Text Summarization using PSO

1Anmol Narang, 2Dr. Neelam R. Prakash, 3Amit Arora
1, 2Department of Electronics and Communication Engineering, PEC University Of Technology, Chandigarh, India
3Develop Tech

DOI: 10.23956/ijarcsse/V7I6/0242

Abstract: In this paper we have proposed Text Summarization using genetic algorithm. According to the required summary size by the user, we have evaluated all possible subsets of that size using genetic algorithm. We have made an attempt to create a summary that is less redundant and covers all the topics in the document. The summary generated was evaluated by comparing it with the user summary and the corresponding rouge score was calculated. In the end, the influence of varying summary size is analyzed and is compared with the human summary.

Keywords: Text Summarization, Genetic Algorithm, Rouge Score, Less Redundant

I. INTRODUCTION

In this modern world of internet there is tremendous amount of data available on the web. But most of it is repetitive or is not relevant for us. In this fast world of internet no one has some much of time to go through the entire data and extract the relevant and important information from it. If we could develop a system that extracts the important information from the text and summarize it into a format that is readable and understandable to the user could reduce the effort of mankind to a great extent. Text Summarization are of two kinds:(i) Abstractive and (ii) Extractive

In abstractive text summarization the system learns about the new data depending upon the previous data and then summarizes it. It extracts the key words in the input document. Once the signature words are extracted, system fetches the relevant sentences relevant to the signature words from the database. On the other hand, in extractive summarization the system extracts the key words and sentences and includes them in the summary as it is, no modification in the sentence is done. In extractive summarization no previous data or information about the text is required. It uses the statistical and linguistic information about the text and the most relevant sentences are included in the summary. [1] In this paper we tried to implement extractive summarization using Particle Swarm Optimization. With the help of Particle Swarm Optimization(PSO) summaries are graded and the top most summary is the actual summary.

PSO, inspired by bird flock or fish school, is precise, easier to implement and involves few mathematical operations. PSO is an optimization technique which helps to predict that out of the given set of solutions which is the best solution. After every iteration it updates the particles position as well as velocity. Based upon the updated position and velocity particles personal best solution and correspondingly the global best solution is updated.

II. LITERATURE REVIEW

The main entity of Text summarization are features. And equally treating every feature is the cause of generating poor summary. In [2], the effect of feature structure on selection of features with the utilization of particle swarm optimization are investigated. DUC 2002 data is used for training particle swarm optimization for learning each feature weight. Different features are used in the context of structure, where some features combination having greater priority in comparison to individual or single features. Therefore, determining the effectiveness of each feature can lead to a mechanism for differentiating among features having higher and lower importance. The combined features have higher priority of being selected than the simple features. Some features are selected in each iteration by utilizing particle swarm optimization, then those features corresponding weights are utilized for sentence scoring and summary selects the top ranked sentences. Each best summary have selected features and are utilized in evaluation of final features weights.

In this paper we have proposed Text Summarization using genetic algorithm. According to the required summary size by the user, we have evaluated all possible subsets of that size using genetic algorithm. We have made an attempt to create a summary that is less redundant and covers all the topics in the document. The summary generated was evaluated by comparing it with the user summary and the corresponding rouge score was calculated. In the end, the influence of varying summary size is analyzed and is compared with the human summary.

III. EXPERIMENTAL RESULTS

The Pennsylvania State University trained PSO summaries data corpus for determining the best particle which represents most appropriate structural feature utilizing regular Arab summarizers.

© www.ijarcsse.com, All Rights Reserved
Attentional Encode-Decoder using RNN is an abstractive text summarization method. It models key words, sentence-to-word hierarchy model and emits rare words at the time of training. It proposed a dataset that uses multi-sentence summarizes but has certain performance benchmarks [7]. RNN encoder-decoder breaks text in parts and treats it as a sequence-by-sequence problem. It is quite effective Machine Translation and is way better than the Rush et al's state-of-the-art-model (2015) on the Gigaword dataset without the need of any ad-on tuning [8].

### III. SYSTEM MODEL

![System Model Diagram]

#### 3.1 Preprocessing

It is the first and foremost step of any research in the field of text and image. If our preprocessing is not correct, there is no use of going any further. Incorrect preprocessing will automatically yield to incorrect results. In text preprocessing, we convert text to a numerical data so that we can analyze and observe patterns in the data in a better way. Converting text to a numerical data provides us the advantage of performing mathematical operations like addition, subtraction and many more on our data. Preprocessing further involves various steps like:

#### 3.2 Stop Word Removal

Certain words like 'is', 'am', 'are', 'was', 'were' and many more which doesn't carry any meaning are widely used in text so as to provide a sense of connectivity to the reader but these words are of no use for the machine. They only increase the weight of the text and reduces its importance for analysis. Stop words mainly includes articles, preposition and conjunctions. So, in stop word removal process we remove them and further works with this preprocessed text. Removing the stop words will also reduce the dimensionality of our text.

#### 3.3 Stemming

Words like connected, connecting, connects etc are all derived from the same root word 'connect'. So anytime these words appear they should be given equal weight age as they all convey the same meaning. So, in stemming we replace all the above words by a single word connect, i.e. we reduce the word to it is root word.
3.4 Term Frequency Inverse Document Frequency (TF-IDF) –
TF-IDF is a combination of two words TF and IDF. TF tells us how relevant a word is to a sentence while IDF tells us the relevance of a word with respect to the entire document. TF is given by the formula:
\[ TF(t,S) = 0.5 + 0.5 \cdot \frac{f_{t,s}}{\text{max}(f_{t,s} : t \in S)} \]
where: \( f_{t,s} \) frequency of word t in sentence 's'.
\[ \text{max}(f_{t,s} : t \in S) \] maximum of all the frequencies in the sentence 's'.
IDF is given by:
\[ \text{IDF}(t,d) = \log\left(\frac{N}{f_{t,d}}\right) \]
where N \( \leftarrow \) total number of sentences
\( f_{t,d} \leftarrow \) number of sentences in which the word 't' is present.

3.5 Sentence Score and Subset Score –
Once we have calculated the Tf-Idf of each word in the document, we add the Tf-Idf of each word in the sentence to get the sentence score and in the end the sentences in the subset are combined to get the respective subset score which will further act as our particle's initial position.

3.6 Particle Swarm Optimization for Subset Selection –
Once we have formed all possible subsets of the required summary size, these subsets will act as particles and the subset score will be our particle's initial position. We assume the particle to be initially at rest and until the end condition is not met we update the particle's velocity using the formula:
\[ \text{svel}(l+1) \leftarrow \text{svel}(l) + s_p \cdot (\text{spbst} - \text{spos}(l)) + s_g \cdot (\text{gbst} - \text{spos}(l)) \]
where: \( \text{svel}(l) \leftarrow \) particle current velocity
\( s_p, s_g \leftarrow \) random numbers between 0 and 1
\( \text{spos}(l) \leftarrow \) particle current position
\( \text{spbst} \leftarrow \) particle's personal best position
\( \text{gbst} \leftarrow \) global best position.
Once we have updated the particle velocity we will update it's position using the formula:
\[ \text{spos}(l+1) \leftarrow \text{spos}(l) + \text{svel}(l+1) \]
where: \( \text{svel}(l+1) \leftarrow \) updated velocity of particle
After the end of every iteration we compare the particle's current position with its personal best position. If it's better we updated the particle's pbest. Similarly, we compare the pbest's of all particles with the gbest. If any particle's pbest is better than the gbest update the gbest. This process goes on until the termination condition is not met or we have reached the maximum number of iterations.
The subset with pbest equal to gbest is our required summary.

3.7 Evaluation Function –
Once the system has summarized the document, it is very important to evaluate it. If we don't know how much effective our summary is, we would never be able to improve it. So, to find the effectiveness of our summary we use the ROUGE method. In this, we compare the system summary with the user summary and calculate how much they differ. We find all the possible n-grams of the system summary and the summary made by the user. The word n-grams means that we combine n continuous words and represent them as a single word. Here, we have only used monogram, bigram and trigram. We then calculate the number of common n-grams between the system summary and the user summary.
Rouge Score = Number of common n-grams / total number of n-grams in the user summary.

**ALGORITHM**

input: subsets with their respective score
output: subset with best optimized score.
for each subset or particle \( j = 1, ..., S \) 
do
  subset initial position = subset score
  subset's best known position = initial position: \( \text{spbst} \leftarrow \text{spos}_j \)
if \( Q(\text{spbst}) < Q(\text{gbst}) \) then
  update subset's best known position: \( \text{gbst} \leftarrow \text{spbst}_j \)
Let the subset initial velocity be zero
while end condition not met do:
  for each subset \( j = 1, ..., S \) 
do
    Pick random numbers:
    \( s_p, s_g \leftarrow G(0,1) \)
    Update subset's velocity: \( \text{svel}(l+1) \leftarrow \text{svel}(l) + s_p \cdot (\text{spbst}_j - \text{spos}(l)) + s_g \cdot (\text{gbst}-\text{spos}(l)) \)
    Update subset's position: \( \text{pos}_j(l+1) \leftarrow \text{pos}_j + \text{vel}_j(l+1) \)
if $Q(s_{pos}(l+1)) < Q(s_{pbst})$ then 
    Update subset’s best known position: $s_{pbst} \leftarrow s_{pos}(l+1)$
if $Q(s_{pbst}) < Q(gbst)$ then 
    Update the subset’s best known position: $gbst \leftarrow s_{pbst}$
end for
end while
if $(s_{pbst} = gbst)$ then 
    subset is the required summary

IV. RESULTS
Once we have developed a system, it’s very important to test it and compare results of various testing cases. Fig 1 shows the rouge score for various summary sizes. Rouge score as explained above compares the system summary with the user summary. From the graph we can conclude that summary size of 5% is short and precise. As we increase the summary size lot of unwanted data gets included in the summary and if our summary size is too long some of the important information might not be included in the summary.

![Fig 1. Comparison of Rouge Score for various summary sizes](image)

V. CONCLUSION
In the end we can conclude that proper subset formation and their evaluation is an important criteria for text summarization. For subsets evaluation we have used Particle Swarm Optimization. We extracted best possible summary of various sizes. From the graph we can conclude that summary size around 5% of the original document is the best summary and the subsets are best optimized. Future scope will be to understand the meaning of each sentence and generate summary in own words.

REFERENCES