ISSN No:-2456 –2165

# Bi-Lingual Translation and Multi-Document Text Summarization using Clustering Approach

Dolly Arun Bonde, Gurvinderpalkaur.P. Dhindsa, Damini Rangnath Landge, Prerna Didwania,
Department of Computer Engineering,
K.K.Wagh Institute of Engineering Education and Research, K.K.W.I.E.E.R,
Nashik-422003, India.

***Abstract:-*Agriculture is the back-bone of economy in India. In order to enhance development in this field knowing only basics of agriculture is not enough. Researcher or framer should be aware of agriculture practices throughout the world. As English is global language most of the information is available in English, so when one browses agriculture information, it becomes difficult for some farmers to get the correct meaning of the information and quite difficult to go through each and every information. This result in language barrier and information overload that either leads to wastage of significant time browsing all information or else useful information is missed out. Hence text summarization in native language in agriculture field is very essential for user to get concise information about new technology. The proposed methodology comprises of machine translation, data pre-processing and automatic text summarization. The machine translation phase translates documents which are either in Hindi or Marathi language to the English document. After that data pre-processing takes place. Data pre-processing step involves noise removal, tokenization, stop word removal and stemming. On the pre-processed data automatic text summarization is performed using clustering approach. Then according to the user choice the summary will be translated to Hindi or Marathi language.**

*Keywords:- Data mining, text summarization, extractive summarization, K-means clustering, Machine translation*

## I. INTRODUCTION

In recent years there is tremendous increase in the data generated by various sources. It is difficult for the user to comb through such huge amount of data to find the relevant information. To find relevant information user needs to search entire documents. If the document is irrelevant then user wastes a lot of time in reading that document. Hence current age of information overload demands text summarization system. Text summarization aims to provide abridged version of the document which gives a proper gist to reader or user. In past several years multi-document text summarization has been area of interest which extracts important sentences from multiple documents and formats them into summary. Such summarization provides more information and also saves time of the user. Globalization is reducing the significance of national borders in terms of trade and information exchange. Hence information on web is also available in regional languages like Hindi, Marathi etc. In order to achieve frequent exchange of information machine translation helps to avoid cultural, language and technological barriers. There are many methods used for automatic text summarization and machine translation.

There are generally two main approaches for automatic text summarizations are abstractive and extractive summarization.[1] Abstractive Summarization is performed by building a semantic representation of original document and then applying some natural language processing techniques to generate the summary. It provides more generalized summary but it is difficult to compute. It works in the way humans used to generate the summary. Extractive Summarization is performed by extracting some selected words or phrases or sentences as it is from the original document to generate the summary. Extractive summarization is easier than abstractive summarization. The general methods used for extractive summarization are TF/IDF method, cluster based method, graph theoretic approach, LSA (Latent Semantic Analysis) method, text summarization with neural networks, automatic text summarization based on fuzzy logic, machine learning approach, query based extractive text summarization, text summarization using regression for estimating feature weights, multilingual extractive text summarization, topic-driven summarization, Maximal Marginal Relevance and centroid-based summarization, etc..

General approaches used in machine translation are Direct Translation, Interlingua –Based Translation, Transfer –Based Translation, Statistical –Based Approach, Example-Based Translation and Hybrid Translation.[2][3]

## II. RELATED WORK

The work in the area of automatic text summarization and machine translation has been going on for last few decades. This section presents a concise note on the existing work carried out in automatic text summarization and machine translation.

- *Summarization*

Automatic summarization is process of extracting the useful content of the document and creating the summary of the document.

In 2005 J. Zhang presented an extractive multi-document text summarization using Cue-based hub authority. It is graph based approach and uses K-nearest neighbor (KNN) for sentence clustering to detect sub-topics.

In 2009 Zhang Pei-ying and LI Cun-he proposed an approach for automatic text summarization based on sentences clustering and extraction. In this method sentence clusters are first formed based on the semantic distances among sentences and then accumulative sentence similarity on each cluster is calculated. And at last the topic sentences are chosen by using some extraction rules. In the same year K. Sarkar presented summarization of multi-documents based on sentence clustering approach using similarity histogram to identify multiple subtopics from input documents (related documents) and selects the representative sentences from cluster to form summary.[4]

In 2010 A. Kogilvani presented multi-document summarization using clustering and feature specific sentence extraction.[5]

In 2012 V. K. Gupta proposed extractive text summarization system based on query approach. In this summaries of single text documents are combined using sentence clustering method to generate multi-document summary.[6]

A. Agrawal and U. Gupta proposed clustering based extractive text summarization system in 2014. It uses k-means clustering algorithm to generate summary of single document in English.[7]

In 2015 Malliros and Skians used graph based approach in which node centrality is used indicate importance of word in the document. Global and local centralities are considered for term scoring to generate summary.

In 2016, an approach for summarization of multiple documents was performed on the basis of the query provided by the user [8]. According to the query provided by the user relevant documents are extracted from the document corpora. A similarity matrix of n x n is generated from all the extracted documents. Then Markov clustering algorithm is used to form the clusters of similar sentences. The sentences in each cluster are sorted according to the weight and top ranked sentences are extracted. The drawback of this approach is that if user does not provide a proper query then the summary generated may not be meaningful.

In 2016, Linjing Lang has developed a method based on fuzzy ontology extraction technique. This method only focused on providing multi-document summarization for only Gabonese documents [9].Fuzzy logic is a way of reasoning similar to human reasoning based on the degree of truth. In this approach, with the help of fuzzy rules and membership function, a fuzzy system is designed. The values from zero to one are assigned to each sentence based on the rules and functions defined in fuzzy system. The top ranked sentences are then used to generate the summary

In 2017 , a system was proposed that could generate an efficient summary for single or multiple Bengali documents by using K-means clustering[1].This approach produces effective summary in case of multiple documents in comparatively less time than other approaches.

- *Machine Translation*

Machine translation is the field of natural language processing which translates one natural language to another natural language using computer system without any human interference.

In 1995 ,Anusaaraka project started at IIT Kanpur for translating Telugu, Kannada, Bengali, Punjabi and Marathi to the target language Hindi using Direct Machine Translation Approach. Initially the system was tested on children's' stories..[3]

In 1995 ,Anubharti is developed using a hybridized example-based machine translation approach
In 1999, Hemant Darbari and Mahendra Kumar Pandey developed a tool MANTRA using Transfer based approach. It translates English text into Hindi. The system is developed for the RajyaSabha Secretariat, the Upper House of Parliament of India. Other systems such as Anubaad, Vaasaanubaa ,Anglabharti , Mat, Shakti were developed .[3]

Goyal V and Lehal G S developed a Hindi to Punjabi system that uses direct word to word translation approach
In 2009, A consortium of 11 institutions in India have developed a multipart machine translation system (Sampark System) for Indian Language to India Language Machine Translation (ILMT) funded by TDIL program of Department of Electronics and Information Technology (DeitY), Govt. of India.[3]

In 2011, Sanjay Chatterji, Praveen Sonare, Sudeshna Sarkar, and AnupamBasu proposed a method for making proper lexical translation in Bengali to Hindi Machine Translation Framework by using a transfer based MT approach.[3]

In 2015 , Pankaj Kumar, Sheetal Srivastava and Monica Joshi presented the Syntax Directed Translator for translation of English language to Hindi Language. It takes text as input in English ,then store that input in file. After that it extract words and punctuations and then store it in array. Finally it find their meanings in context to the sentence and convert it to the target language i.e. Hindi.[10]

## III. PROPOSED METHOD

The document summarization is an application of natural language processing and data mining which deals with extraction of relevant information from the lengthy document. The proposed system works on Hindi, Marathi and English documents and provides summary preferred by user. The input to the system is a single or multiple word documents. The system uses Clustering approach based on sentence scores to perform summarization.[1] The proposed framework has three phases namely translation, data pre-processing, summarization phase. The translation phase [10] includes translating the given document into English only if the document is in Hindi or Marathi. If the document is in English then no translation is performed. In data pre-processing phase stop words and noise is removed from the document/s and tokenization is

performed. In summarization phase word scoring and sentence scoring is done based on TF-IDF and then clustering is applied to get the proper summary of the document.

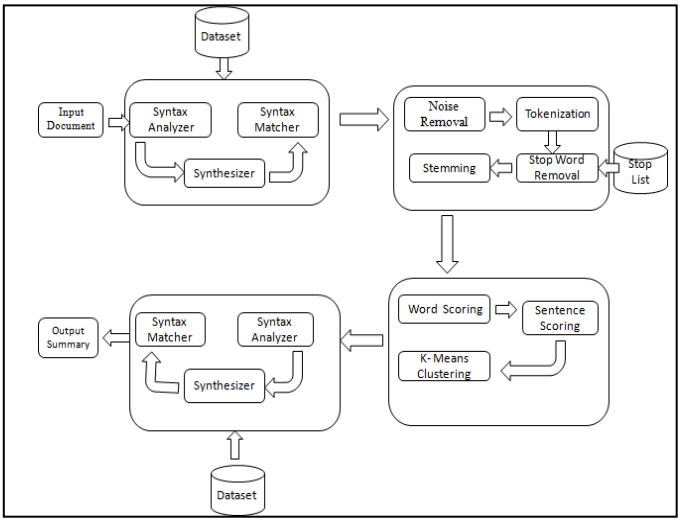


Fig 1:-Proposed System Architecture

*A. Translation Phase*

This phase is further divided into following sub phases:
a)Syntax Analyzer
Syntax analyzer is used for token, word and sentence formation. Syntax analyzer is also used to check if the formation of the words in the sentences is correct or not. Firstly syntax analyzer extracts the sentences then stream of words is generated from these extracted sentences. According to the similarities these words are grouped together. Later syntax of these grouped words is checked. Number or alphanumeric words need not be considered for processing.[10]

*B. Synthesizer*

Synthesizer is mainly used for annotation. The grouped words with the correct syntax are sent to synthesizer to find their actual translation. So to improve the efficiency of the synthesizer phase noun compound translator (NCT) is used. NCT selects correct sense of noun compound by finding all possible translation in corpus and then performs ranking and selects the appropriate translation for corresponding word.[10]

*C. Syntax Matcher*

Syntax matcher is mainly used for matching words with dataset. It places the translated words in the sentence as their position. Firstly it will scan the sentence and replace words with its corresponding translation.[10]

- *Data Pre-Processing Phase*

*A. Noise Removal*

All the documents contain header and footer which is not necessary for summary. Hence they are removed in this phase.[1]

*B. Tokenization*
Each and every word in the document is separated to form tokens. For example: crops, irrigation, pesticides etc.[1]

*C. Stop word removal*

Documents consists of function words like a, the, that, are etc. These stop words are not so relevant while generating summary. So they are removed in this phase.[1]

*D. Stemming*

Words can be present in different forms in the same document. Hence it becomes necessary to convert them into their original form. For example: happiness, happiest, happier all these words can be converted to single word happy.[1]

- *Summarization Phase*

*A. Word Scoring*

Score is calculated for each word in the sentence using TF-IDF. It is considered that more the unique words in the sentence more are the importance of the sentence.[1]

*B. Sentence Scoring*

Sentence score is calculated by the total sum of the words. If the sentence contains any Cue word or skeleton word, then the score is increased by 1. The sentences and their corresponding scores are stored according to decreasing order in a separate file for further processing.[1]

*C. K-means clustering*

From stored file the highest score is considered as the centroid 1and the lowest score as the centroid 2 to apply the K-means clustering algorithm. Then, top K sentences are extracted from each cluster and the final summary is generated. Here, K sentences can be measured as 30% of sentences of the original merged document.[1]

## IV. CONCLUSION

In this paper, translating multiple documents and an extractive-based text summarization has been proposed. This summarization technique uses only relevant sentences from original documents based on their scores. Measuring similarities between sentences based on the sentence score is crucial part. So an improved similarity measuring technique may generate a better summary in future. The relevance among sentences can be measured using semantic and syntactic similarities in future. Proposed system accepts only word documents in three languages namely Hindi, English, Marathi only. And the final summary generated is also in any one of these three languages. In India there are 22 languages

so in future more generalized system can be made to provide summary in many languages.

## REFERENCES

[1] Sumya Akter, Aysa Siddika Asa, Md. Palash Uddin,Md. Delowar Hossain,Shikhor Kumer Roy and Masud Ibn Afjal, "An extractive text summarization technique for bengali document(s) using K-means clustering algorithm," Hajee Mohammad Danesh Science and Technology University(HSTU),IEEE-2017.

[2] Sindhu D.V, Dr. B.M.Sagar and Mr RajaShekar Murthy S,"Survey on Machine Translation and Its Approaches",Department of ISE,R V College of Engineering,Karnataka,India,IJARCCE,June-2014.

[3] G V Garje and G K Kharate, "Survey of Machine Translation Systems In India," Department of Computer Engineering and Information Technology, PVG's College of Engineering and Technology,Pune,India,IJNLC-2013.

[4] Zhang Pei-ying and LI Cun-he,"Automatic text summarization based on sentence clustering and extraction",College of Computer and Communication Engineering, China University of Petroleum,IEEE-2009

[5] A. Kogilvani, "Multi-document summarization using clustering and feature specific sentence extraction.",2010

[6] V. K. Gupta, "Extractive text summarization system based on query approach." ,2012

[7] A. Agrawal and U. Gupta,"Clustering based extractive text summarization system" , 2014

[8] Deepak Sahoo and Rakesh Balabantaray, "Aspect Based Multi document Summarization," Department of Computer Science and Engineering,IIIT, Bhubaneswar,ICCCA-2016

[9] Evrard Stency Larys TSOUMOU , Shichong YANG,Linjing LAI and MBEMBO LOUNDOU VARUS,"An Extractive Multi-Document Summarization Technique Based on Fuzzy Logic Approach," ICNISC-2016

[10] Pankaj Kumar,Sheetal Shrivastava and Monica Joshi,"Syntax Directed Translator for English to Hindi Language," ICRCICN-2015