CLASSY Query-Based 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 Stewart
Department of Defense
jade44@gmail.com

1 Introduction

Our summarizer is 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, since we began participating in DUC in 2001, we have modified the features used by the HMM and have added linguistic capabilities in order to improve the summaries we generate. Our system, called "CLASSY" (Clustering, Linguistics, And Statistics for Summarization Yield), preprocesses each document, applying word- and phrase-elimination techniques. With this year's DUC challenge, we focused on query based methods of summarization.

The overall results indicate our method scored within the top group of systems for both ROUGE and pyramid evaluation. This paper discusses the design of CLASSY, variants adapted to each task, and new linguistic endeavors, in particular the method of query word generation from the topic description along with further experiments in using named entity extraction. In addition, we describe the modifications made to the hidden Markov model to incorporate a query term feature. An analysis of the results of our efforts, using both Rouge and pyramid scoring evaluations, is also included.

2 CLASSY Linguistics

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. We initially used some full sentence eliminations along with the phrase eliminations itemized below; analysis of DUC 03 results, however, demonstrated that the full sentence eliminations were not useful.

The following phrase eliminations were made, when appropriate:

- gerund clauses;
- restricted relative-clause appositives;
- intra-sentential attribution;
- lead adverbs.

See [4] for the specific rules used for these eliminations. Comparison of two runs in DUC 04 convinced us of the benefit of applying these phrase eliminations on the full documents, prior to summarization, rather than on the selected sentences after summarization had been performed. See [3] for details on this comparison.

For DUC 2005, we retained, with minimal change, the pre-summarization phrase elimination as established in DUC 04. Instead, we focused on issues specific to the DUC 05 task, namely identifying query terms for each document set and attempting to focus the generated summaries toward the questions asked in the topic descriptions.

2.1 Query Term Identification

In previous versions of the HMM, we had used a query term feature. It did not help the summary generation as we had expected and so we eliminated it. However, due to the question-answering nature of this year's DUC task, we decided to once again use query terms to help focus our summaries.

To do this, we analyzed the topic descriptions. We pulled individual words and phrases from both the <title> tagged paragraph as well as whichever of the <narr>, <event>, and <explic> tagged paragraphs occurred in the topic description. Any words that were tagged NN (noun), VB (verb), JJ (adjective), or RB (adverb) were included in a list of words to use as query terms. Also, any multi-word groupings of proper nouns (NNP) were also used. The number of query terms extracted in this way ranged from a low of 3 terms for document set d360f to 20 terms for document set d324e.

As can be seen from the examples of query terms shown in Table 1, the lists, as generated, are not very sophisticated. Yet, from the testing we did, they did help determine the "best" sentences to select. Based on our system's performance, relative to other systems, this seems to have been borne out.

Set d324e
argentine
british
argentine british
british relations
relations
argentina
great
great britain
war
falkland
islands
falkland islands
diplomatic
economic
military
military relations
restored
differences
status

Set d333g
welsh
welsh devolution
british
parliament
british parliament
status
re-structuring
welsh government
separation
british rule
treated
british legislators

Figure 1: Examples of Extracted Query Terms

2.2 Summary Focusing

We realized that even with query terms derived from the topic descriptions, our method of summarizing was not geared to answering questions. In order to partially rectify this, we decided to use a named entity identifier, specifically, BBN's Identifinder, to help answer questions about people, places, organizations, etc., when such a query occurred in a topic description.

As a separate pre-processing step, we ran Identifinder on all document sets, generating lists of entities for the categories LOCATION, PERSON, DATE, and ORGANIZATION, among others. We also automatically evaluated each topic description, looking for keywords such as "what", "which", "countries", "people", etc., to identify those data sets whose summaries might be enhanced by the named entity information.

To use this information, we selected sentences for the summary in the usual manner. Then, before generating the summary, if the document set was one that we had already determined could benefit, we selected some number of items from the list generated by Identifinder. First selected were items that had a score greater than 1. If there were none, or too few (determined unscientifically), items were selected by order in the Identifinder list, i.e., by the order in which they were found by Identifinder. A last sentence of $4 + thenumberofitemsselected$ words was generated and placed at the end of the summary which was reduced the necessary number of words. The four words added were "Other XXX mentioned were:", where XXX was "countries", "people", etc., depending on the word found in the topic definition. This was followed by a comma-separated list of the selected items.

We were *very* disappointed with the results from this effort, especially since we are still convinced that this should add great value to a summary. The results were disappointing enough that we did not submit these summaries.

Several factors impacted the disappointing results and need to be addressed. Our primary problem was that we did not get a fine-tuned list from Identifinder. For example, LOCATION included cities, states/provinces/etc., countries, geographic features, etc. This meant that we could not be sure that the items we selected from the list were the items we were looking for. We need to explore the use of the tool, or other named entity tools, to see if we can specify the subset we need.

Another problem is that the named entity tool makes mistakes. In a list of PERSONS, for example, we got "Operation Pisces". While that is a NAME, it is not a PERSON. Other errors of this type include "Guernica" and "Exit" as a PERSON and "D-Wis" as a LOCATION. Additional mistakes include splitting multi-word phrases so they show up as two items in the list. For example, "Los Angeles International" and "Airport" appear as two items in a LOCATION list followed immediately by "Los Angeles International Airport". Also, upper and lower case differences are not resolved by Identifinder and are considered to mean different entities.

Lastly, we were only able to identify questions that could be answered by Identifinder categories for 18 of the 50 document sets with just two of these triggering more than one topic. After a quick review of the topic descriptions, it is clear that more should be included. So, if we are to pursue this approach, it would be useful to have a better way to identify when this information can be used.

3 CLASSY Sentence Scoring

After all linguistic processing is completed and query terms are generated for each data set, we use our hidden Markov model (HMM) to score the individual sentences in a document and then a pivoted QR to select a minimally redundant subset of sentences. We highlight this approach and note the modifications made to incorporate query terms for DUC 05.

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 ([6], [1]), allows the probability that sentence i is in the summary to be dependent on whether sentence $i-1$ is in the summary.

Our DUC 05 HMM used two types of observations (features). The first is related to the number of *signature tokens* in each sentence, where a token is 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. These signature tokens, are identified using the log-likelihood statistic suggested by [5] and used first in summarization by Lin and Hovy ([7]).

This feature 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.

The second type of observation used by the HMM was the $\log(number\_of\_query_tokens + 1)$. The query tokens were generated as described in Section 2.1, based on the topic descriptions. This observation is also normalized to be mean 0 and variance 1 for each document. The HMM was trained using 3 clusters from AQUAINT—110, 132, and 138. We used these data since it was already tagged for summaries and we had previously hand selected query tokens. Furthermore, unlike the DUC 03 data we have used, we found that the HMM benefited from the added observation of query terms on this data.

Finally, the training data determined the number of states for the HMM, which was empirically chosen to be 5: 2 summary states and 3 non-summary states.

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

4 Results

This year's DUC provided a myriad of evaluation methods. We will highlight three of these methods here: ROUGE-1, the linguistic questions, and the pyramid evaluation. This year there were 10 human summarizers (labeled A-J), one baseline (labeled 1), and 31 machine systems. CLASSY is identified as system 7.

Table 1 gives the ROUGE-1 scores for the top 13 peer systems, the 10 humans, and the baseline. CLASSY ranked 10th out of the 31 machine systems. The performance of CLASSY, and all the other systems listed, was approximately midway between the baseline and human performance.

Figure 2 gives a box plot of system performances as measured by the modified pyramid score. The scores are sorted by medians. CLASSY ranked third among the peer systems. Humans A and B, who were evaluated on 25 and 24 summaries, respectively, in lieu of the 26 for the machine systems (25 peers + baseline), had median pyramid scores of 0.46 and 0.44, respectively, which is far above the best peer performance of 0.20.

In addition, the Kruskal-Wallis statistic, which tests for equalities of medians, was run on the 25 peer systems and baseline. The test gives a p-value of 1.4e-6, which gives overwhelming evidence that we can reject the null hypothesis that the medians are the same. However, when the test is run on the top 22 systems, the p-value increases to 0.08 and thus we cannot be 95% confident that the first 22 systems do not have the same median performance as measured by the modified pyramid score.

To further illustrate how close the systems score, the p-value for the top 5 scoring systems is 0.78, which indicates that observed performance of the top 5 systems is what one would expect to see 78% of

Submission	Mean	95% CI Lower	95% CI Upper
C	0.45853	0.44324	0.47210
A	0.45567	0.44369	0.46759
E	0.44885	0.43254	0.46450
I	0.44816	0.43067	0.46759
J	0.44330	0.43089	0.45781
B	0.44232	0.42814	0.45673
D	0.43973	0.42724	0.45305
G	0.43959	0.42694	0.45186
F	0.42580	0.40723	0.44368
H	0.41501	0.39971	0.43161
15	0.37978	0.37468	0.38470
4	0.37517	0.37020	0.38047
5	0.37029	0.36463	0.37563
17	0.36925	0.36324	0.37478
8	0.36495	0.35979	0.37008
10	0.36146	0.35647	0.36626
19	0.36128	0.35607	0.36656
6	0.36114	0.35588	0.36602
14	0.36066	0.35386	0.36710
7	0.35782	0.35261	0.36312
11	0.35715	0.35187	0.36249
24	0.35402	0.34883	0.35878
25	0.35297	0.34788	0.35840
1	0.29243	0.28329	0.30122

Table 1: Average F score of ROUGE 1 Scores

the time if the systems had the same median performance. Finally, we give 5 box plots (Figures 3, 4, 5, 6, and 7) corresponding to the 5 linguistic quality questions. Overall, the systems typically had about the same median scores for each of the questions so they were sub-sorted by their means. It is striking that all the systems had a median of 5 on non-redundancy, although there were some differences in the mean, as shown. Lead sentence selection, i.e., the baseline, does best on referential clarity which is most likely due to the fact that pronouns, if used in a lead sentence, would not be ambiguous or misleading. CLASSY, as with almost all the systems, can be greatly improved in the area of structure and coherence.

5 Conclusion and Future Efforts

We are very pleased with both our system's performance and the performance of all systems at DUC. Consistent with the trend seen last year, systems are regularly outperforming the baseline. However, unlike last year, systems are far from human summarizers on this more focused task. Clearly, the human capability to combine information from many documents into a succinct summarizing statement is still beyond the capability of the automated systems.

Anaphora resolution has been a goal for CLASSY for several years although we have not yet given it the attention needed. We believe that our performance on the linguistic quality questions will greatly improve from this one task. We also intend to proceed with additional work with named entity identifiers to continue the effort begun this year.

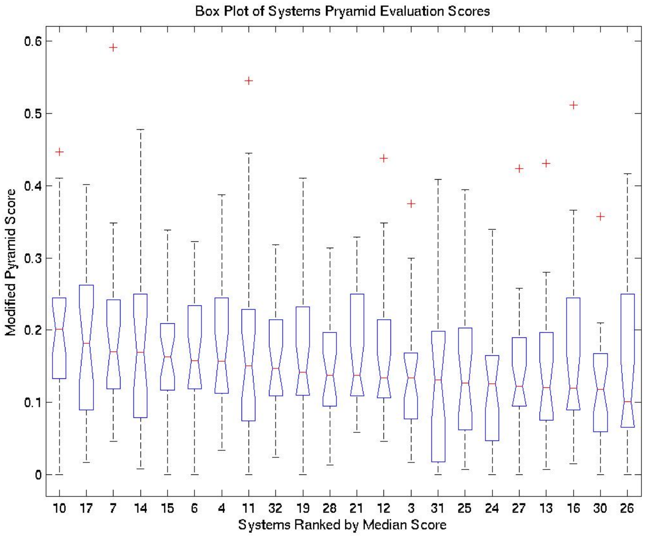

Figure 2: Box Plots for Pyramid Scores

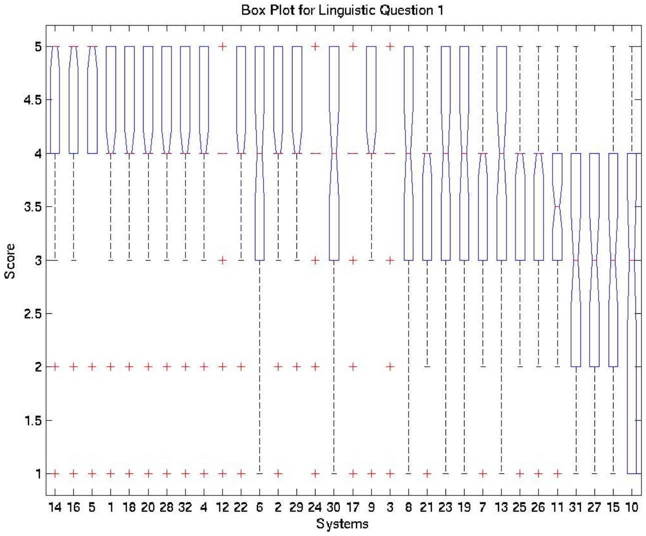

Figure 3: Grammaticality: Linguistic Question 1

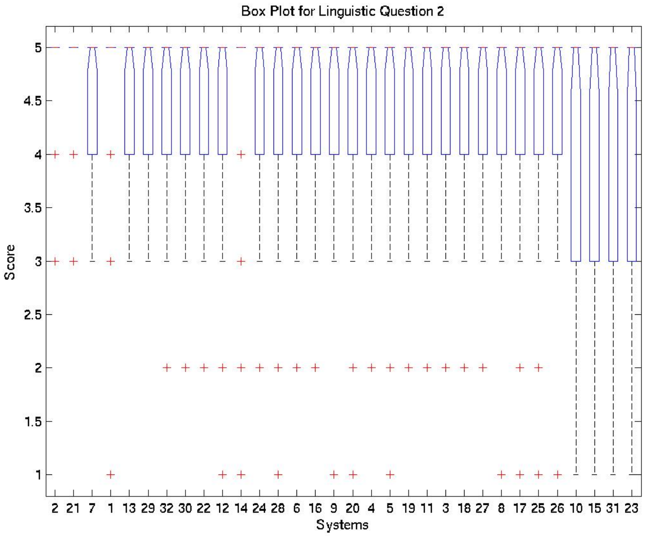

Figure 4: Non-redundancy: Linguistic Question 2

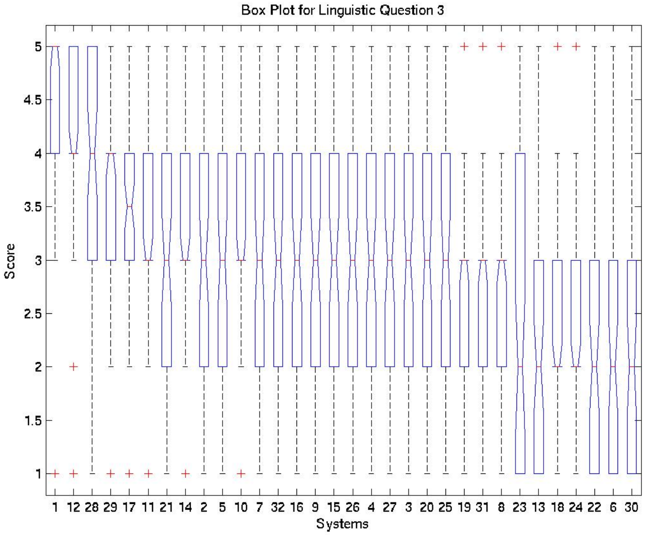

Figure 5: Referential Clarity: Linguistic Question 3

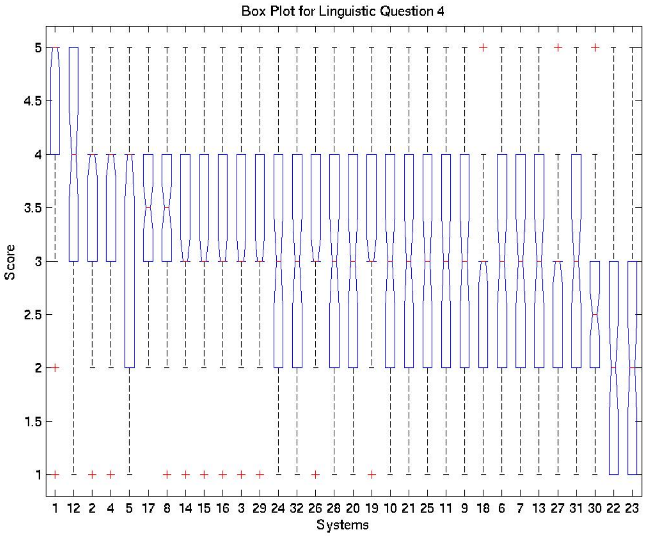

Figure 6: Focus: Linguistic Question 4

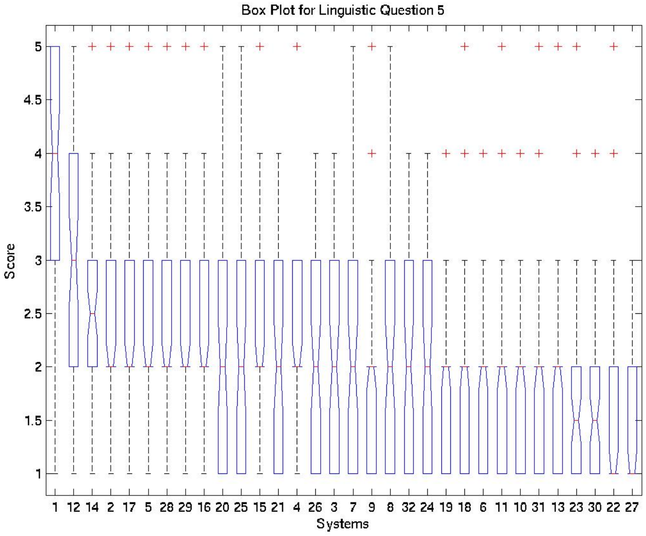

Figure 7: Structure and Coherence Linguistic Question 5

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] J.M. Conroy, J.D. Schlesinger, J. Goldstein, and D.P. O'Leary. Left-brain right-brain multi-document summarization. In *DUC 04 Conference Proceedings*, 2004. `http://duc.nist.gov/`.

[4] 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/`.

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

[6] 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.

[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.