International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 10 | Oct 2019 www.irjet.net p-ISSN: 2395-0072

Semantic based Automatic Text Summarization based on Soft Computing

Janit Chadha[1]

[1]*Student, Dept. of Computer Science Engineering, BNMIT, Karnataka, India*

Abstract - *Automated Summarizer is a tool which extracts lines from a text file and generates a brief information in a proper manner. Even though many approaches have been developed, some important aspects of summaries, such as grammar, responsiveness are still evaluated manually by experts. In the Semantic based Automatic Text Summarization using soft computing, initial the text pre-processing is completed that's the removal of stop words, stemming, lemmatization. The title is chosen for the document mechanically victimization resource description framework. Repetition references are resolved, and text bunch is performed word meaning clarification is completed using NLP-parser, the linguistics similarity, title and its characteristics are known. N-gram Co-occurrences relations are found. Finally, the tag-based coaching is completed, and the final outline is produced.*

Key Words: Text summarization, Text mining, Resource description framework (RDF), Natural Language Processing (NLP), Soft Computing.

1. INTRODUCTION

In today's world voluminous data is getting generated every year and is still growing exponentially. Data is the most precious thing for an organization and every year they spend a huge amount in keeping as it provides a competitive edge. As the new technology advancement and innovation, data is what oil was used to be. Manual data processing is very costly and time consuming. Data processing should be an automated process that is a cost effective and time efficient process figure 1.

Figure 1: Data flow diagram of SATSSC

Dataset: The documents (DUC 2007) for summarization are taken from the AQUAINT corpus, comprising newswire articles from the Associated Press and New York Times (1998-2000) and Xinhua News Agency (1996-2000).

2. RELATED WORK

Syntactic parsing manages grammar pattern in a line. The target of grammar investigation is mainly to relate grammar patterns that is often portrayed as a tree. Recognizing the grammar pattern gives the importance of a sentence. Traditional language making could be a field of software system engineering moreover, phonetics, disquieted regarding the dealings among PCs and people dialects. It forms the knowledge through lexical investigation, Syntax examination, linguistics investigation, speak making ready, Pragmatic investigation. The calculation elements country sentences into elements utilizing POS tagger, and acknowledges the kind of sentence (Facts, dynamic, latent then forth.) and at that time parses these sentences utilizing language principles of linguistic communication [1].

Printed definition of multiple reviews may be practiced by utilizing theoretical ways that specifically specific, for each viewpoint, the rating dissemination over the total review set and, moreover, choose content or disengage scraps from the reviews to point out this opinion distribution. In any case, keen on investigation however way will get in utilizing extractive techniques to accumulate substantiating sentences that mirror the standard read over the survey set. Moreover, extractive ways square measure less complex, have incontestable terribly effective in several territories of automatic report, and need less manual area adjustment than theoretical ways.

*With this objective in mind, separate the general methodology into 3 noteworthy advances: getting ready rating expectation along with n-gram language models; utilizing these models to disengage highlights from every information sentence; and utilizing A*search to find a perfect set of sentences from the information records to form summary. A* obtain may be a methodology to effectively investigate a considerable area of alternatives (for our state of affairs, the challenger sentences for the target rundown) and choose to ideal resolution supported the least-cost method (the best mix of sentences for the target synopsis)[2].*

Different sorts of information that's accessible on an issue electronically has munificently distended over the previous years. It's driven the information road to a circumstance known as "data over-burden" issue. Programmed content summation system in the main addresses this issue by the extraction of an abbreviated rendition of information from writings expounded on the same theme. A couple of mathematical decrease techniques area unit used to tell apart and separate the semantically important messages in an exceedingly report back to define it consequently.

Uncommon center is given to the foremost generally used mathematical ways known as Singular price Decomposition (SVD) and Non-negative Matrix factorization (NMF)[3].

Looking for bits of valuable information from a knowledge on the net remains a hard and tedious endeavor for a large scope of people for instance, understudies, journalists, and various totally different sorts of specialists. The issue needs to analysis higher approaches to alter and method information, that has to be sent during a somewhat very little house, recovered during a temporary span, and spoke to as exactly as would be prudent. This can be positively a standout amongst the foremost important reasons for seeking cheap and effective summation ways suited "refining" the foremost useful things of assortment from an coherently connected origin, because it came back from exemplary net crawlers, therefore on deliver a brief, compact and lingually necessary adaptation of information unfolded in pages and pages of writings. A summarizer framework, called as iWIN (data in net during a Nutshell), which will play out a programmed defined of various records through: a linguistics examination of the content, a positioning strategy wont to assess the importance of the info for the actual consumer, a grouping strategy keen about the archive portrayal as way as set of triplets (subject, action word, object)[4].

Different sorts of information that's accessible on an issue electronically has munificently distended over the previous years. It's driven the information road to a circumstance known as "data over-burden" issue. Programmed content summation system in the main addresses this issue by the extraction of an abbreviated rendition of information from writings expounded on the same theme. A couple of mathematical decrease techniques area unit used to tell apart and separate the semantically important messages in an exceedingly report back to define it consequently. Uncommon center is given to the foremost generally used mathematical ways known as Singular price Decomposition (SVD) and Non-negative Matrix factorization (NMF)[5].

Text summarization may be a method of extracting or accumulating essential facts from the authentic matter content and presents that statistics within the form of outline. Text summarization has return to be the requirement for several applications as an example program, business analysis, market value. Summarization helps to achieve the specified knowledge in less time. The approach deployed for summarization degrees from dependent to linguistic. In Indian several languages conjointly the paintings are applied, however presently, they're within the infancy degree. Text summarization methods could also be extensively divided into 2 groups: extractive summarization and theoretical summarization. Extractive summarizations extract very important sentences or terms from the distinctive files and organize them to supply an explicit while not ever-changing the distinctive text. An extractive text summarization machine is planned supported pos tagging through wondering hidden Andrei Markov model the usage of the corpus to extract crucial terms to create as an explicit.

Theoretical summarization includes experience the supply matter content by suggests that of employing a linguistic approach to interpret and examine the text. Theoretical strategies would like a deeper analysis of the matter content. Those strategies have the potential to come up with new sentences, that improves the main target of a outline, scale back its redundancy and keeps a awfully smart compression fee . [6]

Records on internet square measure growing every minute. Redundancy in information is growing fleetly. data processing is that the approach accustomed extract these records as keep with the person's question. Technically info mining analyzing and summarizing it into helpful information. Keyword obtain could be a crucial tool for exploring and looking out huge statistics corpora whose structure is each unknown, or ceaselessly dynamical. So, keyword obtain has already been studied within the context of relative databases XML documents and a lot of currently over graphs and RDF info. Linguistics internet mining aims to mix linguistics internet and net mining. Linguistics net mining is that they would like of those days' redundant records. On this paper, the foremost necessary consciousness is on minimizing extraction of a variety of pages through the ranking methodology. Thanks to that the extraction of knowledge is performed real as question pink- slipped and therefore the pinnacle graded pages square measure shown to the buyer. Here for these three necessary regions square measure reaching to apply that embody linguistics internet, metaphysics and RDF facts. The difficulty of ascendible keyword obtain on huge RDF records and projected a brand new summary-primarily based mostly answer.

analysis offers a terse outline at the kind level from RDF info within the course of question analysis, this leverage the precis to prune away an outsized a part of RDF information from the hunt space, and formulate SPARQL queries for with efficiency having to access to facts. Moreover, the projected precis is also incrementally up to now because the records get updated. Experiments on each RDF benchmark and real RDF datasets confirmed that the solution is inexperienced, scalable, and moveable across RDF engines. [7]

3. METHODOLOGY

In the figure 2 proposed model for semantic based automated text summarizer is shown. The xml/text file is taken as the input, text preprocessing is performed. The input for the anaphoric resolution is the preprocessed text and produces a filtered text output. The word disambiguation takes the pronomial input and gives the filtered output. Then, the resource description framework takes the preprocessed text and provides RDF triples. N-gram co-occurrence measure is done. At last, the sentence

combination tis done and the output is the brief information generated by the summarizer.

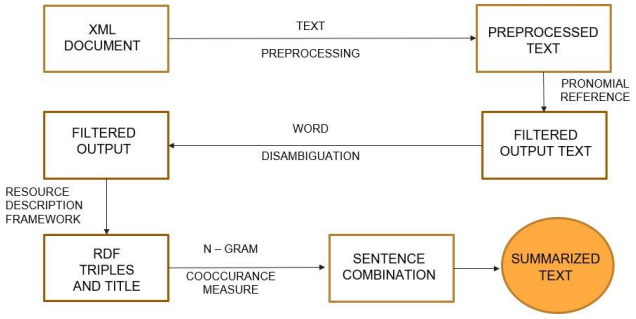


Figure 2: *Proposed model for semantic based automated text summarizer*

Automated text summarization with soft computing 1. For Text/XML file it analyses the relevant topic or the heading.

2. *Anaphoric references are cleared-up for growth of the results.*

3. *Parser find out the syntactical errors in each line and removes tag-based ambiguity from each line.*

4. *The measure of line devaluation is performed using semantic similarity of line score, n-gram co-occurrence score of lines in the file.*

5. *Finally, brief information is achieved according to prescribed percentage.*

Text Pre-processing Algorithm

Algorithm 1: Text Pre-processing

Input: Document (.scu, .txt)

Output: Structured text document

Description: Text pre-processing filtering the text before summarization

```
If xml file
        Remove tags using Beautiful Soup tool
        Remove only the tags make it a structured text

    If txt file
        Take the text as it is
    For all sentences in the text
            Remove special characters
            Remove brackets and delimiters
        Stop word removal
        Tokenize and lemmatize words
```

Algorithm-1, it filters the text for further summarization using the data-preprocessing techniques.

Title Identification Algorithm

Algorithm 2: Title Identification

Input: Pre-processed text

Output: Title for the summarized document

Description: Title identification detects a title for a text document using fundamentals of language specific grammar structure i.e. RDF

```
For all sentences in the processed Text
        Identify the RDF of each sentences
If n gram co-occurrence values are less choose those words
        Calculate semantic distance using those words
        Find the semantic distance with each RDF
Identified Title according the semantic distance
```

Algorithm-2, it picks a complete line within the existing file. After this step, it parses these selected lines into RDF. Computer program is used to recover matched documents for the RDFs. Last, it accepts the title for the present file.

Anaphoric reference Algorithm

Algorithm 3: Anaphoric Reference resolution (Pronominal)

Input: Filtered text

Output: Text document without "they", "there", "those"," these", "it", "this", "that" replaced with the subject it refers to.

Description: Anaphoric reference means that a word in a text refers back to it's subject in the text for its meaning.

```
for all noun_phrase do
  best_antecedent = nil
  for all antecedent < noun_phrase do
    if antecedent is more likely coreferent than best_antecedent then
      best_antecedent = antecedent
    end if
  end for
  if we have confidence in best_antecedent
    make coreferent (noun_phrase, antecedent)
  end if
end for
```

Algorithm-3, In order to create meaningful illustration of a text document, it ought to have the connected lines. Reference may be a means that to link a referring expression to a different referring expression within the close text, as shown within the following Example:

Sachin and Rahul plays cricket and tennis. They also play football.

Here, 'They' refers to associate degree entity Sachin.

Algorithm-4, Word disambiguation using NLP- computer program disambiguates incorrect tags given by the computer program. It corrects them and gives the correct tags as needed.

Word Disambiguation Using NLP – Parser Algorithm

Algorithm 4: Word Sense Disambiguation using NLP-Parser

Input: Pronomial reference filtered text

Output: Text without ambiguity based on Noun, Verb, Adjective etc.

Description: Word sense disambiguates incorrect tags given by the parser.

```
best_sense <– most frequent sense for word
max_overlap <– 0
context <– set of words
Signature <– set of words in the gloss and example of senses
for each sense in senses of words do
    overlap = COMPUTEOVERLAP (signature, context)
    if overlap > max_overlap
        max_overlap <- overlap
        best_sense <- sense
end return (best_sense)
```

4. RESULTS

Automatic text summarizerusingsoft computingapproaches provides the result in a very time efficient manner and is cost effective.

Interface

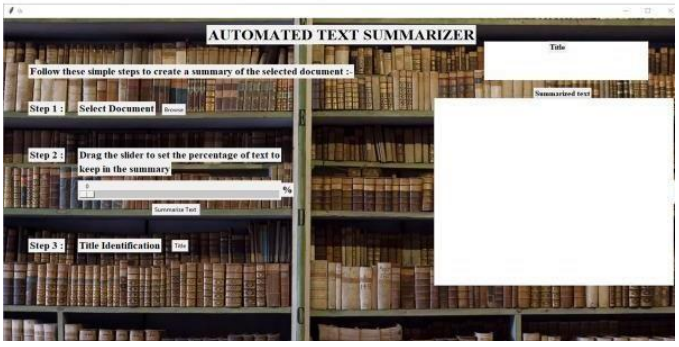


Figure 4: Interface for automated text summarizer

Select document for summarization

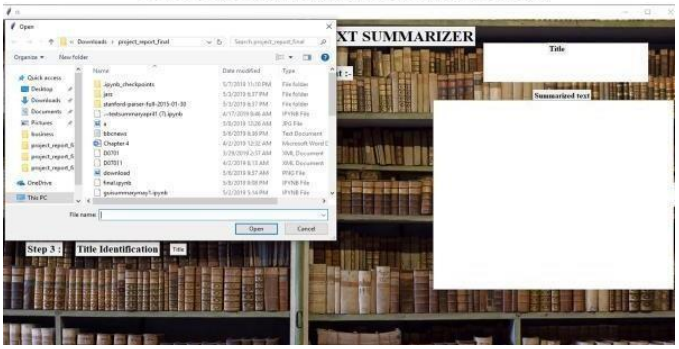


Figure 5: Document selection

Confirmation after selecting a valid document

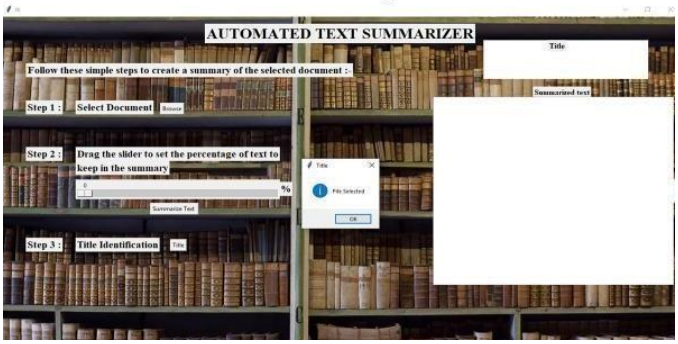


Figure 5: Confirmation message

Message displayed while processing

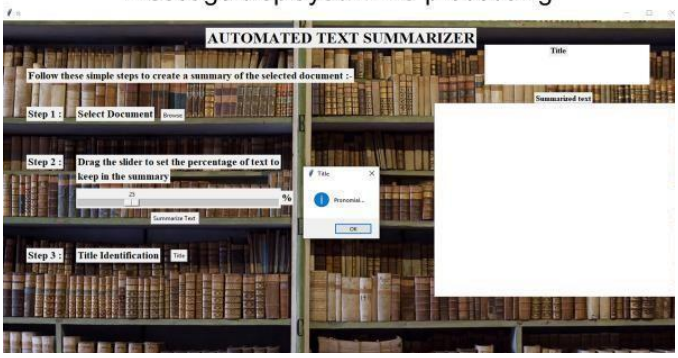


Figure 6: Processing

Summary of selected document

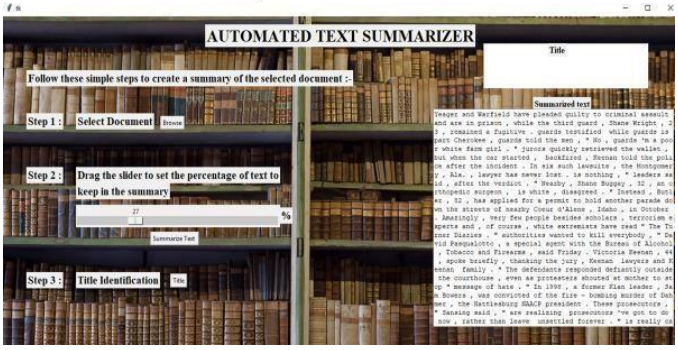


Figure 7: Summary of selected document

Title for summarized document

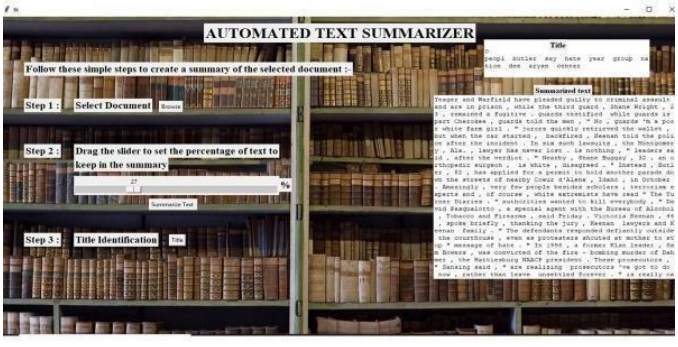


Figure 8: Title for selected document

5. COMPARISION GRAPH XML v/s TEXT

Figure 9 shows the comparison graph for xml v/s text file. The results for text file are much efficient than xml file.

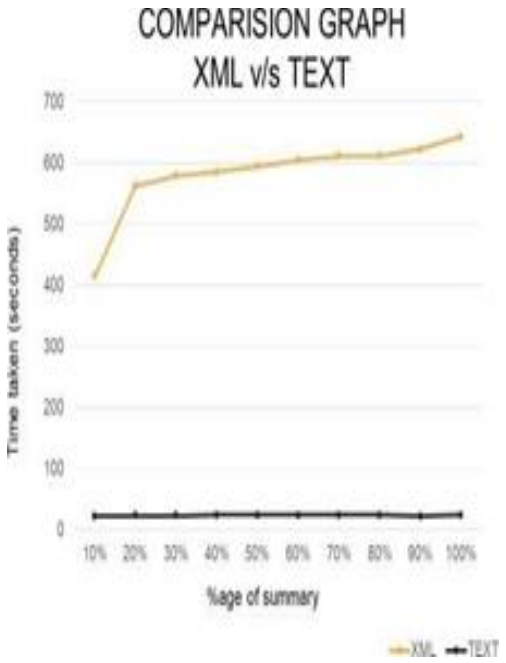
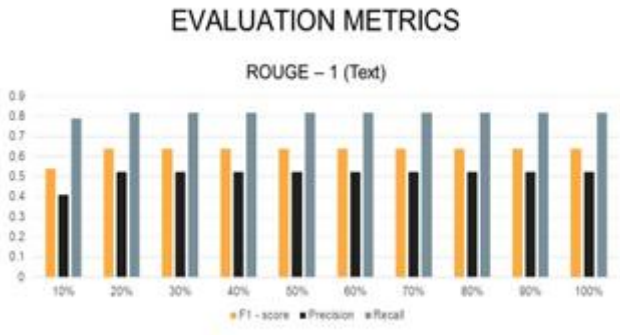
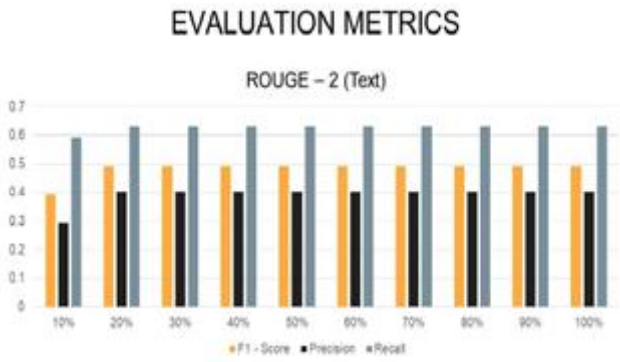
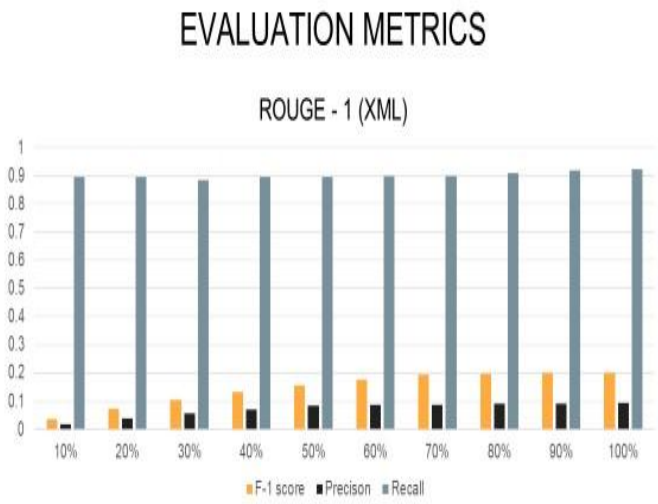
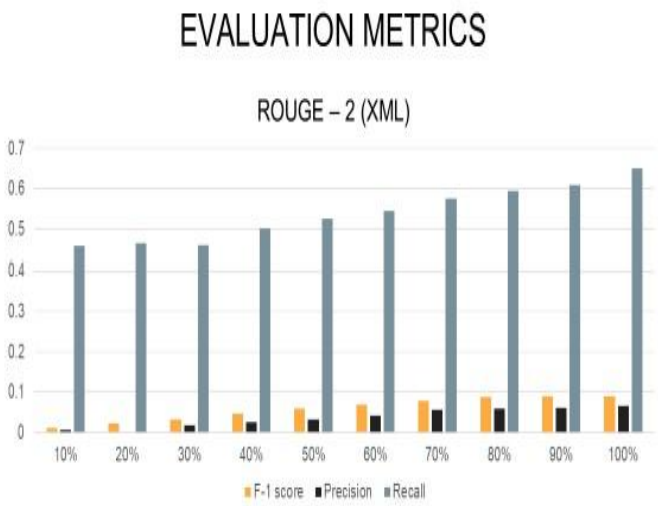


Figure 9: COMPARISION GRAPH XML v/s TEXT









EVALUATION METRICS (ROUGE 1 and 2)

7. CONCLUSION

Automated summary generates and gives the summary as per the required percentage. In future, some more types of references can be resolved for improvement of the performance of the SBATSSC method. The performance can be further improved through more adept methods for reducing and combining the sentences. The automatic text summarizer can be further modified for generating summaries of PDF documents.

ACKNOWLEDGEMENT

I want to thank god and my parents for educating me.

REFERENCES

1) Madhuri A. Tayal, Dr. M. M. Raghuwanshi, Dr. Latesh Malik." Syntax Parsing: Implementation using Grammar-Rules for English Language". In *IEEE. International Conference on Electronic Systems, Signal Processing and Computing Technologies*, IEEE (2014), pp. 376–381.

2) Di Fabbrizio, G., Aker, A., Gaizauskas, R. "Summarizing online reviews using aspect rating distributions and language modeling". *IEEE* (2013) *Intell.Syst*. 28–37. R. Nicole

3) Azmi, A.M., Al-Thanyyan, S., "A text summarizer for Arabic". *Comput. Speech Language* (2012) 260–273.

4) dAcierno, A., Moscato, V., Persia, F., Picariello, A., Pento, A., "iWIN: A Summarizer System Based on a Semantic Analysis of Web Documents" *IEEE Sixth International Conference on Semantic Computing. (2012.)*

5) Eduard Hovy and Chin-Yew Lin "Automated text summarization and the summarist system" *Springer International Publishing AG 2018*.

6) Deepali K. Gaikwad and C. Namrata Mahender "A review paper on text summarization" *International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.*

7) Roshna Chettri, Udit Kr. Chakraborty "Automatic Text Summarization" *International Journal of Computer Applications (0975 – 8887) Volume 161 – No 1, March 2017.*

BIOGRAPHIES

M.Tech fresher enthusiastic about data science.