Text Summarization for Multi Documents using Machine Learning

Sandip Patil[1] Anuja Patil[2] Jinesh Gajjar[3] Suverna Satkar[4]
[1,2,3,4]G.H.Raisoni College of Engineering and Management Wagholi, Pune, India

Abstract— Automatic text summarization is one of the important challenges of natural language tasks. It will help the readers save time to get the important information from a lengthy document automatically. Automatic Text Summarization techniques aim to extract the fundamental information in documents. Multi-document summarization is useful when a user deals with a group of heterogeneous documents and wants to compile the important information present in the collection, or there is a group of homogeneous documents, taken out from a large corpus as a result of a query Summarization reduces the complexity of a document while retaining its important features The aim of multi-document summarization is to produce an abridged version which contains important information from a set of documents on the same topic. Multi-document summarization has gained popularity in many real world applications because significant information can be obtained within a short time.

Keywords: Artificial Intelligence, Sequence-To-Sequence, Automatic Text Summarization, Long Short-Term Memory, Recurrent Neural Network, Machine Learning, Deep Learning

I. INTRODUCTION

People widely use internet to find information through efficient information retrieval (IR) tools such as Google, Yahoo, AltaVista, and so on. However the abstraction of information from the results of the IR becomes necessary for user to find out really information. So the huge number of information returned by IR system need to be summarized. Text with the sharply growth of the information on the internet, summarization has become an important and timely tool for helping and interpreting text information in today's fast-growing information age.

This Software Requirements Specification provides a complete description of all the functions and constraints of the "An Adaptive Method for Text Summarization for Multiple Documents". The document describes the issues related to the system and what actions are to be performed by the development team in order to come up with a better solution.

The basic idea of idea of text summarization comes from the fact that due to high increase of the digital data, in depth study of the same always takes more time than of estimation. So to decrease this time of analysis for this kind of data there is an always a need of summarizer system.

So proposed system put forwards an idea of text summarization using feature extraction by applying strong NLP protocols and then these features are classified by using fuzzy logic to get the best document summary.

A. K-Mean

It is clustering algorithm in the field of text summarization. this algorithm is also called as fintering algorithm this algorithm is easy to implement, requiring a kd-tree as the only measure data structure

B. ANN

ANN stands for artificial neural networks. Ann are one of the commonly applied machine learning algorithm. Ann are the parts of computing system. Design to simulate the way human brain analyses and process information.

C. Random forest

Radom forests are and ensemble learing method for Classification regression and other task the first algorithm for random forest was created by" Tin Kam HO" using random subspace method

II. LITERATURE SURVEY

Rachel TszWai, Lo, Ben He, Iadh Ounis [1] proposed a new method for automatically generating a stopword list for a given collection. The approach, called term-based random sampling approach, is based on how informative a given term is. They investigated the effectiveness and and the robustness of this new approach using various standard collections. The new approach was compared to four variant baselines approaches inspired by Zipf's law. The results show that the proposed novel approach achieves a comparable performance to the baseline approaches, while having a lower computational overhead. In addition, using the proposed approach, the optimal threshold setting is easier to obtain. Moreover, the experimental results demonstrate that a more effective stopword list could be derived by merging Fox's classical stopword list with the stopword list produced by either the baselines or the proposed approach.

Murphy Choy [2] proposed a new method for automatically generating a stop word list for a given collection of tweets. The approach is based on the combinatorial nature of the words in speeches. They investigated the effectiveness and robustness of the approach by testing it against 9 collections of tweets from different periods. The approach is also compared with the existing approaches using TD*IDF and variants. The results indicated that the new approach is comparable to existing approaches if not better in certain cases.

A. Alajmi, E. M. Saad, and R. R. Darwish [3] introduces statistical approach is presented to extract Arabic stop- words list. The extracted list was compared to a general list. The comparison yield an improvement in an ANN based classifier using the generated stop-words list over the general list. Relationships can be regarded as finding maximum matchings in bipartite graphs. A comprehensive study of the semantic roles of the words AND and OR within labels/names was also carried out. Extensive experiments were performed to quantitatively assess the performance of their solutions.

Arun R., Saradha R., V. Suresh and M. Narasimha Murty [5] applied LDA on stopword streams and demonstrated its efficacy in author and author gender identification over a database of novels that span a wide time-line and a good mix of genres. Their results indicate that stopword in conjunction with LDA are robust features for

stylometric purposes. Since their approach uses stopwords as features, it places minimal demands on the number of words required for authorship attribution. The abstract topic distributions over stopwords assigned by the LDA mechanism seems to be as meaningful as the intuitive and semantically relatable topic distribution over content words. As seen from the table 1 the data used spans across genres and has a good mix of gender; their approach performs well across this mix. Identifying features that are sensitive towards genre and the time-period of the documents would be among our topics of interests for future studies.

Hassan Saif, Miriam Fernandez and Harith Alani[6] proposed a semantic approach to automatically identify and remove stopwords from Twitter data. Unlike most existing approaches, which rely on outdated and context- insensitive stopword lists, their proposed approach considers the contextual semantics and sentiment of words in order to measure their discrimination power.

Evaluation results on 6 Twitter datasets show that, removing our semantically identified stop words from tweets, increases the binary sentiment classification performance over the classic pre-complied stopword list by 0.42% and 0.94% in accuracy and F-measure respectively. Also, our approach reduces the sentiment classifier's feature space by 48.34% and the dataset Sparsity by 1.17%, on average, compared to the classic method.

Hassan Saif,Miriam Fernandez, Yulan He, Harith Alani [7] studied how six different stopword removal methods affect the sentiment polarity classification on Twitter. Their observations indicated that, despite its popular use in Twitter sentiment analysis, the use of pre-compiled (classic) stoplist has a negative impact on the classification performance. They also observed that, although the MI stopword generation method obtains the best classification performance, it has a low impact on both the size of the feature space and the dataset sparsity degree.

S.Santhana Megala, Dr.A.Kavitha and Dr. A.Marimuthu [8] introduced an improvised stemming algorithm to produce a clear and meaningful stem, which is based on the famous Porters stemming algorithm. A Slight modification is done without compromising the efficiency and simplicity of porter‟s algorithm. The experimental result shows that the improvised algorithm shows a better accuracy in generating a meaning full stems comparing to standard porters algorithm, which reduces the error rate (Over stemming and Under stemming) to a bare minimum one. The improvised stemming algorithm is further applied to the research work on summarization and classification of textual data in future to utilize the efficiency and simplicity.

Ms. Anjali Ganesh Jivani [9] proposed purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. They can say that the goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. In this paper they have discussed different methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The basic difference between stemming and lemmatization is also discussed.

David A. Hull [10] presented a detailed analysis of stemming algorithms. The goal will not be to present any dramatic new approaches to the problem, rather it will be to demonstrate a comprehensive and rigorous strategy for analyzing information retrieval experiments.

III. CONCLUSION

By analyzing all the past work in section 2 this research article come to conclusion that still many things in text summarization So, as a solution to this ones coming additions mainly focuses on To enforce the summarization of the multi-documents of different extensions based on the natural language processing protocols which is catalyzed by fuzzy logic powered with the normal distribution factors.

IV. FUTURE SCOPE

The scope of this project includes project developer assisted by project guide. The scope thus far has been the completion of the basic interfaces that will be used to build the system. The text of the multi documents are giving as the input to the system in desired. doc., pdf or in .txt format. The constraints felt thus far by the developer have only been our weekly story cards, the end-to-end side of the interface, and time to time brushing on methodology of implementation which schedule the completion of the project in March 2019.

The major scope of this project is as follows

- Preprocessing of the text to get rid of redundant data
- Extraction of the best features from the text

Providing meaningful summary for the documents

REFERENCES

[1] Rachel TszWai Lo, Ben He, Iadh Ounis, "Automatically Building a Stopword List for an Information Retrieval System", 5th DutchBelgium Information Retrieval Workshop (DIR) '05 2005.
[2] Murphy Choy, "Effective Listings of Function Stop words for Twitter",. 2012.
[3] Alajmi, E. M. Saad and R. R. Darwish "Toward an ARABIC Stop-Words List Generation", International Journal of Computer Applications (0975 – 8887, Volume 46– No.8, May 2012.
[4] Eduard Dragut, Fang Fang, Prasad, Sistla, Clement Yu and Weiyi Meng, "Stop Word and Related Problems in Web Interface Integration," VLDB '09, August 24-28, 2009, Lyon, France, 2009.
[5] Arun R., Saradha R., V.Suresh, M. Narasimha Murty and C. E. Veni Madhavan" Stopwords and Stylometry: A Latent Dirichlet Allocation Approach." Department of Computer Science and Automation, Indian Institute of Science, Bangalore.
[6] Hassan Saif, Miriam Fernandez and Harith Alani," Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter", Knowledge Media Institute, The Open University, United Kingdom.
[7] Hassan Saif, Miriam Fernandez, Yulan He, Harith Alani," On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter ",Knowledge Media Institute, The Open University.

[8] S.Santhana Megala, Dr.A.Kavitha, and Dr. A.Marimuthu", Improvised Stemming Algorithm - TWIG", Volume 3, Issue 7, July 2013, 2013.

