International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2380-2384
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2380-2384

# Unsupervised Automatic Text Summarization of Konkani Texts using K-means with Elbow Method

**Jovi D'Silva[1] and Dr. Uzzal Sharma[2]**

[1]*Research Scholar, Computer Science Engineering, Assam Don Bosco University, Assam, India.*

[2]*Assistant Professor, Computer Science Engineering, Assam Don Bosco University, Assam, India.*

*ORCID ID: 0000-0002-7222-7111 (Jovi D Silva)*

**Abstract**

Text Summarization is an emerging field of research in Natural Language Processing (NLP). A bulk of the work is related to texts in English and other popular languages. This paper presents some of the early works attempted at performing single document extractive Automatic Text Summarization on Konkani language documents, which is an under-research language in the domain of Automatic Text Summarization (ATS). The input documents need to be cleaned of punctuation and then sentence scores are calculated for each sentence in the document. The scores for each sentence are computed using Term-Frequency/Inverse Document Frequency (TF-IDF) of constituent words and overlap with the title of the story and its positional value. K-means algorithm is applied to determine clusters of sentences for the formation of the final summary. The value of 'K' is determined using the Elbow method. The dataset employed was specially designed by the authors of the paper to perform the experiments. It consists of folk tales derived from books on Konkani literature. The performance assessment of the output summaries indicated that the summaries obtained by using three clusters were better than the ones obtained using two clusters. The proposed system exhibited promising outcome, considering, no language-dependent domain knowledge or any training corpora was utilized.

**Keywords:** K-means; Unsupervised; Text Summarization; Konkani; Elbow method

## I. INTRODUCTION

Automatic Text summarization has become a widely researched domain in the past couple of years. Data on the internet is expanding at an exponential pace on the internet and this is one of the primary reasons why a number of researchers are interested in exploring the area of Automatic Text Summarization. It has become the desideratum of the hour because, in an expeditious paced world like today people infrequently have the time to read all of the content available on the web [1], [2], [3]. It would be preferable to read a gist of the document and then establish whether it is worth the time to look at the whole article. Automatic Text Summarization avails in providing concise summaries that highlight pivotal points of a document [1], [3]. The summaries engendered by Automatic Text Summarization methods can either be 'Extractive Summaries' or 'Abstractive Summaries' [1], [2]. In Extractive summarization, certain critical phrases and sentences are identified to constitute a summary [4], [6], [7]. Abstractive summary, on the contrary, is the result of comprehension and interpretation of the contents of a source document. Thus, the summary that is generated may or may not contain the sentences or phrases from the original document [4], [6], [7].

Substantial research has been done for automatic text summarization in English language since it is one of the most widely used languages on the internet. However, a plethora of digital content is available in other languages too. An abundance of summarization tools has been developed in assorted languages antecedently. The authors of this paper have made an endeavor at exploring unsupervised techniques for automatic text summarization. The technique presented in this paper has been experimented on a Konkani language dataset which was specially constructed for promoting research pertinent to Konkani language [2]. Konkani language is spoken in some regions along the west coast of India. The dataset is explicit to the dialect spoken in the state of Goa and it is available in Devanagari script [2].

There are various approaches that could be employed to realize the summarization task viz. Graph-based, Statistical and Machine Learning approaches [8]. Machine learning approaches can either learn dynamically or operate with the support of a dataset. Machine learning techniques further bifurcate into supervised and unsupervised techniques [6]. In this paper, an unsupervised machine learning technique has been presented, using K-means clustering algorithm for automatic text summarization using Konkani language dataset.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2380-2384
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2380-2384

## II. RELATED WORK

García-Hernández et. al. presented a domain and language independent text summarization methodology using unsupervised machine learning. The algorithm for unsupervised approach clusters the sentences representing similar ideas, following which the most reflective of the sentences are chosen from each cluster to make up the summary [9]. Khan et. al. demonstrated a K-means clustering approach for generating extractive summaries using TF (Term Frequency)-IDF (Inverse Document Frequency). They have also suggested using the concept of the value of 'true K' that helps in splitting the input document sentences to generate the eventual output summary [10].

Nomoto and Matsumoto put forth a unique perspective called "Information-Centric Approach" to summary evaluation. According to this concept, the summary quality is not resolute by the overlap of terms or phrases between the system-generated summary and the human-generated summary; rather, its quality is determined based on the efficacy of the generated summary in representing the source document in IR related tasks. The algorithm for summarization has been built upon K-means clustering method with an added extension of 'Minimum Description Length Principle' [11].

Agrawal and Gupta presented an algorithm using K-means clustering technique, along with TF-IDF and tokenization, for producing extractive summaries [12]. Shetty and Kallimani illustrated a different technique using K-means clustering method for extractive text summarization. The whole process of summary generation was split in different stages, i.e. first being the pre-processing stage where cleaning of the source document is done, where it is put through lemmatization, stop word removal and tokenization. In the next stage, feature extraction was done and TF-IDF values were computed to form the TF-IDF matrix. Sentences were associated with clusters based on cosine similarity. More clusters imply higher precision of the generated summary. Pivotal sentences were picked from every cluster to constitute a summary in the last stage. Verification of the effectiveness of summary was done using 'recall' and 'precision' measures [13].

Chou et. al. proposed an automatic extractive summarization technique using 'sentence-level' K-means clustering algorithm to divide sentences based on the topic and then determine the relevance of each sentence. ROUGE toolkit was used to establish the quality of the summaries generated [14]. Agarwal et.al. demonstrated a novel technique for generating extractive summaries using 'sentence embeddings' with the assistance of K-means clustering algorithm. The sentence embeddings were clustered as per the summary size requirements using K-means method. Thereafter, 'Ridge Regression' sentence scoring model was used to pick the most relevant sentences from each cluster to formulate a summary [15].

Akter et. al. employed K-means clustering algorithm for extractive summarization on text in Bengali language. The mentioned approach worked with single or multiple Bengali documents. The input documents were passed through lemmatization and stemming phases. Thereafter, for every term the TF-IDF scores were calculated. Finally, K-means clustering technique aided in generating the document summary [16].

## III. PROPOSED METHODOLOGY

### A. Block Diagram

A visual representation of the summary generation process using K-means clustering has been illustrated in Fig. 1.

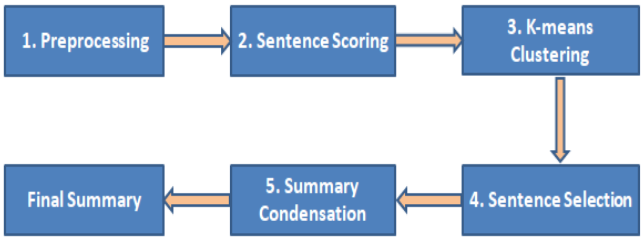


**Fig. 1.** Summary Generation Using K-means Clustering

### B. Algorithm

The steps of the extractive summary generation process using K-means clustering are as follows:

#### 1) Pre-processing

The pre-processing step involves two phases i.e. Sentence segmentation, wherein every sentence has been placed on a new line; and Punctuation removal, wherein all the punctuations from the source text are eliminated.

#### 2) Calculating Sentence Scores

The scores for each of the sentences is calculated using language independent features as described below.

- TF-IDF: Also called "Term frequency", quantifies the recurrence of a word appearing in a text document [17]. TF-IDF is calculated using (1) and (2) [16],

$$TF\text{-}IDF = tfw_i * idfw_i \quad (1)$$

$$idfw_i = log\left(\frac{N}{n_i} + 1\right) \quad (2)$$

Where,

'$tfw_i$' is frequency of occurrence of a word $w_i$ in a sentence

'$idfw_i$' is Inverse document frequency

'$N$' is the sum of the total count of sentences in the text

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 9 (2020), pp. 2380-2384
 https://dx.doi.org/10.37624/IJERT/13.9.2020.2380-2384

*'$n_i$'* is the total count of the sentences in which word *'$w_i$'* occurs

- Sentence Scoring: Sentence score can be calculated using (3) and (4) [16],

$$Sentence\ Score = \sum TF - IDF + PV + TW \quad (3)$$

$$PV(Position\ Value) = \frac{1}{\sqrt{SP}} \quad (4)$$

*'SP'* is sentence position

*'TW'* is Title Word; the value of *'TW'* will be 1 if there is an overlap of words between the sentence under consideration and the tile of the document. If there is no overlap, then the *'TW'* value is taken as 0.

### 3) *K-means Clustering*

The K-means clustering algorithm is then applied to the results obtained in step 2 to get distinct clusters.The time complexity for K-means to execute would be $O(n \times K \times S)$ where, $'n'$ is the total count of the iterations, $'K'$ is the sum total of the number of clusters and $'S'$ is the number sentences in each document. The corresponding space complexity would be $O(S + K)$.

### 4) *Sentence Selection*

The top highest scoring 50% sentences from each cluster are then picked to constitute the summary.

### 5) *Summary Condensation*

A threshold is set to 300 words; if the resultant summary exceeds 300 words then it is condensed to limit its length. The human generated summaries against which the system generated summaries will be checked are made of 300 words.

## IV. DATASET

Konkani is spoken in the state of Goa, India. In the field of text summarization, not much research was previously extended to include Konkani, especially considering the small fraction of population that speaks the language. Therefore, a suitable database for the said experimentation was not previously available. The preliminary task was to spawn a dataset in Konkani [3]. Five books, featuring a sum total of 71 folk tales, written in Konkani were hand-picked to constitute the dataset. The precise technique of designing the dataset is presented in [3].

## V. DETERMINING VALUE OF PARAMETERS FOR K-MEANS

One of the fundamental and important steps in unsupervised learning using K-means is determining an optimal value of 'K'. For the purpose of this analysis, we put forth the adoption of the "Elbow method" to estimate an optimal value of 'K'.

The Elbow methods objective is to find the smallest value of 'K' that still has a low value of inertia. We also note that from this point as the value of 'K' increases we start to have diminishing returns.

As seen in the Fig. 2, the plot formed an elbow. Plots were generated for all 71 folk tales and it was observed that a value of 3 would yield an optimal cluster formation. Experiments were performed for K = 2 as well as K = 3.

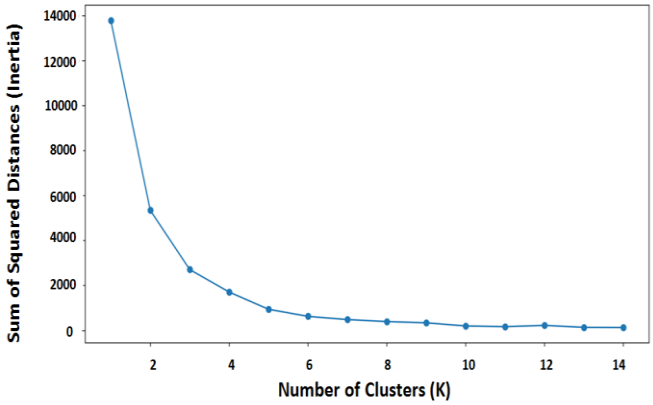


**Fig. 2.** Elbow plot of clusters

The Value of K determines the number of clusters of sentences that are formed, finding an optimal value of K is imperative to produce a summary with sentences that ideally represent the document at hand. Increasing the value of K does not guarantee a good summary formation rather choosing a value of K that takes into account the inertia is imperative.

Once the value of K has been identified we need to find values for centroid initialization. In this experiment we compared manual centroid initialization versus random centroid initialization. In random centroid initialization, 'K' values are randomly selected to be the initial centroid value. To ensure a deterministic outcome we set the random seed to 1. When setting the values manually if K = 2 then the values would be the largest sentence score value (max) and smallest sentence score value (min). If K = 3 then we would additional take the average as the third value ((max + min) / 2).

## VI. EVALUATIONS AND RESULTS

'ROUGE Toolkit' was employed to execute the performance evaluations. ROUGE has been ascertained to work precisely with human assessments and adopts N-gram statistics [18]. The 300 words output summary was then compared with the two reference

summaries generated by the human annotators [2]. ROUGE metric was utilized for the purport of quantifying overlap of uni-grams, bi-grams and 'Longest Common Subsequence' (LCS) between the reference summaries and the automatically created summary.

ROUGE-1 depicts uni-gram scores, ROUGE-2 denotes bi-gram scores and ROUGE-L mirrors LCS scores. The perception behind using the above measures was to verify the eloquence of the summaries generated, with growing levels of granularity shifting through uni-gram unto bi-gram and then to LCS. If the arrangement of words in the resultant summary is akin to the human-summarized reference summary, then, it is an implication of a rather eloquent system generated output. The specifics of the evaluation measures employed are underlined by Lin [18]. The Tables 1, 2 and 3, below outline the records of different cluster configurations acquired through assessment of the corresponding human summaries with the system-generated summaries. We also introduce a new benchmark which consists of a 300-word summary constructed using sentences indicated as critical sentences by subject experts that are to be included in a summary.

Table 1 shows ROUGE-1 (uni-gram) scores. Table 2 gives an account of ROUGE-2 (bi-gram) and Table 3 illustrates ROUGE-L (LCS) points. In the tables displayed below, we have estimated the equivalent ROUGE metrics, where the quantitative interpretation of the overlay between the system-generated summary and the human-generated summaries are provided by 'Precision' and 'Recall'.

'Precision' aims to determine to what extent the contents of a system generated summary are relevant, this is critical as a generated summary may contain unimportant portions of the original text. 'Recall' intends to analyse to what extent the content of the 'gold-standard' human-generated reference summary was adequately captured by the automatically generated summary. The 'F1-Score' takes into consideration the Precision and Recall and thereafter, provides a combined report of the two measures as a unique score. The scores are depicted as percentage of overlap. Each of the systems represents how many clusters were used and how the centroid values were initialized, where Random indicates that the centroid values were randomly initialized.

**Table 1.** ROUGE-1 Uni-gram scores

| System | ROUGE-1 (uni-gram) | | |
|---|---|---|---|
| | ***Recall*** | ***Precision*** | ***F1-Score*** |
| kMeans-2Clusters | 0.31276 | 0.31324 | 0.31298 |
| kMeans-2ClustersRandom | 0.31259 | 0.31295 | 0.31275 |
| kMeans-3Clusters | 0.31408 | 0.31373 | 0.31388 |
| kMeans-3ClustersRandom | 0.31190 | 0.31167 | 0.31175 |
| HumanAnnotated-Benchmark | 0.35844 | 0.35460 | 0.35608 |

**Table 2.** ROUGE-2 Bi-gram scores

| System | ROUGE-2 (bi-gram) | | |
|---|---|---|---|
| | ***Recall*** | ***Precision*** | ***F1-Score*** |
| kMeans-2Clusters | 0.07984 | 0.07997 | 0.07989 |
| kMeans-2ClustersRandom | 0.08030 | 0.08040 | 0.08035 |
| kMeans-3Clusters | 0.07942 | 0.07927 | 0.07934 |
| kMeans-3ClustersRandom | 0.07872 | 0.07856 | 0.07863 |
| HumanAnnotated-Benchmark | 0.11088 | 0.10908 | 0.10977 |

**Table 3.** ROUGE-L LCS scores

| System | ROUGE-L (LCS) | | |
|---|---|---|---|
| | ***Recall*** | ***Precision*** | ***F1-Score*** |
| kMeans-2Clusters | 0.30524 | 0.30572 | 0.30546 |
| kMeans-2ClustersRandom | 0.30474 | 0.30510 | 0.30490 |
| kMeans-3Clusters | 0.30680 | 0.30644 | 0.30659 |
| kMeans-3ClustersRandom | 0.30434 | 0.30412 | 0.30420 |
| HumanAnnotated-Benchmark | 0.35228 | 0.34847 | 0.34994 |

## VII. CONCLUSION

This paper exhibits a distinct illustration of automatic text summarization approach on a Konkani literature dataset using unsupervised machine learning approach. The scores obtained through ROUGE toolkit indicate that the summaries procured by three clusters were found to be better than the ones procured using two clusters. Furthermore, it was noted that manually setting centroid values mostly yielded better outcomes. Although the system was incapable of outperforming the Human Annotated Benchmark, yet the outcome was promising taking into consideration that system does not utilize any language dependent domain knowledge like stop-words, stemming and lemmatization or any other training corpora. In the absence of any training data, our summarization based on K-means has provided promising results.

## REFERENCES

[1] Moratanch N, Chitrakala S. A survey on extractive text summarization. IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP). 10-11 Jan. 2017. doi:10.1109/icccsp.2017.7944061.

[2] D'Silva J, Sharma U. Development of a Konkani Language Dataset for Automatic Text Summarization and its Challenges. International Journal of Engineering Research and Technology. International Research Publication House. ISSN 0974-3154. 2019: 12(10): 1813-18917.

[3] Andhale N, Bewoor LA. An overview of Text Summarization techniques. International Conference on Computing Communication Control and Automation (ICCUBEA), 2016. doi:10.1109/iccubea.2016.7860024

[4] Gupta V. A Survey of Text Summarizers for Indian. Languages and Comparison of their Performance. Journal of Emerging Technologies in Web Intelligence. November 2013: 5(4): 361-366.

[5] Allahyari M, Pouriyeh S, Assef M, Safaei S, Trippe ED, Gutierrez JB, Kochut K. Text Summarization Techniques: A Brief Survey. arXiv:1707.02268, USA, July 2017.

[6] Nimavat K, Joshiara HA. Query-Based Summarization Methods for Conversational Agents: An Overview. International Journal of Advanced Research in Computer Science. ISSN No. 0976-5697, September-October 2017: 8(8).

[7] Verberne S, Krahmer E, Hendrickx I et al. Creating a Reference Data Set for the Summarization of Discussion Forum Threads. Lang Resources & Evaluation, Springer Netherlands. 52: 461, 2018. Available: https://doi.org/10.1007/s10579-017-9389-4.

[8] Al-Taani AT. Automatic Text Summarization Approaches. International Conference on Infocom Technologies and Unmanned Systems, Trends and Future Directions (ICTUS). doi:10.1109/ictus.2017.8285983

[9] García-Hernández RA, Montiel R, Ledeneva Y, Rendón E, Abukh A, Cruz R. Text Summarization by Sentence Extraction Using Unsupervised Learning. Lecture Notes in Computer Science, pp. 133–143. doi:10.1007/978-3-540-88636-5_12.

[10] Khan R, Qian Y, Naeem S. Extractive based Text Summarization Using K-Means and TF-IDF. I.J. Information Engineering and Electronic Business, 2019: 3: 33-44. Published Online May 2019 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2019.03.05.

[11] Nomoto T, Matsumoto Y. A New Approach to Unsupervised Text Summarization. Conference: SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA. 2001: 26-34. 10.1145/383952.383956.

[12] Agrawal A, Gupta U. Extraction Based Approach for Text Summarization Using K-means Clustering. International Journal of Scientific and Research Publications. November 2014: 4 (11). ISSN 2250-3153.

[13] Shetty K, Kallimani JS. Automatic Extractive Text Summarization Using K-means Clustering. International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017. doi:10.1109/iceeccot.2017.8284627

[14] Chou VT, Kent L, Góngora JA, Ballerini S, Hoover CD. Towards Automatic Extractive Text Summarization of A-133 Single Audit Reports with Machine Learning. 2019. arXiv:1911.06197

[15] Agarwal S, Singh NK, Meel P. Single-Document Summarization Using Sentence Embeddings and K-Means Clustering, International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), ISBN: 978-1-5386-4119-4/18/. 2018 IEEE, Greater Noida (UP), India, 12-13 October 2018.

[16] Akter S, Asa AS, Uddin MP, Hossain MD, Roy SK, Afjal MI. An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm. IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Dhaka, Bangladesh, 13-14 Feb. 2017. 10.1109/ICIVPR.2017.7890883.

[17] Ferreira R et al. Assessing Sentence Scoring Techniques for Extractive Text Summarization. ELSIVIER International Journal of Expert systems with Applications, Netherlands, 2013: 40 (14): 5755-5764.

[18] Lin CY. ROUGE: a Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona, Spain. July 2004: 74–81.