# A Heuristic Approach for Telugu Text Summarization with Improved Sentence Ranking

**Kishore Kumar Mamidala[a]*, Suresh Kumar Sanampudi[b]**

[a]Department of CSE.Research Scholar, Jawaharla Nehru Technological University, Hyderabad, Hyderabad, INDIA
[b]Department of Information Technology. Assistant Professor and Head, JNTUH College of Engg Jagtial, Telangana, INDIA

**Article History**: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

**Abstract:** Extracting/abstracting the condensed form of original text document by retaining its information and complete meaning is known as text summarization. The creation of manual summaries from large text documents is difficult and time-consuming for humans. Text summarization has become an important and challenging area in natural language processing. This paper presents a heuristic appraoch to extract a summary of e-news articles of the Telugu language. The method proposes new lexical parameter-based information extraction (IE) rules for scoring the sentences. Event score and Named Entity Score is a novel part in sentence scoring to identify the essential information in the text. Depending on the frequency of occurrence of event/named entites in the sentence and document, sentences are selected for summary. Data is collected from online news sources (i.e., Eenadu, Sakshi,Andhra Jyothi, Namaste Telangana) to experiment. The proposed method is compared with other techniques developed for Telugu text summarization. Evaluation metrics like precision, recall, and F1 score is used to measure the proposed method's performance. An extensive statistical and qualitative evaluation of the system's summaries has been conducted using Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a standard summary evaluation tool. The results showed improved performance compared to other methods.

**Keywords:** Telugu Text Summarization, Extractive Summarization, Natural language Processing, Information Extraction, Events, Named Entities

## 1. Introduction

For English, several advancements are made in the field of Text Summarization but not for Indian languages. Telugu is the 2nd famous language in India and the 15th most popularly speaking world language[4]. Telugu is an agglutinative language, due to which text summarizations developed for other Indian languages like Hindi, Bengali does not support Telugu. Text summarization for Telugu obtained little attention because of non-availability of Telugu resources like data sets, dictionaries, wordnet, etc. Nowadays, Telugu e-newspapers (Ennadu, Sakshi, Andhrajyothi, Namaste Telangana) is freely available online. Extraction of important information from these newspapers is a time-consuming task. Text summarization plays a role in mining the significant sentences to generate the summary of the entire document.

The automatic text summarization method provides the original text document's condensed form by retaining the meaning and information. The summary helps the readers to understand the content quickly without reading the entire text. Depending on the type of summary, Text summarization methods are broadly classified into extractive/abstractive. Extractive summarization retrieves selected sentences from the source text. Sentences are extracted depending on the statistical and linguistic features in the input text [16]. Abstractive summarization methods interpret the source document and rewrite the sentences to obtain summaries. This paper proposed an improved sentence ranking approach to generates effective summaries for Telugu text socuments based on occurrences of events and named entities in the text.

The rest paper's sequencing is as follows: Section 2 explains the literature of various summarization techniques developed for Indian languages. The framework of Text Summarization approach developed for Telugu are described in section 3. Section 4 illustrates the dataset and experimental results of the work. Section 5 provide conclusion of the paper.

## 2.Related Work

In the literature, Automatic text summarization systems are available for English and other foreign languages in maximum but less for Indian languages. This section explains various text summarizers developed for Indian languages.

Several researchers developed summarization techniques using extractive methods for Indian languages. In [8], the sentence scoring mechanism is used to obtain a summary for Hindi text. The rules are built depending on the features like cue words, nouns, title words, sentence length, position, numerical data, inverted commas, etc., to obtain different sentence scores. Lexical rule-based text summarization is developed for Hindi [12]. Word-level features such as word frequency, word length, word occurrences, and sentence level features such as sentence length and a similarity score of sentences are used in rule formation.

In [14], vectors space term weighing is used to rank the sentences in the document. Query words are given importance in sentence scoring. Topic-based opinion text summaries for Bengali are developed that consolidates the sentiment information in the given input text document [4]. Extractive summarization for Bengali is created using the thematic term and the word's position as features [5]. In [11], the multi-document text summarization for Bengali is explained. Statistical methods like term frequency are used to score the sentences and extract the relevant information from multiple documents.

In [7], proposed a text summarization for Tamil. In this method, semantic graphs are built for the source text document. By analysing these semantic graphs, humans' experts obtain the summary of the text. Statistical methods such as word frequency, word position, number of named entities in sentences are used to score the sentences, highest-ranked sentences are retrieved to generate a summary for online sports news in Tamil [15].

For Kannada language, Extraction based text summarization developed depending on key term scores [10][13]. Sentences are scored using the key terms obtained based on term frequency and inverse document frequency measures. In [2], relevant sentences are extracted by computing sentence scores in Malayalam text document. The term frequency and position of words are used to find the score.

For Telugu, keyword-based approaches are used to generate the summaries [9]. The probability distribution of tags is used to identify the keywords, which helps to score the sentences. Human intervention is needeed to some extent at annotation level to identify the keywords. In[3] neural network based appraoch is used to genrate the summaries, but are not evaluated for their performance. A literature study has shown that all the Indian language text summarization is Extraction based. Statistical and Lexical features are used to rank the sentences. This paper presents a complete automated heuristic approach of text summarization with an improved sentence ranking mechanism.

## 3.Proposed Summarization Method

Text Summarization for Telugu is one of the vital applications in Natural Language Processing (NLP). This section proposes a heuristic approach for automatic text summarization of Telugu documents. An improved sentence scoring method is used to rank the sentences. ISentence scoring mechanism is based on the event and named entity scores.

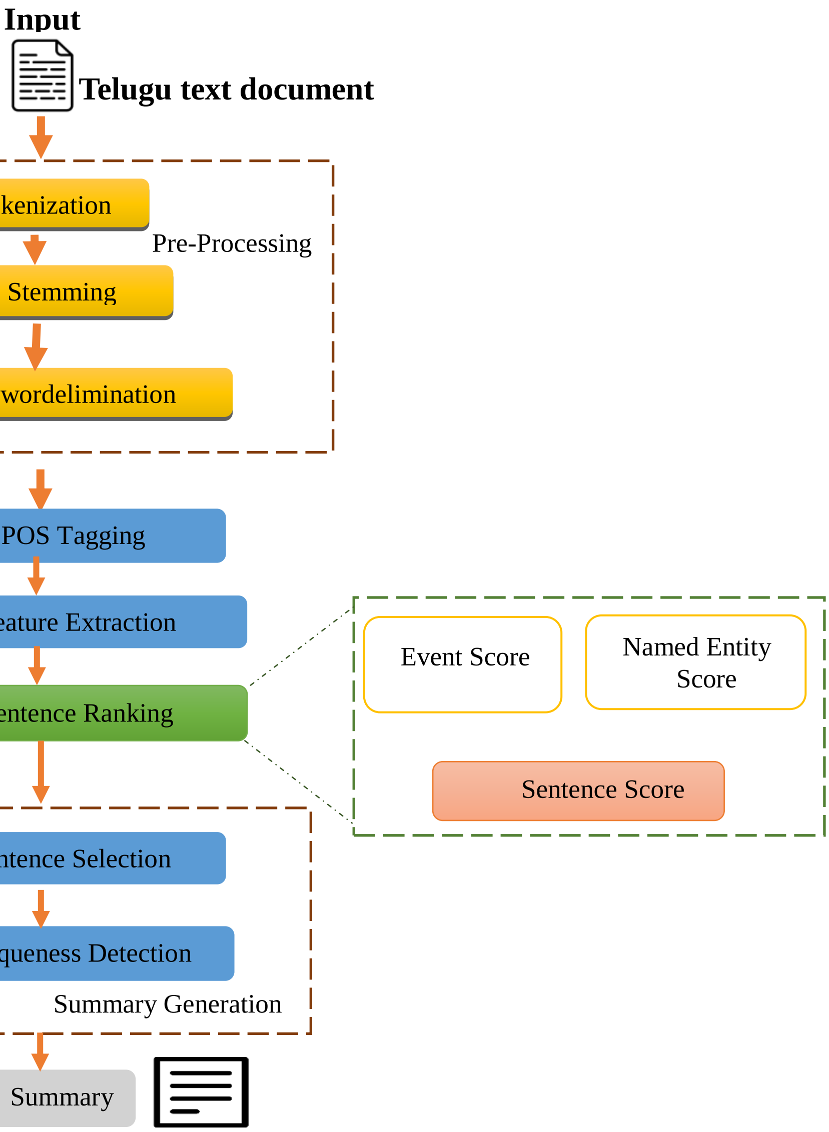

**Fig 1**: Architecture of  Heuristic based Approach for Telugu Text Summarization

An event is defined as a happening/occurrence of any situation in the real-world scenario. The named entity is defined as the people, place, things involved in an event happening. The statistical-based lexical rules of extraction are developed for scoring the events and named entities. The scores are further used to identify the sentence scores. In the proposed method, the Telugu text document is taken as Input. Pre-processing steps such as tokenization and stemming are performed. Tokenization performs the splitting of a text document into a sequence of words. Using Stemming, the term is divided into stem and suffix. The stemmer algorithm removes the suffixes utilizing a set of frequent suffixes. For example, in words, దేశంలో, and దేశం the letter లో is removed, and both the terms are treated the same. The stop words are extracted from the document. There are 228 stopwords built for Telugu. Stop words such as లో, ఒక,మార్పు,పేజీ,ఈ,కు, etc. are removed from the text. The remaining terms are sent for tagging the Parts of Speech (POS).

"Events" and "Named Entities" are linguistic features used in the proposed method. Events are terms that indicate happenings in the real-world. The verbs in the text explain the actions. They form an essential role in scoring the sentence relevance for a summary. Named Entities are the name of a person, place, thing, and animal involved in the occurrence of this action. Nouns are the POS tagged to such words in the language. The available events and named entities in each sentence are retrieved by feature extraction part in proposed method and sent  to perform statistical analysis on them.

Sentence scoring is done by applying statistical measures on Events and Named Entities obtained. The number of event/named entity occurrences is used to find the word frequency score. The correlation between the number of events/named entities in the document with that of total events/named entities is determined as word frequency score. Equation 1 is used to calculate the word frequency score. The number of sentences in which the event/named entity occurred helps to find the inverse sentence frequency. Equation 2 explains the calculation of inverse sentence frequency of events/named entities. The word occurrence in many sentences gets the least significance to be included in the summary. The Product of word frequency and inverse sentence frequency obtains the wf_isf score of term t. Equation 3 finds the term's significance to be included in the summary based on the score of wf_isf.

$$Wf_t(D)_{t\in\{Event\ /NamedEntity\ \}} = \frac{Occurences\ \ of\ t\ in\ document\ \ D}{Total\ \ number\ \ of\ terms\ (t)\ in\ document\ \ D} \tag{1}$$

$$ISf_t(D)_{t\in\{Event\ /NamedEntity\ \}} = log_e\left(\frac{Total\ \ number\ \ of\ \ sentenc\ es\ in\ document\ \ D}{Number\ \ of\ Sentences\ \ containing\ \ \ term\ \ t}\right) \tag{2}$$

$$WF\_ISf_t(D)_{t\in\{Event\ /NamedEntity\ \}} = WF_t(D) * ISF_t(D) \tag{3}$$

The summation of each event's or named entity "wf_isf" score in the sentence is done. Sentence score is obtained by finding a correlation between this value concerning the number of events and named entities in the entire sentence. Equation 4 shows the calculation of sentence score.

$$Sentence\ Score(S_i) = \frac{\sum_{t=1}^{N_{E\_NE}} WF\_ISf_t(D)}{N_{E\_NE}} \tag{4}$$

$$N_{E\_NE} = Total\ number\ of\ ef events\ and\ named\ entities\ in\ Sentence\ S_i$$

The sentence ranking step arranges the sentences in the chronological order of sentence scores. The average score of these sentence scores is used to fix the threshold for sentence selection. In the proposed method, sentences are selected for summary only if the sentence score is greater than the threshold. Sometimes the sentences retrieved for summary may contain duplicate content.

Uniqueness detection in the proposed method identifies whether the sentence selected contains unique information or not. The sentence similarity measure is used to compare whether two sentences are similar. The sentences are converted to vectors, and the similarity between the two sentences Si and Sj is computed using equation 5. If the similarity score between two sentences is greater than 80%, then the sentence with the less scored sentence is eliminated by retaining the highest score sentence in summary.

$$\boldsymbol{Similarity}(\boldsymbol{S_{i,}S_j}) = \frac{s_i \cdot s_j}{\|s_i\| \times \|s_j\|} \tag{5}$$

For example, consider the sentences

$S_1$: తరువాతహైబ్రిడీకరణద్వారాఊదా , ముదురుగులాబీరంగుచుక్కకనిపించేది .

$S_2$: తరువాతహైబ్రిడీకరణద్వారాఊదా , ముదురుగులాబీ , గులాబీ , కాషాయం , తెలుపు , ఎరువు , పీచ్రంగులుబాగాఆదరణపొందాయి .

Sentence $S_1$ is 95% similar when compared with $S_2$ using the sentence similarity metric. Sentence scoring of $S_1$ is 0.24, and $S_2$ score is computed as 0.73. Out of these two sentences, $S_1$ is eliminated to form the summary sentence since it has a low sentence score when compared to $S_2$. Summary generation part of the proposed method extract the highest score unique sentences to form the summary.

## 4.Expreriment Results and Disucssion

This section evaluates the quality of summaries obtained by the proposed algorithm. The experimentation starts with data collection by scraping the content from popular e-newspapers like Ennadu, Sakshi, Andhra Jyothi, Namaste Telangana, etc. The dataset contains 90 articles from each newspaper collected for 30 days. A total of 360 articles were collected. Each document contains around 50 to 60 sentences. Human-generated summaries for these documents are developed by Telugu linguists and are termed model summaries. These summaries are used to compare the system summaries for measuring the performance.

To compare the results of the proposed method, the precision, recall, and F-score are calculated using the Recall-Oriented Understudy for Gisting *Evaluation (ROUGE) 1.5.5 tool[6]. It is a standard summary evaluation tool to access summaries generated by systems. ROUGE tool returns three evaluation metrics, namely "average precision, average recall, and average f-score," to determine the performance of the system.* Precision is defined as the number of sentences comparative in both model and system summaries to the number of sentences in the system summary. *Recall metric plays a crucial role in identifying the number of sentences identical in both model and system-generated- summaries. F-score is defined as the harmonic mean of precision and recall scores.*

Table 1 compares the experiments conducted on the created dataset of 360 documents. The results are compared with that of keyword-based text summarizer[9] and neural text summarizer[3] developed for Telugu in the literature. The result shows that the proposed work beats the other methods considering the "average precision, average recall and average f-score" values. Figure 2 gives the comparative chart for average scores of three evaluation metrics – precision, recall, and f-score obtained by different summarization methods.

**Table 1:** Comparison of average scores of precision, recall and f-score of proposed method with keyword based and neural based summarizer

| **Approach** | **Avg-Precision** | **Average recall** | **Average F Score** |
|---|---|---|---|
| **Heuristic based (Proposed Method)** | **0.883** | **0.865** | **0.873** |
| **Keyword based-Summarizer** | 0.826 | 0.784 | 0.766 |
| **Neural summarizer** | 0.845 | 0.812 | 0.828 |

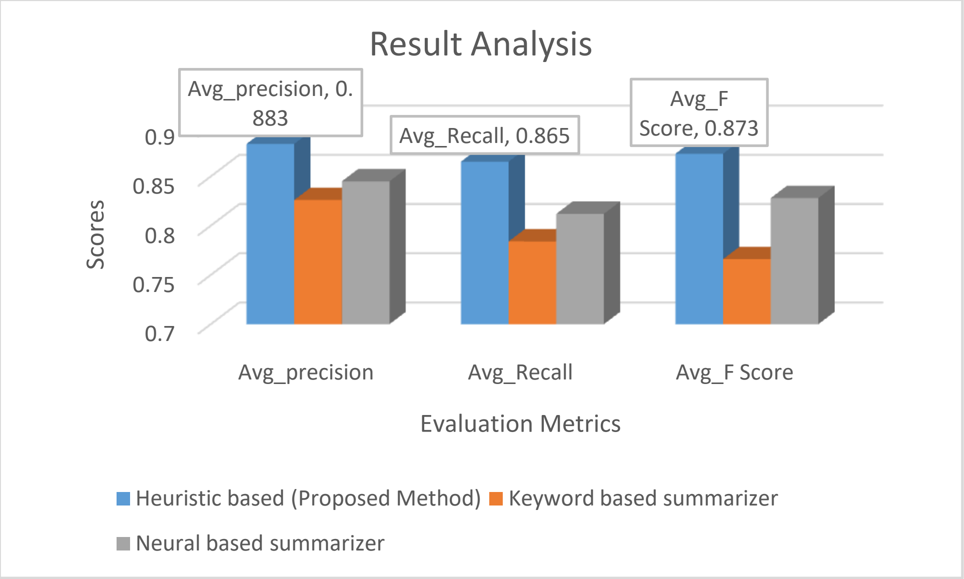

**Fig. 2.** Comparative chart for average values of precision, recall & f-score of Heuristic approach (proposed method) with other summarization approaches.

## 5.Conclusion

This paper proposed a heuristic-based method of extractive text summarization with an improved sentence ranking mechanism for Telugu text documents. Events and named entities are linguistic parameters used to identify the significant sentences in the text. Sentence scoring is computed using events and named entities occurrences in the text. The highest-ranked unique sentences are selected to generate the summary. Three hundred sixty articles are collected from various Telugu e-newspapers, which are used to evaluate the experiments in the proposed method. Standard evaluation metrics – "precision, recall, and f-score" are used to measure the proposed method's performance. ROUGE evaluation tool is used to find these scores. The results obtained for the proposed method are compared with other approaches such as keyword-based and neural-based approaches. The proposed method has shown an average precision of 0.883, average recall of 0.865, and average f-score of 0.873. On Comparision, proposed Heuristic based approach showed the improved performance over the other methods.

## References

A. Das and S. Bandyopadhyay. (2010), Topic-based Bengali opinion summarization‖, In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 232–240, 2010.

Ajmal E.B, Posna P Haron, (2015) "Summarization of Malayalam Document Using Relevance of Sentences" International Journal of Latest Research in Engineering and Technology, Volume I Issue 6 pp 08-13.

B, Mohan & B, Aravindh & M, Akhil. (2021). Neural Abstractive Text Summarizer for Telugu Language. Third International Conference on Soft Computing and Signal Processing (ICSCSP 2020).

http://www.ethnologue.com/statistics/size.

K. Sarkar, (2012), An approach to summarizing Bengalinews documents‖. In proceedings of the InternationalConference on Advances in Computing, Communicationsand Informatics, Pp. 857-862, 2012.

Lin, C.Y., (2004). Rouge: A package for automatic evaluation of summaries. Text Summarization.

M. Banu, C. Karthika, P. Sudarmani and T.V. Geetha, (2007), Tamil Document Summarization Using Semantic Graph Method‖, Proceedings of International Conference on Computational Intelligence and Multimedia Applications, Vol. 2, Pp. 128-134, 2007.

M. Gupta and N. K. Garg, (2016), "Text Summarization of Hindi Documents Using Rule Based Approach," 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, 2016, pp. 366-370, doi: 10.1109/ICMETE.2016.104.

Naidu R., Bharti S.K, Babu K,S, Mohapatra R.K, (2018), Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers. A conference on Smart Computing and Informatics. Pp 555-564. https://doi.org/10.1007/978-981-10-5544-7_54.

R. Jayashree, (2012), Categorized Text Document Summarization in the Kannada Language by Sentence Ranking‖, Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA), Pp.776-781, 2012.

Ratul Sikder, Md. Monowar Hossain, F.M. Rahat Hasan Robi, (2019) Automatic Text Summarization For Bengali Language Including Grammatical Analysis, International Journal of Scientific & Technology Research. Volume 8, Issue 06, June 2019.pp-288-292.

S. Vijay, V. Rai, S. Gupta, A. Vijayvargia and D. M. Sharma, (2017),"Extractive text summarisation in hindi," 2017 International Conference on Asian Language Processing (IALP), Singapore, 2017, pp. 318-321, doi: 10.1109/IALP.2017.8300607.

Srinath S, A. S. (2019). A Fused Extractive Summarization Approach for Kannada text Documents. International Journal of Advanced Science and Technology, 28(18), 565 - 580.

T Eslam and S.M.A. Masum ,"Bhasha, (2004),A Corpus Based Information Retrieval and Summarizer for Bengali Text"(2004).

T. Priyadharshan and S. Sumathipala, (2018), "Text Summarization for Tamil Online Sports News Using NLP," 2018 3rd International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2018, pp. 1-5, doi: 10.1109/ICITR.2018.8736154

Virat V. Giri, Dr.M.M. Math and Dr.U.P. Kulkarni ,(2016), "A Survey of Automatic Text Summarization System for Different Regional Language in India"(2016).