International Journal of Future Generation Communication and Networking
Vol. 13, No. 04, (2020), pp. 1329-1337

Automatic Text Summarization using Extractive Techniques and Attribute Tagger Algorithm

Dr. D.N.V.S.L.S. Indira[1]

[1]*Associate Professor, Department of Information Technology,*

Gudlavalleru Engineering College, Gudlavalleru, Krishna District, Andhra Pradesh, India. -521356

indiragamini@gmail.com

Abstract

With the emotional development of the Internet, individuals are overpowered by the enormous measure of online data, archives and documents. This growing accessibility documents or records have requested comprehensive exploration in the region of automatic text summarization. Text Summarizer is a method of shortening long pieces of text. The expectation is to make a cognizant and fluent summary having just the primary concerns delineated in the document. Automatic Text Summarization techniques are classified in to two, first one Abstraction based approach and second one Extraction based approach. In this paper we used extractive text summarization technique along with our novel algorithm attribute tagger to outline the document to present key information in it. Extraction-based summarization model takes an input that encapsulates some paragraphs and returns a text summary that represents the outline information or message in the input text. Attribute Tagger algorithm reduces the work by identifying keywords or important attributes in the input using NER technique. The output from the Attribute Tagger algorithm is given as input to TextRank and SentenceRank algorithms. The test results show that our proposed approach can improve the performance compared to sate-of-the-art summarization approaches.

Keywords: *Big Data, Text Summarization, Extractive Techniques, Attribute Tagger, Named Entity Recognizer*

1. INTRODUCTION

Automatic Text Summarization is one of the most testing and intriguing issues concerning the field of Natural Language Processing (NLP)[3] [4] .It is an example of making a limited and enormous outline of text from different substance assets, for example, books, reports, blog areas, research papers, messages, and tweets. Preceding embarking to the Text rundown, first we have to understand that what a synopsis is[1]. A framework or outline is a book that is made from in any event one message that passes on noteworthy information in the primary substance and it is of a shorter structure. At last the content outline diminishes understanding time, quickens the way toward investigating for data, and builds the measure of data that can fit in a zone.

Today, our reality is dropped by the social event and dispersal of gigantic measures of information. Indeed, the International Data Corporation (IDC) ventures that the aggregate sum of advanced information circling every year around the globe would grow from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025 .With such a major measure of information coursing in the computerized space, there is have to create calculations that can consequently abbreviate longer messages and convey precise outlines that can fluidly pass the planned messages. Moreover, applying text summary lessens understanding time, quickens the way toward exploring for data, and builds the measure of data that can fit in a territory.

ISSN: 2233-7857 IJFGCN

There are two main types of summarization techniques in NLP:

1) **Extraction-based Summarization. [10][6][2]**
   The extraction strategy includes pulling key expressions from the source archive and joining them to make a synopsis. The extraction is made by the characterized measurement without rolling out any improvements to the writings. Consider it a highlighter, which chooses the fundamental data from a source text.

2) **Abstraction-based Summarization. [10]**
   The deliberation strategy involves rewording and shortening portions of the source record. The abstractive content summary calculations make new expressions and sentences that hand-off the most valuable data from the first content, much the same as people do. Consider it a pen, which produces novel sentences that may not be a piece of the source report.

   In spite of the fact that reflection performs better at text outline, building up its calculations requires convoluted profound learning methods and advanced language demonstrating. To create conceivable yields, reflection based synopsis approaches must address a wide assortment of NLP issues, for example, normal language age, semantic portrayal, and derivation change. In that capacity, extractive content summary approaches are still broadly main stream. In this paper, we'll be concentrating on extraction-based techniques.

2. RELATED WORK

Now, the whole world of Natural Language Processing dedicated to summarization created, covering a grouping of ordinary positions, for example, Headlines ,Outlines ,Minutes of a gathering, Previews of films, Synopses , Reviews of a book, CD, film, and so forth., , Biography , Bulletins like climate conjectures/financial exchange reports, and so on. Text outline or summary issue and the solutions can be depicted along different dimensions: [10][6]

- Input – Single/Multiple
- Context- Domain Specific/Query
- Output- Abstractive/Extractive
- Machine learning solution approaches- Supervised/Unsupervised

Different methods of text summary are: [5]

- Term Frequency Method (TF-IDF)
- Time Based Method
- Graph Based Method
- Clustering Based Method (Separation and Merging)
- Semantic Dependency based Method

All the Existing algorithms suffer from the following problems.

1) These algorithms just cut down a maximum of 40% manual effort only.

1) Applying multiple filters on documents would leave a confusion on which filter to be applied first to summarize the text.

2) When the required number of documents that needs to be scanned goes beyond a limit, application fails and it does give accurate result.

Above all else writing's and exploration some of the papers were inspected for getting more data about the issue and realizing which kind of work was finished by others on this point and by which technique.

ISSN: 2233-7857 IJFGCN

Xiaojun Wan et al. [1] proposes the Cluster-based Conditional Markov Random Walk Model (ClusterCMRW) and the Cluster-based HITS Model (ClusterHITS) to fully leverage the cluster-level information for exploited for multi-document summarization by making use of the link relationships between sentences in the document set.

Ledeneva et al. [2] proposed a language-and space free measurable based strategy for single record extractive synopsis to deliver a book outline by extricating a few sentences from the given content.

Aramaki et al. [3] attempted to do a fundamental arrangement of clinical content outline with regulated discovering that recognized negative occasions and furthermore examined what sorts of data assisted with distinguishing negative occasions. In recognizing negative occasions from different occasions, he utilizes the SVM Classification.

D.N.V.S.L.S.Indira et al. [4] proposed a novel concept of Attribute Tagger for identifying Named entities in the documents using NLP techniques.

3. PROPOSED WORK

This paper describes a system for the summarization of the given text input. For this we are using Extractive text summarization, which means pulling key phrases from source document. Here we are using two extractive based summarization techniques. In this we used another algorithm Attribute Tagger [4] for identifying keywords in the input. We are giving these keywords as input to the sentence rank and text rank algorithms. We observed good results from the above said algorithms when we supplied output of attribute tagger algorithm as input before text summarization. This paper works on two phases. In the first phase we used attribute tagger algorithm and identified some keywords or attributes in the documents. And in the next phase we supplied these keywords for shortening the document for better understanding the text quickly.

3.1 Phase 1: Data pre-processing

In this phase of processing we made use of Stanford NLP POS Tagger, Parser and Named Entity Recognizer. Phase 1 of this processing talk about initial screening of the text data algorithm which works on retrieving Nouns, Places or Locations, Numbers (Ex: Phone numbers, IDs) etc.. in the given document. Utilizing Stanford NLP, we recover Parts of Speech (POS) labeling from the content record. Header segment of content document is searched for Nouns and it would be labeled as NN, NNP. Parts of discourse labeling are the bundle of Stanford NLP toolbox which helps in labeling content information to be handled. The library gives you "a chance to tag" the words in your string. That is, for each word, the "tagger" gets whether it's a thing, a verb. and so forth.

This pre-processing phase is same for both the extractive summarization techniques.

- Tokenization.
- Stop word removals
- Lemmatization

3.1.1. Proposed Algorithms for Phase 1:

Algorithm 1: NER (N) [4]

//Let us assume N: An Unlabeled Documents //Documents to be trained. NER – Named Entity Recognizer
Output:
A: Set of Attributes
Assumptions:
T: A set of trained data

ISSN: 2233-7857 IJFGCN

C: MaxEnt_Classifier
for i=1 to N in steps of 1 do
//MaxEnt_Classifier C trains N, based on T for a given label using Maximum entropy model. Extract required attributes A, based on C.
A= C(T)
end for

Algorithm 2: Attribute Tagger (D,N)

Let
D: Text data of a Document
N: Number of Documents
A: Set of attributes obtained from NER algorithm
W: Weight of each attribute
k : Number of attributes in A
Let y1, y2, y3, - - - yk be the attribute set in A
Let
Wi= W(yi) where i = 1 to k
Score(Di)= POS(A)= ∑_(i=1)^kWi

3.2. Phase 2: Text Summarization using Extractive Techniques

Algorithm 3: Sentence Rank(D, ST)

Sentences are situated by consigning weights and they are positioned subject to their heaps. Significantly situated sentences are isolated from the information record so it eliminates critical sentences which facilitate to an incredible once-over of the information text.[11]

D: Input variable which is a document consists of text.

A: Set of Attributes or Key phrases obtained from Attribute Tagger which has highest score.

ST: Output variable contains summarized text generated from the document which is shorter compared to D.

1. Read the document and the content is tokenized.
2. Apply POS algorithm and Attribute Tagger algorithm to get the important keywords.
3. Lemmatize each token.
4. Calculate frequency of individual token.
5. Calculate weighted frequency of token by dividing frequency with maximum one.
6. Calculate the weights of keywords obtained from AT algorithm.
7. Calculate weights of each sentence by substituting weighted frequency of token in sentence and sum up the keywords weight in the sentence.
8. At last, summarizer will separate the weighted recurrence sentences whose worth is more noteworthy than or equivalent to average of totals of sentences so as to discover outline of text.

Algorithm 4: TextRank(D, ST)[11]

D: Input variable which is a document consists of text.

A: Set of Attributes or Key phrases obtained from Attribute Tagger which has highest score.

ST: Output variable contains summarized text generated from the document which is shorter compared to D.

1. Perusing the given content and given content is tokenized into sentences.

ISSN: 2233-7857 IJFGCN

2. Find significant catchphrases utilizing Attribute Tagger Algorithm.

3. Ascertain likenesses between sentence vectors and put away in a framework.

4. The similitude grid is then changed over into a diagram, with sentences as vertices and likeness scores as edges, for sentence rank figuring.

5. At long last, a specific number of positioned sentences whose rank is more prominent than or equivalent to the normal of all sentences structure the last summary.

4. RESULTS AND DISCUSSION:

Consider the following paragraph as source text:
Ex: So, Keep working. Keep striving. Never give up. Fall down seven times, get up eight. Ease is a greater threat to progress than hardship. Ease is a greater threat to progress than hardship. So, Keep moving, keep growing, keep learning. See you at work. By using sentence rank algorithm, calculate weighted frequency of each sentence.

Sentence	Sum of weighted frequency
So, Keep working	1+0.2=1.2
Keep striving	1+0.2=1.2
Never give up	0.2+0.2=0.4
Fall down seven times, get up eight	0.2+0.2+0.2+0.2+0.2=1.0
Ease is a greater threat to progress than hardship	0.4+0.4+0.4+0.4+0.4=2.0
Ease is a greater threat to progress than hardship	0.4+0.4+0.4+0.4+0.4=2.0
So, Keep moving, keep growing, keep learning	1.0+0.2+1.0+0.2+1.0+0.2=3.6
See you at work	0.2+0.2=0.4

Table 1: Weights of the Sentences

Now select sentences of whose sum is greater than the threshold. Here threshold is the average of the sentence weights. These sentences together called a summary.

Therefore the output summary for the above document from sentence rank is :

Ease is a greater threat to progress than hardship. So, Keep moving, keep growing, keep learning

ISSN: 2233-7857 IJFGCN

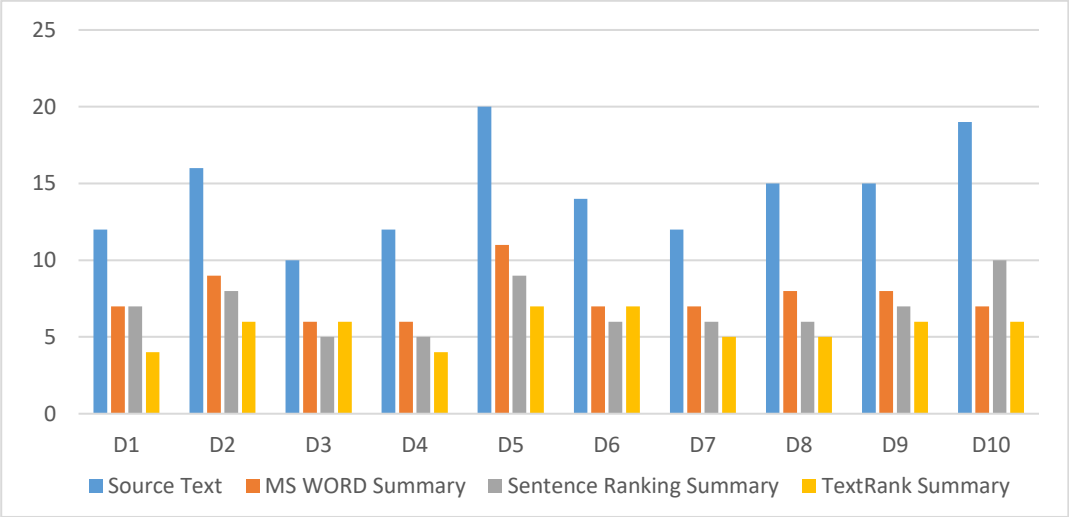


Fig 1: Comparison between SentenceRank and TextRank algorithms with total number of sentences present in summary document. Results show that TextRank algorithms gives short text with good content.

Source Text	Algorithm	Precision	Recall	F - measure
D1	SR	0.52	0.77	0.62
	TR	0.80	0.45	0.58
D2	SR	0.68	0.86	0.76
	TR	0.78	0.76	0.77
D3	SR	0.87	0.84	0.85
	TR	0.93	0.88	0.90
D4	SR	0.88	0.79	0.83
	TR	0.67	0.41	0.51
D5	SR	0.96	0.88	0.92
	TR	0.86	0.47	0.61
D6	SR	0.76	0.80	0.78
	TR	0.73	0.76	0.74
D7	SR	0.70	0.87	0.78
	TR	0.87	0.73	0.79
D8	SR	0.78	0.78	0.78
	TR	0.79	0.48	0.60
D9	SR	1	0.79	0.88
	TR	0.66	0.41	0.51
D10	SR	0.8	1	0.89
	TR	0.79	0.52	0.63

Table 2: A Table of Data obtained by calculating Precision, Recall and F-measure on sample documents

ISSN: 2233-7857 IJFGCN

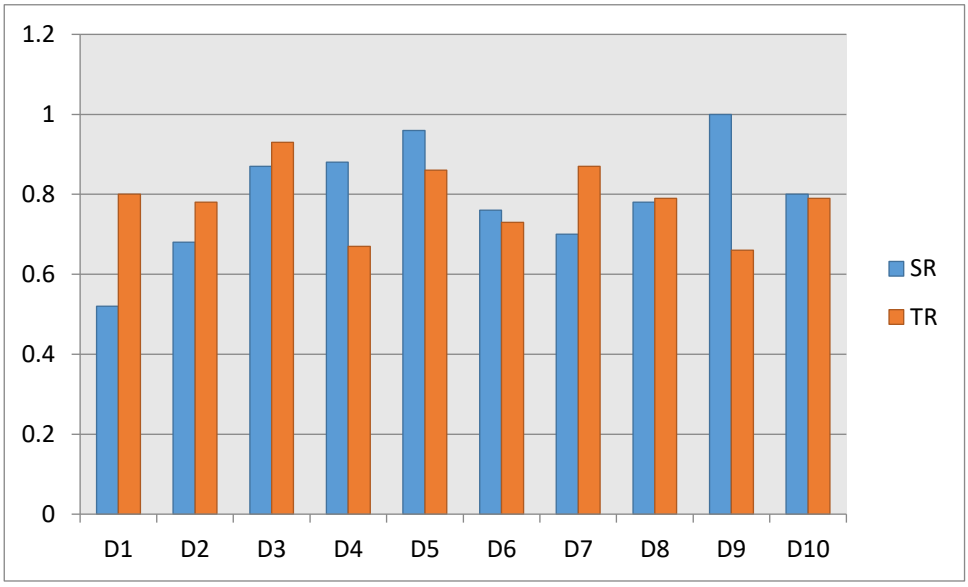


Fig 2: Precision values of Sentence Rank and Text Rank algorithms
X-Axis – Documents, Y-Axis- Precision Value

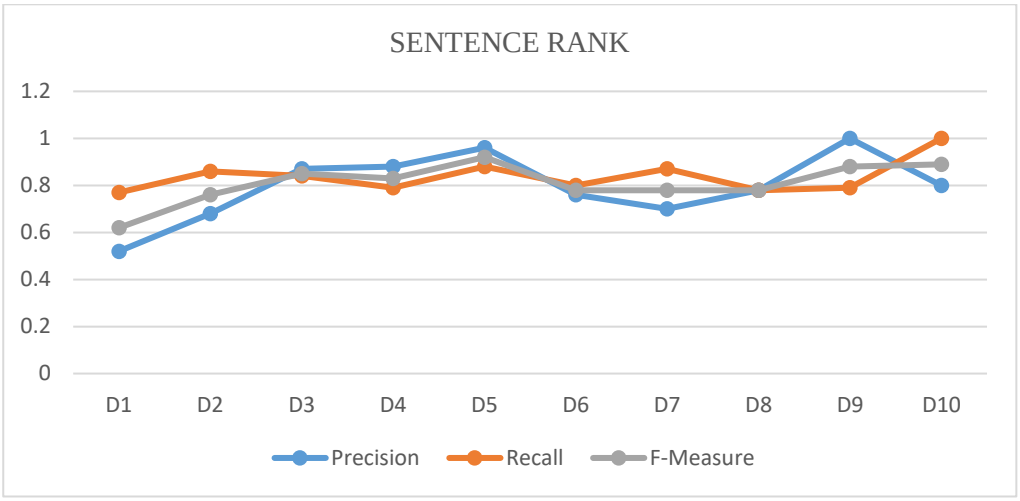
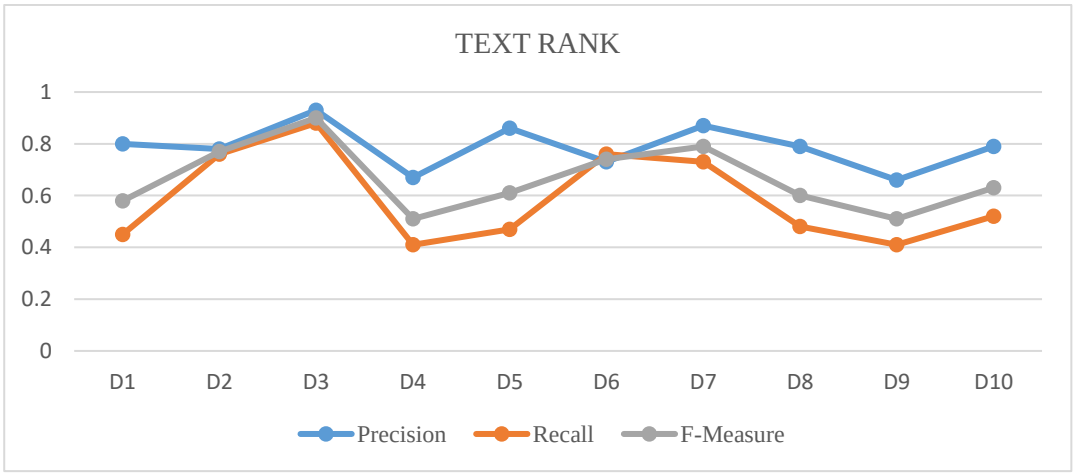


Fig 3: Precision, Recall and F-measure values using SentenceRank algorithm



ISSN: 2233-7857 IJFGCN

Fig 4: Precision, Recall and F-measure values using Text Rank Algorithm

5. CONCLUSION

Text summarizer is a perplexing endeavor which contains many sub-tasks in it. Each sub task has an ability to get extraordinary quality rundowns. The huge part in extractive substance summary is perceiving basic sentences from the given sentences. In this assignment we proposed extractive based substance once-over by using sentence situating and text rank. The sentences which are removed from input given by User Interface are conveyed as a summarized book and it is appeared in new page. We utilized NER and Attribute Tagger calculations for finding significant words and given as contribution to sentence and text rank calculations. The test results show that our proposed approach can improve the introduction diverged from satisfy of-the-craftsmanship diagram moves close.

6. FUTURE SCOPE

The proposed technique is fundamentally an extraction based methodology. As extractive content rundown approaches are still generally mainstream there is have to create reflection based methodologies which performs better than extraction based. Our undertaking expands upon single archive rundown. Future work incorporates joining of multi-archive rundown with report bunching to give synopses and development of intelligent interfaces so clients can viably utilize multi document outline to peruse and investigate huge record sets. This work can be improved to find opinions of people reviews in social media in different linguists.

7. REFERENCES

[1] Xiaojun Wan and Jianwu Yang, “Multi-document summarization using cluster-based link analysis”. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 299–306, 2008.
[2] Yulia Ledeneva, Alexander Gelbukh, and René Arnulfo García-Hernández, "Terms Derived from Frequent Sequences for Extractive Text Summarization", CICLing 2008, LNCS 4919, pp. 593–604, 2008.
[3] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, Kazuhiko Ohe, “TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification”, Proceedings of the BioNLP Workshop, pp. 185–192, June 2009.
[4] Mrs. D.N.V.S.L.S.Indira, Dr. R. Kiran Kumar “Profile Screening and Recommending using Natural Language Processing (NLP) and leverage Hadoop framework for Bigdata”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 6, June 2016
[5] Neelima Bhatia, Arunima Jaiswal “Automatic text summarization and its methods – a review”, 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016 IEEE
[6] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assef “Text Summarization Techniques: A Brief Survey”, ACM ISBN 978, July 2017.
[7] Pavan Kartheek Rachabathuni, “A survey on abstractive summarization techniques”, Published in: IEEE International Conference on Inventive Computing and Informatics (ICICI) , 2017.
[8] Shashi Narayan Shay B. Cohen Mirella Lapata, “Ranking Sentences for Extractive Summarization with Reinforcement Learning”, Proceedings of NAACL-HLT , pages 1747–1759, 2018
[9] J.N.Madhuri, Ganesh Kumar.R “Extractive Text Summarization Using Sentence Ranking”, 978-1-5386-9319-3/19- 2019 IEEE.
[10] Adhika Pramita Widyassari, Edy Noersasongko, Abdul Syukur “Literature Review of Automatic Text Summarization: Research Trend, Dataset and Method” 978-1-7281-1655-6/19, 2019 IEEE.

ISSN: 2233-7857 IJFGCN

[11] Ming Zhong,, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, Xuanjing Huang "Extractive Summarization as Text Matching" , [cs.CL] 19 Apr 2020.

ISSN: 2233-7857 IJFGCN