Hybrid Document Summarization using NLP

[1]Chandramma R, [2]Naveen P Pandurangi, [2]Srivatsa V Jamadagni, [2]Nikhil Chandran, [2]Mohammed Abu Talha Ahmed

[1] Associate Professor of Computer Science and Engineering Vivekananda Institute OF Technology Bangalore, India
[2]Students of Computer Science and Engineering Vivekananda Institute OF Technology Bangalore, India

***Abstract*— Hybrid Document Summarization is the technique by which the huge parts of content are retrieved. The Hybrid Document Summarization plays out the summarization task by unsupervised learning system. The significance of a sentence in info content is assessed by the assistance of 3 algorithms. As an online semantic lexicon WordNet is utilized. Word Sense Disambiguation (WSD) is a critical and testing system in the territory of characteristic dialect handling (NLP). A specific word may have distinctive significance in various setting. So, the principle task of word sense disambiguation is to decide the right feeling of a word utilized as a part of a specific setting. To begin with, Document Summarization assesses the weights, keyword and parts of speech of the considerable number of sentences of a content independently utilizing the algorithms and orchestrates them in diminishing request as indicated by their weights. Next, as indicated by the given level of rundown, a specific number of sentences are chosen from that requested rundown.**

Keywords- Document Summarization; Natural Language Processing; Word Net ; NLTK.

I. INTRODUCTION

Hybrid Document Summarization, is the plan to get an important data from a huge amount of information. The amount of data accessible on internet is increasing every day so it turns space and time expanding matter to deal with such huge amount of information. So, managing that large amount of data is makes a major problem in different and real data taking care of uses. The Hybrid Document Summarization undertaking makes the users simpler for various Natural Language applications, like, Data Recovery, Question Answering or content decreasing etc. Document Summarization assumes an inescapable part by creating significant and particular data from a lot of information. Filtering from heaps of reports can be troublesome and tedious. Without a summary or rundown, it can take minutes just to make sense of what the people will discuss in a paper or report. So, the Document Summarization that concentrates a sentence from a content record, figures out which are the most imperative, and returns them in a readable and organized way. Document Summarization is a piece of the field natural language processing, which is the manner by which the PCs can break down, and get importance from human dialect.

Document Summarization that uses the classifier structure and its rundown modules to look over huge amount of reports and returns the sentences that are helpful for producing a summary. Programmed outline of content works by taking the overlapping sentences and synonymous or sense from wordnet most overlapping sentences are considered as high score words. The higher recurrence words are considering most worth. And the top most worth words and are taking from the content and sorted according to its recurrence and generate a summary.

II. PROPOSED SYSTEM

In the Hybrid Document Summarization, we are using a solitary or single input content is going to outlined by the given rate of summarization utilizing unsupervised learning. In any case, the streamlined lesk's computation is associated with each of the sentences to find the guarantees of each sentence. After that, sentences with induced weights are composed in sliding solicitation concerning their weights. Presently as per a particular rate of summarization at a specific occurrence, certain quantities of sentences are chosen as an outline. The proposed computations, abridges solitary or single report content utilizing unsupervised learning approach. Here, the heaviness of every sentence in a substance is resolved using streamlined Lesk's computation and wordnet. After that, summarization procedure is performed as indicated by the given rate of synopsis. In which, we are taking solitary info content and display summarization as yield. First info content is passed, to the lesk computation and wordnet, where the weights of each sentences of the content are inferred utilizing and semantic investigation of the concentrates are performed. Next, weight doled out sentences is passed to derive the final summary according to the percentage of synopsis, where the last abridged outcome is assessed as and showed.

Where, input document will be in the form of a word document file or a pdf file. The pre-processing includes the data cleaning and data abstraction. The data will be the input to the lesk algorithm with the weights given to the words. Wordnet acts as a dictionary for comparing the importance of the word given is the input. Once the data is processed in the lesk algorithm it gives the output values which will be further converted into the summarized document format.

1. System Architectures of Proposed System

The proposed system depicts the three stages for Automatic Text Summarization and they are listed below.

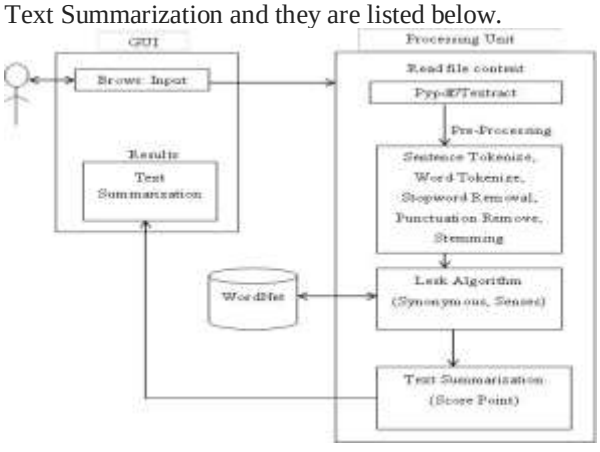

Figure 1: System Architecture for Automatic Text Summarization Using Common Handling Dialect.

Stage 1: Data Pre-Processing
Programmed record outline generator is for clearing the undesirable things which exist in the substance. Henceforth it will additionally process it will be performing sentence part, tokenization, empty stop word, clear accentuation and perform stemming.

Stage 2: Evaluation of weights
This stage processes the repeat of the sentences of a substance utilizing lesk count and wordnet. In the first place finding the total number of spreads between a particular and the radiance this philosophy is performed for the all n number of sentences. By then once-over a particular sentence of the substance is set up for each of the sentences. A sentence is snatched from the once-over. Stop words are removing from the sentence as they don't take an intrigue particularly in sense task method. Sparkles of each vital word removed using wordnet. Union is performed between the sparkles and the data content itself. Once-over of all the intersection guide comes to fruition talks toward the largeness of the sentence.

Stage 3: Summarization
This stage evaluates the last outline of a substance and the introductions the yield, which is surveyed at the period of arranging the sentences. In the first place it selects the once-over of weight named sentences are planned in jumping demand concerning their weights. Pined for number of sentences is picked by the rate of summary. Picked sentences are re-composed by their genuine gathering in the information content. The modified substance summary will gather a substance without depending upon the association of the substance, rather than the semantic information lying in the sentence. Modified substance once-over is without vernacular. To remove the semantic information from a sentence, only a semantic word reference in the last vernacular is required.

III. WORKING OF OUR MODEL

There are basic three phases of document summarization as follows:

1) **Pre-processing**

2) **Processing**

3) **Summary Generation**

Pre-processing

Pre-processing is defined here as cleaning the data. For cleaning unwanted characters, symbols, extra spaces, hyperlinks etc. are removed. The stop words like ‘a’, ‘the’, ’and’. Stemming is done where the words are reduced to their word root for example ‘playing’ would be reduced to ‘play’. Moreover the parsing of the above data is done where the words are bifurcated into nouns, adjectives, verbs etc. The pre-processing is done as per the requirement using one or combination of the above defined techniques.

Processing

This is the next phase of document summarization. After the Text data is cleaned and pre-processed, the processing techniques are applied in which the scores, frequency of words are calculated. The data is grouped on the basis of similarity, dissimilarity so that the summary generation can be made efficiently. For this different clustering techniques are used.

Summary Generation

After the data is processed, the summary is to be generated based on the requirement. Extraction of sentences is done from the processed data and the summary is processed. The words to added to sentences, reducing the sentences based on score etc. is done in this step to produce summary. The representation of the summary can be in the form of words, sentences, paragraphs, graphs etc. for the summary to be generated different summarizers are used.

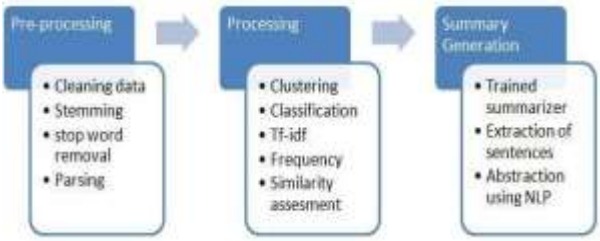

Figure 2: Model Processing

IV. ALGORITHMS

1. Automatic summarization based on user query

Text summarization systems usually provide the user with a generic summary that highlights the most salient information in a text. A Question-Answering (QA) system, however, tries to find an exact answer to the user’s query and generate a suitable response to the query. In our system we are using three different techniques to generate a query-based summarization and present the best summary, according to a modified version of the evaluation measure to the user.

We are using a query modification technique to add extra information to the query. The algorithm of this approach is as follows:

1. Generate a Document Graph (DG) for each sentence in the input documents,
2. Generate a DG for the query (topic),
3. Measure the similarity between each sentence and the query (topic),
4. Search for and add the best sentence to the summary,
5. If the summary’s length restriction is met or there are no more sentences to add then finish and report the target summary; otherwise add the DG for the chosen sentence to the query graph,
6. Repeat from step 3 until no more sentences can be added to the summary.

In this approach, every time we add a sentence to the target summary, we extend the query graph by adding the sentence's DG to it. We called this new summarizer the Q Inc-summarizer. The "Inc" stands for increment, since in this approach we increment the query graph every time we add a sentence to the summary.

2. Automatic summarization based on weighting

This calculation compresses multiple report content utilizing unsupervised learning approach. In This approach, the heaviness of each sentence in a content is determined utilizing Improved Lesk calculation and WordNet. The summarization procedure is performed as indicated by the given level of summarization

Info: Multiple-report input content.
Yield: Summarized content.
Step 1: The list of distinct sentences of the content is prepared.
Step 2: Repeat steps 3 to 7 for each of the sentences.
Step 3: A sentence is gotten from the list.
Step 4: Stop words are expelled from the sentence as they don't take an interest straightforwardly in sense assessment system.
Step 5: Glosses (dictionary definitions) of all the important words are extricated utilizing the WordNet.
Step 6: Intersection is performed between the sparkles and the information content itself.
Step 7: Summation of all the crossing point comes about speaks to the heaviness of the sentence.
Step 8: Weight appointed sentences are arranged in descending request concerning their weights.
Step 9: Desired number of sentences are chosen by the level of summarization.
Step 10: Selected sentences are re-orchestrated by their real sequence in the info content.
Step 11: Stop.

3. Information Extraction algorithm

1. Introduce a method to extract the merited keyphrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the keyphrases.
2. Gather text documents with positively labeled keyphrases. The keyphrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled keyphrases.
3. Train a binary machine learning classifier to make the text summarization. Some of the features you can use include:
    - Length of the keyphrase
    - Frequency of the keyphrase
    - The most recurring word in the keyphrase
    - Number of characters in the keyphrase
4. Finally, in the test phrase, create all the keyphrase words and sentences and carry out classification for them.

V. RESULTS

1. Summarization using query-based algorithm

Input:
Please enter a query: friend
Output:
Processed query: [u'friend']
[("She was ransacking the stores for Jim's present.", 1.324313323278029), ('Jim stepped inside the door, as immovable as a setter at the scent of quail.', 1.3260330166909116), ('"Give it to me quick" said Della.', 1.337516806722689), ('Jim had not yet seen his beautiful present.', 1.3436049382716049), ('"You needn\'t look for it," said Della.', 1.3593439153439153), ('And then Della leaped up like a little singed cat and cried, "Oh, oh!"', 1.3629012660542072), ('Only $1.87 to buy a present for Jim.', 1.4183846153846156), ('It surely had been made for Jim and no one else.', 1.444810744810745), ('"Jim, darling," she cried, "don\'t look at me that way.', 1.5813651999874911), ("One was Jim's gold watch that had been his father's and his grandfather's.", 1.6168950552674828)]

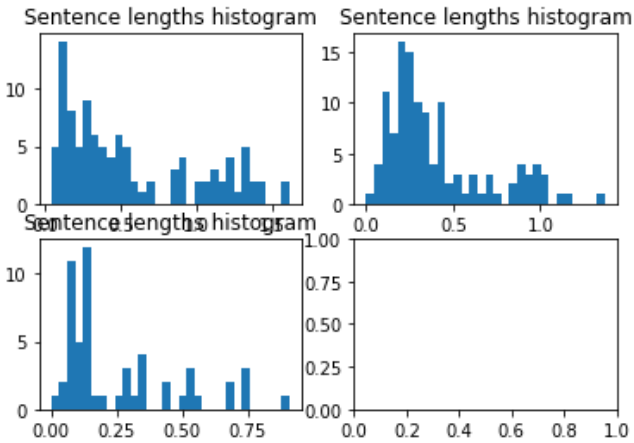

Figure 3: Sentence length histogram for query-based output

2. Automatic summarization based on weighting

Output:
[('Jim drew a package from his overcoat pocket and threw it upon the table.', 16.0), ('"You needn\'t look for it," said Della.', 16.0), ('The door opened and Jim stepped in and closed it.', 16.0), ('It surely had been made for Jim and no one else.', 17.0), ('Jim had not yet seen his beautiful present.', 17.0), ('And then Della leaped up like a little singed cat and cried, "Oh, oh!"', 17.0), ('Only $1.87 to buy a present for Jim.', 17.0), ('Jim looked about the room curiously.', 17.0), ('"Jim, darling," she cried, "don\'t look at me that way.', 18.0), ("One was Jim's gold watch that had been his father's and his grandfather's.", 18.0)]

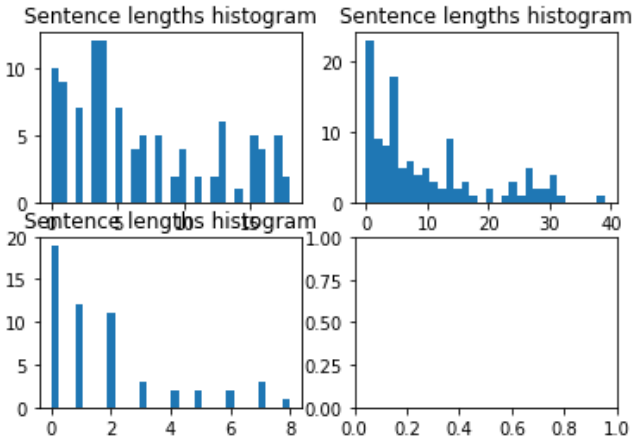

Figure 4: Sentence length histogram for Weight-based output

3. Summarization Information Extraction

Output: Separation between Subject verb Object

(NX ONE/CD DOLLAR/NN AND/CC EIGHTY-SEVEN/NNP CENTS/NNP NX) ./.
(NX THAT/WDT NX) (VX WAS/VBD VX) ALL/PDT ./.
(NX AND/CC SIXTY/CD CENTS/NNP NX) of/IN (NX it/PRP NX) (VX was/VBD VX) in/IN (NX pennies/NNS NX) ./.
(NX Pennies/NNP NX) (VX saved/VBD VX) (NX one/CD and/CC two/CD NX) at/IN (NX a/DT time/NN NX) by/IN (VX bulldozing/VBG VX) (NX the/DT grocer/NN NX) and/CC (NX the/DT vegetable/NN man/NN NX) and/CC (NX the/DT butcher/NN NX) until/IN (NX one/CD 's/POS cheeks/NNS NX) (VX burned/VBN VX) with/IN (NX the/DT silent/JJ imputation/NN NX) of/IN (NX parsimony/NN NX) (NX that/WDT such/JJ close/NN NX) (VX dealing/VBG implied/VBN VX) ./.
(NX Three/CD times/NNS Della/NNP NX) (VX counted/VBD VX) (NX it/PRP NX) ./.
(NX One/CD dollar/NN and/CC eighty-seven/JJ cents/NNS NX) ./.
And/CC (NX the/DT next/JJ day/NN NX) (VX would/MD be/VB VX) (NX Christmas/NNP NX)

Output: Summery

THAT" BE. It was in penny saved one and two at time. Bulldozed grocer. One 's cheek burnt with silent imputation of parsimony. That such close" implied. Three-time Dellum counted it. Next day was Christmas.

Input: Article 1st sentence	Model-written headline
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

Figure 5: Model testing on other articles

VI. Conclusion

We have worked on three algorithms and by comparing those algorithms, query-based algorithm gives the least accurate summary. The weight-based algorithm gives the summery according to the score of the sentence but it is not accurate in the formation of the sentences. Sometimes the sentences might not a right meaning. The information extraction algorithm uses parts of speech tagging which gives the summary having proper meaning for the sentences. Hence, we like to conclude that information extraction algorithm is more accurate than query and weight-based algorithms.

Acknowledgment

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible and under whose constant guidance and encouragement the task was completed. We would like to thank Visvesvaraya Technological University, Belagavi, for having this dissertation as a part of Curriculum, which has given me this wonderful opportunity. We express our sincere regards and thank the project guide Mrs. Chandramma R, Associate Professor and HOD, Department of Computer Science and Engineering for her valuable guidance, keen interest and encouragement at various stages of our project.

References

[1] Alexander M., Rush Sumit Chopra, Jason Weston- A Neural Attention Model for Abstractive Sentence Summarization arXiv:1509.00685v2 [cs.CL] 3 Sep 2015

[2] Çaglar˘Gulçehre˙ Bing Xiang, Ramesh Nallapati, Bowen Zhou, Cicerodos Santos - Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, arXiv:1602.06023v5 [cs.CL]

[3] Santosh Kumar Bharti, Korra Sathya Babu, Sanjay Kumar Jena - Automatic *Keyword* Extraction for Text Summarization: A Survey, National Institute of Technology, Rourkela, Odisha 769008 India e-mail@nitrkl.ac.in 08-February-2017

[4] Abigail See, Peter J. Liu, Christopher D. Manning - Get To The Point: Summarization with Pointer-Generator NetworksarXiv:1704.04368v2 [cs.CL] 25 Apr 2017

[5] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut - Text Summarization Techniques: A Brief Survey, arXiv:1707.02268v3 [cs.CL] 28 Jul 2017

[6] Shashi Narayan, Shay B. Cohen,Mirella Lapata-Ranking Sentences for Extractive Summarization with Reinforcement Learning arXiv:1802.08636v2 [cs.CL] 16 Apr 2018

[7] Qingyu Zhou[y], Nan Yang[z], Furu Wei[z], Shaohan Huang[z], Ming Zhou[z], Tiejun Zhao[y] [y]Harbin Institute of Technology, Neural Document Summarization by JointlyLearning to Score and Select SentencesarXiv:1807.02305v1 [cs.CL] 6 Jul 2018

[8] Feny Mehta - Machine Learning Techniques for Document Summarization: A Survey, 2016 IJEDR | Volume 4, Issue 2 |

[9] Ziqiang Cao˙ Furu Wei˙Li Dong˙Sujian Li˙Ming Zhou-Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence 2153

[10] Alexander M. Rush Sumit Chopra Jason Weston A Neural Attention Model for Sentence Summarization, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 379–389, Lisbon, Portugal, 17-21 September 2015. c 2015 Association for Computational Linguistics

[11] Deepali K. Gaikwad[1] and C. Namrata Mahender A Review Paper on Text Summarization, International Journal of Advanced Research in Computer and Communication Engineering

[12] elima Bhatia, Arunima jaiswal – Automatic text summarization and its methods- A review , 978-1-4673-8203/16/$31.00 IEEE 2016

[13] Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Sowmya Kamath S, A Novel Technique for Efficient Text Document Summarization as a Service 2013 Third International Conference on Advances in Computing and Communications, 978-0-7695-5033-6/13 $26.00 © 2013 IEEE,

[14] Pratibha Devihosur1, Naseer R - Automatic Text Summarization Using Natural Language Processing, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056,

[15] Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah- A Review on Automatic Text Summarization Approaches, 2016 Yogan Jaya Kumar, Ong Sing Goh, Halizah Basiron, Ngo Hea Choon and Puspalata C Suppiah. This open access article is distributed under a Creative Commons Attribution (CC-BY) 3.0 license.

[16] Nenkova, Ani, and Kathleen McKeown. Automatic summarization. Now Publishers Inc, 2011.

[17] Mani, Inderjeet, and Mark T. Maybury. Advances in automatic text summarization. the MIT Press, 1999.

[18] Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. "Multi-document summarization by sentence extraction." In Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4, pp. 40-48.

[19] Lal, Partha. "Text Summarization." (2002).

[20] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management (1988)

[21] Kumar Nagwani, Naresh, and Shrish Verma. "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm." International Journal of Computer Applications (2011)

[22] Yang, Guangbing, Wen, Nian-Shing, and Sutinen. "Personalized Text Content Summarizer for Mobile Learning: An Automatic Text Summarization System with Relevance Based Language Model." InTechnology for Education (T4E), 2012 IEEE Fourth International Conference on, pp. 90-97. IEEE, 2012.

[23] Aksoy, Bugdayci, Gur, Uysal, and Can. "Semantic argument frequency-based multi-document summarization." In Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on, pp. 460-464. IEEE, 2009.

[24] Shams, Rushdi, M. M. A. Hashem, Suraiya Rumana Akter, and Monika Gope. "Corpus-based web document summarization using statistical and linguistic approach." In Computer and Communication Engineering (ICCCE), 2010 International Conference on, IEEE, 2010.

[25] Foong, Oi-Mean, and Alan Oxley. "A hybrid PSO model in Extractive Text Summarizer." In Computers & Informatics (ISCI), 2011 IEEE Symposium on, pp. 130-134. IEEE, 2011.

[26] Resnik, Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." arXiv preprint arXiv:1105.5444 (2011)