Using Background Information for Multi-document Summarization and Summaries in Response to a Question

Atefeh Farzindar
Guy Lapalme

RALI
Département d'informatique et de recherche opérationnelle
Université de Montréal
C.P.6128, Succ Centre-Ville
Montréal, Québec, Canada, H3C 3J7
{farzinda,lapalme}@iro.umontreal.ca

Abstract

In our second participation to the DUC evaluation, we used the SumUM system for the multi-document summarization focused by events task and summaries in response to a question task. Our multi-document summarization algorithm is based on the use of background information gathered by summarizing previous texts, which is then combined with a new document of the cluster. For producing the summaries to answer a question, we used a similar technique applied on passages retrieval using Okapi.

1 Introduction

The main purpose of a summary is to give the reader an accurate and complete idea of the contents of the source and to present it to the user in a condensed form in natural language. In the new world of information, especially for the newswire or paper stories domain, producing a single-abstract is seldom enough to process a vast collection of information. This is why NIST is now considering multi-document summarization whose goal is to identify what is common and what differs in a variety of the related documents and to remove redundant information from the summary

For DUC 2003 NIST (National Institute of Standards and Technology of U.S.) produced 90 document sets as test data. For this purpose NIST used the documents from the TDT and TREC collections and incorporated focus of various sorts to reduce variability and to better model real tasks.

Four different tasks were evaluated at DUC 2003 (we decided to take part in the second and fourth tasks):

- Very short summary of a single document. (~10 words)
- Short summaries focused by events: given a document cluster and the associated TDT topic, create a short summary of the cluster (~100 words).
- Short summaries focused by viewpoints: given a document cluster and a viewpoint description, create a short summary of the cluster from the point of view specified (~100 words).
- Short summaries in response to a question: given a document cluster, a question, and the set of sentences in each document deemed relevant to the question, create a short summary of the cluster that answers the question (~100 words).

At DUC2002, the results of the evaluation of our system SumUM [1] (Summarization at

Université de Montréal) for single-document abstraction had been very encouraging: between 13 participant systems, SumUM was ranked first according to the mean score for questions about the quality of the produced output; for the mean length-adjusted coverage SumUM was ranked second. So we decided to participate to DUC evaluation for a second time by starting from the same system, but we concentrated our efforts in the development of a multi-document capability based on the strengths of SumUM for single-document summarization. We also added some new additional functions for generating the short summaries in response to a question.

2 Description of SumUM

SumUM was used last year for producing single-document abstracts of scientific and technical papers. SumUM is based on the identification of the structure of the text (words, groups of words, titles, sentences, paragraphs, sections etc.) and the detection of specific concepts and relations. For example, in a scientific article, the first section is often the one where most information about the subject can be found, this is why SumUM ranks higher information given at the start of the article. SumUM identifies concepts and relations of an article by means of templates containing patterns of words most often encountered in the scientific domain. From the information gathered in the templates a short (10-15 lines) indicative summary is generated to identify the topics of the document; an informative phase can then be used to elaborate on some topics according to the interest of the reader.

3 Adaptation of SumUM for multi-document summarization

The current version of our summarizer is a continuation of last year's system but with enhancements suggested in our analysis of last year's DUC results and by experimentation with various new components to produce multi-document summaries as required by the specifications of DUC 2003 (summaries focused by events and summaries in response to a question).

3.1 Generation the short summaries focused by events

Our approach to multi-document summarization results from our observations on a corpus of newspapers to present a series of events for human readers. We have studied different types of newspapers to understand what is considered to be a good method for presenting a collection of information about a specific concept to a reader and how the reader can combine the new information of the current article with previous information to manage to follow a series of events.

One common way for following a succession of events in a story is to keep some background about the subject in his mind and to search for novel information about this topic in an article. Figure 1 shows an example taken from the first page of *The Montreal Gazette* where can be found the **Headline** of event, the **Background**, **News** and **Next part** with page reference number to find more details about this event. This organization will help the reader through the reading and understanding of the news. Given a short background about each event, a reader can easily find the needed information in the related documents. Over time, a reader gets more information in the background for a new document about a precise concept and is helped in searching for new information while avoiding some inherent redundancy that appears in a series of events.

IN FOCUS

Credit-card numbers stolen

Background: Most credit-card holders are protected against unauthorized transactions on their accounts.
New: An "unauthorized intruder" gains access to 8 million credit-card account numbers – including Visa, MasterCard and American Express – by breaching the security of a company that processes transactions for merchants.
Next: Online merchants and security experts fear the stolen numbers will be used on Web sites and at call centres. **Page B1**

Figure 1: Use of background information for keeping track of a series of events in a newspaper article (Montreal Gazette 02/19/2003)

We perform text summarization by using a similar process to extract information from a multi-document cluster, the background being given by the *indicative summary* that can be obtained from SumUM. The documents of a cluster are ordered in increasing chronological order (regular expressions are used as patterns of dates) and summarized by considering the previous summary as background for the current text. SumUM regards the first section as an *introduction* and thus considers its content as quite important, so it will keep the main information in the background for the resulting summary. For the first document, the oldest, we use the topic of multi-document as its background.

Each background summary plays the role of a topic-oriented text for a new document, which is summarized according to the important topic of its background. With this procedure we can find and cluster the important topics of document set and the common information that might be relevant for a multi-document summary.

For implementing this algorithm within SumUM, two *tricks* are used:

- SumUM is *cautious* to the information found in the first section of the text, which is considered as the introduction;
- Sentences that contain word coming from titles and subtitles of a document have more weight and more relevance about the subject.

As the documents in a cluster are ordered in ascending order of their date, the important topics of the background are inserted as titles for the new text and the summary is added as first section of the new text. Figure2 shows an example of some of the results of this process. Figure 3 shows how the length of summaries varies along this iterative process..

Given Topic of document by NIST, for cluster d30042t:

```
PanAm Lockerbie Bombing Trial
Seminal Event
     WHAT: Kofi Annan visits Libya to appeal for surrender of
           PanAm bombing suspects
     WHERE: Tripoli, Libya
     WHO: U.N. Secretary-General Kofi Annan; Libyan leader
          Moammar Gadhafi
     WHEN: December, 1998
```

Summary, output of system:

Summary of the text of 11/23/1998 is used as background for the one of 11/25/1998, which was summarized according to the important topics of its background.

<DOCNO>APW19981123.1112</DOCNO>
<DATE_TIME> **11/23/1998 15:49:00** </DATE_TIME>
African countries voted in June to ignore the U.N. flight ban which was imposed in 1992 to try and force Libya to hand over for trial two suspects wanted in the 1988 bombing of an American airliner over Lockerbie, Scotland.
<DOCNO>APW19981125.0279</DOCNO>

<DATE_TIME> **11/25/1998 06:13:00** </DATE_TIME>
CAIRO, Egypt _ Newspapers reported Wednesday that three top Libyan officials have been tried and jailed in the Lockerbie case, but Libyan dissidents said the reports appeared to be a political ploy by Libya's leader, Col. The reported jailing of the three officials comes as Gadhafi is under pressure to accept a plan to turn over for trial two other Libyans wanted for the 1988 bombing of Pan am flight 103 over Lockerbie, Scotland, that led to 270 deaths. Two london newspapers, the Guardian and the leading Arabic daily Al-Hayat, reported Wednesday that three top intelligence chiefs at the time of the airliner bombing had been convicted and imprisoned in Libya. After refusing for years to turn over the two men for trial in the United States or Britain, Libya recently accepted in principle the proposed trial in The Hague, Netherlands, but has delayed in actually turning over the two suspects. Libya has been under U.N. sanctions since 1992 to force it to hand over the two defendants in the Lockerbie case.

Final summary of multi-document generated by system:

African countries voted in June to ignore the U.N. flight ban which was imposed in 1992 to try and force Libya to hand over for trial two suspects wanted in the 1988 bombing of an American airliner over Lockerbie, Scotland. The reported jailing of the three officials comes as Gadhafi is under pressure to accept a plan to turn over for trial two other Libyans wanted for the 1988 bombing of Pan am flight 103 over Lockerbie, Scotland, that led to 270 deaths. The visit was Farrakhan's fifth to Libya in the past three years. The leader of the U.S.-based Nation of Islam most recently visited in December 1997.

Figure 2: An example of using background information for multi-document summarization

Date	Document	Source words	Summ words
10/20/1998 16:24	APW19981020.1108	521	38
10/21/1998 18:30	NYT19981021.0303	777	10
10/28/1998 08:34	APW19981028.0445	319	45
10/31/1998 13:18	NYT19981031.0088	1424	120
11/23/1998 15:49	APW19981123.1112	122	40
11/25/1998 06:13	APW19981125.0279	475	172
11/25/1998 12:18	APW19981125.0886	375	58
11/29/1998 12:54	APW19981129.0652	271	64

Figure 3 : Number of the words of source documents and their summaries in the cluster of Figure 2: There were only 8 different texts in this cluster of 10.

3.1.1 Discussion

In this section, we have presented an algorithm for automatic multi-document summarization based on gathering more information in the background for a new document about a specific concept. The fact in this method, we rely on ordering the documents of a cluster in increasing chronological order for generating the backgrounds,. We also tried this algorithm in decreasing date order in which the first text was the newest one, which would seem to contain more information and thus be a good candidate for background information for the other documents in this cluster. Unfortunately, we noted that the quality of the final summary worsened because over time the abstract of the newest document gets compressed more as more and more summary information from old documents are added. This is the reason why we have chosen the increasing chronological order for the documents in our summarization procedure.

3.2 Generating short summaries in response to a question

We now describe how we adapted SumUM to produce a summary as a response to a question in a cluster of documents. For a given a cluster and a question, we produce a short summary (~100 words) which answers the question.

Before summarizing, we select the most relevant passages of the document using Okapi [2] to retrieve variable length passages. Okapi is an information retrieval engine that has the ability to return relevant paragraphs instead of whole documents. We feed it with the question as a query and we set it up so that it returns all passages responsive in a cluster. The relevant passages are the paragraphs that could be the candidate for the summary of multi-document; a paragraph has an average of 5 sentences.

Passages are ordered according to the ranked list of the topic, which we consider as one *long* document, which is given to the summarization function to produce a short summary. According the important features in a question and the topics of document set, SumUM summarizes the ordered passages of relevant sentences to create a summary of about 100-words. The output for each cluster is the summary with complete re-generated sentences in response to a question.

For this task, we used essentially the same *tricks* as for the previous section: we summarize passages using the headlines and the question as titles. So SumUM looks for the words of the title in the text and produces a reasonable summary. Given that we were aiming for a 100 words summary, if the resulting summary is more than 130 words long we only consider the most recent headlines. If the summary is too short (less than 80 words), we add the *old* headlines as topics for producing the longer summaries.

3.2.1 Discussion

Although, in this task, NIST had provided a set of relevant sentences in each document for the question, we did not use them because their use did not seem to serve any useful purpose. We imagine well how to create a summary directly from the cluster and the question but it is not yet clear to us why. somebody, who has already relevant sentences about a cluster, would want another summary

4 NIST Evaluation

NIST evaluated the summaries intrinsically for **quality** and **length-adjusted coverage** using a modified SEE[3]. DUC measured coverage - how much of a model summary's content was expressed by a system-generated peer summary. It also tried to look at the ability of a system to produce a summary shorter than the predefined target length and to devise a combined measure of coverage and compression.

Although we did not into deep modifications into the SumUM system, we achieved better numeric scores this year than last year on mean coverage, precision and recall. On the other hand, the mean score for quality questions were better last year because we produced shorter text (20 words average instead of about 120) which might explain we had more errors this year.

5 Conclusion

We have shown how we managed to adapt the summarization engine of SumUM, which had been designed for single documents, in order to participate successfully in the multi-document DUC competition of this year. Using the fact that SumUM ranks higher information in the first paragraphs of a text, we used previous outputs of as background for new summaries. By iteratively summarizing texts in chronological order, we were able to achieve good multi-document summaries that characterize well the context of a cluster of documents.

6 References

[1] Saggion, H. and G. Lapalme, (2002), “Generating Informative and Indicative Summaries with SumUM”, *Computational Linguistics*, Special Issue on Automatic Summarization

[2] Okapi-Pack: <http://www.soi.city.ac.uk/~andym/OKAPI-PACK>

[3] DUC 2002: Length-Adjusted Coverage <http://duc.nist.gov/duc2002/covbrev.html>.