Left-Brain/Right-Brain Multi-Document Summarization

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

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

Jade Goldstein
Department of Defense
jgstewa@afterlife.ncsc.mil

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

1 Introduction

Since we began participating in DUC in 2001, our summarizer has been based on an HMM (Hidden Markov Model) for sentence selection within a document and a pivoted QR algorithm to generate a multi-document summary. Each year, however, we have modified the features used by the HMM and added added linguistic capabilities in order to improve the summaries we generate. This year's entry, called "CLASSY" (Clustering, Linguistics, And Statistics for Summarization Yield), was designed to evaluate phrase elimination and co-reference resolution in pre- and post-processing.

Our participation in DUC 2004 was limited to Tasks 2, 4, and 5; i.e., we did not do any headline generation. We are very pleased with this year's results, although the effort to improve summarization performance goes on. This paper discusses the design of CLASSY, variants adapted to each task, and new linguistic endeavors. An analysis of the results of our efforts using both Rouge and SEE evaluations is also discussed.

2 CLASSY Sentence Scoring

The HMM used in CLASSY contains two kinds of states, corresponding to summary and non-summary sentences. An HMM, in contrast to a naive Bayesian approach ([5], [1]), allows the probability that sentence i is in the summary to be dependent on whether sentence $i-1$ is in the summary.

Our HMM used just one feature, related to the number of *signature tokens* in each sentence. All text was first converted to lower case, and then a token was defined to be a white-space-delimited string consisting of the letters a-z, minus a stop list. The signature tokens are the tokens that are more likely to occur in the document (or document set) than in the corpus at large. To identify these tokens, we used the log-likelihood statistic suggested by [4] and used first in summarization by Lin and Hovy ([6]). The statistic is equivalent to a mutual information statistic and is based on a 2 by 2 contingency table of counts for each token. The value of the feature was $\log(number\_of\_signature\_tokens + 1)$. This lone feature (observation) for the HMM was normalized component-wise to be mean zero and variance one. In addition, the features for both "junk sentences" (e.g., bylines, dates, etc.) and "subject" sentences (e.g., headlines, picture captions, titles, etc.) were forced to be -1, which had the effect of making them have an extremely low probability of being selected as a summary sentence.

For DUC03 we used "subject tokens" in addition to the signature tokens as features. The subject tokens are a special subset of the signature tokens that occur in headline and subject heading sentences. We opted not to use them this year since headlines were not available in most of the data. Fortunately, this feature is not very strong, so losing it was not a big disadvantage.

The HMM was trained using the NIST DUC03, Task 5 novelty data. We focused on only the novel sentences in this set. To strengthen the model further, we sorted the novel sentences by hand for 24 of the document sets, removing many sentences which were no longer relevant in isolation. These data were then used to build an HMM to score the sentences and determine which features should be included. In particular, the training data helped determine the number of states for the HMM, which was empirically chosen to be 13: 7 summary states and 6 non-summary states.

For more details of the HMM and how it is used in conjunction with a pivoted QR algorithm for sentence scoring and selection, please see [2].

3 CLASSY Linguistics

As mentioned earlier, much of our current effort is focused on linguistic processing to improve the quality of our summaries. This section describes the linguistic work we did for DUC 2004, beginning with a brief recap of the 2003 work.

3.1 Sentence Manipulation

For DUC 2003, we *postprocessed* the sentences selected by the summarization algorithms in order to 1) shorten chosen sentences so we could include additional information within the allotted summary size, and 2) improve summary readability and flow.

We developed patterns using "shallow parsing" techniques, keying off of lexical cues in the sentences after processing them with a part-of-speech (POS) tagger. The following eliminations were made, when appropriate:

- Sentence Eliminations:
  - sentences that begin with an imperative;
  - sentences that contain a personal pronoun at or near the start;
- Phrase Eliminations:
  - gerund clauses;
  - restricted relative-clause appositives;
  - intra-sentential attribution;
  - lead adverbs.

See [3] for details on this work.

These eliminations improved the quality of our summaries from 2002 to 2003. But, a natural question arose: since the summarization algorithms use signature (and subject, when available) tokens to help select sentences, wouldn't it be better to make the eliminations *before* selection is made rather than after? In other words, wouldn't it be preferable to *preprocess* all the documents before attempting to summarize?

For DUC 2004, we took advantage of being able to submit more than one run for a task. For Task 2, we submitted one run (#65, CLASSY-pre) in which we preprocessed and a second (#66, CLASSY-baseline) in which we postprocessed. We did not perform either type of *sentence* elimination (see list above) for these two runs. Based on the 2003 data, very few imperatives occur in the selected sentences, either because they don't occur in the text or the summarizer doesn't select them. Elimination of imperatives is, therefore, an unnecessary effort. Eliminating sentences with lead pronominals improves summary readability but often at the cost of reducing the information content. Since the summaries were scored automatically this year, making readability less of an issue, we decided to focus on information content. And, of course, it would not be wise to eliminate all sentences with lead pronominals when preprocessing!

Table 1 shows the ROUGE-1 mean and 95% confidence intervals for our pre- and postprocessing runs (CLASSY-pre and CLASSY-baseline, respectively), as calculated by NIST. From these, we determined that CLASSY-pre has a standard deviation of approximately 0.0064 $((95\%\_\mathit{CI\_upper} - \mathit{mean\_of\_run}\_65)/2)$. Thus, the CLASSY-pre is in excess of 2.4 standard deviations from the CLASSY-baseline mean $((\mathit{mean\_of\_run}\_65 - \mathit{mean\_of\_run}\_66)/.0064))$. This means that we can be 99% certain (using a 1-tail z-score for normalcy) that preprocessing is better than postprocessing.

Run	Mean	95% CI Lower	95% CI Upper
65: CLASSY-pre	0.38224	0.36941	0.39507
66: CLASSY-baseline	0.36671	0.35329	0.38013

Table 1: ROUGE-1 NIST Results for Task 2, Runs 65 and 66

In anticipation of the results we see for Task 2, we used preprocessing for Task 5 as well. However, since we were trying to answer the question, "Who is xxxx?", we decided to *not* eliminate relative-clause appositives that began with "who," specifically because they usually are exactly the kind of information we're looking for in this case. See Sections 3.2 and 3.3 for more on our Task 5 processing.

3.2 Co-Reference Resolution with Serif

Yet another experiment was to utilize Serif ([7]), BBN's co-reference resolution system, to preprocess the documents before summarizing and postprocessing. This generated run #67 (Task 2, CLASSY-Serif), which could then be compared to run #66, CLASSY-baseline.

Serif was used only to influence the selection of summary sentences. For names (nominals), we took the entity chain links provided by the output of Serif and looked for the longest entity mention, which we defined as the number of words/characters beginning with a capital letter. We then substituted this entity string into the original documents. For example, "King Norodom Sihanouk" was used for mentions such as "Norodom Sihanouk", "Sihanouk", "the king (for all occurrences in the entity sequence), and pronouns ("he", "his", etc.) that Serif determined referred to "King Norodom Sihanouk".

We then ran the summarization algorithms over these revised, expanded documents. Sentences that would have received a lower score from the internal HMM scoring could now receive a higher score due to the full entity mention in the sentence and could potentially be selected as a summary sentence.

Since Serif is somewhat prone to errors, after selecting the summary sentences, the original document sentences were used to create the final summaries after postprocessing (see Section 3.1) was applied.

For Task 2, ROUGE-4 scores, CLASSY-Serif outscored, with statistic significance, the CLASSY-baseline results with a certainty of 93% (using a 1-tail z-score for normalcy). CLASSY-Serif also scored higher than CLASSY-baseline for the other metrics (with the exception of Rouge-L), but these results were not statistically significant. Table 2 shows the ROUGE-4 scores for CLASSY-Serif and CLASSY-baseline.

Run	Mean	95% CI Lower	95% CI Upper
67: CLASSY-Serif	0.01658	0.01300	0.02016
66: CLASSY-baseline	0.01394	0.01087	0.01701

Table 2: ROUGE-4 NIST Results for Task 2, Runs 66 and 67

For Task 5, we tried a variant of the algorithm above (Task 5, run 73, CLASSY-Serif5). We substituted the full entity string found by Serif for *only* the name for which the summary was to refer and only if we could find a match for that name (there were a few misspelled names in the query list). For example "Robert Rubin" became "U.S. Treasury Secretary Robert Rubin".

We used these long strings to influence the HMM for sentence selection. For output sentences, we used a slightly modified version of the original documents. For output sentences, instead of just taking the original sentences of the document as we did in Task 2, we decided to use the entity strings for sentence initial pronouns referring to the query subject. This would allow sentences with pronouns to avoid being discarded by the sentence postprocessing. The entity strings were required to be four words or less and begin with capital letters. Initials were eliminated from these word sequences as the summarization algorithms ignore single characters.

CLASSY-Serif5 (run #73) did not perform well on this task compared to either of the preprocessed versions (Task 5, run #71, CLASSY-pre5, run #72, CLASSY-whois). We believe this is because of the inherent strength of preprocessing over postprocessing but additional analysis is required to verify that that is the only thing contributing to the difference.

3.3 Signature Word List Extension

Task 5 required that the generated summaries should be directed by the question "Who is xxxx?", where a specific name was given for each document set. In order to focus on the named individual, we tried to identify words, other than those in the signature term list, that were closely associated with that person.

Using a perl script developed for other purposes, we were able to break each sentence into its verbs and the likely "subject" and "object" of each of those verbs. We then looked for the required name in the subject phrases. If found, we added the associated verb, non-lead-word gerunds, nouns, both proper and common, and adverbs that occurred in that subject phrase to the signature term list.

Table 4 shows the rank results for CLASSY-pre5 and CLASSY-whois. Clearly, we did not get any benefit from this process. Looking at the signature word lists generated shows that we didn't add many words to the signature term list. Since we had promising results with this process with the training data, we need to analyze why it did not perform as well with the real DUC data.

4 Results

We participated in all of the multi-document summarization Tasks for DUC 2004. In each case we submitted three priority runs. See Sections 3.1 and 3.2 for a description of the different runs.

For Task 2, the multi-document summary of TDT data, we wanted to measure the effect of using the linguistics techniques and Serif upon our HMM/QR sentence selection method. Our priority 1 run was to perform linguistic preprocessing before sentence selection. Priority 2 was to do the sentence selection first and then use the linguistics methods as a postprocessing. For priority 3 we used Serif as a preprocessing step to resolve pronouns, then the sentence selection and linguistic postprocessing was done. Table 3 shows our results, as compared to 8 humans and 35 systems scored by the automatic evaluation system ROUGE.

For Task 4, multi-document summaries for machine translated Arabic documents, our priority 1 run was very similar to the approach of our Task 2 priority 1 run. Based on the training data provided by NIST, we judged the ISI translations to be more readable than the IBM translations, so our submissions for priorities 1 and 3 were based solely upon the ISI machine translations. For the priority 3 run, where a selection of topic-related English documents were given, we computed signature tokens based on the background documents and used these tokens to select sentences from the ISI machine translated Arabic. For priority 2 of Task 4, we were given the human translated document sets. Our submission here used the same algorithm used in Task 2, priority 1, i.e., linguistic preprocessing followed by HMM/QR sentence selection. Table 4 shows our results, as compared to 4 humans and 29 systems. Note that the ranking listed include all three priorities. We note, for example, that our ranking with ROUGE-1 for our priorities 1 and 3 runs were 4 and 2 respectively when compared with other priority 1 and 3 runs.

In Task 5, we were presented with document sets about an individual. Our priority 1 submission used the method of preprocessing followed by the HMM/QR. In the training data we found this approach always included as tokens the names of the subject of the "Who is" question. In addition to this submission we decided to experiment with two additional approaches to "steer" the signature tokens toward tokens related to the subject of the question. To this end, shallow processing, as described in Section 3.3, was used to generate a list of tokens which were likely associated with the person. Table 5 shows our results, as compared to 8 humans and 23 systems

Run 65: CLASSY-pre	linguistic preprocessing
Run 66: CLASSY-baseline	linguistic postprocessing
Run 67: CLASSY-Serif	Serif preprocessing and linguistic postprocessing

Metric	Rank 65	Rank 66	Rank 67	# Humans Scoring Higher
ROUGE-1	1	11	7	8
ROUGE-2	1	3	2	5
ROUGE-3	2	3	1	1
ROUGE-4	2	4	1	0
ROUGE-L	3	6	7	8
ROUGE-W	2	6	5	8

Table 3: Task 2 - Short multi-document summaries focused by TDT events

Run 68	using ISI MT Arabic translations
Run 69	using human translations
Run 70	using ISI and related documents

Metric	Rank 68	Rank 69	Rank 70	# Humans Scoring Higher
ROUGE-1	13	5	12	4
ROUGE-2	14	5	11	3
ROUGE-3	13	5	12	1
ROUGE-4	12	5	13	0
ROUGE-L	14	4	9	4
ROUGE-W	14	4	10	4

Table 4: Task 4 - Short cross-lingual multi-document summaries focused by TDT events

Run 71: CLASSY-pre5	linguistic preprocessing
Run 72: CLASSY-baseline	linguistic preprocessing, extended signature tokens
Run 73: CLASSY-Serif5	Serif preprocessing and linguistic postprocessing

Metric	Rank 71	Rank 72	Rank 73	#Humans Scoring Higher
ROUGE-1	3	4	16	8
ROUGE-2	2	3	17	8
ROUGE-3	6	7	17	8
ROUGE-4	9	8	19	8
ROUGE-L	4	5	16	8
ROUGE-W	4	5	16	8

Table 5: Task 5 - Short summaries focused by questions

Overall our system did comparable with the top systems as rated by mean coverage, the SEE/human evaluation. The box plots of Figures 1 and 2, give the ranking of the systems: human, machine, and baseline for Tasks 2 and 5 respectively. Our entry, 65 finished first in Task 2 for mean coverage, which is consistent with the ROUGE evaluation. For Task 5, our entry 71 ranked 9th out of the machine systems. We used a Kruskal-Wallis test, a non-parametric test whose null hypothesis is that the medians are equal, on the top 9 systems. The test measures just how close the top systems are. Figure 3 gives the plot for these systems. A small p-value say 0.05 would mean that we could be 95% certain that the medians are different. However, the test gives a $p-$value of 0.78, which means we cannot reject the null hypothesis that the medians are equal. Conversely, $p-$value for Task 2 for the top 9 systems is 0.12 and 0.16 for the top two systems.

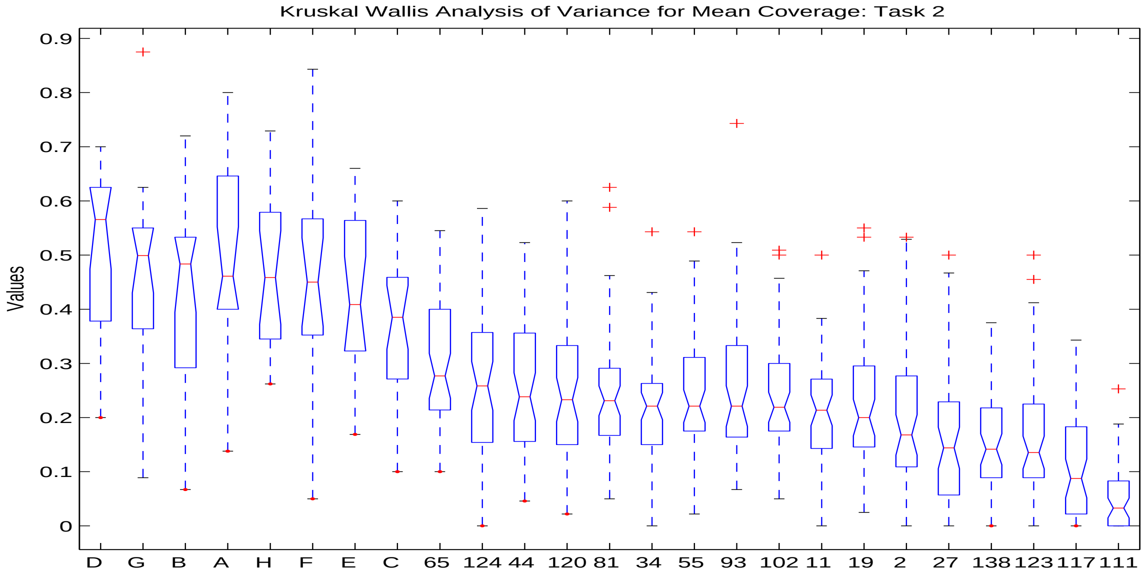

Figure 1: Mean Coverage Results for Task 2.

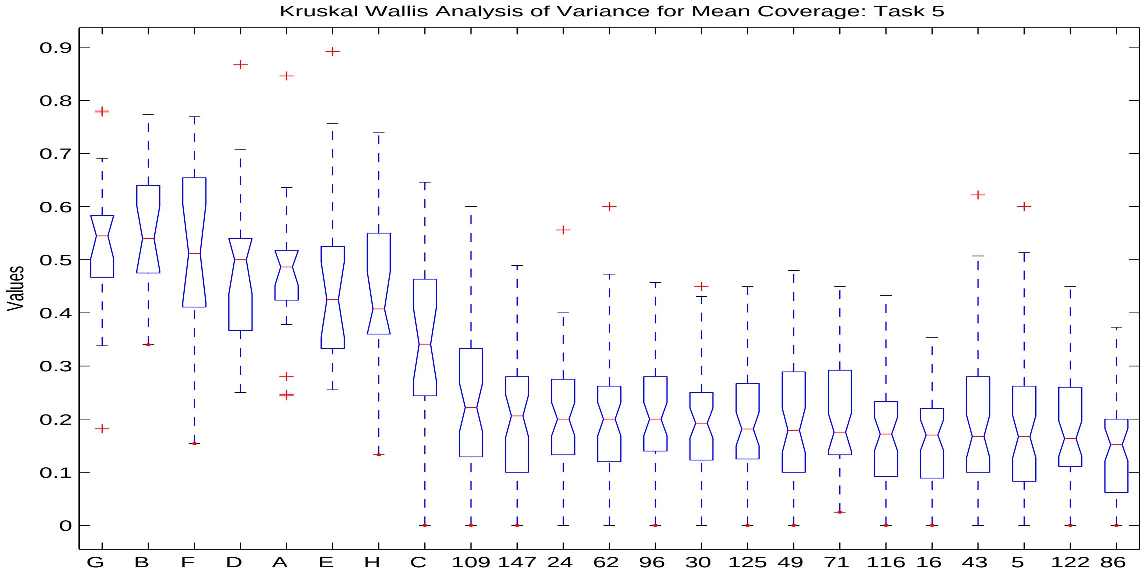

Figure 2: Mean Coverage Results for Task 5.

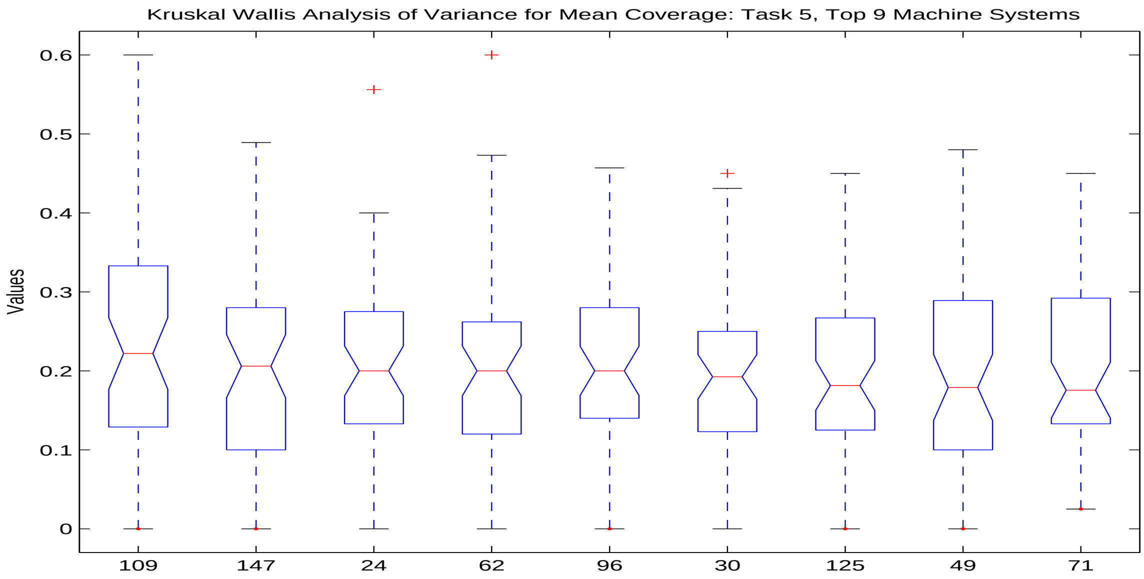

Figure 3: Mean Coverage Results for Task 5, Top 9 Machine Systems.

5 Conclusion and Future Efforts

We're very pleased with both our system's performance and the performance at DUC in general. While systems were hard put to beat the baselines last year, this year, it was more the norm. Systems were even beating human summarizers in some cases.

However, there is still more to do. We need to understand why we did not get the results from Serif, especially on Task 5, that we expected. We also need to understand why we didn't improve the signature term list for Task 5.

We have a list of additional linguistic modifications that we would like to apply. These include additional adverb elimination, more attribute elimination, and identification and removal of unnecessary parentheticals. While we were hoping that Serif would solve the pronominalization problem, so far, we were disappointed in its performance. Therefore, pronoun resolution remains to be done.

We also need to better understand the different scores that are generated. ROUGE provides 6 different scores and the results differ, sometimes significantly, from one to the next. The ROUGE scores are different from the SEE scores. It would seem that each of the scores has something to provide, but we need better understanding of each.

We would like to thank J.K. Davis as well as Lance Ramshaw and Ralph Weischedel for making a copy of Serif available for us to use in this evaluation. We would also like to thank Jessica Stevens of BBN for her assistance with the installation and use of Serif.

References

[1] C. Aone, M.E. Okurowski, J. Gorlinsky, and B. Larsen. “A Scalable Summarization System Using Robust NLP”. In *Proceedings of the ACL’97/EACL’97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73, 1997.

[2] J.M. Conroy and D.P. O’Leary. “Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition”. Technical report, University of Maryland, College Park, Maryland, March, 2001.

[3] D.M. Dunlavy, J.M. Conroy, J.D. Schlesinger, S̃.A. Goodman, M.E. Okurowski, D.P. O’Leary, and H̃. van Halteren. “Performance of a Three-Stage System for Multi-Document Summarization”. In *DUC 03 Conference Proceedings*, 2003. `http://duc.nist.gov/`.

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

[5] J. Kupiec, J. Pedersen, and F. Chen. “A Trainable Document Summarizer”. In *Proceedings of the 18th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, 1995.

[6] C.Y. Lin and E. Hovy. “The Automatic Acquisition of Topic Signatures for Text Summarization”. In *DUC 02 Conference Proceedings*, 2002. `http://duc.nist.gov/`.

[7] L. Ramshaw, E. Boschee, S. Bratus, S. Miller R. Stone, R. Weischedel, and A. Zamanian. “Experiments in Multi-Modal Automatic Content Extraction”. In *Proceedings of Human Language Technology Conference*, San Diego, CA, 2001.