Summarization for Q&A at Columbia University for DUC 2007

Barry Schiffman
Department of Computer Science
Columbia University
New York, NY 10027
bschiff@cs.columbia.edu

Abstract

This paper describes an experimental system for the 2007 Document Understanding Conference that sought to include contextual information in query-focused summarization. In addition, the system tried to incorporate corpus-driven semantic information in the selection of passages to respond to the query.

1 Introduction

Our system for the 2007 Document Understanding Conference represents a departure from our past entries. We took part in both the main task and the update task, and sought to build a unified approach that would be able to save the processed information of clusters of documents so that the information in future clusters of information could be tested against what the system had seen before. Regrettably, the system remains a work in progress. We tried to construct the system in a very limited time frame, and the results suffered for that. Performance in the main task was quite poor; performance in the update task was stronger relative to the other systems, than in the main task, but much more work needs to be done.

We have observed in related tasks, including the defunct TREC novelty track, the Opinion Pilot and the GALE[1] program, and in prior work that relevant material tends to appear in runs of sentences. Rather than being independent units of text that can be measured against a query or topic statement on its own, sentences exist in a coherent discourse. Often to interpret a sentence, one needs to resolve complex references of several kinds. State-of-the-art pronomial reference resolution procedures operate at about 70% to 75% accuracy, and that task is one of the more straightforward of reference resolution tasks, especially since we are dealing with news text, professionally written, well-formed text.

In addition, in the query-focused DUC tasks of recent years, the questions are varied, and often require answers that do not simply echo the key words. To deal with this, we tried to use an existing database of expansion terms that we had extracted for an earlier project from part of the AQUAINT data, the collection used for the DUC evaluation.

Together, the heart of our main task response was formed by an attempt to extract runs of sentences rather than single sentences, and to base the selection of these on an expansion of terms with which we computed relevance to the topic. These two elements were the only parts of the system we envisioned that were complete at the deadline for the main task submission.

We also computed the probability of finding a key word or associated work appeared in each document, and eliminated a proportion of the documents based on their quantiles in order to estimate whole document relevance. While the documents in the sets were generally on topic, there were often some that more or less tangential. This strategy was intended to work with the selection of runs of sentences, especially since the answers were limited to 250 words in the main task.

[1]The DARPA program called "Global Autonomous Language Exploitation"

For the update task, we first had to build some mechanism to process the partial clusters in turn. For each update question, there were three sequential subclusters that had to be considered in turn. For the first subcluster, the *A* subcluster, we used the same logic as for the main task, and for the subsequent subclusters *B* and *C*, we sought to select information that was novel with respect to the all the information in the previous sets, regardless of what had been selected for inclusion in the first answers, or the computed relevance of those sentences that had not been selected.

In the 10 days between the main task deadline and the update task deadline, we add a number of features, including named entities, dependency paths and the probability that a word is in the given document cluster.

For both the main task and the update task, we submitted results from an untuned system, and therefore it is not surprising that performance was poor. In a few experiments after the deadline, tuning the parameters of the system and discarding some strategies that were not beneficial lifted the scores to above average for the evaluation.

2 Main Task

2.1 Context

Over all, our attempt to include the notion of context was not successful. By context, we mean some sequence of sentences that combine into a cohesive, topical segment. In the Novelty Track at TREC in 2004 (Schiffman and McKeown, 2005), we found that the classification of a sentence S_i was a useful predictor of the classification of sentence S_{i+1}. We participated in the opinion pilot conducted by the National Institute of Standards and Technology (NIST) in 2005, and saw some value in applying a topic-segmentation algorithm. We continued to experiment with this in our work in GALE. For DUC, we decided to eliminate the separate segmentation routine, and try to locate runs of sentences.

Because of the lack of time, we set the size of a segment at three, and considered all overlapping in the document cluster. Normalization was problem because of the bias toward shorter passages when a count of the words, or of content words, is used. We decided to use the absolute count of the expanded key words unadjusted by length as the value.

2.2 Term Expansion

We expanded the content words in the title section of the questions with terms drawn from a arge study of document co-occurrences that had been compiled for other purposes. A large table of words that occurred at least 100 times in the Associated Press portion of the Aquaint corpus, a 52 million-word collection of English newswire from 1998 through 2000. The statistic used to measure the strength of the co-occurrences, or association, between two words was the log likelihood ratio (LLR) (Dunning, 1993), using a binomial distribution.

$$\lambda = \frac{max_p(L(p,k_1,n_1)L(p,k_2,n_2)}{max_{p_1,p_2}L(p_1,k_1,n_1)L(p_2,k_2,n_2)}),$$

where the binomial gives the following likelihood with the parameters of k successes out of n trials, with a probability of p.

$$L(p,k,n) = p^k(1-p)^{n-k}.$$

If the occurrences of the two words are independent, $p == p_1 == p_2$, and the ratio will be small. the LLR has the desirable quality that $-2log\lambda$ is asymptotically χ^2 distributed. Likelihood ratio tests do not depend on the assumption of normality as do many other statistical tests, but the χ^2 critical values can be used with the degrees of freedom set at the difference in the number of parameters, here $df = 1$, which has a critical value of about 6.6.

The procedure described here follows what Lin and Hovy did (2000) in their work on topic signatures and others. When our data had been collected, it was not intended to be used in this way in summarization, but it was something we had readily available.

Table 1 shows the top scoring word associations with disease. A concentration of these in a segment of text would be strong evidence that disease is the topic, with the appearance of the word disease. In addition, the table we built gives a high score to numerous diseases, particularly those that were in the news in the late 1990s.

In the 2006 DUC evaluation, topic 625 asked for the types of diseases in Kenya. The narrative section asked, "What are the most prevalent diseases in

study	12362.439
patient	9622.507
heart	8385.765
cell	7899.654
cancer	7733.667
blood	6840.572
risk	6742.717
researcher	6625.2
drug	6261.981
gene	5891.802
virus	5857.0
health	5839.683
vaccine	5579.67
doctor	5248.808
research	4900.125
treatment	4845.574
brain	3879.055
infection	3561.688
symptom	3300.541
scientist	3245.623
diabetes	3161.762
infect	3132.148
cholesterol	2986.598
medical	2615.85
age	2463.863

Table 1: The list above gives the 25 highest scoring associations with the word disease and the log likelihood ratios for the association between them and disease. Among the words scoring at least 500 are epidemic, tuberculosis, asthma, arthritis, fever, polio, hepatitis, dengue, malaria, smallpox, sclerosis, syphilis, measles, encephalitis, cholera.

Kenya and how are they affecting the population? What is being done to combat them?" The models for that topic mention several diseases, including HIV, malaria, typhoid, cholera, dysentery and heart disease. Clearly several of those are found in our table, while they are not easily recoverable from WordNet, like HIV or malaria, which are given as instances of "infection", which in turn is given as a "health problem". Unfortunately, when we created this table, we excluded proper names, and therefore miss the association of HIV and disease.

3 Update Task

For the update task, we added a number of features we had hoped to use for the main task, including named entities, dependency paths, and cluster probabilities. After marking named entities in both the question and the documents in the cluster, we counted matches in the documents. Similarly, we parsed both question and documents with a dependency parser and recorded exact matches. In both these cases, it was clear in the four training examples we were given that neither named entities or parses would be consistently useful. In the larger main task, and in the new update task, names were not always in the questions. As far as the parses, we used exact matches because of the time constraints, and these, too, were few and far between. In addition, we added a feature to show the relative position of the sentences.

The largest difference in our system for the two tasks was that we dropped the multi-sentence judgments because of the length requirements. In the main tasks, we had 250 words for summaries, and could touch on three or four areas in the summaries. But with the incremental summaries reduced to 100 words each, the larger segments seemed very risky.

To consider the later clusters, the *B* and *C* clusters, we relied on the dependency paths. We considered all the information in the earlier clusters when making the decision on whether something was novel or not. So that if a fact from a prior cluster did not make it into the prior summary, it would still disqualify a new realization of in the later clusters. Partly because the dependency paths were infrequently repeated exactly, and partly because we were dealing with news, we gave a large amount of weight to location feature. We also split the exact key words and the expanded terms into separate features.

4 Evaluation

In the main tasks, our scores were well below average, falling 28th out of the 31 systems, including the two baselines, in the Rouge2 scores. We did better in the update task, placing 17th out of the 23 systems in the Rouge2 scores. Our performance was similar in the other metrics and we focus on Rouge2 so that we could measure the effect of tuning the system.

After the official evaluation, we sought to test

Jan. 31 version	0.06201
Feb. 11 version	0.07174
Bug fixes	0.07203
More weight to term expansion	0.07012
More weight to key words	0.07636
Zero weight to term expansion	0.07669
Reduced weight for names	0.07859
25% document relevance	0.08016
No document relevance	0.07984

Table 2: Selected results from experiments after the official evaluation on the DUC 2006 data. The Jan. 31 results was the system that we used in the main task in 2007. Feb. 11 was the system used in the update task. From there, a greater and greater emphasis was placed on the key words alone.

the different features and the weights we had begun with. We experimented with the 2006 main task, and focused on the Rouge2 scores since they did not require any manual evaluation. Table 2 shows the results of various efforts to tune the system.

From these experiments, it showed that the approximation we used for topical segmentation did not work well, nor did the original emphasis on term expansion. When we look at the performance of our main task system on the 2006 data, we see that it performed well below the average of all systems, and that the sentence-based, system used in the update task was far better. As we increased the proportion of the weight on the key words, that is the content words in the both the title and narrative sections of the questions, the scores improved on average. In the main task system, we accepted on the documents in the top quantile in a calculation based on the probability of finding any of the expanded terms in them. As we relaxed this constraint, the scores grew steadily, until we eliminated it altogether. At that point there was a small decline, suggesting there was some value in using the expanded terms.

5 Conclusion

While our results were disappointing, we were able to test some ideas in the 2007 DUC evaluation, although we admit that we had begun too late to explore the task. Our experiments after the official evaluation show that how a very straightforward system based on key words turns in a respectable performance. Looking back, it seems that the segmentation approach is clearly inappropriate for short summaries, and perhaps has some value as an intermediate step. Our version of term expansion was also inadequate to the task. A number of problems are apparent. For one thing, it was based only on the Associated Press portion of the corpus. For another it is too coarse grained to be used directly in a task like this.

References

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Chin-Yew Lin and E.H. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics*.

Barry Schiffman and Kathleen R. McKeown. 2005. Columbia university in the novelty track at trec 2004. In *The Thirteenth Text Retrieval Conference Proceedings*.