MULTI-DOCUMENT BI-LINGUAL NEWS SUMMARIZATION

Sai Teja Alubelli

Department of cse, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

Dr. K. S. Deepthi

(Assistant Professor)

Department of cse, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

Janni Ramdhar Thanmayi Sharon

Department of cse, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

ABSTRACT

Summarization of text is one of most challenging task in the field of natural language processing. The need to be summarized from the large datasets which contains the knowledge discovery such as information extraction and retrieval. When the information is extracted it consists of specific domain knowledge data. So in this paper the data from English and Telugu newspapers is extracted, now compare the missing news and then summarize the bi-language data where implementation of sentence extraction and weightage strategy to mine the data form multiple documents is done.

***Keywords*— newspaper, document extraction, term frequency, inverse document frequency**

I. INTRODUCTION

Information is of two types informative and non-informative. Which is mostly relevant in query processing for single and multiple documents and it is monolingual and multilingual based on languages used. Most of the summarization tools work on selecting the portion of input document such are sentence, words and paragraphs but getting the relevance information from the large volume of the data is the most important part. So that we can identify the information and can get the deeper content from it. There is a lot of difference between the data extraction from the single and multi-document. Multi-document requires the high redunency and then extraction of the data from two different language newspaper and summarizing them is more complicated. So it has to give better knowledge discovery there are no true bi-language summation system implemented yet Intrinsic Measures which finds the similarity of the document with one or more models is used for summary of tasks such as document retrieval and text classification. These approaches are mostly applied in news and story data extraction which are mostly in high level structures. The text summarization consist of two task, summarizing the single document and understanding the pattern matching such as characters, numbers and identifying their rate of character addresses to topics and keywords.

Information retrieval is the most popular technique for extraction of relevant information. So it is more complicated to get the relevant information on query based processing. When the user enters the query it should return the appropriate data. So we need a proper approach, one of the best approach is clustering technique to get the information form the potential clusters and find the most relevant data from clusters by implementation of co-efficient correlation approach. So that the user can get the relevant pages and then summaries the data.

TYPES OF SUMMARIZATION [2]

The Summary of the document is reduced and made precise, representation of the text which seeks to render the exact idea of its contents. Its principal objective is to give information and provide privileged access to the source documents. Summarization is automatic when it is generated by software or by an algorithm. The main types of automatic summarization include extraction-based, abstraction-based and maximum entropy-based.

EXTRACTION-BASED SUMMARIZATION

Extraction consists of selecting units of text (sentences, segments of sentences, paragraphs or passages), taken to contain a document's essential information, and assembling these units in an adequate manner. According to Radev et al. , algorithms for automatic summarization by extraction can be divided into three types: surface-level, intermediate-level and deep parsing techniques.

i) SURFACE-LEVEL ALGORITHMS:

Surface-level algorithms do not go through the linguistic depths of a text; rather they use certain linguistic elements to identify the most relevant segments of a document. Which are in practice from the very first studies on summarization, surface-level techniques use the occurrences of words to weight sentences.

ii) INTERMEDIATE-LEVEL ALGORITHMS:

Intermediate-level algorithms use linguistic information that is more sophisticated than the surface-level algorithms, but less sophisticated than deep parsing. One intermediate-level technique is lexical chain recognition. Lexical chains follow lexical semantic relations in which words are connected in sequences.

iii) DEEP PARSING ALGORITHMS:

Deep parsing approaches are based on the idea that is necessary to use in depth linguistic techniques. Which exploit the discursive structure of texts. Some of these approaches are based on rhetorical structure theory (RST), which aims to finely exploit the structure of the discourse to generate abstracts, or on meaning-text theory (MTT).

ABSTRACTIVE SUMMARIZATION

Extraction techniques merely copy the information taken to be most significant by the system to the summary (for example, key clauses, sentences or paragraphs), while abstraction involves paraphrasing sections of the source document. In general, abstraction can reduces a text more efficiently than extraction, but the programs that can do this are harder to develop as they require the use of natural language generation technology, which itself is a booming field. Systems that produce summaries by abstraction are based on text understanding and seek to generate a grammatically correct, concise and coherent text. Very few abstract summarization systems have been created. FRUMP (Fast Reading Understanding and Memory Program) is one such system and was the first to use semantic interpretation for English texts to produce their summaries.

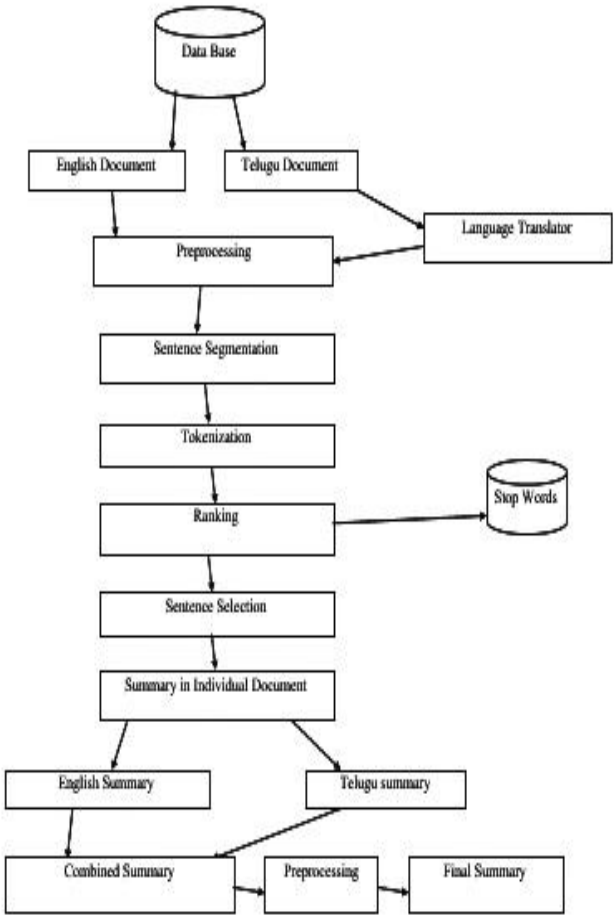
II RELATED WORK:

The volume of electronic data obtainable on web is increasing day by day. Alok Ranjan Pal et al. [1] suggested as a result, dealing with such huge volume of data is creating a big problem in different real life data handling applications. A.R.Kulkarni et al., [2] proposed that Natural language processing automates the translation process between

computers and humans. A.Kogilavani et al., [3] presented an approach to cluster multiple documents using cluster approach. BenHachey et al., [4] presented a novel representation based on genetic relation extraction. It's main motto is to build systems for relation identification and characterization. Bhavesh Pandya et al.,[5] proposed novel approach to infer user search goals by analyzing the whole search engine query logs. DasariAmarendra et al., [6] proposed approach that uses k-means clustering with meaningful words and relationship using TF-IDF, giving more information related to document. Fang Chen et al., [7] presented the extractive techniques for text summarization. H. Chen et a.,[8] presented the user interface that organizes web search results into hierarchical categories. H.-J Zeng et al.,[9] organized web search results into clusters and facilitated users quick browsing through search results. H. Cao et al., [10] provided the rank algorithm and the correlated algorithm. Harshada P. Bhambure et al.,[11] proposed the concept of pseudo document which, at the end clusters the pseudo documents to infer the user search goals which presents them the keywords. I. Mele et al., [12] presented graph based approach that uses the user web browsing log to maximum extent. P. Sudhakar et al., [13] proposed a novel approach using a weighted technique to mine the web contents catering to the user needs. R. Baeza-Yates et al., [14] proposed the method based on query clustering process in which similar queries are identified and the process uses the content of historical preferences of users. X. Wang et al., [15] proposed clustering search results which is an effective way of organizing search results, it allows the user to navigate into the relevant documents quickly.

III PROPOSED METHOD

Automatic text summarization and implementation using bi-language , needs high concise documents. So that they can express the relevant information with proper meanings. Document summary is from one or more documents. So the methodology is to find the prevalent keyword and extract the plain text which consists of images and different symbols we need to clean the data and then extract only the text data and form the subset of each document summary.



We implement the process using python scripts for processing of text and also we scrap the data from English and Telugu news articles. The data is preprocessed by using word and sentence tokenization, removing the irrelevant data and then we implement the text rank approach using term frequency and inverse document frequency approach. Then construct the step of news summarization. Implementing the measures like word overload,TF_IDF statics and consider the article titles for calculating the similarity scores for better accuracy. The obtained set of scores for each sentences by taking the length of summary and take the sentence which has the maximum rank and contains the significant information. Translate the Telgu article to English and again implement the bi-language summarization. Now combine both

relevant and non-matched news together form English and Telugu article and then summarize the data again to get the relevant information.

Module Description

Preprocessing

Data pre-processing is very crucial step in the data mining process. The process of data gathering methods are often loosely controlled. Analyzing the data that has not been carefully screened for such problems can produce misleading results. So quality of the data is most important factor for implementation of analysis.

Stemming Words

Stemming is the term used for information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. In morphological operations we need to relate the words with mapping the root stem with the related words to make it a valid root.

Data Cleaning

In this module the noisy data and the inconsistent data are removed so it has classify into two parts as correct and incorrect data which is implemented using the data preprocessing methods. Then prepare the data for further analysis which is stored in database systems.

TF-IDF weighting

This approach is used to eliminate the common words form the document , then take the stop word list and remove the repellant words the TF*IDF weights process. The data from the multiple document's and then the meaningful words are taken into the consideration. The only additional information besides the term frequency c(w) that we need in order to evaluate the weight of a word w which appears c(w) times in the input for summarization is the number of documents, d(w), a background corpus of D documents that contains the word. This allows us to evaluate the inverse document frequency.

TF*IDF=C (w) *log (D/d (w))

TF:

Term Frequency, which measures how repeatedly a term occurs in a document. Since every document varies in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is divided often by the document length as a way of normalization:

TF (t) = (Number of times term t occurs in a document) / (Total number of terms appear in the document).

IDF:

Inverse Document Frequency, which measures how vital a term is. While calculating TF, all terms are considered with same important. However it is known that certain terms, such as "is", "of", and "that", may occur many times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the IDF for t, the number of terms is given below:

IDF (t) = log e(Total number of documents / Number of documents which are having term t in it).

IV RESULTS

Indian cricketer Mahendra Singh Dhoni adds another record to his credit in the current series with Newzealand. This match becomes the 300 T20 match he is playing and he is the first Indian to achieve this feat. MSDhoni played a total of 96 matches for India, 175 matches in Indian Premier League, 24 in the Champions T20 League, 4 for Jharkhand, 1 for the first class. Standing first in the list is Windies all-rounder Pollard with 446 matches and others before captain cool include Chris Gayle, Dwayne Bravo, Shoaib Malik. Rohit Sharma with 298 matches and Suresh Raina with 296 matches are standing next to Dhoni. The former Indian team captain is performing well with his mark batting and wicket keeping the Newzealand series. Talking about Dhoni in an interview another star player Yuvraj Singh said, "I feel his presence is very important in the decision-making. He has had a fantastic tournament in Australia, and it is good to see him hit the ball the way he used to do. I wish him all the best," he said. Dhoni is regarded as one of the most successful captains in Indian cricket, winning the 50-over World Cup in 2011 and the inaugural World Twenty20 in 2007.

Fig1:- English News article Extracted

An English news article from The Hindu newspaper is extracted and shown in Fig1.

న్యూజిలాండ్‌తో జరుగుతున్న సిరీస్ నిర్ణయాత్మక మ్యాచ్‌లో టీమిండియా వెటరన్ క్రికెటర్ మహేంద్ర సింగ్ ధోని అరుదైన ఘనతను సొంతం చేసుకున్నాడు. ఈ మ్యాచ్ ధోనికి ఓవరాల్‌గా 300వ టీ20 కాగా. ఈ ఘనతను అందుకున్న తొలి భారత క్రికెటర్‌గా ఈ మిస్టర్ కూల్ నిలిచాడు. తద్వారా 300 అంతర్జాతీయ టీ20లు ఆడిన ఆటగాళ్ల జాబితాలో ధోని చేరాడు.ఈ జాబితాలో విండీస్ ఆల్‌రౌండర్ కీరన్ పొలార్డ్ 446 మ్యాచ్‌లతో అగ్రస్థానంలో ఉండగా. క్రిస్‌గేల్, డ్వేన్ బ్రేవో, షోయబ్ మాలిక్‌లు ధోని కన్నా ముందున్నారు. ఇక ఐపీఎల్‌లో చెన్నై సూపర్ కింగ్స్, రైజింగ్ పుణె జట్ల తరపున ధోని ఆడిన విషయం తెలిసిందే. భారత్ తరపున రోహిత్ శర్మ 298, సురేశ్ రైనా 296 మ్యాచ్‌లతో ధోని తర్వాతి స్థానంలో ఉన్నారు. ధోని ఈ ఫీట్ అందుకున్న సందర్భంగా సోషల్‌మీడియా వేదికగా ప్రశంసల జల్లు కురుస్తోంది. ఇక ఆస్ట్రేలియా పర్యటనలో బ్యాట్‌తో రప్పాడించిన ధోని. కివీస్ పర్యటనలో కూడా తన మార్క్ కీపింగ్, బ్యాటింగ్‌తో ఆకట్టుకుంటున్నాడు.

Fig2:- Telugu News article Extracted

A Telugu news article from Sakshi newspaper is extracted and shown in Fig2.

న్యూజిలాండ్‌తో జరుగుతున్న సిరీస్ నిర్ణయాత్మక మ్యాచ్‌లో టీమిండియా వెటరన్ క్రికెటర్ మహేంద్ర సింగ్ ధోని అరుదైన ఘనతను సొంతం చేసుకున్నాడు. ఈ మ్యాచ్ ధోనికి ఓవరాల్‌గా 300వ టీ20 కాగా. ఈ ఘనతను అందుకున్న తొలి భారత క్రికెటర్‌గా ఈ మిస్టర్ కూల్ నిలిచాడు. తద్వారా 300 అంతర్జాతీయ టీ20లు ఆడిన ఆటగాళ్ల జాబితాలో ధోని చేరాడు.ఈ జాబితాలో విండీస్ ఆల్‌రౌండర్ కీరన్ పొలార్డ్ 446 మ్యాచ్‌లతో అగ్రస్థానంలో ఉండగా. క్రిస్‌గేల్, డ్వేన్ బ్రేవో, షోయబ్ మాలిక్‌లు ధోని కన్నా ముందున్నారు. ఇక ఐపీఎల్‌లో చెన్నై సూపర్ కింగ్స్, రైజింగ్ పుణె జట్ల తరపున ధోని ఆడిన విషయం తెలిసిందే. భారత్ తరపున రోహిత్ శర్మ 298, సురేశ్ రైనా 296 మ్యాచ్‌లతో ధోని తర్వాతి స్థానంలో ఉన్నారు. ధోని ఈ ఫీట్ అందుకున్న సందర్భంగా సోషల్‌మీడియా వేదికగా ప్రశంసల జల్లు కురుస్తోంది. ఇక ఆస్ట్రేలియా పర్యటనలో బ్యాట్‌తో రప్పాడించిన ధోని. కివీస్ పర్యటనలో కూడా తన మార్క్ కీపింగ్, బ్యాటింగ్‌తో ఆకట్టుకుంటున్నాడు.
In the decisive match against New Zealand, India's veteran cricketer Mahendra Singh Dhoni has gained a rare distinction. The match is over 300 per T20. Mr. Cool is the first Indian cricketer to receive this feat. Dhoni has joined the list of 300 international T20s. Windies all-rounder Kieran Pollard topped the list with 446 matches. Chrisgale, Dwayne Bravo and Shoaib Malik are ahead of Dhoni. Dhoni has played for Chennai Super Kings and Rising Pune teams in the IPL. India's Rohit Sharma's 298 and Suresh Raina are at the top of the table with 296 matches. Dhoni has been praising her for this feat. Dhoni, who was bowled by a bat in the Australia tour. Kevis also impresses with his mark keeping and batting on tour

Fig3:- Translated Telugu Text to English

The extracted Telugu news article is translated into English using translator tool, it is observed in Fig3.

first summary

MSDhoni played a total of 96 matches for India, 175 matches in Indian Premier League, 24 in the Champions T20 League, 4 for Jharkhand, 1 for the first class. Standing first in the list is Windies all-rounder Pollard with 446 matches and others before captain cool include Chris Gayle, Dwayne Bravo, Shoaib Malik. Rohit Sharma with 298 matches and Suresh Raina with 296 matches are standing next to Dhoni.

second summary

In the decisive match against New Zealand, India's veteran cricketer Mahendra Singh Dhoni has gained a rare distinction. Kevis also impresses with his mark keeping and batting on tour. Cool is the first Indian cricketer to receive this feat. Windies all-rounder Kieran Pollard topped the list with 446 matches.

final summary

Rohit Sharma with 298 matches and Suresh Raina with 296 matches are standing next to Dhoni.In the decisive match against New Zealand, India's veteran cricketer Mahendra Singh Dhoni has gained a rare distinction. MSDhoni played a total of 96 matches for India, 175 matches in Indian Premier League, 24 in the Champions T20 League, 4 for Jharkhand, 1 for the first class. Standing first in the list is Windies all-rounder Pollard with 446 matches and others before captain cool include Chris Gayle, Dwayne Bravo, Shoaib Malik.

0.151162790698

0.757142857143

0.325581395349

Fig4:- Bi_language **Summary with Relevance Score Factor**

The English text is summarized first (first summary) and then the Telugu text is translated into English. The translated text is then summarized (second summary) now both the summaries are combined and then summarization technique is applied which results in the final summary (Fig4).

V. CONCLUSION

This paper presents a machine learning approach using NLP for mining the newspaper information by using bi-language and summaries the meaning full information. So that the end user can get the information for the news articles which are been missed by the other papers and in this paper our approach gives the better accuracy since there are very less work done on this platform.

VI FUTURE ENHANCEMENTS

In future this can be implemented by using un-supervised learning approach with multi language summarization with large corpus which can give better accuracy level.

REFERENCES

1. Alok Ranjan Pal, Projjwal Kumar Maiti and Dig anta Saha, "An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And WordNet", International Computer Modelling (IJCTCM)Journal of Control Theory and Vol.3, No.4/5, September 2013
2. A.R.Kulkarni, S.S.Apte, an automatic text summarization using lexical cohesion and correlation of sentences, International Journal of Research in Engineering and Technology
3. A.Kogilavani and Dr.P.Balasubramani, "Clustering and Feature Specific Sentence Extraction Based Summarization of Multiple Documents", International Journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010.
4. BenHachey, "Multi-Document Summarization Using Generic Relation Extraction", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 420-429, 2009.
5. Bhavesh Pandya et al., "A New Algorithm for Inferring User Search Goals with Feedback Sessions", Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 5, Issue 8, (Part - 2), pp.30-33, August 2015.
6. DasariAmarendra, KavetiKiran Kumar, "Inferring User Search Goals with Feedback Sessions using K-means clustering algorithm", Volume 2, Issue 11,pp. 780-784,November-2015.
7. Fang Chen, Kesong Han and Guilin Chen, "An Approach to Sentence Selection Based Text Summarization", In the Proceedings of IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Volume: l,pp.489-493,2002.
8. H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search
9. H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search
10. H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875883, 2008.
11. Harshada P. Bhambure, Mandar Mokashi, "Inferring User Search Goals Using Feedback Session" Conf. Research paper, pp.2319-7064 and 2013.
12. I. Mele, " Web Usage Mining for Enhancing Search –Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '13), pp. 765-769, 2013
13. P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, "Content Based Ranking for Search Engines," Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.
14. R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, and 2004.

Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

15. X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.