Evolving XML Summarization Strategies in DUC 2005

Kenneth C. Litkowski
CL Research
9208 Gue Road
Damascus, MD 20872
ken@clres.com

Abstract

In the Document Understanding Conference for 2005, CL Research made some improvements in its summarization routines based on the use of massively XML-tagged documents containing increasingly richer characterizations of texts. We extended the Knowledge Management System to include an improved capability for identifying redundancy when adding extracted sentences to the summary and for improving the organization of the sentences. These minor improvements increased our ROUGE-1 scores from 0.30 to 0.35. However, we did not make further modifications to our routines, pending a further examination of the utility of the summarization task. We used DUC 2005 to assess the value of various metrics, concluding that the ROUGE metrics provided the most guidance for improving our system. We found the linguistic quality and responsiveness metrics to be of use. However, we found the pyramid method to be very difficult to use, although promising in theory. Overall, the metrics still beg the question of analyzing the crucial summarization questions of establishing semantic equivalence in the ways that content is expressed. Based on other summarization experiments performed in a demonstration project, we suggest that the solution to establishing equivalence and more specifically, when a concept represents an instantiation of a concept in a topic description, may be found in a richer ontological representation of documents.

1 Introduction

CL Research made only minor changes in its summarization algorithms for the Document Understanding Conference (DUC) for 2005, primarily because of uncertainty about the usefulness of one-time summaries in real-world environments. Summarization is a component of CL Research's Knowledge Management System (KMS), which contains several other components used for investigating the content of document collections. While we generated and submitted summaries for the DUC task, we focused more on our underlying technology, rather than trying to optimize KMS to perform this year's task. Notwithstanding, we were able to improve our performance substantially over our results for earlier years (Litkowski, 2004 and Litkowski, 2003).

In DUC 2005, a primary objective was the consideration of alternate mechanisms and measures for evaluating summarization performance. We assess these measures in light of the DUC 2005 task, but also in conjunction with other summarization efforts in a real-world environment where a biologist has the task of creating a report summarizing what is known about the mechanisms of action of a biological toxin that can be used by terrorists. As a result, we suggest that the task posed in DUC 2005, in its overall design and in the manner in which each topic is characterized, may not reflect real-world needs and processes.

Section 2 presents a description of the DUC 2005 task. Section 3 provides an overview of KMS, with an emphasis on the extensions made during our preparations for DUC 2005 and the procedures used to perform the DUC task. Section 4 describes the KMS summarization procedures as used in DUC 2005. Section 5 presents and analyzes the DUC 2005 results, particularly noting our experience with metrics used in assessing results. Section 6 describes the companion project that involves several summarization components. Section 6 presents our observations about the summarization task.

2 DUC 2005 Task Description

DUC 2005 consisted of one task, to create a 250 word summary for each of 50 topics from about 30 articles from the Financial Times of London and the Los Angeles Times from the early 1990s. The 50 document clusters were constructed by NIST assessors based on topics of interest. The assessors looked for aspects of a topic of interest and created a DUC topic. The topic was specified with a topic number, a title of a few

words, a narrative, and a granularity. Table 1 shows one topic and the information provided.

Table 1. Topic Description	
Number	d311i
Title	VW/GM Industrial Espionage
Description	Explain the industrial espionage case involving VW and GM. Identify the issues, charges, people, and government involvement. Report the progress and resolution of the case. Include any other relevant factors or effects of the case on the industry.
Granularity	Specific

This information provides a "user profile". The description and granularity are intended to model real-world complex question answering. Granularity was either general or specific; these terms were not further defined, but presumably are intended to elicit either general statements about the topic or specific facts and events pertaining to the topic. In the topic descriptions, two types of words are present: (1) retrieval task words (*explain*, *identify*, *report*) and (2) content specific words (*issues*, *VW*, *people*, *progress*). Some of the content words (*factors*) are general.

The human assessors hand-generated four summaries for 30 of the topics and ten summaries for the remaining 20 topics. These summaries were used as the reference points for assessing system performance.

Submissions were judged with four sets of scores: (1) linguistic quality (using a 5-point scale, on grammaticality, non-redundancy, referential clarity, focus without extraneous information, and structure and coherence); (2) responsiveness to the information need expressed in the description (using a 5-point scale from unresponsive to fully responsive); (3) automatic scoring using ngram analysis; and (4) semi-automatic scoring measuring summarization content units.

The automatic ngram scoring used a Perl script, ROUGE (Recall-Oriented Understudy for Gisting Evaluation).[1] ROUGE compares a submitted summary with a manual summary, after stemming each word in the summaries, counting the proportion of words in submission with the words in the manual summaries. In addition to ngram matching, ROUGE was extended to count the "longest common substring", a weighted form of the longest common substring, and bigrams allowing for skipping words with a maximum skip distance of 4 words. Official scores returned to participants were the ROUGE bigram and skip bigrams scores.

The pyramid method is a manual method for summarization evaluation, developed in an attempt to address the fact that different humans choose different words when writing summaries. The pyramid method uses multiple human summaries to create a gold standard of summarization content units (SCUs) deemed equivalent in meaning. The frequency of SCUs in the human summaries is used to assign importance to different facts.[2] DUC participants used an interface to annotate system summaries against the gold standards, from which a score was then computed and returned. The pyramid score for the summary equals the weight of the summary content units normalized by the weight of an ideally informative summary consisting of the same number of content units as the peer. This score resemble precision, because it directly reflects how many of the chosen content units are as highly weighted as possible.

3 System Description

CL Research's Knowledge Management System consists of three main components: (1) conversion of documents in various formats to a standard format identifying text portions; (2) parsing and processing the text into an XML-tagged representation, and (3) document querying, involving use of the XML-tagged representation for NLP applications such as text summarization, question answering, information extraction, and other analyses. The overall architecture of the system is shown in Figure 1 and is described in detail in Litkowski (2004), with only a broad overview provided here.

The DUC 2005 documents for each topic cluster were combined into a single XML file. The 50 files (of total size 7 MB) were then parsed and processed into an XML representation (approximately 68 MB, or 10 times the size of the original files). The parsing and processing component consists of three modules: (1) a parser producing a parse tree containing the constituents of the sentence; (2) a parse tree analyzer that adds to a growing discourse representation of the entire text and identifies key elements of the sentence (clauses, discourse entities, verbs and prepositions) and captures various syntactic and semantic attributes of the elements (including anaphora resolution and WordNet lookup); and (3) an XML generator that uses

[1] Available from http://www.isi.edu/~cyl/ROUGE.

[2] Described in detail and with scoring available at http://www1.cs.columbia.edu/~ani/DUC2005/.

the lists developed in the previous phase to tag each element of each sentence in creating the XML-tagged version of the document.

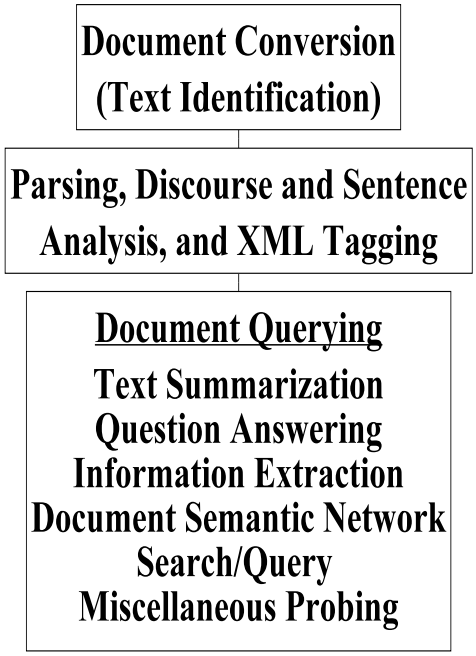

Figure 1. Architecture of Knowledge Management System

The processed files are then identified to KMS as a repository (named DUC 2005), from which any functionality incorporated in KMS can be used to query the individual files. Broadly, this component consists of a graphical user interface that enables a user to generate summaries, answer questions, extract information, or probe the content of the documents. The XML files can be viewed (with retention of the nested structure) in Microsoft's Internet Explorer, but this does not allow any systematic examination of the data.

In KMS, a user can explore the contents of a repository along several dimensions. Initially, the KMS interface only identifies the documents contained in a repository. A usual first step in examining the documents is to create a keyword list and a headline describing each document. The user can select all documents in a repository and create these "short" summaries in about 10 seconds (for documents of the size used in DUC). KMS remembers these summaries in an XML file, so that they can be redisplayed immediately as a user switches back and forth among repositories.

The user can then explore the contents of a repository, either one document at a time or by selecting multiple or all documents. KMS includes three main methods of exploration: (1) asking fact-based questions, (2) summarizing either generally or topic-based, and (3) probing the contents by the semantic types of entities, relations, and events. Each of these tasks is implemented by using XPath expressions to query the document (i.e., select and manipulate nodes of the XML tree).

In general, each KMS task selects particular node sets (e.g., sentences meeting particular criteria, all discourse entities labeled as persons, all discourse segments labeled as subordinate clauses, or all prepositions labeled as locational). The node sets are then subjected to analysis to produce final output corresponding to the task (e.g., summaries or answers to questions).

In addition to the document sets, the DUC 2005 topic descriptions (contained in an XML file) were also processed as if they were ordinary texts. Within KMS, the topic descriptions were identified as "topic groups" that could then be used as the basis for topic-based summarization. This mechanism allows a user to prepare an ordinary text description of topics of interest, without the need to create boolean search queries. Each topic group thus acts as a filter that can be used to query document sets.

4 Summarization for DUC 2005

KMS provides several summarization alternatives. As mentioned above, these include keyword and headline generation. The user identifies the repository and the documents within that repository to be summarized. Summaries can be generated for each document or for multiple documents (including all documents within a file, as in DUC 2005). The user specifies the summary length in characters, words, or sentences. The user can choose to create a general summary or a topic-based summary. The topic-based summary can be based on a set of keywords (treated without syntactic and semantic analysis) or a topic description (of any length, such as a couple of paragraphs). Once the specifications are entered, the summary is produced in a few seconds with the click of a button. In addition to displaying the summary, all summaries are saved to an XML file which includes the specifications as node attributes and a list of each sentence included in the summary, with its source, sentence number, and score.

In general, all summarization in KMS begins with a frequency analysis of discourse entities. A simple XPath expression retrieves all discourse entities and these are then examined in turn to develop a frequency count of the words in them. However, the KMS

method of counting is somewhat different from traditional methods used in information retrieval. First, the traditional use of the stop list is employed to remove frequent words (like articles). Next, the entity is examined to determine whether it is a referring expression, i.e., whether it has an antecedent (pronouns, co-referring expressions, or definite noun phrases). For referring expressions, the words in the antecedent are counted instead of the words in the referring expression.

Except for keyword generation, summarization is based on extraction of sentences from the document cluster. Sentences for all documents are ranked, weighted either on the word frequency analysis described above (for a general summary) or the occurrence of words in the topic or viewpoint specification. Sentences are added to the summary in the order of their scores and as long as their addition does not exceed the specified length. Before a sentence is added, it is compared to sentences already added to determine whether the new information duplicates information already present (based primarily on an analysis of the noun phrases). As sentences are added, the set may be reordered so that sentences from the same document appear in the summary in the order they appear in the source documents. The last sentence was truncated if it contained more than 10 words and was not redundant, potentially interleaving a partial sentence in the summary.

At this time, there is no smoothing of a summary; sentences are included exactly as given. Each sentence included in the summary is present in its full XML form, as represented in the document. In order words, all information about the discourse, syntactic, and semantic structure is available, including identification of discourse markers and antecedents for anaphors and other referring expressions. Pending further analysis, we have not yet implemented routines to make use of the available information to make the summary more readable, such as replacing referring expressions by their antecedents or removing certain types of discourse markers.

Summaries generated using KMS for submission usually required only a few seconds for each. Total processing time for the entire DUC submission was about thirty minutes. The actual submission was created from the XML files generated by KMS using a Perl script.

In preparation for DUC 2005, we used our DUC 2004 submissions, particularly Task 2 where 665 character general summaries were created, as the basis for making changes to KMS. Specifically, we used ROUGE to identify where to focus our efforts. We modified slightly an earlier version of ROUGE so that it would generate ngram scores for each document cluster. In particular, we ordered our performance on the unigram scores and used those document clusters for which we obtained the lowest scores as the basis for making changes. After making changes, we were then able to rescore our results and note any improvements. While we did not complete our efforts using this approach, we improved our results from 0.30057 (30^{th} out of 35 systems) to 0.34026 (20^{th} out of 35).

5 Results and Analysis

Table 1 show CL Research's ROUGE unigram macroaverage recall scores by granularity and overall. The top score for all participating teams was 0.38036. While this result appears to be statistically better than our result, the difference is not considerable. Our results are slightly higher than that achieved during early modifications to our summarization routines, but seem to show that KMS is performing at a consistent level. The results show that KMS did not perform differently according to the granularity of the topic, i.e., if the assessors summaries were in fact different in some essential property, that difference was not reflected in our system, where generation of the summaries was the same for both types of granularity.

Table 2. DUC 2005 ROUGE-1 Recall

Granularity	Score	Rank
General	0.35685	
Specific	0.34183	
Overall	0.34849	16/32

We do not show the official results (for bigrams or skip bigrams), since they do not permit direct comparison with last year's scores. We also do not show the precision scores and the f-scores. Perhaps surprisingly, our precision scores were almost identical to the recall scores, and hence to the f-scores. Our scores on the other ROUGE metrics were essentially the same relative to other participants, with our ranks varying between 13^{th} and 16^{th}.

Table 3 shows the performance of our system on the five measures of linguistic quality. The scaled scores show the average over the 50 topics. These scores are consistent with expectations. We attribute the lower score on grammaticality to the presence of truncated sentences; otherwise, since sentences were taken directly from the source documents, we would have expected them to be grammatical. The score on non-redundancy suggests that our assessment of redundancy was generally successful. Our scores on the other three measures can be attributed to the fact

that we have as yet not attempted any smoothing of the summary.

Table 3. DUC 2005 Linguistic Quality	
Quality Measure	**Scaled Score (1-5)**
Grammaticality	3.82
Non-redundancy	4.36
Referential clarity	2.94
Focus	3.30
Structure/coherence	2.20

On the measure of responsiveness, CL Research was ranked 24th out of 32 participating teams. This suggests that KMS does not as yet have the capability for moving from general terms expressed in the topic description to sentences that best satisfy these terms.

CL Research received an unmodified pyramid score of 0.187158 and a modified pyramid score of 0.143923, with both scores achieving a rank of 17th out of 25 teams participating in the pyramid annotation. These scores differ somewhat from the median performance suggested by the ROUGE metrics. However, in examining the results for the individual 26 topics for which pyramid scores were determined, we were unable to discern any pattern between the pyramid scores and the measures of linguistic quality and responsiveness. The pyramid scores and the ROUGE-1 recall scores also seemed to be unrelated. For example, on topic 632, we obtained a ROUGE score of 0.43577 (the highest of all topics) and a pyramid score of 0.1099 (the 4th lowest of all topics). We have not examined the pyramid scoring in more detail at this time.

In general, we believe that the ROUGE metrics provide the best information for assessing overall performance and identifying where we need to make improvements in our summarization algorithms. The measures of linguistic quality also provide significant insights into the performance of our system and accord well with intuitions. We are sympathetic to the goals of the pyramid method of analysis, i.e., to the identification of semantically equivalent phraseology. However, the method is very difficult to employ. CL Research's annotation was judged as requiring major revision (along with the annotations of at least 7 other teams). Overall, we find that the metrics still do not solve the problem of assessing semantic equivalence.

6 Assessing the Summarization Task

CL Research has been engaged in a demonstration project for a client, assessing the utility of KMS functionality in a real-world environment.[3] In this project, we were provided with (1) a document collection of 200 documents describing the hazards of to a biological toxin (mostly from the primary literature, but including some summary papers and reports), (2) a spreadsheet characterizing each document, including extensive notes prepared by a biologist to identify material of importance in the document, and (3) a final report summarizing the hazards for an intelligence agency. All material used and generated in the process of producing the report is unclassified.

The overall objective of the demonstration project was to determine whether KMS could generate the kind of information developed by the biologist in preparing the final report. With the wide range of functionality available in KMS, the project permitted some evaluation of alternative techniques for querying the document collection and extracting relevant information. The intermediate spreadsheet and final report provided a corpus of documentation implicitly encapsulating a biologist's thought processes of defining the problem, accumulating documentary evidence, analyzing that evidence, and synthesizing the findings into a coherent summary document of considerable length.

The document collection had already been converted into PDF format and from there into an XML format, following a DTD that separated the documents into specific fields. The documents were thus broken down into bibliographic fields, the abstract (if any), text fields, tables and figures, and bibliographic citations. From this raw form, we parsed and processed the documents into XML representations for use in KMS. We developed separate repositories for the abstracts and the bodies of the documents for purposes of the project.

The document collection was also indexed using Lucene. The collection was therefore accessible using Lucene's search technology. This enabled us to examine the relationship between using standard search technology and KMS functionality.

Working backward from the final report, we developed a series of questions that were, in essence, answered by different sections of the report. Except in some minor questions, the questions were not of the type that could be answered by fact-based question-

[3]Details of the project, such as the name of the client, are business-sensitive and cannot be made known at this time. Discussions are under way to determine if materials used in the project can be made publicly available.

answering of the type used in TREC (e.g, see Litkowski, 2005). Examples of questions include: *What is the incidence of diseases caused naturally by the toxin?, What animal models are best for modeling the effect of the toxin in humans?, What animals are most susceptible to the toxin?,* and *What nerve terminals are most affected by the toxin?*

The availability of an intermediate spreadsheet characterizing the document collection also provided a source of answers. For example, a column in the spreadsheet was used to identify the animals used in experiments involving the toxin. Thus, a standard question (or filter) of the document collection could ask what animals were studied.

In addition to answering questions such as the above, the questions themselves could be used as "topic" descriptions to produce topic-based summaries. KMS could also be used to explore the document content via its functionality to identify all noun phrases in the collection that had been semantically tagged during disambiguation with the relevant WordNet category for animals.

Finally, the biologist who wrote the final report (as well as other biologists) was available to assess results that were generated during the study. He also prepared a 200 word topic description when requested to provide a characterization of what kind of content would be relevant to addressing the general topic of *absorption* (i.e., how was the toxin absorbed into bodily organs in producing its toxic effects).

Thirty questions were submitted to KMS and to the Lucene database (the latter with hand-crafted boolean search expressions). KMS was not tailored to the more general questions described above, but rather attempted to answer the questions as fact-based. KMS and Lucene answers were randomly ordered and presented to two biologists who assessed the answers. Comparable and low results were obtained with both systems, each returning good answers in about five percent of the cases.

Several of the questions were then examined in more detail with a view toward constructing an information extraction query using XPath expressions. For many of the questions, good sentence answers could be obtained much more reliably than with the KMS question answering or with boolean search. For these questions, the sentences identified were significantly better than the results provided; the biologist felt that if the technique for creating appropriate XPath expressions could be automated, the results would be much superior to answers achieved by other methods. For many of the questions, however, this strategy was not effective (e.g., identifying which animals were most susceptible to the toxin). The issue is how to identify ways in which susceptibility are likely to be expressed and capturing these in a set of XPath expressions, making the process transparent to the user.

Several experiments were performed with summaries. First, we compared the notes prepared by the biologist in the spreadsheet (averaging about 145 words) with the abstracts for the documents where available (averaging 200 words), using them as the models. We used ROUGE to perform this analysis. The ROUGE-1 recall was 0.33617, indicating that the biologist was not capturing a considerable portion of what was in the abstracts (accounted for in part by the average size). However, the precision was 0.56511, indicating that the biologist was quite accurate in capturing the content of the documents, perhaps frequently cutting and pasting. The fact that precision was not higher suggests that the biologist was adding comments specific to his assessment of the document content and relevance for his purposes.

We next compared 200-word general summaries produced by KMS against the abstracts, against the notes, and against both together (as the model summaries). The results were considerably better than our performance in DUC 2005. Against the abstracts, the KMS summaries received a recall of 0.46353 and a precision of 0.42492. Against the notes, the recall was 0.47002, but the precision was 0.27369. These results suggest that for the task of producing a summary that is close to an abstract, KMS can perform relatively well. However, for matching the particular needs of the biologist, the precision indicates that matching a user's needs is much more difficult.

We next used the topic description on *absorption* provided by the biologist to identify and rank all sentences in the collection. As mentioned above, the topic description was employed in the same method as used for creating the DUC 2005 summaries. In this case, we added to KMS the capability for using a cutoff score to limit the number of answers that would be obtained. As a result, summaries were produced for only 162 of the 200 documents in the collection with 587 sentences identified. Increasing the cutoff score by 1 would have eliminated 215 of these sentences. After creating the summaries, the XML output files were processed with a Perl script to produce a list of the sentences in rank order. Although the biologist has yet to evaluate this output, it seems as if there is still a large number of sentences that are not specifically responsive to the biologist's requirements, but rather encapsulate what the biologist is looking for. More is still required to recognize sentences that are instantiations of what the biologist specified.

7 Conclusions and Future Developments

The primary objective of DUC 2005 was to assess the utility of various evaluation metrics. As described above, we found ROUGE to be the most useful for identifying improvements in our summarization routines. The linguistic quality and responsiveness metrics are very useful in pinpointing certain kinds of problems with the summaries that are created, although they are not reusable in the same way that ROUGE is. The pyramid method does not seem to be very workable in practice, since the rules for identifying summarization content units are not yet well-defined. Notwithstanding, the data provided by this year's annotations may be of great assistance in further characterizing and sharpening procedures for identifying semantic equivalence.

As a result of DUC 2005, we feel that the main problem of summarization is still the ability to recognize semantic equivalence and whether document content may be characterized as being an instantiation of what has been specified in the topic description. KMS currently has a nascent capability for creating ontological representations of each document. Currently, discourse entities (primarily noun phrases) in a document are analyzed into a taxonomy. This functionality is being extended to create a richer characterization by examining synonymic and other WordNet-like semantic relations. The relation set will be further extended to incorporate verb and other semantic relations present in a document into an output file that follows the Web Ontology Language (OWL). However, initial indications are that basing the reasoning system for an OWL representation will be too rigid. We expect that the reasoning system will have to be implemented in a tailored fashion allowing relaxed methods, such as abduction.

References

Litkowski, K. C. (2003). Text Summarization Using XML-Tagged Documents. Available: http://www-nlpir.nist.gov/projects/duc/pubs.html.

Litkowski, K. C. (2004). Summarization Experiments in DUC 2004. Available: http://www-nlpir.nist.gov/projects/duc/pubs.html.

Litkowski, K. C. (2005). "Evolving XML and Dictionary Strategies for Question Answering and Novelty Tasks. In E. M. Voorhees & L. P. Buckland (Eds.), *Information Technology: The Thirteenth Text REtrieval Conference (TREC 2004)*, NIST Special Publication. Gaithersburg, MD: National Institute of Standards and Technology. Available: http://trec.nist.gov/pubs.html.