The University of Lethbridge Text Summarizer at DUC 2003

Yllias Chali Maheedhar Kolla Nanak Singh Zhenshuan Zhang
Department of Mathematics and Computer Science
University of Lethbridge
4401 University Drive
Lethbridge, Alberta, Canada, T1K 3M4
E-mail: {chali, kolla, singh, zhangj9}@cs.uleth.ca

Abstract

This paper presents the techniques implemented in the University of Lethbridge Text Summarizer for producing very short and short summaries from both single and multiple documents. These techniques show that good quality result may be generated when using topical clues identified by the lexical information. We present these techniques in the context of DUC 2003.

1 Introduction

For DUC 2003, the National Institute for Standards and Technology (NIST) produced 30 document sets/clusters from the TDT collection, 30 document sets/clusters from the TREC collection, and 30 document sets/clusters from the TREC Novelty collection as test data. Each cluster from TDT or TREC collection contains on average 10 documents. Each cluster from the TREC Novelty collection contains on average 22 documents. Additionally, NIST provided topics for the 30 TDT sets, and viewpoint descriptions for the 30 TREC sets. Also, NIST provided questions for the 30 TREC Novelty clusters and the set of sentences deemed relevant to the question. Four tasks were defined in DUC 2003:

Task 1 – Using the 30 TDT clusters and the 30 TREC clusters, and given each document, create a very short summary (approximately 10 words) of the document.

Task 2 – Using the 30 TDT clusters, and given each document cluster and the associated TDT topic, create a short summary (approximately 100 words) of the cluster.

Task 3 – Using the 30 TREC clusters, and given each document cluster and a viewpoint description, create a short summary (approximately 100 words) of the cluster from the point of view specified.

Task 4 – Using the 30 TREC clusters, and given each document cluster, a question, and the set of sentences in each document deemed relevant to the question, create a short summary (approximately 100 words) of the cluster that answers the question.

We participated in the four tasks. We describe below the algorithms and the results for each task.

2 Task 1

2.1 Algorithm

1. Segment the original document into segments that address the same topic (Choi, 2000).
2. Tag the words in the segment according to the part of speech they represent (Ratnaparkhi, 1996).
3. Parse the tagged words into their syntactic structure (Collins, 1997).
4. Filter the nouns from the parsed segment using the idea that nouns contained within subordinate clauses are less useful for topic detection than those contained within main clauses.
5. Compute the lexical chains (Morris and Hirst, 1991) for each segment.
6. Rank the segments based on:

$$score(seg_j) = \sum_{i=1}^{m} \frac{score(chainMember_i, seg_j)}{s_i} \tag{1}$$

   where $score(chainMember_i, seg_j)$ is the number of occurrences of a $chainMember_i$ in seg_j, m is their number, and s_i is the number of segments in which $chainMember_i$ occurs.

   The top n segments - with the highest scores - are chosen for the process of sentence extraction. Note that the process of segment selection can substantially limit the size of the important portions of the text.

7. Extract the relevant sentences using a scoring process which considers the number of words shared by the sentence and the chains that have been considered in the segment selection phase.
8. Generate the 10-word summary using some heuristics to do some surface repairs on the summaries. Such heuristics were eliminate some function words, and when we cut the sentence, we make sure that it ends properly.

2.2 Results

Average of mean coverage	0.40
Average of median coverage	0.40
Average of mean length-adjusted coverage	0.28
Average of median length-adjusted coverage	0.28
Average of usefulness	2.38

3 Task 2 and 3

3.1 Algorithm

1. Order the set of documents in the cluster according to their chronological order using patterns of dates in the documents.

2. Concatenate the set of documents according to their chronological order.

3. Segment the concatenated document into segments that address the same topic (Choi, 2000).

4. Tag the words in the segment according to the part of speech they represent (Ratnaparkhi, 1996).

5. Parse the tagged words into their syntactic structure (Collins, 1997).

6. Filter the nouns from the parsed segment using the idea that nouns contained within subordinate clauses are less useful for topic detection than those contained within main clauses.

7. Compute the lexical chains (Morris and Hirst, 1991) for each segment.

8. Rank the segments based on:

$$score(seg_j) = \sum_{i=1}^{m} \frac{score(term_i, seg_j)}{s_i} \tag{2}$$

   where $term_i$ is a chain member and/or a term from the TDT topic in the case of task 2. For task 3, $term_i$ is a chain member and/or a term from the viewpoint description. $score(term_i, seg_j)$ is the number of occurrences of a $term_i$ in seg_j, m is their number, and s_i is the number of segments in which $term_i$ occurs.

   The top n segments - with the highest scores - are chosen for the process of sentence extraction. Note that the process of segment selection can substantially limit the size of the important portions of the text.

9. Extract the relevant sentences using a scoring process which considers the number of words shared by the sentence and the chains that have been considered in the segment selection phase, or the TDT topic associated with the cluster, or the viewpoint description associated with the cluster.

10. Generate the 100-word summary using some heuristics to do some surface repairs on the summaries.

3.2 Task 2 Results

Average of mean of quality questions	0.92
Average of mean coverage	0.17
Average of median coverage	0.03
Average of mean length-adjusted coverage	0.12
Average of median length-adjusted coverage	0.03

3.3 Task 3 Results

Average of mean of quality questions	0.73
Average of mean coverage	0.19
Average of median coverage	0.07
Average of mean length-adjusted coverage	0.13
Average of median length-adjusted coverage	0.05

4 Task 4

4.1 Algorithm

1. For each document cluster, consider the set of sentences deemed relevant to the question.

2. Extract terms from the question.

3. Expand the terms by adding semantically related words to the terms.

4. For each sentence deemed relevant to the question, compute the number of matches between the expanded terms and the sentence. This gives us $score(sentence_i)$ for each sentence.

5. Extract the top ranked sentences.

6. Generate the 100-word summary using some heuristics to do some surface repairs on the summaries.

4.2 Results

Average of mean of quality questions	0.89
Average of mean coverage	0.14
Average of median coverage	0.05
Average of mean length-adjusted coverage	0.10
Average of median length-adjusted coverage	0.04
Average of responsiveness	1.88

5 Conclusion

In this paper we have shown that short summary for single and multi-document can be obtained using topical clues. We have presented the summarization procedure that produces summary of variable size (i.e., very short and short). The DUC 2003 evaluations show that this technique produces reasonable results when considering single document, however, some improvements are still required when considering multi-document summarization.

Acknowledgments

This work was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada and Research Excellence Envelope funding from the Alberta Heritage Foundation for Science and Engineering Research.

References

Barzilay, R. and M. Elhadad. (1997). Using lexical chains for text summarization. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter Meeting of the Association for Computational Linguistics, Workshop on Intelligent Scalable Text Summarization*, pages 10-17, Madrid.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics*, pages 26 - 33, Seattle, Washington.

Collins, M. (1997). Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter Meeting of the Association for Computational Linguistics*, Spain.

Lin, C. Y. and E. Hovy. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of 18th International Conference in Computational Linguistics*, Saarbrücken, Germany.

Mani, I. (2001). *Automatic Summarization*. John Benjamins.

Mani, I. and M. Maybury. (1999). *Advances in Automatic Text Summarization*. MIT Press.

Morris, J. and G. Hirst. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21-48.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.

Silber, H. G. and K. F. McCoy. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496.