SCISSOR- Auto-Extraction Tool to Boost Document Scanning

Jaini Vora
Ex-student
D.J. Sanghvi College of Engineering
Mumbai, India.

Chetashri Deshmukh
Assistant Professor
D.J. Sanghvi College of Engineering
Mumbai, India.

ABSTRACT
Text summarization has become an important tool for assisting and interpreting text information in today's fast-growing information age. It is very difficult and time consuming for human beings to manually summarize large documents of text. This paper discusses 'Scissor', the simple yet reliable tool, which allows the creation of a shortened version of a text by a computer program. The output of this tool contains only the most important points of the original text.

General Terms
Pattern Recognition and Algorithms

Keywords
Scissor, Extraction, Summary, Sentences, Cardinality

1. INTRODUCTION
There is a huge amount of text data available on the internet. Due to the information overload, access to precise and correctly-developed summaries is important. As access to data has increased, so has interest in summarizing the relevant data. Technologies that can make a useful summary of any kind of text, need to take into account several factors such as length, writing-style and syntax.

Summarization has several applications like summarizing the search-engine results, providing briefs of big documents that do not have an abstract etc. There are two categories of summarizers, extractive and abstractive. Extractive summarizers operate by finding the important sentences using statistical methods (like frequency of a particular word etc). Abstractive ones use linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text. Extractive summarizers normally do not use any linguistic information.

2. PROBLEM DEFINITION AND SCOPE
In this work, an auto-summarization tool is developed using statistical technique. The technique involves finding the frequency of words, scoring the sentences (based on the cardinality), ranking the sentences etc. The summary is obtained by selecting a particular number of sentences (specified by the user) from the top of the list. It operates on a single document (but can be made to work on multiple documents by choosing proper algorithms for integration) and provides a summary of the document. The size of the summary can be specified by the user when invoking the tool.

A document can be viewed as a cluster of sentences which may or may not be related to each other. Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. Similar sentences usually refer to a common subject. If we represent each sentence with a node, then a link between two sentences represents the similarity between them.

Thus, nodes with maximum cardinality in a given sub-graph represent sentence with more relevant information. These sentences become important candidates for the summary. Also the disjoint sub graphs represent different sections or topics addressed by the document.

The summarization tool enables user to specify the amount of summarization. The amount of summarization indicates the number of sentences that will appear in the output summary. Some users may require just an overview of the document whereas some other users may have some prior knowledge about the document and might require more elaborate summary.

3. PROPOSED SYSTEM
The tool will shorten the input text by extracting important information or sentences from it. The amount of text to be shortened can be defined by the user as per his/her requirements. Then the sentences will be filtered depending on the threshold that is set depending on the amount of text user wants as the summary.

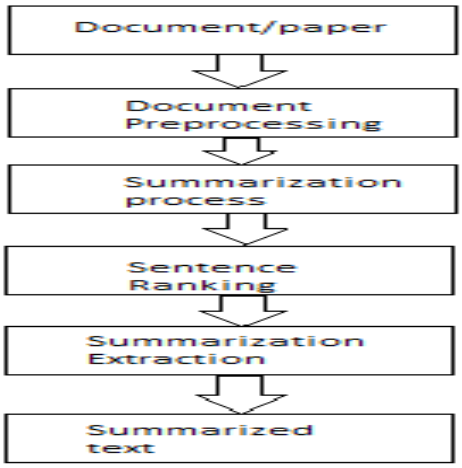

Figure 1: Proposed System

3.1 Document Pre-Processing
In the pre-processing step, the input text will undergo some processing to make it suitable for the summarization process. It includes converting all capital letters to lowercase; removal of stop words like 'a', 'an', 'the' etc. It also helps in concentrating on the main subjects of the input text.

3.2 Query Based Summarization
The summarization can also be done using query inputted by the user. The query can comprise of various topics separated by comma. In this case the summary will be generated in such a way that it will focus on the topic/subject given by the user.

The other sub-topics in that paragraph will be ignored during summarization.

3.3 Summarization Process

The tool will be built using a graph structure in which all the sentences will be represented using nodes and the related sentences will be linked to each other. The sentences that are closely related to each other have several links joining them. This representation yields two results: The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub graphs) form distinct topics covered in the documents. This allows a choice of coverage in the summary. For query-specific summaries, sentences may be selected only from the pertinent sub graph, while for generic summaries, representative sentences may be chosen from each of the sub-graphs. User can also specify the amount of summary required.

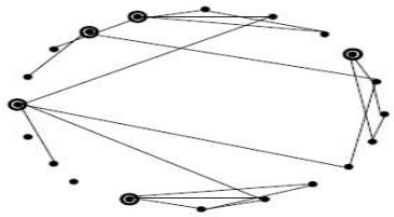

Figure 2: Graph Theoretic Approach

3.4 Sentence Ranking

The second result yielded by the graph-theoretic method is the identification of the important sentences in the document. The nodes with high cardinality (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference for inclusion in the summary.

3.5 Summarization Extraction

At the end the important sentences will be extracted from the rest of the paragraph.

4. SYSTEM ARCHITECTURE

The components of system architecture include:

4.1 Graphic User Interface

User uploads the document and accesses the summarized text.

4.2 Stop Words Remover

The document is first pre processed to remove all stop words like is, a, the etc.

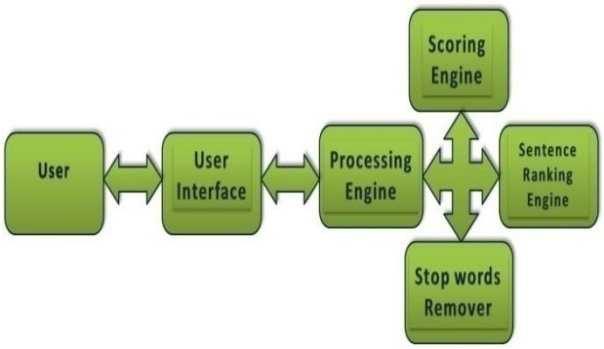

Figure 3: System Architecture

4.3 Processing and Sentence Ranking Engine

A graph is created for each document where node corresponds to each sentence. Nodes with maximum cardinality are selected for summarized text.

5. CONTROL FLOW

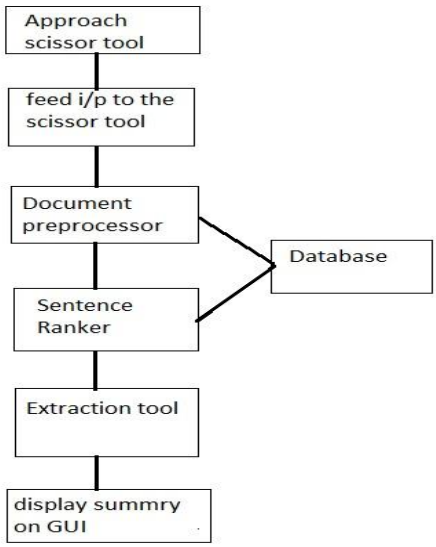

Figure 4: Control Flow

The user gives input to the summarization tool by pasting the text in the textbox or by browsing for a document stored in the PC. The summarization tool gets the control and it scans through the input data. Stop-words are eliminated and each sentence is analyzed for the score. Each sentence is given a score by the scoring algorithm using graph-theoretic approach. The sentences with high score are extracted and given as summarized output.

6. EXPERIMENTAL RESULTS

The Graphical User Interface of the tool is as shown below

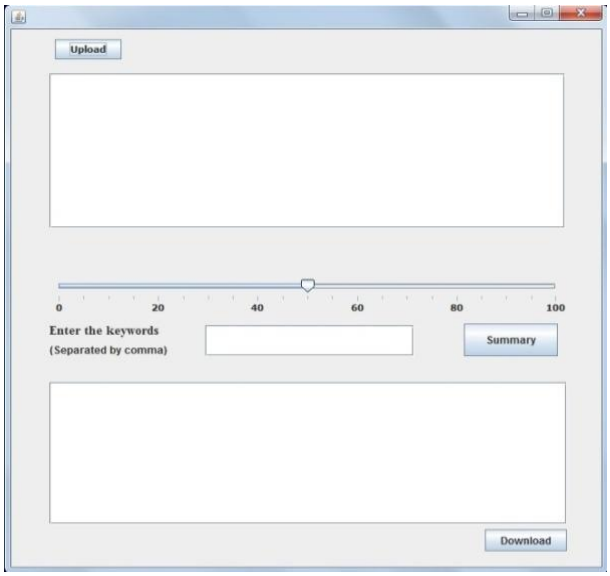

Figure 5: GUI

Consider the text given below as our sample document, which is a ten-sentence long document.

"A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun. This configuration can only occur during a new moon, when the Sun and the Moon are in conjunction as seen from the Earth. In ancient times, and in some cultures today, solar eclipses are attributed to mythical properties. Total solar eclipses can be frightening events for people unaware of their astronomical nature. The Sun suddenly disappears in the middle of the day and the sky darkens in a matter of minutes. However, the spiritual attribution of solar eclipses is now largely disregarded. Total solar eclipses are very rare events for any given place on Earth. Totality is only seen where the Moon's umbra touches Earth's surface. A total solar eclipse is a spectacular natural phenomenon and many people consider travel to remote locations in order to observe one. The 1999 total eclipse in Europe, said by some to be the most-watched eclipse in human history, helped to increase public awareness of the phenomenon."

The summary generated by our summarizer:

10% summary

A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun.

25% summary

A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun. The Sun suddenly disappears in the middle of the day and the sky darkens in a matter of minutes. Total solar eclipses are very rare events for any given place on Earth.

50% summary

A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun. Total solar eclipses can be frightening events for people unaware of their astronomical nature, as the Sun suddenly disappears in the middle of the day and the sky darkens in a matter of minutes. A total solar eclipse is a spectacular natural phenomenon and many people consider travel to remote locations in order to observe one. The 1999 total eclipse in Europe, said by some to be the most-watched eclipse in human history, helped to increase public awareness of the phenomenon.

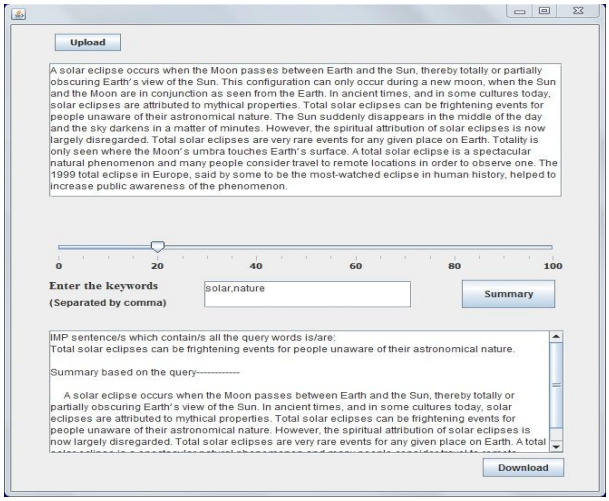

Figure 6: 50% Summary

The above output shows 50% summary of the above input text.

Figure 7: Query-Based Output

The above output shows the query based summary. User can enter keywords or topics of his interest eg. Solar, nature in the Figure. 7. The above document will be summarized in such a way that it will contain only the information that is of interest to the user in the summary. This helps in extracting a particular subject or relevant information from the large source text document. It also displays the sentence which has all the keywords as the most important sentence.

Let us compare the summary generated by Scissor to any other available tools.

The summary generated by Copernic:

10% summary

A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun.

25% summary

A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun. A total solar eclipse is a spectacular natural phenomenon and many people consider travel to remote locations in order to observe one.

50% summary

A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun. Total solar eclipses can be frightening events for people unaware of their astronomical nature, as the Sun suddenly disappears in the middle of the day and the sky darkens in a matter of minutes. A total solar eclipse is a spectacular natural phenomenon and many people consider travel to remote locations in order to observe one. The 1999 total eclipse in Europe, said by some to be the most-watched eclipse in human history, helped to increase public awareness of the phenomenon.

The observations clearly indicate that the summaries generated by our method are closer to human generated summary than the summaries produced by Copernic.

7. CONCLUSION

In this summarization tool, we have mainly focused on the statistical similarity between sentences. It is a summarization tool which involves mapping of the words and sentences onto a semantic space and exploiting their similarities to remove the less important sentences containing redundant information. Using Graph theoretic approach we ensure that all the topics are covered in the summary. The summary includes atleast some important information about all the topics present in the input document. The results obtained showed that our approach can be suitable for selecting important sentences of a document, and therefore can be a good idea to take this feature into account when building a summarization system.

8. REFERENCES

[1] Gunes Erkan and Dragomir R.Radev. 2011. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.

[2] Dipanjay Das. 2007. A Survey on Automatic Text Summarization

[3] Elena Lloret. Text Summarization:An Overview.

[4] Rafeeq Al-Hashemi. 2010. Text Summarization Extraction System (TSES) Using Extracted Keywords.

[5] Dalianis. 2010. SweSum - A Text Summarizer for Swedish Technical report

[6] Martin Hassel. 2004. Evaluation of Automatic Text Summarization Licentiate Thesis Stockholm

[7] Ani Nenkova, Lucy Vanderwende and Kathleen McKeown. 2006. A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization.