# The University of Lethbridge Text Summarizer at DUC 2002

Meru Brunn Yllias Chali Barbara Dufour
Department of Mathematics and Computer Science
University of Lethbridge
4401 University Drive
Lethbridge, Alberta, Canada, T1K 3M4
E-mail: {brunnm9, chali, dufourb9}@cs.uleth.ca

August 30, 2002

**Abstract**

Text summarization addresses both the problem of selecting the most important portions of text and the problem of generating coherent summaries. We present in this paper the summarizer of the University of Lethbridge at DUC 2002, which is based on an efficient use of topical clues. The method we present addresses the problem of producing summaries, in the context of single and multiple documents.

## 1 Introduction

We present a system for identifying the most important portions of the text which are *topically* best suited to represent the source text according to the author's views in the context of summarizing single-document and in the context of summarizing sets of documents that all describe the same event. This identification must also take into consideration the degree of *connectiveness* among the chosen text portions so as to minimize the danger of producing summaries which contain poorly linked sentences. These objectives can be achieved through an efficient use of *topical clues.*

The overall architecture of the system is outlined in *Figure 1.* It consists of several modules organized as a pipeline.

## 2 Summarization Algorithm

1. Preprocessing

   (a) Segmentation: the original text is first divided into segments that address the same topic (Choi, 2000).

   (b) Tagging: which is essential for using the parser, involves classifying words in the segment according to the part of speech they represent (Ratnaparkhi, 1996).

   (c) Parsing: tagged words are collected and organized into their syntactic structure (Collins, 1997).

2. Noun Filtering: selectively removes nouns from the parsed text. We designed heuristics using the idea that nouns contained within subordinate clauses are less useful for topic detection than those contained within main clauses. For our system, we selected a relatively simple heuristics. Such heuristics are:

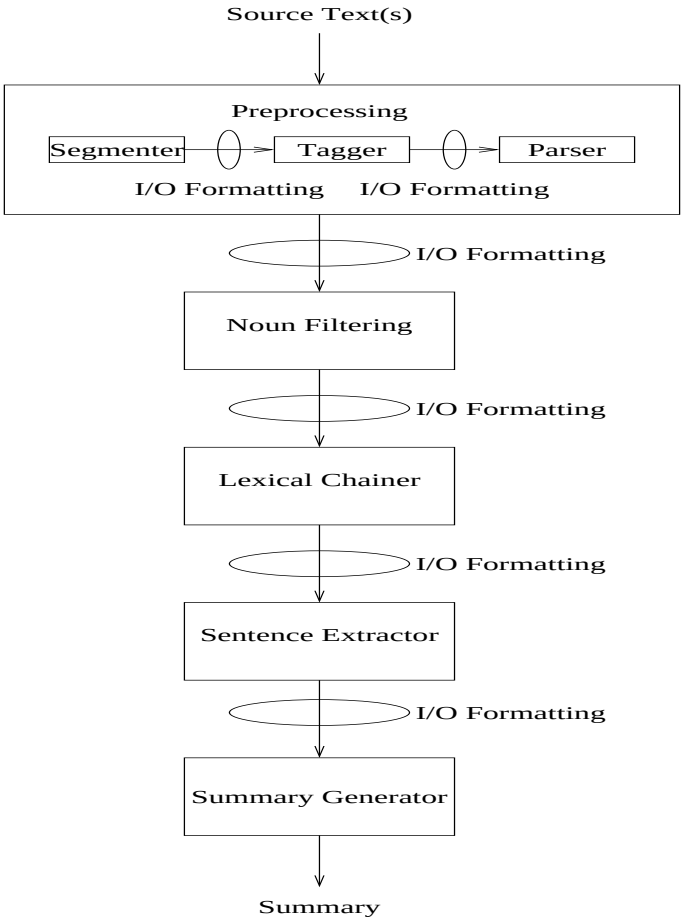


Figure 1: System Overview

(a) *remove noun phrases of complement clause,*

(b) *remove noun phrases of relative clause,*

(c) *remove noun phrases of abbreviated clause,*

(d) *remove noun phrases of adverbial clause,*

(e) *remove noun phrases of infinitival clause,* and

(f) *remove noun phrases of present participial clause.*

3. Lexical chaining: computes the lexical chains (Morris and Hirst, 1991) for each segment.

4. Sentence extraction

   (a) Segment selection:

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

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.

(b) Sentence extraction: the score for $sentence_i$ is the number of words that belong to $sentence_i$ and also to those chains that have been considered in the segment selection phase.

5. Summary generation: we use some heuristics to do some surface repairs on the summaries. Such heuristics are:

   (a) *add previous sentence to a sentence containing a dangling anaphora*,

   (b) *remove sentences with less than N words*,

   (c) *remove sentences with quotation marks*, and

   (d) *remove sentences with question marks*.

   The summary consists of the ranked list of top-scoring sentences, according to the desired compression ratio, and ordered in accordance with their appearance in the source text.

## 3 Multi-Document Summarization

Summarizing sets of documents that all describe the same event is handled by putting the segments of documents stemming from the same set in the same pool of segments, that is, contrary to single document summarization where the pool of segments are issued from the single document, in multi-document summarization the pool of segments contains the segments of all documents in the set. However, to solve the problem of detection the chronological order of the documents issued from the same set, we implement a procedure that assigns time-stamps to each document from the same set according to their chronological order. This procedure is based on finding patterns of dates in the documents using regular expressions, and ordering the documents according to the detected patterns.

## 4 Example

Below is a sample of the single document summaries for each document in the set d061j, respectively.

- AP880911-0016 Hurricane Gilbert Heads Toward Dominican Coast
- AP880912-0095 Gilbert Reaches Jamaican Capital With 110 Mph Winds
- AP880912-0137 Hurricane Hits Jamaica With 115 mph Winds; Communications Disrupted
- AP880915-0003 Storms Batter Yucatan; Thousands Flee
- AP880916-0060 What Makes Gilbert So Strong?
- WSJ880912-0064 Hurricane Gilbert Heading for Jamaica With 100 MPH Winds

(1) a. Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a “broad area of cloudiness and heavy weather” rotating around the center of the storm. The

weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.

b. Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic. Forecasters say Gilbert was expected to lash Jamaica throughout the day and was on track to later strike the Cayman Islands, a small British dependency northwest of Jamaica. Hurricane warnings were posted for the Cayman Islands, Cuba and Haiti. Cuba's official Prensa Latina news agency said a state of alert was declared at midday in the Cuban provinces of Guantanamo, Holguin, Santiago de Cuba and Granma. Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane's center passed just south of the Barahona peninsula, then less than 100 miles from neighboring Haiti.

c. Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines. Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island. Sheets said Gilbert was expected next to sweep over the Cayman Islands, on its westward track, and in two to three days veer northwest into the southern Gulf of Mexico. The National Weather Service warned that the Caymans could expect high waters and large waves "which may undermine buildings along the beaches."

d. Hurricane Gilbert, one of the strongest storms ever, slammed into the Yucatan Peninsula Wednesday and leveled thatched homes, tore off roofs, uprooted trees and cut off the Caribbean resorts of Cancun and Cozumel. Officials were checking low-lying areas of Cancun for stranded residents. As Gilbert moved away from the Yucatan Peninsula Wednesday night, the hurricane formed a double eye, two concentric circles of thunderstorms often characteristic of a strong storm that has crossed land and is moving over the water again. Only two Category 5 hurricanes have hit the United States - the 1935 storm that killed 408 people in Florida and Hurricane Camille that devastated the Mississippi coast in 1969, killing 256 people. On Sunday, Monday and Tuesday, Gilbert pounded the Dominican Republic, Jamaica and the Cayman Islands.

e. Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. It's certainly one of the larger systems we've seen in the Caribbean for a long time," said Hal Gerrish, forecaster at the National Hurricane Center in Coral Gables, Fla. At 3 p.m. EDT, the center of the hurricane was about 100 miles south of the Dominican Republic and 425 miles east of Kingston, Jamaica. In Puerto Rico, besides tearing off several roofs, the storm caused coastal flooding and brought down power lines and trees along roads and highways in the west and southwestern regions.

Below is a sample of the multidoc summary, both the 200 word and 400 word summaries, respectively. It is from the set d061j.

(2) Hurricane Gilbert's growth from a harmless low pressure zone off Africa to a ferocious killer in the Gulf of Mexico was fueled by a combination of heat, moisture and wind that baffles forecasters. Ham radio operators around Cancun said Gilbert knocked down a radio and television communications tower, uprooted trees and blew the roofs off buildings. As Gilbert moved away from the Yucatan Peninsula Wednesday night, the hurricane formed a double eye, two concentric circles of thunderstorms often characteristic of a strong storm that has crossed land and is moving over the water again. At 3 p.m. EDT, the center of the hurricane was about 100 miles south of the Dominican Republic

and 425 miles east of Kingston, Jamaica. Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island. Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane's center passed just south of the Barahona peninsula, then less than 100 miles from neighboring Haiti. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.

(3) Hurricane Gilbert's growth from a harmless low pressure zone off Africa to a ferocious killer in the Gulf of Mexico was fueled by a combination of heat, moisture and wind that baffles forecasters. Prime Minister Edward Seaga of Jamaica said Wednesday the storm destroyed an estimated 100,000 of Jamaica's 500,000 homes when it throttled the island Monday. Ham radio operators around Cancun said Gilbert knocked down a radio and television communications tower, uprooted trees and blew the roofs off buildings. There was heavy damage visible to the humble wood and thatch homes typical of the Yucatan. As Gilbert moved away from the Yucatan Peninsula Wednesday night, the hurricane formed a double eye, two concentric circles of thunderstorms often characteristic of a strong storm that has crossed land and is moving over the water again. On Sunday, Monday and Tuesday, Gilbert pounded the Dominican Republic, Jamaica and the Cayman Islands. Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba. At 3 p.m. EDT, the center of the hurricane was about 100 miles south of the Dominican Republic and 425 miles east of Kingston, Jamaica. In Puerto Rico, besides tearing off several roofs, the storm caused coastal flooding and brought down power lines and trees along roads and highways in the west and southwestern regions. Hurricane warnings were issued Monday for the south coast of Cuba east of Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued for the Dominican Republic. Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island. Sheets said Gilbert was expected next to sweep over the Cayman Islands, on its westward track, and in two to three days veer northwest into the southern Gulf of Mexico. Cuba's official Prensa Latina news agency said a state of alert was declared at midday in the Cuban provinces of Guantanamo, Holguin, Santiago de Cuba and Granma. Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane's center passed just south of the Barahona peninsula, then less than 100 miles from neighboring Haiti. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.

## 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

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 (ACL-97)*, Spain.

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.