Natural Language Processing – Project Report – Fall 2015
Automatic Summarization

Sri Ram Kannan – sxk138130

ABSTRACT

Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text. The project is about implementing two different kinds of summarization. One is about building a model called fractal summarization based on the fractal theory. Other method is traditional summarization based on word frequency.

INTRODUCTION

As the Internet is growing exponentially, huge amount of information are available online. It is difficult to identify the relevant information. The information-overloading problem can be reduced by automatic summarization.

There are basically two methods to do the summarization. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

The project is about algorithms related to extractive methods and comparing their efficiency, especially the importance of the fractal summarization model and how it outperforms the traditional summarization.

PRIOR RESEARCH

Most of the traditional summarization is not based on document structure and they do not take into account of the fact that the human abstractors extract sentences according to the hierarchical document structure. Document structure can be described as fractals that are some mathematical objects with high degree of redundancy. Fractal summarization can be related to the document structure by describing each structure of the documents as fractals that are some mathematical object with high degree of redundancy. Fractal summarization controls the amount of information displayed. The fractal summarization highly improves the divergence of information coverage of summary and it is robust and transparent, the user can easily control the compression ratio, and the system generates a summary that maximize the information coverage and minimize the distance of summary from the source document.

SYSTEM DESIGN

Two models have been implemented here. One is traditional summarization approach while the other is fractal summarization based on fractal theory.

A. Traditional summarization:

Traditional automatic text summarization is selection of sentences from source document based on the salient features of document, such as thematic, location, title, and cue features. Totally 3 different algorithms are implemented and they are listed below.

Word Frequency algorithm: A basic algorithm that selects sentences based on the sum of their words weights relative to the document.

Sin Transform Frequency algorithm: An improvised version of word frequency algorithm which gives more weight to the beginning/end of the text and less weight to the center. This algorithm works best for scientific papers and articles based around an introduction and conclusion.

Sin Transform Word Frequency: An update of the Sin Transform Frequency algorithm, that gives more weight to the words located at the beginning and end of the text, and less weight to the words in the middle. This weight is compounded every time when the word is found, without the value based on its location. As a result, the algorithm is able to better grasp the main subject of the text, without requiring any machine learning.

A detailed step of implementation is given below.

- The entire text is tokenized and the weight of the text is calculated using term-frequency method.
- Sentence Boundary is drawn from the given passage of text using NLTK library.
- Each sentence is marked as a vertex of the graph and the stop words and punctuations are removed
- Weight Calculation:

  Word Frequency: Relative weight of each of the sentence is calculated using tf-idf.

  Sin Transform Frequency: Relative weight of each of the sentence is calculated using tf-idf and multiplied with a sin similarity formula:
  (1-sin(sentence index*(PI/sentence total)))/2+1

Sin Transform Word Frequency: Relative weight of each of the words is calculated using the formula: (1-sin(sentence index*(PI/sentence total)))/2+1 . This overall text document weight is updated with the sum of the word weight.

- From the weights, the importance of each sentence is calculated and stored.
- The top most important sentences are combined in the same order as in the original passage to form the summary.

B. Fractal Summarization:

The Fractal Summarization is developed based on the models of fractal view and fractal tree. The source document is partitioned into range-blocks according to document structure and represented as a fractal tree. The degree of importance of each node in the fractal tree is represented by its fractal value. The fractal value of focus is set to 1. The fractal value of each node is calculated as the sum of sentence weights under the range-block. User may choose a compression ratio to specify the ratio of sentences to be extracted as the summary. The sentence quota of the summary can be calculated accordingly and it will be propagated to the child-nodes directly proportional to their fractal values.

Fractal Summarization Algorithm steps:

1. Choose a Compression Ratio and Threshold Value.
2. Get the total Sentence Quota of the summary.
3. Partition the document into range blocks.
4. Transform the document into fractal tree.
5. Set the current node to the root of the fractal tree.
6. Repeat
   6.1 For each child node under current node,
   Calculate the fractal value of child node.
   6.2 Allocate Quota to child nodes in proportion to fractal values
   6.3 For each child nodes,
      If the quota is less than threshold value
         Select the sentences in the range block by extraction
      Else
         Set the current node to the child node
      Repeat Step 6.1, 6.2, 6.3
7. until all the child nodes under current node are processed.

Major differences between traditional and fractal summarization:

Traditional	Fractal
It adopts the traditional salient features, but they consider the document as a sequence of sentences.	It adopts the traditional salient features and hierarchical fractal structure.
In tf-idf weight calculation the document structure is not taken into account and weight of the term remains same over the entire document.	The tfidf of a term in a range block is defined as proportional to the term frequency within a range-block and inversely proportional to the frequency of range-block containing the term.
It assumes that the location weight of a sentence is static.	It calculates the location weight based on which document-level we are looking at.

EXPERIMENTAL RESULTS

Experimental result shows that the fractal summarization has more precision than the traditional summarization. It is inferred that fractal summarization model extracts the contents distributed in all chapters/paragraph whereas traditional summarization model extracts most sentences mainly from few chapters/paragraph.

INPUT/OUTPUT:

- The input to both the models is a passage of text and the output will be the topic summary.
- The sample input/output for both the models is given below.

INPUT:

The brain of one monkey has been used to control the movements of another, "avatar", monkey, US scientists report. Brain scans read the master monkey's mind and were used to electrically stimulate the avatar's spinal cord, resulting in controlled movement. The team hopes the method can be refined to allow paralyzed people to regain control of their own body.

The findings, published in Nature Communications, have been described as "a key step forward". Damage to the spinal cord can stop the flow of information from the brain to the body, leaving people unable to walk or feed themselves. The researchers are aiming to bridge the damage with machinery.

Match electrical activity. The scientists at Harvard Medical School said they could not justify paralyzing a monkey. Instead, two were used - a master monkey and a sedated avatar. The master had a brain chip implanted that could monitor the activity of up to 100 neurons.

During training, the physical actions of the monkey were matched up with the patterns of electrical activity in the neurons.The avatar had 36 electrodes implanted in the spinal cord and tests were performed to see how stimulating different combinations of electrodes affected movement.

The two monkeys were then hooked up so that the brain scans in one controlled movements in real time in the other.

The sedated avatar held a joystick, while the master had to think about moving a cursor up or down. In 98% of tests, the master could correctly control the avatar's arm.

One of the researchers, Dr Ziv Williams, told the BBC: "The goal is to take people with brain stem or spinal cord paralysis and bypass the injury." The hope is ultimately to get completely natural movement, I think it's theoretically possible, but it will require an exponential additional effort to get to that point."He said that giving paralysed people even a small

amount of movement could dramatically alter their quality of life.

OUTPUT:

Fractal Summarization:

Brain scans read the master monkey's mind and were used to electrically stimulate the avatar's spinal cord, resulting in controlled movement.

Damage to the spinal cord can stop the flow of information from the brain to the body, leaving people unable to walk or feed themselves.

The scientists at Harvard Medical School said they could not justify paralyzing a monkey. Instead, two were used - a master monkey and a sedated avatar.

The avatar had 36 electrodes implanted in the spinal cord and tests were performed to see how stimulating different combinations of electrodes affected movement.

In 98% of tests, the master could correctly control the avatar's arm.

One of the researchers, Dr Ziv Williams, told the BBC: "The goal is to take people with brain stem or spinal cord paralysis and bypass the injury.

Traditional Summarization (Sin Transform Frequency algorithm):

The brain of one monkey has been used to control the movements of another, "avatar", monkey, US scientists report.

Brain scans read the master monkey's mind and were used to electrically stimulate the avatar's spinal cord, resulting in controlled movement.

Damage to the spinal cord can stop the flow of information from the brain to the body, leaving people unable to walk or feed themselves.

The scientists at Harvard Medical School said they could not justify paralyzing a monkey. Instead, two were used - a master monkey and a sedated avatar.

One of the researchers, Dr Ziv Williams, told the BBC: "The goal is to take people with brain stem or spinal cord paralysis and bypass the injury.

He said that giving paralyzed people even a small amount of movement could dramatically alter their quality of life.

FUTURE IMPROVEMENTS:

- One possible improvement would be the implementation of thematic, location and heading feature all together in the fractal summarization.
- Generating an interactive summary to reduce the computational load in handheld devices and displaying a skeleton of summary at the first stage and the details of the summary on different levels.

CONCLUSION:

Thus 3 different algorithms of traditional summarization and 1 algorithm on fractal summarization has been implemented in this project. Considering the scope of improvement and further enhancements, this implementation will definitely be helpful.

REFERENCES:

[1] Daniel Hurafsky, James H.Martin , “Speech and Language Processing” Second edition

[2] Edmundson H. P., 1968. : *New Method in Automatic Extraction. Journal of the ACM, 16(2) 264-285, 1968.*

[3] Feder J.: *Fractals.*, Plenum, New York, 1988.

[3]]Koike, H.: *Fractal Views: A Fractal-Based Method for Controlling Information Display,*, ACM Transactions on Information Systems, ACM, 13(3) 305-323, 1995.