Clustering technique for Automatic Kannada Text Summarization

Arpitha Swamy

[1]Lecturer, Department of Computer Science & Engineering, Government Polytechnic, KrishnarajaPete, Mandya, Karnataka, India

Abstract

Text summarization is an application of natural language processing in the field of data mining, used to generate the summary of document. It is a process to reduce the contents in the original text to shorter form which contains important information which is useful for the user in different real-life applications. A lot of techniques have been developed to summarize English text documents but only a small number of methods have been developed for Kannada text because of lack of resources and tools available for Kannada language. This paper discusses the extractive text summarization technique which selects main sentences from the Kannada document. In the proposed approach, Term- Frequency/Inverse Sentence Frequency (TF/ISF) is used to compute the sentence score first and then sentences are grouped by means of clustering algorithm called K-means to produce the extractive summary. The results of the proposed model are evaluated using ROUGE toolkit to measure the performance based on F-score of generated summaries. Experimental studies on custom-built dataset containing 50 Kannada text documents shows significantly better performance in producing extractive summaries as compared to human summaries.

Keywords: Extractive, K-means algorithm, ROUGE, Text summarization, TF-ISF.

1. Introduction

Automatic summarization is a method to compress the contents in the original document into shorter form. The shorter form is called the summary and it should contain the vital information of the original document. Due to the large amount of textual data accessible in the form of online documents, there is a requirement for an automatic tool to produce summary of large documents so that it saves effort of users in finding the information they are interested in. The summary of document is created by pre-processing the document then extracting the features followed by sentence scoring step to find out which sentences are important and crucial in the document.

Text summarization techniques are mainly classified as extractive summarization techniques and abstractive summarization techniques. Extractive techniques generate the summary of the document by identifying important sentences using some word and sentence based statistical and linguistic features. In Abstractive techniques, summary is created by selecting and reordering the words in the original document or by adding some new words that are not exist in the original document through linguistic analysis of the text. Text summarization systems are also categorized as single-document summarization systems and multi-document summarization systems based on the count of documents passed as input. Depending on the purpose, summarization systems are classified as generic, domain specific, or query-based.

There are many summarization systems available to summarize the documents in English language and they are producing summaries with satisfactory accuracy. But in Indian languages, there is no accurate and complete document summarization system to produce the summary. Therefore, developing an automated text summarization system for Indian languages can help readers understanding large documents and provides the essential information about the document content in less time. We proposed a method for extractive text summarization in one of the Indian languages –Kannada. Kannada language is spoken mainly in the Karnataka state of country India. The work presented in this paper uses the clustering technique to create the extractive summary of Kannada text document.

The flow of the paper is ordered as follows: previous research works in this area is discussed in section 2. We discuss the developed text summarization method for Kannada documents using clustering technique in Section 3. Section 4 illustrates the experimental results and discussion. Finally, the work is concluded in Section 5.

2. Related Works

Over the past years, only little research works has been carried out to solve the problem of Kannada text document summarization in Natural Language Processing. This section gives a brief overview of previous research works carried out in the area of text summarization using different methods to produce summaries of Kannada text documents. Kallimani et al. [1] developed a text summarizer for Kannada language called Kansum which uses first line, position, keywords, numerical values and simple combination function as parameters to score the sentences to produce summary. An approach based on the keyword extraction for extractive text summarization of Kannada documents is proposed by Jayashree et al. [2]. To extract the keywords, GSS (Galavotti, Sebastiani, Simi) coefficients and the IDF (Inverse Document Frequency) method along with TF (Term Frequency) are used. Another work by Jayashree et al. [3] presented a system to generate the extractive summaries of Kannada documents by ranking the sentences of document by assigning scores. Kallimani et al. [4] designed a method to generate the summary of news articles based on word-scoring techniques and ontology. To compute the scores, parameters such as first line position, keywords, numerical values and simple combination function are used. Jayashree et al. [5] developed an approach which uses artificial neural network to summarize Kannada text documents. To train the model, they used the feed forward neural network with back propagation. The parameters used to train the model are location of the paragraph, length of the sentence, whether paragraph follows title or not, title word ratio, thematic word ratio for each sentence, location of sentence and first sentence in the paragraph. A federated approach for Kannada Language using Text Rank algorithm and Nave Bayesian method is proposed by Ranganatha et al. [6] to generate the extractive summary of the document. The summaries produced by Text Rank algorithm and Nave Bayesian method are combined and sentences with high scores are selected to produce the summary of document. Jayashree et al. [7] designed hybrid methodologies for summarization of Kannada language text documents. The four different methodologies - (a) Keyword based summarization, (b) Sentence ranking based summarization, (c) Jaccards' similarity score based summarization and (d) Neural network approach based summarization are proposed to summarize text documents. An extractive approach using latent semantic analysis for Kannada text summarization is developed by Geetha and Deepamala [8].

3. Proposed Method

Our work proposed the Kannada document summarizer, an application of Natural Language Processing (NLP) extracts the important information from the text document. There are mainly two techniques in automatic summarization - text extraction and text abstraction. The extraction method produces summary by selecting the important words, phrases or sentences from the input document. An abstraction method creates the summary by adding some new words that are not present in the input document. There are mainly three basic steps to produce the extractive summary: pre-processing, feature extraction and the summary generation [9]. The proposed automatic text summarization model for Kannada documents is illustrated in Fig. 1.

The proposed approach uses the K-means algorithm to generate the summary of a single document. K-means clustering is an unsupervised machine learning algorithm which accepts the number of clusters to be formed in the clustering process as input. It is the popular cluster analysis method in the data mining field. The data points are clustered into a k number of clusters which are mutually exclusive [10][11][12][13].

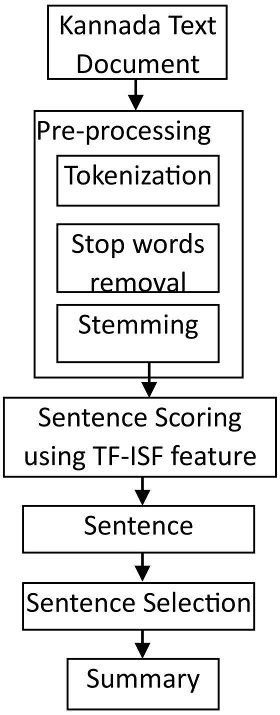

Fig. 1: Text summarization process in the proposed system

The method consists of four stages to create the summary of document – preprocessing, sentence scoring, applying clustering algorithm and sentence selection. The input document is preprocessed and score for each word is computed using the TF-ISF. Then, score of every sentence is calculated by the total sum of the scores of all words present in the sentence. Next, the scores of sentences are arranged in descending order and value of highest scored sentence is designed as centroid-1and the value of lowest scored sentence is considered as centroid-2 to execute the K-means clustering algorithm. Then from each cluster, top K sentences are picked up and the summary is created.

3.1 Pre-processing

In the process of Kannada document summarization, some pre-processing operations are carried out before the sentence scoring algorithm is executed. The pre-processing function prepares the document for ranking of sentences and the generation of summary.

The pre-processing operations performed on the documents are:

1) Tokenization – A document is the arrangement of sentences and every sentence consists of group of words. Each word is treated as a token. Tokenization splits the document into sentences and then sentences into individual words.

2) Stop words removal – The commonly occurring words are called stopwords, which have less importance in the conclusion of document, are discarded for summarization process. In Kannada words like ಮತ್ತು (and), ಅದು (it), ಇದೆ (is), etc. are frequently used stopwords in sentences of the document.

3) Stemming – The document may contain many words with same root but in different forms. All such words are converted to their root form for ease. Transformation of words into their canonical forms is done by the stemming algorithm. For example, the words ಆನೆಗೆ ಆನೆಯ ಆನೆಗಳ ಆನೆಯನ್ನು ಆನೆಗಳಿಗೆ should be converted into their original form ಆನೆ. All inflected words are converted into their root form through stemming operation using a predefined suffix list in this work.

3.2 Scoring Process

The sentences of the document are represented in vector form with *n* dimension, where *n* is the total number of words present in the document vocabulary after removing the stop words and stemming process. Each word is linked with a coefficient which determines the weight of each word present in the document. The combined weight of all words present in the sentence represents the sentence weight. A matrix of size *m* x *n* is used to represent the document where *m* indicates the count of sentences contained in the document and *n* defines the count of words present in the vocabulary of a document after removing the stop words.

$$M = \left\{ \begin{matrix} a_{11} & a_{12} & \ldots\ldots & a_{1n} \\ a_{21} & a_{22} & \ldots\ldots & a_{2n} \\ & & . & \\ & & . & \\ a_{m1} & a_{m2} & \ldots\ldots & a_{mn} \end{matrix} \right\}$$

Term Frequency-Inverse Sentence Frequency (TF-ISF)

Every element of the above sparse matrix is assigned with a value. The TF-ISF rules are used to determine the weights of every word. Term frequency (TF) is defined as the ratio of total occurrences of the term in the entire document to the entire count of terms present in the document as given in (1).

$$TF(t) = N_t / TW \qquad (1)$$

Where,

N_t: number of times a term t present in the document

TW: total count of terms in the document

The inverse sentence frequency *(*ISF) is a measure of the importance of the word based on its unusual occurrence in the document and is calculated as in (2).

$$ISF(t,d) = log(N/ n_{t)} \qquad (2)$$

Where,

n_t: total count of sentences containing the word w

N: total count of sentences present in the document d.

The sentence is relatively more important if the sentence contains more number of unique words, which is measured by *TF*-ISF value. The *TF-ISF* value of a term is computed as the product of both tf and isf values of a word in a sentence, as shown in (3).

$$TF\text{-}ISF(t) = TF(t)*ISF(t,d) \qquad (3)$$

The sentences scores are obtained by adding up the scores of TF-ISF values of each word present in that sentence. Then the scores of sentences are arranged in descending order to make the data available for clustering algorithm required to group the sentences.

3.3 Applying K-means clustering algorithm

After the sentence scoring and sorting process, value of highest scored sentence is designed as centroid-1and the value of lowest scored sentence is designed as centroid-2 for the K-means clustering algorithm. Then from each centroid, the distance to every sentence is calculated. The closest distance from one centroid specifies that cluster. As a result, two clusters are formed and the values of centroids are updated for next iteration. The new values for centroids are assigned by calculating the average value of each cluster respectively. This clustering process is repeated until the two successive iterations ends with the same result.

3.4 Sentence Selection

The sentences in each cluster are ranked based on their scores and top *K* sentences are chosen from every cluster to create the extractive summary. The sentences selected for the summary are reordered to retain the order same as in the original document.

4. Results & Discussion

The dataset is created by collecting 50 articles belonging to different categories from Kannada Webdunia website and articles are saved as text documents using UTF-8 format. The five categories chosen are Astrology, Business, Cricket, Politics and Sandalwood.

The proposed system is evaluated against the text documents belonging to five different categories selected from the dataset. Only one human summary for each document is considered for evaluation. The summary generated by the system is evaluated by comparing it to the human summary using the ROUGE toolkit [14]. There are different ROUGE measures - ROUGE1, ROUGE2, ROUGEL, and ROUGES etc. We have used ROUGE1 measure to evaluate the system generated summaries. The values for the three evaluation metrics: average precision, average recall and average f-score are generated by the ROUGE evaluation toolkit and is used to evaluate the summary.

Precision can be defined as the ratio of count of common sentences present in both system and model summaries over the total count of sentences present in the system summary. Recall is defined as the ratio of number of common sentences present in both system and model summaries and the total count of sentences present in the model summary. F-score is defined as a composite measure that combines recall and precision. It is calculated as the harmonic average of recall and precision.

Fig. 2 shows the comparative chart for average recall, precision and f-score values obtained by the proposed summarization method in five different categories.

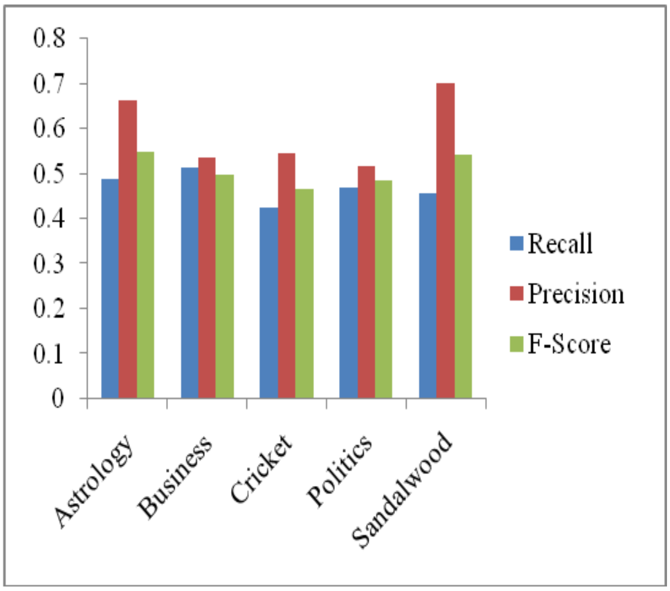

Fig. 2: Average values of recall, precision and f-score for each category of documents

5. Conclusion

An extraction-based Kannada text summarization for a single document using the clustering technique is discussed in this paper. K-means algorithm is used to form the clusters of sentences based on the similarities. The extractive summary of document is created by selecting top ranked sentences from each cluster. The generated summaries are evaluated using ROUGE toolkit with recall, precision and f-score evaluation measures. The performance of this proposed system is acceptable in terms of average recall, average precision and average f-score values. In future, this work can be extended to multi-document summarization to generate the summary of multiple Kannada documents related to same topic.

References

1. Kallimani, J.S. and Srinivasa, K.G., 2010, August. Information retrieval by text summarization for an Indian regional language. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)* (pp. 1-4). IEEE.
2. Jayashree, R., Murthy, S.K. and Sunny, K., 2011. Keyword extraction based summarization of categorized Kannada text documents. *International Journal on Soft Computing*, *2*(4), p.81.
3. Jayashree, R., Murthy, S. and Anami, B.S., 2012, November. Categorized Text Document Summarization in the Kannada Language by Sentence Ranking. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)* (pp. 776-781). IEEE.
4. Kallimani, J.S., Srinivasa, K.G. and Reddy, B.E., 2012. Summarizing news paper articles: experiments with ontology-based, customized, extractive text summary and word scoring. *Cybernetics and Information Technologies*, *12*(2), pp.34-50.
5. Jayashree, R., Murthy, K.S. and Anami, B.S., 2013, December. An artificial neural network approach to text document summarization in the Kannada language. In *13th International Conference on Hybrid Intelligent Systems (HIS 2013)* (pp. 45-48). IEEE.
6. Ranganatha, S., Vinay, S.K. and Bhargava, H.S., 2014. Federated Document Summarization Using Probabilistic Approach for Kannada Language. *International Journal of Innovative Research & Development*, *3*(1), pp.228-233.
7. Jayashree, R., Murthy, K.S. and Anami, B.S., 2014. Hybrid methodologies for summarisation of Kannada language text documents. *International Journal of Knowledge Engineering and Data Mining*, *3*(1), pp.82-114.
8. Geetha, J.K. and Deepamala, N., 2015, August. Kannada text summarization using Latent Semantic Analysis. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1508-1512). IEEE.
9. Rautray, R., Balabantaray, R.C. and Bhardwaj, A., 2015. Document summarization using sentence features. *International Journal of Information Retrieval Research (IJIRR)*, *5*(1), pp.36-47.
10. Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society.Series C (Applied Statistics)*, *28*(1), pp.100-108.
11. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), pp.881-892.
12. Gupta, V. and Lehal, G.S., 2010.A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, *2*(3), pp.258-268.

13. Shetty, K. and Kallimani, J.S., 2017, December. Automatic extractive text summarization using K-means clustering. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)* (pp. 1-9).IEEE.
14. Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.