International Journal For Technological Research In Engineering
Volume 6, Issue 12, August-2019 *ISSN (Online): 2347 - 4718*

DISTANCE BASED IMAGE TEXT EXTRACTION AND SUMMARIZATION APPROACH

Vikas Sharma[1], Rahul Sharma[2]
[1]M. Tech Scholor, [2]Assistant Professor,
Sri Balaji College of Engineering and Technology Jaipur Rajasthan

Abstract: Text Summarization means making the layout of the account. The record can range to the report containing few lines or even the document contains different measures of pages. Considering any field whether identified with the Engineering, Medical or some other stream, the quick overview is normally essential. In like manner, the little charts are snatching centrality as the length of the reports is expanding likewise as the nonappearance of time. This need of the outline drives us to work in the field of the robotized content rundown. Regardless of the manner in which that the different portrayals or sorts of work is generally officially done is this field at any rate the basic still the equivalent , of giving signs of advancement review, that is the exactness of working moreover able like the physically made once-finished. The reports on the filtered picture are extricated and afterward will be tried on the base work and the calculation which is proposed. The reproduction of the base and the proposed work is made in the Java dialects and the IDE which we have utilized for this design is overshadow.
Keywords : Text Summarization, Image Text Extraction.

I. INTRODUCTION

Customized text summarization frameworks can be assembled into two or three specific creates . The unquestionable estimations of text summarization can be commonly requested in context on its information make (single or multi record), reason (non unequivocal, space explicit, or question based) and yield form (extractive or abstractive). [1]

Single archive summarization produces once-over of single information report. Then again, multi archive summarization produces summary of various information record. These various wellsprings of information are associated records examining a tantamount subject. Huge amounts of the early summarization frameworks supervised single archive summarization.

Non express summarization setup is to gather all texts paying little character to its subject or space; i.e., non unequivocal plans make no presumptions about the region of its source information and view all records as homogenous texts. Most of the work that has been done turns around nonexclusive summarization . There have also been updates of summarization structures which are locked in upon different zone of intrigue. For instance, compacting resource articles, biomedical reports, climate news, mental lobbyist occasions and different more . Regularly, this sort of summarization requires zone explicit learning bases to help its sentence choice procedure. Question based summary contains just information which are tended to by the client.

The solicitation are typically trademark dialect solicitation or watchwords that are identified with a specific subject. For example, bits made through web search instruments is a case of request based application [1] Extractive outlines or focuses are passed on by perceiving essential sentences which are expressly perused the archive. A tremendous fragment of the summarization structures that have been made are for extractive sort synopses. In abstractive summarization, the picked archive sentences are joined reasonably and compacted to bar immaterial zones of the sentences. [1]

The fundamental motivation behind this work is to make the modules for Automatic observation structure. This module ought to get a touch of the obtained picture as the information and is depended upon to restore the number in editable sort of tag. Everything considered, the structure is needed to see a wide extent of names. The gathering of them is monstrous. They are of various shapes and shades, letters can be facilitated more than one segment. In ID and affirmation procedure four essential advances are there.

A. Preprocessing
B. Restriction
C. Division
D. Acknowledgment.

II. RELATED WORK

S. Wang, X. Zhao, B. Li, B. Ge and D. Tang [3] With the risky headway of data on the Internet, it winds up being continuously essential to improve the proficiency of data obtaining. Altered text summarization gives a not all that awful plans to vivacious tying down of data through weight and refinement. While existing strategies for tweaked text summarization accomplish flawless execution on short groupings, in any case, they are opposing the inconveniences of low ampleness and exactness while supervising long text.
In this paper, they show a twophase methodology towards long text summarization, explicitly, EA-LTS. In the extraction mastermind, it imagines a crossbreed sentence resemblance measure by joining sentence vector and Levenshtein specific, and bearings it into layout model to remove key sentences. In the reflection organize, it develops a sporadic neural structure based encoder-decoder, and devises pointer and thought parts to make outlines. Producers test the model on a bona fide long text corpora, gathered

from sina.com, starter happens assert the accuracy and validness of the proposed philosophy, which is gave off an impression of being better than top tier frameworks. Dan Cao and Liutong Xu [4] Automatic text summarization is an essential research zone in the space of data frameworks. It expects to make a compacted alteration of reports, which should cover all the essential substance and general noteworthiness. In extractive text summarization, sentences are scored on different of highlights. A broad number of highlights organize based have been proposed by scientists in the past creative works. This paper surveys every last one of the highlights that use estimations and thought of complex structure for scoring sentences. The test happens as expected on single portion and blends of different highlights we proposed are investigated. Quantitative and abstract perspectives were considered in our appraisal performing on the DUC 2002 informational collections. Nimisha Dheer [5] The present progress of revamp text summarization gives a fundamental part in the data recuperation (IR) and text request, and it gives the best response for the data over-stack issue. Text summarization is a procedure for diminishing the cross of a text while secures its data content. While considering the size and number of records which are open on the Internet and from exchange sources, the necessity for an uncommonly capable contraption on which produces usable once-overs is clear. They demonstrate a dominating count using lexical chain estimation &WordNet. The figuring one which makes lexical chains that is computationally attainable for the customer. Using these lexical chains the customer will make a synopsis, which is broadly additionally convincing risen up out of the methodologies open what's all the more closer to the human passed on outline. N. M. Chidiac, P. Damien and C. Yaacoub [6] A ground-breaking count that recognizes text from common scene pictures and focuses them paying little personality to the presentation is proposed. Each and every current procedure are planned to work under a particular impediment, for example, distinguishing text only one way. Maximally Stable Extremal Regions (MSER) pointer is expelled parallel areas since it has wound up being generous to lighting conditions. An update system for MSER pictures is expected to secure clear letter limits. Pictures are then reinforced into a Stroke Width Detector and a couple of heuristics are associated with clear non-text pixels. A brief timeframe later, recognized text areas are reinforced into an Optical Character Recognition module and after that filtered by their sureness measure.
The affirmation of characters isn't a bit of the figuring and the results are just about the disclosure of text. Our computation wound up being convincing on clouded pictures and uproarious pictures as well, in perspective on both abstract and target appraisals. L. Wang, W. Fan, J. Sun, S. Naoi and T. Hiroshi [7] Text line extraction in archive pictures is a basic for a few, content based picture getting applications. In this paper, we propose an exact and fiery procedure for vague text line extraction, which can be associated on enormous groupings of archive pictures, arranged vernaculars, and text lines with different presentations. Immediately, the contender related portions are isolated from archive picture using Maximal Stable Extremal Region (MSER) with the bustles filtered by Adaboost and Convolution Neural Network (CNN).

By then, the coarse text lines are created from dynamic edges proliferation and cut by adjacent linearity of text lines in the record spreading over tree.

Finally, for exact text line extraction, the cut multi-parts are re-related in perspective on text line essentialness minimization with respect to text line consistency and the fitting screw up.

Test comes to fruition on multilingual test dataset display the ampleness and amazing of the proposed procedure, which yields higher execution differentiated and top tier systems.

M. Afsharizadeh, H. Ebrahimpour-Komleh and A. Bagheri [8] Today there is a colossal proportion of data from an extensive proportion of various resources, for instance, World Wide Web, news stories, digital books and messages.

From one point of view, individuals defy an insufficiency of time, and afterward once more, on account of the social and word related necessities, they need to get the most basic data from various resources.

III. PROPOSED WORK

The proposed work for the Text Summarization is explained in the following steps.;

Step 1: Read the document to be summarized.

Step 2: Read the Manually generated summary.

Step 3: Read the Percentage of final generated summary to be matched with the original summary.

Step 4: Segment the document into the single line or the sentences.

Step 5: Remove the punctuation marks to perform the process of the tokenization.

Step 6: Perform the stemming of the line obtained after the tokenization in order to obtain the base form of the word using the wordnet library.

Step 7: Using the MaxentTagger perform the tagging over the stemmed document lines.

Step 8: Process similarly the steps for all lines of the documents , starting from step 4 to step 7.

Step 9: Find the unique words in the document lines and also compute the frequency of each word appearing in the document.

Step 10: Form the lexical chain for the unique works containing the synonyms , Antonyms, etc.. together with that calculate the score of the each chain using the following formula ,

Chain Score = ∑ Frequency of Related Words

Step 11 : Find the Unique Word score of the words present in the chain by making use of the distance based approach.

The concept of the distance base approach is

Position 1 = Position of the Word Preceding the word analyzed in the lexical chain.

Position 2 = Position of the Word Successing the word analyzed in the lexical chain.

Min = (Position 1 < Position) then Position 1 otherwise Position 2

Determine the value of apha=1*(min/total unique words)

Unique Word Score = alpha* chainscore (score of chain in which word is present)

Step 12: Calculate the Score of each line on the basis of the score of each unique word.

Step 13: Arrange the lines on the basis of the line score in the decreasing order such that the highest score line will be on the top.

Step 14: Fetch the percentage number of lines which are specified in the Step 3.

Step 15: Find the array of the words present in the matched lines.

Step 16: Read the Manual Summary file and form the array of the words present in the manual summary.

Step 17: Compare both the arrays for the similarly, the percentage of match occurred is called the recall.

Step 18: Print the Recall value.

Implementation

The implementation is done in java and eclipse IDE

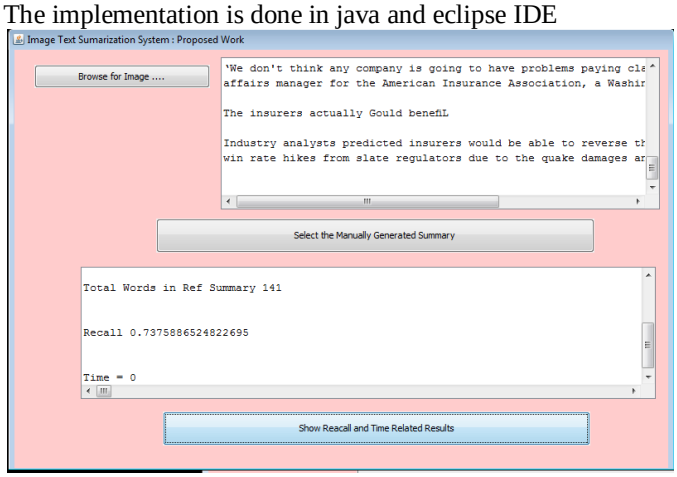
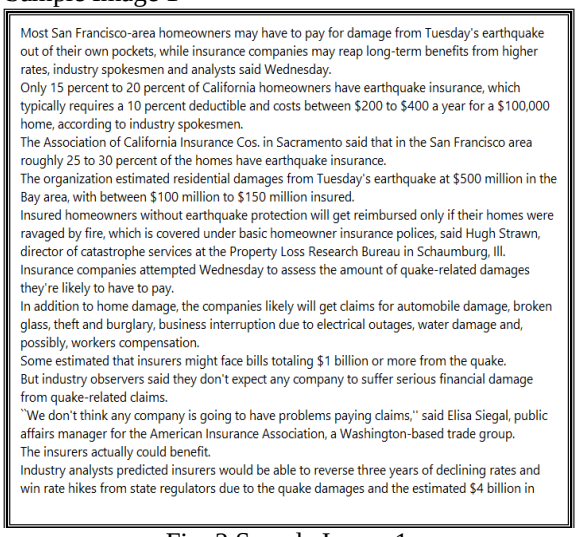


Fig1. Implementation Snapshot

IV. RESULT ANALYSIS

5.1 Sample Image 1



Most San Francisco-area homeowners may have to pay for damage from Tuesday's earthquake out of their own pockets, while insurance companies may reap long-term benefits from higher rates, industry spokesmen and analysts said Wednesday.
Only 15 percent to 20 percent of California homeowners have earthquake insurance, which typically requires a 10 percent deductible and costs between $200 to $400 a year for a $100,000 home, according to industry spokesmen.
The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance.
The organization estimated residential damages from Tuesday's earthquake at $500 million in the Bay area, with between $100 million to $150 million insured.
Insured homeowners without earthquake protection will get reimbursed only if their homes were ravaged by fire, which is covered under basic homeowner insurance polices, said Hugh Strawn, director of catastrophe services at the Property Loss Research Bureau in Schaumburg, Ill.
Insurance companies attempted Wednesday to assess the amount of quake-related damages they're likely to have to pay.
In addition to home damage, the companies likely will get claims for automobile damage, broken glass, theft and burglary, business interruption due to electrical outages, water damage and, possibly, workers compensation.
Some estimated that insurers might face bills totaling $1 billion or more from the quake.
But industry observers said they don't expect any company to suffer serious financial damage from quake-related claims.
``We don't think any company is going to have problems paying claims,'' said Elisa Siegal, public affairs manager for the American Insurance Association, a Washington-based trade group.
The insurers actually could benefit.
Industry analysts predicted insurers would be able to reverse three years of declining rates and win rate hikes from state regulators due to the quake damages and the estimated $4 billion in

Fig 2 Sample Image 1

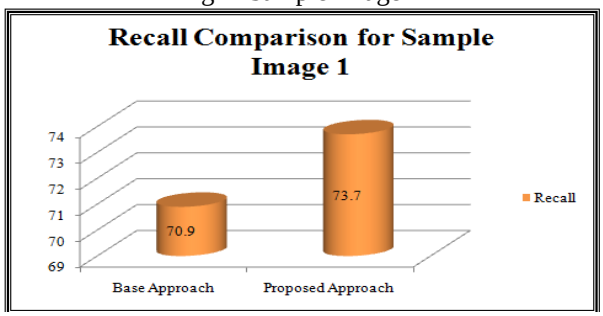
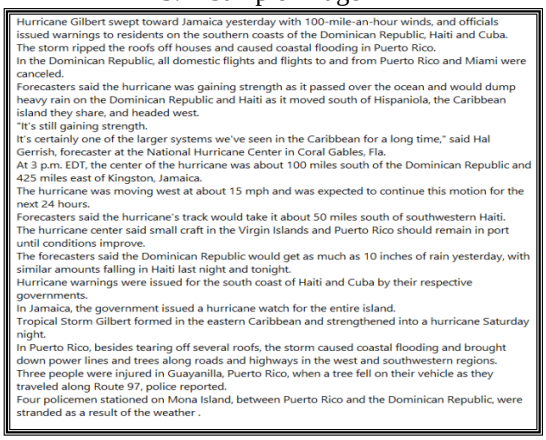


Fig. 3 Result for Sample 1

5.2 Sample Image 2



Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba.
The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico.
In the Dominican Republic, all domestic flights and flights to and from Puerto Rico and Miami were canceled.
Forecasters said the hurricane was gaining strength as it passed over the ocean and would dump heavy rain on the Dominican Republic and Haiti as it moved south of Hispaniola, the Caribbean island they share, and headed west.
"It's still gaining strength.
It's certainly one of the larger systems we've seen in the Caribbean for a long time," said Hal Gerrish, forecaster at the National Hurricane Center in Coral Gables, Fla.
At 3 p.m. EDT, the center of the hurricane was about 100 miles south of the Dominican Republic and 425 miles east of Kingston, Jamaica.
The hurricane was moving west at about 15 mph and was expected to continue this motion for the next 24 hours.
Forecasters said the hurricane's track would take it about 50 miles south of southwestern Haiti.
The hurricane center said small craft in the Virgin Islands and Puerto Rico should remain in port until conditions improve.
The forecasters said the Dominican Republic would get as much as 10 inches of rain yesterday, with similar amounts falling in Haiti last night and tonight.
Hurricane warnings were issued for the south coast of Haiti and Cuba by their respective governments.
In Jamaica, the government issued a hurricane watch for the entire island.
Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
In Puerto Rico, besides tearing off several roofs, the storm caused coastal flooding and brought down power lines and trees along roads and highways in the west and southwestern regions.
Three people were injured in Guayanilla, Puerto Rico, when a tree fell on their vehicle as they traveled along Route 97, police reported.
Four policemen stationed on Mona Island, between Puerto Rico and the Dominican Republic, were stranded as a result of the weather .

Fig 4 Sample Image 2

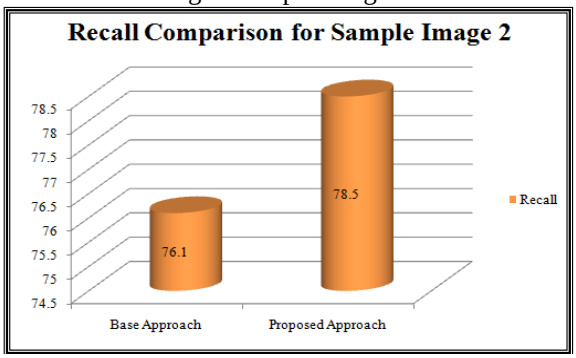


Fig. 5 Sample Image 2

V. CONCLUSION

The administrative work present the unraveled computation which uses the division base technique for choosing the utility of the extraordinary works in the file, which accordingly is used for finding the line score. The results which are gained will clarify the capability of computation in an enormous part of the occasion of the took a gander at documents.

Regardless, this field of the Natural language no work is adequate and still the redesigns is needed in all respects , in like manner for the future work we endeavor to stretch out this in order to give indications of progress results, together with that will endeavor to wear down INDOWORDNET , which is across the board programming interface for the word reference of all Indian language , to make comprehensive stage for summarization of files reliant on the India vernaculars..

REFERENCES

[1] Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar,"Automatic Text Summarization of News Articles",International Conference on Big Data, IoT and Data Science (BID),2017

[2] Mr. Pratik Madhukar Manwatkar,Mr. Shashank H. Yadav, "Text Recognition from Images",IEEE Sponsored 2nd International Conference on Innovations in Information,Embedded and Communication systems (ICIIECS)2015

[3] S. Wang, X. Zhao, B. Li, B. Ge and D. Tang, "Integrating Extractive and Abstractive Models for Long Text Summarization," 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, 2017, pp. 305-312.

[4] Dan Cao and Liutong Xu, "Analysis of complex network methods for extractive automatic text summarization," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2016, pp. 2749-2756.

[5] Nimisha Dheer,Mr. Chetan Kumar ,"Extractive Automatic Text Summarization through Lexical Chain Method using WordNet Dictionary", IEEE 2016

[6] N. M. Chidiac, P. Damien and C. Yaacoub, "A robust algorithm for text extraction from images," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), Vienna, 2016, pp. 493-497.

[7] L. Wang, W. Fan, J. Sun, S. Naoi and T. Hiroshi, "Text line extraction in document images," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, 2015, pp. 191-195.

[8] M. Afsharizadeh, H. Ebrahimpour-Komleh and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th International Conference on Web Research (ICWR), Tehran, 2018, pp. 128-132.

[9] H. A. Chopade and M. Narvekar, "Hybrid auto text summarization using deep neural network and fuzzy logic system," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 52-56.

[10] L. Wan, "Extraction Algorithm of English Text Summarization for English Teaching," 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xiamen, 2018, pp. 307-310.

[11] A. Jain, D. Bhatia and M. K. Thakur, "Extractive Text Summarization Using Word Vector Embedding," 2017 International Conference on Machine Learning and Data Science (MLDS), Noida, 2017, pp. 51-55.

