinguishable from human summarizers! In short, we are optimistic that human level performance can be achieved in other summarization tasks, such as the 100 word update summary.

References

[1] John M. Conroy, Dianne P. O'Leary, and Judith D. Schlesinger. CLASSY Arabic and English multi-document summarization. In *Multi-Lingual Summarization Evaluation 2006*, 2006. `http://www.isi.edu/~cyl/MTSE2006/MSE2006/papers/index.html`.

[2] John M. Conroy, Judith D. Schlesinger, and Jade Goldstein. Three CLASSY ways to perform Arabic and English multi-document summarization. In *Multi-Lingual Summarization Evaluation*, 2005.

[3] 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*, 2006.

[4] John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary, and Jade Goldstein Stewart. Back to basics: Classy 2006. In *DUC 06 Conference Proceedings*, 2006.

[5] T. Dunning. "Accurate Methods for Statistics of Surprise and Coincidence". *Computational Linguistics*, 19:61–74, 1993.

[6] Mirella Lapata and Regina Barzilay. Automatic evaluation of text coherence: Models and representations. In *International Joint Conferences on Artificial Intelligence*, 2005.

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

[8] Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. In *MSR-TR-2005-101*, 2005.

[9] Martin Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.