UNC-CH at DUC 2007: Query Expansion, Lexical Simplification and Sentence Selection Strategies for Multi-Document Summarization

Catherine Blake, Julia Kampov, Andreas K. Orphanides, David West, and Cory Lown

School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27499 USA
`{cablake,jkampov,ako,dwest,lown}@email.unc.edu`

Abstract

This paper describes the approach used in the UNC-CH system to generate a topic-focused summary of information reported in multiple news articles. We explored query expansion, lexical simplification and sentence simplification. Results suggest that cluster membership plays an important role in improving summarization performance, while query expansion does not. The UNC-CH system performed well in both automated and manual evaluations, achieving the 12th highest ROUGE-2 score and a score greater than or equal to the average system responsiveness score for 30 of the 45 DUC 2007 topics.

1 Introduction

The goal of the Document Understanding Conference (DUC) is to advance research in development of automatic document summarization systems. DUC 2007 had two tasks. The main task was to generate a fluent 250-word summary from a given topic, query, and set of 25 documents containing information pertinent to the topic. NIST assessors evaluated each automatically generated summary manually with respect to linguistic quality (grammaticality, non-redundancy, referential clarity, focus, and structure and coherence), and responsiveness (the amount of information that helped to satisfy the expressed information need from the topic). The second task was to provide a user with new information related to an event given news articles over time. We participated in only the main task.

As this is the first year that UNC-CH participated in DUC, we went through the process of preparing manual extractive summaries for two of the 2006 topics before starting any system development. We strategically selected topics that were easy and difficult using the average system performance as a measure of topic difficulty.

Our primary goal in 2007 was to get a baseline system up and running quickly so that we could experiment with different settings. To achieve this goal the first author reviewed DUC 2006 system descriptions and identified components that appeared to work well. The most popular components were query expansion, lexical simplification, sentence selection, sentence generation, clustering, and sentence cohesion. In this paper, we report results on the first three of those components.

In addition to the engineering motivation required for a first time DUC participant, our hypothesis was that lexical simplification using linguistic sentence features would improve system performance. Thus, the UNC-CH system includes a component to prune gerundive clauses, noun appositives, non-restrictive relative clauses, intra-sentential attributions, and lead adverbials.

In this paper, we describe how the UNC-CH system balances query expansion, lexical simplification, and sentence selection, and the experiments we ran to tune parameters. We then compare and contrast the performance of the UNC-CH system with other automated and manual summaries produced as part of DUC 2007.

2 System Architecture

System development and tuning used only the DUC 2006 corpus. The implementation employs an Oracle 10g database manipulated using Java.

2.1 Document Pre-Processing

The UNC-CH system uses a custom Java program to split news articles into sentences. After reviewing the processed DUC 2006 documents, additional abbreviations, such as Sgt. were added to correct erroneous sentence splits. Errors such as mark-up tags that split titles, missing paragraph marks, and the source location, as part of the article text were corrected manually.

As with most linguistically motivated systems, our approach requires a dependency grammar representation of each sentence. Systems in the past have used a variety of tools to generate a dependency grammar, such as Minipar (Lin, 1998), Link grammar (Grinberg, Lafferty, & Sleator, 1995), the Collins Parser (Collins, 1997), and the Stanford Parser (Klein & Manning, 2002; 2003, 2003). These experiments use the type dependency grammar generated from the Stanford Parser (version 1.5) (de Marneffe, McCartney, & Manning., 2006).

2.2 Query Expansion

Query expansion (QE) can be a powerful way to improve the recall of relevant sentences, but often comes at a cost of decreased precision. We explored two forms of query expansion, both based only on nouns. The first experiment used a lexicon to identify the base form of each topic and query term; for example, the term 'communities' was expanded to 'community'. In the second experiment, we passed permutations of the original and base forms to WordNet(Miller, 1995). In both cases, the system did not expand stop words such as 'the', 'a', 'and'. Based on preliminary experimental results we do not use WordNet to expand topic of query terms in the DUC 2007 system.

2.3 Lexical Simplification

Lexical simplification can aid in summarization by removing sections of a sentence that do not contain essential information. Zajic et al have used such simplification to generate an article headings (Zajic et al., 2005). Approaches may be either purely statistical (Knight & Marcu., 2000) or linguistically motivated (see examples below). The UNC-CH system uses the latter and prunes noun appositives, gerundive clauses, part modifiers, as well as attribution and adverbial clauses from a sentence.

Our approach is most similar to (Vanderwende, Suzuki, & Brockett, 2006) who removed noun appositive, gerundive clause, non-restrictive relative clause, intra-sentential attribution, lead adverbials and to (Conroy, Schlesinger, O'Leary, & Goldstein, 2006) who removed extra words, adverbs, attributable information, joining words, gerund phrases. Our approach differs from (Vanderwende et al., 2006) in that we do not consider the distribution of material from the documents as they do when selecting the final sentences. In contrast to (Conroy et al., 2006), the UNC-CH system does not use relative clauses such as 'whom', 'which', and 'when' and has far fewer heuristics than the more developed CLASSY system.

In addition to the linguistic features described above, we added a drastic pruning step, which we call sub-sentences. This step identifies all minimal clauses in a sentence – branches in the dependency tree that comprise both a subject and an object. Although this is a drastic form of pruning, our intuition was that these clauses would contain grammatically correct sentences that have the most concentrated meaning. (See section (1) below for details). As with other tree pruning approaches, we use both heuristics and the dependency tree representation to identify pruned branches.

(1) Sub-Sentences: In addition to the original sentence, we identified syntactically valid sub-sentences by extracting all branches of the dependency that contain a subject and an object. For example, the system would include the original sentence and the two bolded sections shown below.

```
But it went on to say that e-
conomic reform has not
brought political freedom and
that Chinese who try to dis-
sent "live in an environment
filled with repression."
```
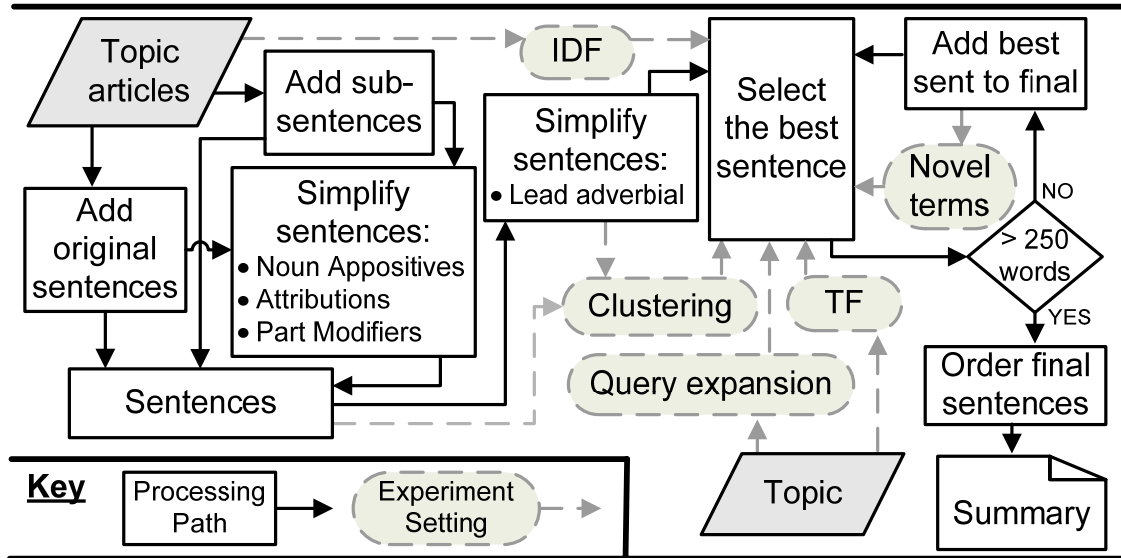
Figure 1. System Architecture

We did explore expanding conjunctive clauses; however, our preliminary analysis with queries suggested that such an approach would not be fruitful. In contrast to all other linguistic features, the system generates sub-sentences before any additional pruning takes place.

(2) Noun appositives: An appositive is a noun phrase used to modify another noun phrase. Fortunately, the Stanford Parser tags such branches in the dependency tree with the `appos` (MacCartney & Galen, 2006), which the system uses to identify appositives. For example, the system would prune the bolded text from the following sentence.

```
For nearly a decade, Queen
Latifah, the first lady of
hip-hop, has been bobbing and
weaving questions about
whether she prefers princesses
to princes in her queendom.
```

(3) Participial Modifier: A participial modifier is a meaningful phrase containing a participle, such as "running", "related" etc that modifies a noun or verb phrase. The system again uses a Stanford Parser tag, in this case `partmod` (MacCartney & Galen, 2006). This class of pruning also identifies gerundive clauses. For example the system would prune the section of bolded underlined text.

```
Indeed, some people reading
this report could get the im-
```

```
pression that Amnesty be-
lieves violence can be a
legitimate instrument, the
statement said.
```

(4) Lead Adverbials: Adverbial phrases, such as 'Also' and 'In fact,' are often used to open a sentence, but typically do not provide important information. In contrast to the other lexical simplification methods, the system prunes lead adverbials from all sentences, regardless of earlier pruning. We did try to identify Stanford Parser tags (such as `ccomp`), but heuristics that from previous DUC papers and by looking at the text in the DUC 2006 corpus achieved better accuracy.

(5) Intra-sentential attributions: A good reporter will cite the source of quoted material, but manual summaries rarely use this information. Thus, our system prunes attribution information from sentences using heuristics to identify branches of the dependency tree. The system would remove 'the statement said' from the example participial modifier sentence in (3), and the bolded text in:

```
Lynn Cutler, the president's
top adviser on Indian issues,
said it's the largest spending
increase ever sought for Indi-
ans and includes new or ex-
panded programs in nearly all
federal agencies.
```

2.4   Sentence Selection

A variety of weighting schemes were explored to identify the best sentences, where performance was measured using the 2006 corpora and ROUGE 1.5. Note that frequencies in 1-3 do not include stop words. Factors that we considered in these experiments:

(1) Query expansion. The final system uses only the original terms in the topic or query, and the base form of the word. (See the section 3.1 for the WordNet query expansion evaluation).

(2) Percentage of terms (%WdTopic). The number of stemmed terms in the topic and query divided by the number of stemmed terms in the sentence.

(3) Percentage of unique terms (%WdNew). The number of stemmed terms in the topic and query that are not already in the summary, divided by the number of stemmed terms in the sentence.

(4) Weighted Term frequency (wtf). The following weighting scheme to favor topic and query terms from a sentence that differed from the terms already in the summary. The weight of a sentence is the sum of the weighted term frequency of all words within that sentence.

Feature	Weight
Stopword or punctuation	0
Topic/Query ∧ ¬Summary	1
Topic/Query ∧ Summary	0.5
¬ Topic/Query ∧ ¬Summary	0.01
¬ Topic/Query ∧ Summary	0.001

(5) Weighted Term Frequency x IDF (wtfidf). In information retrieval systems, the inverse document frequency (IDF) is an effective way to decrease the weight of terms that appear throughout the entire corpus (Spärck Jones, 1972). Combining the term frequency with the IDF weight thus results in a higher weight for terms with more discriminative power. IDF is calculated the for the entire corpus (as apposed to calculating IDF for each topic) and used the combination of weighted term frequency outlined above and IDF.

(6) Clustering (CW). Several sentences in the corpus contain very similar information. To reduce redundancy, the system clusters both the original and pruned sentences using a K-means clustering algorithm with 100 clusters, and 1000 iterations to reach equilibrium. Our preliminary experiments with sentence clustering revealed that stop words and punctuation dominated the clusters, so we exclude all non-content terms including determiners, prepositions, auxiliary verbs, conjunctions and dependency terms. The system orders the sentences by the cluster membership. We wanted to bias the system towards selecting non-redundant sentences. To achieve this the system clusters sentences and favors larger clusters by first ranking clusters with respect to the number of sentences (clusterRank). Each sentence has a cluster score from 0 to 100 that reflects how well the sentence captures the cluster centroid (clusterScore). Lastly, to favor sentence drawn from different clusters, so the system keeps a boolean value that indicates if the cluster has already been selected (newCluster). The overall cluster weight, (depicted as CW in Figure 3 and Table 3) is newCluster * [(1/clusterRank) * clusterScore].

3   DUC 2006 Evaluation

The following experiments were conducted to evaluate query expansion and sentence selection parameters independently of the other system components.

3.1   Query Expansion

We designed the following experiment to measure the impact of QE in DUC 2006 collection and tune our system for 2007. Manually written summaries in DUC provide a gold standard for evaluating the summarization task, but they are problematic for query expansion evaluations because the manual summaries are not extractive, i.e. human summarizers re-order and re-word information in the original articles. To alleviate this difference, we developed an alternative gold standard based on the DUC 2006 corpus. Three annotators, read the 25 documents related to nine different topics. The annotators identified sentences that contained information pertinent for each topic.

		CL		Tot
		rel	¬rel	
	rel	55	25	80
JK	¬rel	36	448	484
	Tot	91	473	564

Kappa 0.58 (moderate)

		TB		Tot
		rel	¬rel	
	rel	42	38	80
JK	¬rel	61	423	484
	Tot	103	461	564

Kappa 0.36 (fair)

		CL		Tot
		rel	¬rel	
	rel	41	62	103
TB	¬rel	50	411	461
	Tot	91	473	564

Kappa 0.30 (fair)

Figure 2a. Inter-Rater Reliability for Topic 6

		CL		Tot
		rel	¬rel	
	rel	50	48	98
JK	¬rel	36	653	689
	Tot	86	701	787

Kappa 0.48 (moderate)

		TB		Tot
		rel	¬rel	
	rel	43	55	98
JK	¬rel	29	660	689
	Tot	72	715	787

Kappa 0.45 (moderate)

		CL		Tot
		rel	¬rel	
	rel	41	31	72
TB	¬rel	45	670	715
	Tot	86	701	787

Kappa 0.47 (moderate)

Figure 2b. Inter-Rater Reliability for Topic 34

To measure inter-rater reliability, all three annotators reviewed two topics (6 and 34). Figure 2a and 2b show that there was moderate agreement between annotators regarding relevance of specific sentences to the topic query. Although annotators did not completely agree on relevancy, agreement was higher for topic 34 than for topic 6. Annotators reviewed topic 6 first, so their increased familiarity with the review task may be responsible for the higher agreement in topic 34.

Once each annotator had independently identified relevant sentences, they reached consensus for topics 6 and 34, and then reviewed other topics shown in Table 1.

With the gold standard in place, we developed four query expansion mechanisms: (A) Any word in the topic or query; (B) Original or base form of words in the topic or query that are not stop words; (C) Limited WordNet query expansion including any term from the same synset as WordNet terms that were identified from the original or base form of the topic or query terms and (D) WordNet expansion where terms from any synset that was a synonym of C were included. Both query expansions C and D use WordNet version 3.0 (Miller, 1995). Our experiments before submitting the 2007 responses used the JWord 2.0 interface (http://home.gwu.edu/~kjohar/), but the results shown in Table 1 reflect data from the WordNet 3.0 data files.

The precision (P) shown in Table 1 is the proportion of relevant vs. non-relevant sentences retrieved and recall (R) is the proportion of relevant sentences retrieved vs. relevant sentences that the

		A				B				C				D			
T	rel	ret	rel	P	R	ret	rel	P	R	ret	rel	P	R	ret	rel	P	R
07	74	426	73	0.17	0.99	209	46	0.22	0.62	209	46	0.22	0.62	209	46	0.22	0.62
34	101	759	101	0.13	1.00	355	70	0.20	0.69	353	70	0.20	0.69	354	70	0.20	0.69
23	119	330	95	0.29	0.80	231	83	0.36	0.70	217	81	0.37	0.68	230	83	0.36	0.70
48	35	687	34	0.05	0.97	144	26	0.18	0.74	114	21	0.18	0.60	118	21	0.18	0.60
12	77	974	77	0.08	1.00	351	43	0.12	0.56	329	42	0.13	0.55	333	42	0.13	0.55
13	70	814	59	0.07	0.84	113	28	0.25	0.40	105	27	0.26	0.39	105	27	0.26	0.39
06	84	529	82	0.16	0.98	155	38	0.25	0.45	155	38	0.25	0.45	155	38	0.25	0.45
10	69	1020	67	0.07	0.97	280	38	0.14	0.55	257	36	0.14	0.52	257	36	0.14	0.52
03	77	367	73	0.20	0.95	181	63	0.35	0.82	31	13	0.42	0.17	48	16	0.33	0.21
Avg	78	656	73	0.13	0.94	224	48	0.23	0.61	197	42	0.24	0.52	201	42	0.23	0.53

Table 1. Query Expansion Results

system failed to identify. Relevance is the consensus sentences for topics 6 and 34 and the independently identified sentences for the remaining topics. The topics in Table 1 are listed in increasing order of difficulty, i.e. the total manual scores given to system generated summaries for topic 7 (109) were much better than for topic 3 (54).

Table 1 suggests that although using any terms in the topic or query provides excellent recall (94%), the precision is very low (13%). The recall performance dropped from experimental condition B to C, which suggests that several of the terms used in the topic and query from DUC 2006 are not in WordNet. Although expanding sentences to include synonyms improved recall from condition C, the system achieved better recall performance without using query expansion. In contrast to our hypothesis that query expansion would have a negative impact on precision, these results show little change in precision performance, and a slight increase 23 to 24%. Based on these findings, our DUC 2007 system uses only the original and base form of terms found in the topic and query, and does not use any form of query expansion.

3.2 Sentence Selection Evaluation

We explored several combinations of the sentence selection strategies outlined in section 2.4 and measured the performance of each using ROUGE 1.5.5 with the following settings: ROUGE-1.5.5.pl -e data -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a –d.
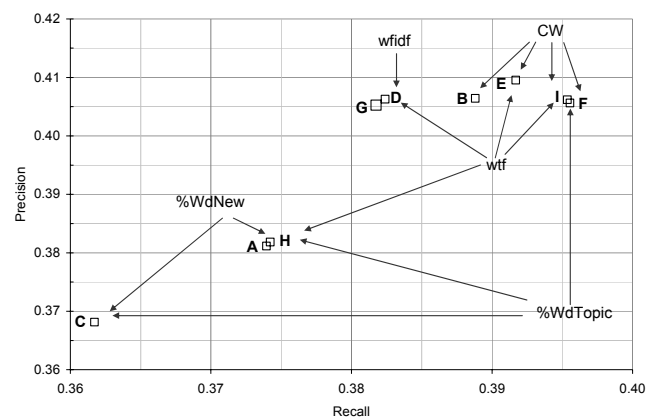

Figure 3. ROUGE-1 scores for alternative sentence selection strategies in DUC 2006

Figure 3 illustrates the trade-off between precision and recall. Table 3 shows the sentence selection strategies that we experimented with for this

paper and the average ROUGE-1 F-measures for both DUC 2006 and 2007. The UNC-CH system uses best strategy, I. To put these results in context, consider that the average F-measure for all systems that competed in DUC 2006 was 0.37791 and the average human score was 0.45766. The labels show different weighting schemes and suggest that including a clustering component improves sentence selection, which is consistent with previous summarization systems such as Mead (Radev et al., 2004).

4 DUC 2007 Evaluation

NIST provides three main evaluations of each system in DUC; one automated evaluation, ROUGE, which compares the system generated summary with the four manually written summaries from NIST evaluators; and two manual evaluations of responsiveness and linguistic quality. Two baseline systems are included for comparison, the first 250 words of the most recent document, and a high-performance generic single document summarizer (CLASSY04).

With respect to automated evaluation, the UNC-CH system received the 12th highest ROUGE-2 score of 0.10329 (95%-conf.int.0.09933-0.10725). Given that this is the first time UNC-CH has participated in the document summarization conference, and that we designed, developed and evaluated the system over the semester break, we found these results particularly encouraging.
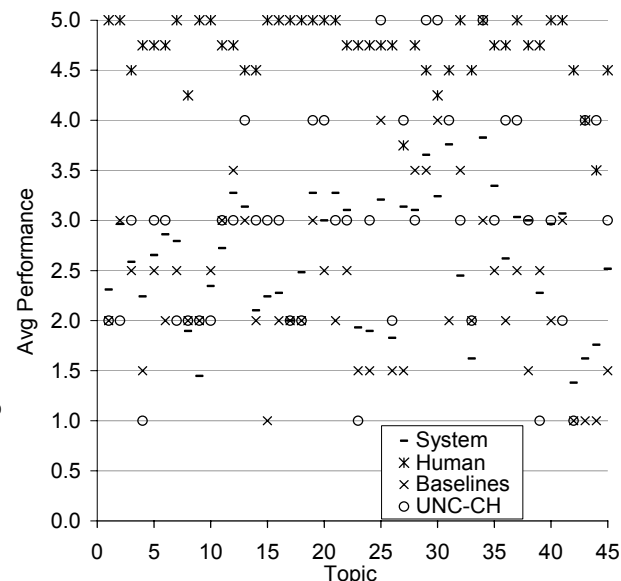

Figure 4. DUC 2007 Manual Responsiveness

Responsiveness (also called content in DUC 2007) captures the amount of information in the summary that helps to satisfy the information need expressed in the topic and query. The average topic content score for systems in DUC 2007 other than UNC-CH was 2.6276 (ranging from 1.3793 to 3.8276); the average for baseline systems was 2.2889 (1.0-4.0) and the average for human summaries was 4.7111 (3.5 to 5.0). The UNC-CH system achieved an average topic content of 2.9556 (1.0 to 5.0), which placed the system at 7th out of 32. Figure 4 shows the average responsiveness per DUC 2007 topic and shows that the UNC-CH responsiveness score was greater than or equal to the average system responsiveness score for 30 of the 45 topics.
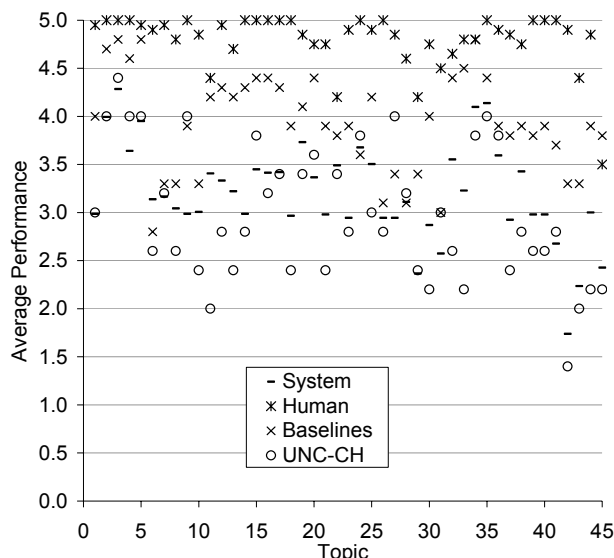

Figure 5. DUC 2007 Linguistic Quality

The DUC linguistic quality has five dimensions: 1.Grammaticality, 2.Non-redundancy, 3. Referential clarity, 4.Focus, 5.Structure and Coherence. The average linguistic quality per topic for systems other than UNC-CH was 3.1972 (1.7379 to 4.2828), the average for the baselines was 3.9289 (2.8-4.8) and the average for human summaries was 4.8022 (3.5 to 5.0). The UNC-CH system average was 2.9867 (1.4 to 4.4). In contrast to responsiveness, our system performed no worse in only 16 of the 45 topics (see Figure 5).

To understand the linguistic results we looked at each of the five dimensions and the UNC-CH system constantly scored below average. The only dimension where the system performed almost as well as the other systems was with grammaticality.

Our subsequent analysis of the generated summaries revealed that the pruning applied to generate sub-sentences was too severe and produced sentences that really did not stand alone. We are currently exploring methods to resolve this issue.

ID	Description	DUC06	DUC07
I	(tottf/numWdSent*CW	0.3981	0.4212
F	%WdTopic*CW	0.3979	0.4171
E	tottf*CW	0.3977	0.4183
B	CW	0.3947	0.4169
D	Tfidf	0.3912	0.4086
G	Tottf	0.3904	0.4109
H	tottf/numWdSent	0.3754	0.3913
A	%WdTopic+%WdNew+CW	0.3749	0.3963
C	%WdTopic+%WdNew	0.3623	0.3786

Table 3. ROUGE-1 F-Measures for alternative Sentence Selection Strategies

In addition to the overall system evaluations, we were curious to see if the performance of the sentence selection strategies that we developed using DUC 2006 correlated with performance with 2007 results. Table 3 shows the result of each sentence selection strategy described in section 3.2 and the subsequent F-measure. The results suggest that the topics in 2007 were easier than 2006, and that cluster membership continued to play an important role in achieving good ROUGE performance.

5   Conclusion

The most intriguing finding from our DUC 2007 participation was the importance of a good sentence selection strategy. We explored several features to favor sentences with a high proportion of topic and query terms, in particular topic and query terms that had not yet been included in the summary. This lead to the **weighted term frequency** described in section 2.4. To reduce redundancy, the system clustered both the original and pruned sentences using a K-means clustering algorithm. The **cluster weight**, which favors sentences in large new clusters, where the sentence best reflects the new clusters' centroid had the strongest impact on system performance. The UNC-CH system uses the optimal sentence selection strategy (I), which combines the weighted term frequency, the number of words in the sentence, and the cluster weight.

Our primary hypothesis in DUC 2007 was that linguistically motivated pruning would produce a pool of grammatically valid sentences that the system could compare with the topic and query statements. Using a combination of heuristics and the dependency grammar representation produced from the Stanford Parser, the system removed gerundive clauses, noun appositives, non-restrictive relative clauses, intra-sentential attributions, and lead adverbials. We also included a drastic sentence pruning strategy that identified syntactically valid sub-sentences – dependency tree branches that contained both a subject and an object. A subsequent review of the sub-sentence pruning was responsible for the many of the ungrammatical sentences and we are working on methods that produce pruned sentences with more context.

Our secondary hypothesis was that query expansion would play an important role in summarization performance. To test this hypothesis, three annotators, read the 25 documents related to nine different topics and identified relevant sentences. We ranked the DUC 2006 topics in descending order of system performance and deliberately selected topics along the spectrum from easy to most difficult. Inter-rater reliability, for two of the topics ranged between fair and moderate. The annotators reached consensus on the relevant sentences, which we used to calculate precision and recall for four different query expansion methods. Our results showed that query expansion had negligible system improvement and thus we removed this component from the UNC-CH system.

Given that 2007 is UNC-CH's first participation in DUC, we are particularly encouraged by achieving the 12th highest ROUGE-2 score and score greater than or equal to the average system responsiveness score for 30 of the 45 DUC 2007 topics.

Acknowledgements

References

Collins, M. (1997). *Semantic tagging using a probabilistic context free grammar.* Paper presented at the Proceedings of 6th Workshop on Very Large Corpora.

Conroy, J. M., Schlesinger, J. D., O'Leary, D. P., & Goldstein, J. (2006). *Back to Basics: CLASSY 2006.* Paper presented at the Document Understanding Workshop, Brooklyn, New York USA.

de Marneffe, M.-C., McCartney, B., & Manning., C. D. (2006). *Generating Typed Dependency Parses from Phrase Structure Parses.* Paper presented at the LREC 2006.

Grinberg, D., Lafferty, J., & Sleator, D. (1995). *A robust parsing algorithm for link grammars.* Paper presented at the Proceedings of the Fourth International Workshop on Parsing Technologies, Prague.

Klein, D., & Manning, C. (2002). *Fast Exact Inference with a Factored Model for Natural Language Parsing.* Paper presented at the Advances in Neural Information Processing Systems.

Klein, D., & Manning., C. (2003). *Accurate Unlexicalized Parsing.* Paper presented at the In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003).

Knight, K., & Marcu., D. (2000). *Statistics-based summarization -- step one: Sentence compression.* Paper presented at The 17th National Conference of the American Association for Artificial Intelligence.

Lin, D. (1998). *Dependency-based Evaluation of MINIPAR.* Paper presented at the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation, Granada, Spain.

MacCartney, B., & Galen, A. (2006). Class English-GrammaticalRelations. JavaNLP Parser API Documentation. Stanford: Stanford NLP Group.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), 39-41.

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., et al. (2004). *MEAD - A platform for multidocument multilingual text summarization.* Paper presented at the LREC 2004, Lisbon, Portugal.

Karen Spärck Jones (1972) *A statistical interpretation of term specificity and its application in retrieval.* Journal of Documentation, 28:11-21.

Vanderwende, L., Suzuki, H., & Brockett, C. (2006). *Microsoft Research at DUC 2006: Task-Focused Summarization with Sentence Simplification and Lexical Expansion.* Paper presented at the Document Understanding Workshop, HLT-NAACL 2006, Brooklyn, New York USA.

Zajic, D., Dorr, B., Schwartz, R., Monz, C., & Lin, J. (2005). *A Sentence-Trimming Approach to Multi-Document Summarization.* Paper presented at the HLT-EMNLP Workshop on Text Summarization.