CL Research Summarization in DUC 2006: An Easier Task, An Easier Method?

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

Abstract

In the Document Understanding Conference (DUC) for 2006, CL Research made a basic change in the method for assessing the significance of sentences in its Knowledge Management Systems summarization routines. This change led to an apparently significant improvement in scores compared to results for DUC 2005, with ROUGE-1 increasing from 0.348 to 0.388. After further detailed comparisons of the DUC 2006 results with those of other participants and with the effect of the change on DUC 2005 summaries, however, the improvement was not as significant as initially thought. Further analysis suggests that the DUC 2006 task might have been somewhat easier, perhaps because of more detailed topic descriptions. Notwithstanding, the change in the sentence scoring method simplifies the selection of key sentences by focusing on adjective and noun roots for sentence selection. It is suggested that documents can be rapidly scanned to identify significant sentences, which can then be examined in more detail with methods for detecting sentence similarity (or entailment) and overlap.

1 Introduction

CL Research made a basic change in its method for scoring sentences for the Document Understanding Conference (DUC) for 2005. 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. We were able to improve our performance substantially over our results for earlier years (Litkowski, 2005; Litkowski, 2004; and Litkowski, 2003). However, it appears that performance by other participants in DUC 2006 also improved dramatically. We suggest that this improvement is somewhat illusory and may be a result of improved topic descriptions in DUC 2006.

Section 2 presents a description of the DUC 2006 task. Section 3 provides an overview of KMS, with an emphasis on the extensions made during our preparations for DUC 2006 and the procedures used to perform the DUC task. Section 4 describes the KMS summarization procedures as used in DUC 2006. Section 5 presents and analyzes the DUC 2006 results, particularly characterizing attributes of the KMS summaries and comparing these attributes with results from other participants in DUC 2006 and with results from rerunning our system on the DUC 2005 task. Section 6 provides conclusions about our results and suggests next steps that can be taken to build upon the changed scoring method for assessment of sentence similarity and overlap.

2 DUC 2005 Task Description

DUC 2006 consisted of one task, to create a 250 word summary for each of 50 topics from 25 newswire articles in the AQUAINT corpus, from the *Associated Press Newswire*, *New York Times Newswire*, and *Xinhua News Agency*.. 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, and a narrative. Table 1 shows one topic and the information provided.

Table 1. Topic Description

Number	d0609I
Title	Israeli West Bank settlements
Description	What impact have Israeli settlements in the West Bank had on the Israeli/Palestinian peace process? What are the reactions of both parties and of the international community?

In the topic descriptions for DUC 2005 and earlier, two types of words were present: (1) retrieval task words (*explain*, *identify*, *report*) and (2) content specific words (*settlements*, *West Bank*, *peace process*). Some of the content words (*reasctions*) are general. In DUC 2006, the topic descriptions generally do not contain retrieval task words.

The human assessors hand-generated four summaries for each of the 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. 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. CL Research did not participate in this aspect of the DUC 2006 evaluation.

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 5.3 MB) were then parsed and processed into an XML representation (approximately 55.2 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 the lists developed in the previous phase to tag each element of each sentence in creating the XML-tagged version of the document.

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

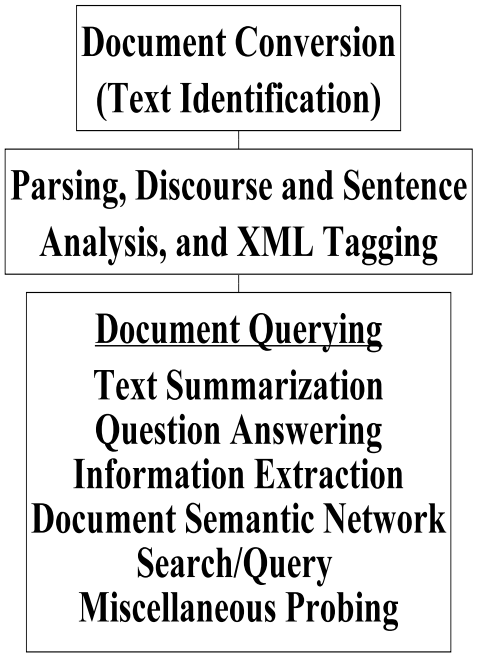

Figure 1. Architecture of Knowledge Management System

During the last year, a significant change was introduced into the characterization of discourse entities. Although the basic content of an XML representation was largely unchanged (i.e., consisting of the same attributes in the XML node), child nodes were added to break the discourse entity into its constituents. These child nodes are comparable to leaf nodes in a parse tree, and for the most part, consist of adjectives, adverbial modifiers, and nouns. The leaf nodes contain various attributes, most notably WordNet sense, other dictionary disambiguation sense identifiers, and root forms when the constituent is inflected.

The processed files are identified to KMS as a repository, 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 2006 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 2006

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 2006). 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 other 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.

5 Results and Analysis

Table 1 show CL Research's results for ROUGE-1, ROUGE-2, and ROUGE-SU4. The top score for all participating teams was 0.40488, up from 0.38036 in DUC 2005. 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.

Table 2. DUC 2006 ROUGE Recall

Granularity	Score	Rank
ROUGE-1	0.38803	10/34
ROUGE-2	0.07741	17/34
ROUGE-SU4	0.13318	17/34

In DUC 2005, our official ROUGE-1 score was 0.34849. Similarly, our ROUGE-2 score was significantly better. The ROUGE-2 score for DUC 2006 was at a level better than any participating team in DUC 2005. In fact, the level was within

only a short distance of the lowest score for a human summarizer. In spite of what appears to have been a significant improvement in performance, our overall rank was essentially the same, generally about the median value over all participating systems. When we reran DUC 2005 to take into account the modification to our scoring routine, the ROUGE-1 recall was improved by only 0.006, in contrast to the apparent improvement of 0.040, suggested above.

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	3.60
Non-redundancy	4.34
Referential clarity	3.16
Focus	3.80
Structure/coherence	2.48

On the measure of responsiveness, CL Research had an average score of 2.54 for content (18th of 34) and 2.18 overall, 17th of 34. For DUC 2005, our scores on linguistic quality and responsiveness were virtually the same as this year's performance. Thus, issues pertaining to these measures, as discussed in last year's report, still remain unresolved. 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.

To examine the performance of our system in more detail, we first examined the sentence number and the scores of the sentences selected for the summaries. As indicated above, in creating the XML output of the summaries that KMS generates, each sentence is identified specifically as to its source document, the sentence number within that document, and the score that was computed for that sentence.

Figure 2 shows a histogram of the sentence number frequencies. As expected for newswire texts, a large preponderance of the selected sentence (80 of 377 sentences in total) were the first sentences. However, a substantial number were selected from later positions in the source document. In general, while the first sentence may not present capsule statements about a topic, significant sentences can be expected within the first several sentences. However, sentence numbers lower than 10 only accounted for 60 percent of the

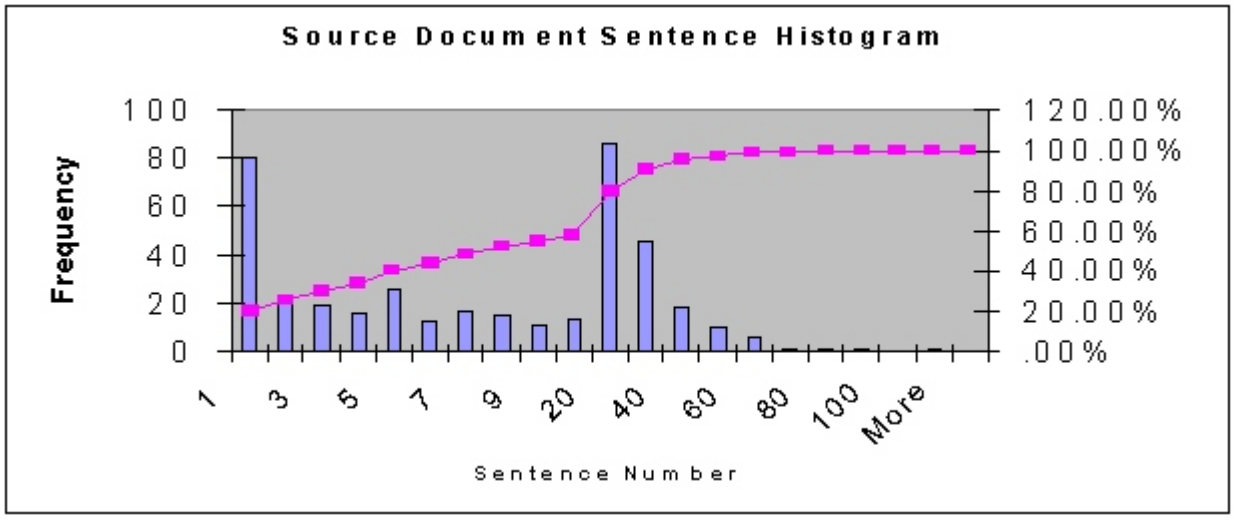

Figure 2. Sentence Number of Source Document Sentences in Summary

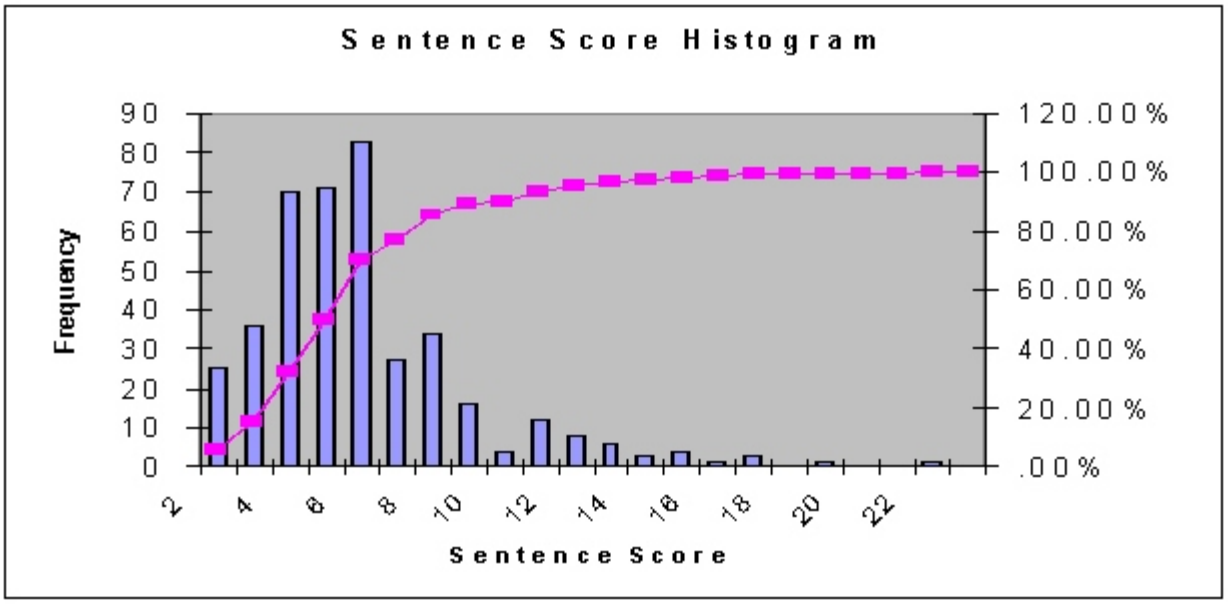

Figure 3. Sentence Scores of Sentences in Summary

selected sentences. Some newswire articles are actually compilations of several short pieces, only one of which is relevant to the topic at hand, and these can account for some of the higher sentence numbers. In general, though, Figure 2 indicates that KMS is not selecting the first sentence automatically. The average sentence position for DUC 2006 is 12.80. This contrasts with an average position of 17.24 for DUC 2005 documents. Although DUC 2005 used a different set of documents (from the *Financial Times* and the *Los Angeles Times*), it seems unlikely that the

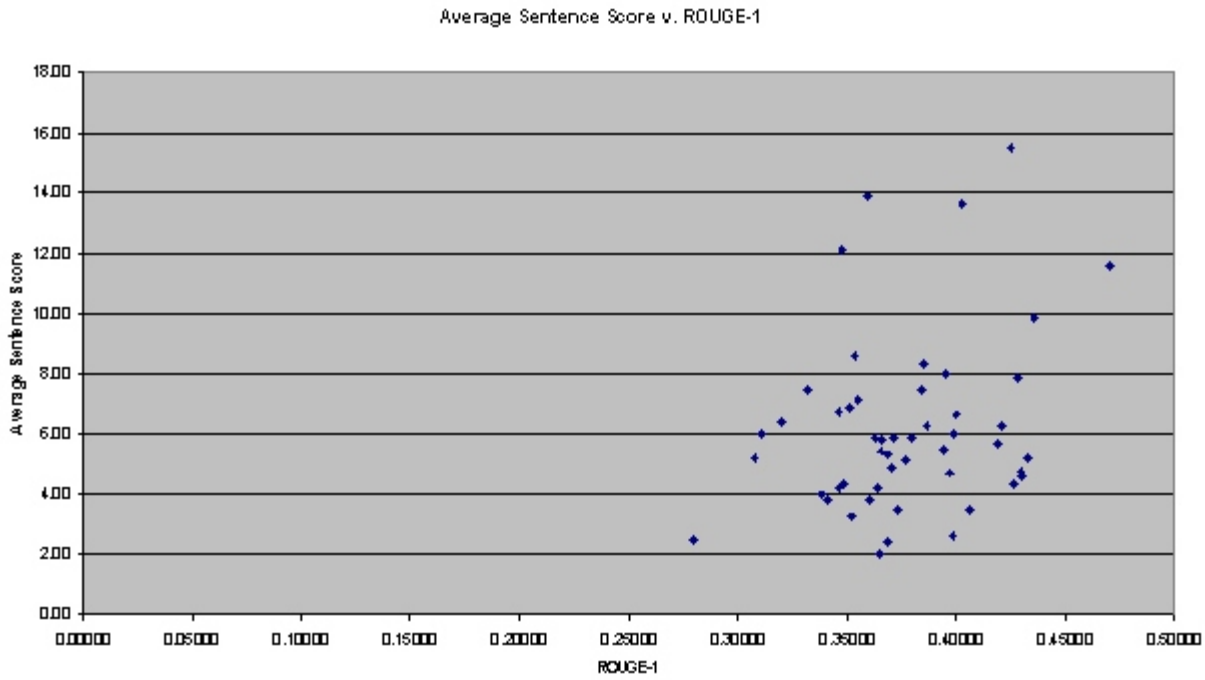

Figure 4. Average Sentence Score vs. ROUGE-1 Score

source documents account for the significant difference in sentence number.

Figure 3 shows a histogram of the scores for the sentences selected for the summaries. As described above, increments to the score for a sentence is based primarily on the presence of (the base form of) words in the topic description in the sentence. Generally, only one point is given for each match, with capitalized words given an additional point. The minimum score for a sentence to be selected is 2, so that is the lowest point in the histogram.

As can be seen in the histogram, the modal value is 7, with large numbers of sentence having scores of 5 or 6. The cumulative curve shows that about 90 percent of all selected sentences have scores of 10 or lower. The average score for a selected sentence in DUC 2006 is 6.05. For DUC 2005, the average score was 5.16. As pointed out above, this difference agrees with the intuition that DUC 2005 questions contained more general words, which were unlikely to be used in sentences in the documents.

We next correlated the ROUGE-1 scores by the average sentence scores. Figure 4 shows a scatter plot of ROUGE-1 against the average sentence score for each topic. As can be seen, there is only a weak correlation between the two; the correlation coefficient is 0.27. By contrast, for DUC 2005, the correlation coefficient between the ROUGE-1 scores and the sentence scores was only 0.156, suggesting that there is a real difference between the tasks in the two years. Since KMS was essentially unchanged from 2005 to 2006, the increased correlation seems lie in the way that the topic descriptions were constructed.

6 Conclusions and Future Developments

The improvement in our results stems from both the change in the XML representation (i.e., to include leaf nodes) as well as the modification in scoring that looks at the base forms of nouns and adjectives. This suggests that a potentially useful and efficient method for identifying important sentences can be first to scan texts for nouns and adjectives and to obtain their base forms. These base forms can then be used to look for synonyms and hyponyms. When a candidate set of sentences has thus been identified, they can be subjected to further more detailed analysis of similarity, paraphrase detection, and entailment.

As discussed in Litkowski (2006) and Dagan et al. (2006), many methods are available for recognizing textual entailment. In the second PASCAL challenge for Recognizing Textual Entailment (RTE-2), considerable advances have been made in this area, particularly for summarization. The methods developed in KMS provide a basic step for efficiently identifying sentences that can then be subjected to procedures used for recognizing textual entailment.

References

Bar-Haim, R., I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini and I. Szpektor. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Trento, Italy. Available: http://www.pascal-network.org/Challenges/RTE2.

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

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 Summarization Strategies in DUC 2005. Available: http://www-nlpir.nist.gov/projects/duc/pubs.html.

Litkowski, K. C. (2006). Componential Analysis for Recognizing Textual Entailment. In Bar-Haim, R., I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini and I. Szpektor, *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Trento, Italy.

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.