# Multi-Document Abstractive Summarization Based on Ontology

Harsha Dave
M.E Student
Computer Department
St.Francis Institute of Technology
Mumbai,India

Shree jaswal
Associate Professor
Computer Department
St.Francis Institute of Technology
Mumbai,India

***Abstract*—Ina era of Internet, readers are overloaded with a very lengthy text document and selecting the most important sentence from the document is very tedious, difficult and time consuming task. So for this purpose a new technique is evolved which is called as Summarization. It is a technique which is used to solve the problems of readers. The summary is the short and precise form of the original document which helps users to handle the vast and complicated document easily. Extractive and abstractive are two types of summarization. Extractive summary picks the sentences from document and combines them to form a summary and Abstractive summary generate summary with the help of sentence compression and fusion). Here we proposed a new system which is used to generate abstractive summary of multi-document text which is based on ontology. Multi-document means input is multiple documents and output is single document, the multi-document is the collection of same topic or relevant documents, and ontology is the specification of conceptualized knowledge. These ontologies are used to share a common vocabulary to user.**

**Keywords—** ***Summarization, Extractive, Abstractive, Multi-document, Ontology.***

## I. INTRODUCTION

Now days World Wide Web is the most favorable source for getting information and discovering the important and relevant data from the web is very challenging, difficult, time consuming, and tedious task. Users have to read the whole document to understand the concept which is waste of time. Because some time the document that user have to read is not relevant or it may contain irrelevant data. To tackle this issue a new technique is paid attention to user called as text summarization. Text summarization is a system which clusters the important sentences and presents a summary. The important sentences are selected using sentences scoring and sentence ranking method [1]. The generated summary is short and precise which conveys the essence of the document, which helps in finding relevant information swiftly. Summarization can be of two types [2] Extractive and Abstractive. Extractive summaries are generated by reusing the portion (i.e. sentences) of original document without changing the source text. For example the summary generated by Microsoft 2007 professional is an extractive summary. Whereas Abstractive summaries are created by constructing new sentences using statistical natural language processing and generate a summary in concise way. Early workstarted with single document summarization where input and output is single document, as the research proceeded and web contains a huge amount of information the requirement for multi-document is come in picture.

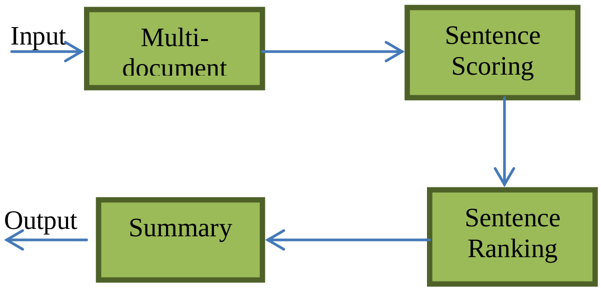

Fig.1 Multi-document Summarization technique [3]

As shown in fig.1 Summarization process consists of two important steps they are sentence scoring and sentence ranking. Sentence Scoring[3][4] used to score the important sentences and using that reference Sentence Ranking ranks the sentences[3] [4] and from this a summary is generated. Multi-document[2] [4] means the input is multiple documents where the content is about similar topic or they are relevant documents and the output is single document. There are various methods for finding the summary using extractive type but getting abstractive summary is difficult task. Most of the research work had been done on extractive summary, whereas abstractive summary is still an ongoing research work. Here we provide a new technique in which the ontology has been used for summarization tasks. According to Tom Gruber, he defines Ontology is a specification of a conceptualization[5]. Ontology is important because Ontological analysis clarifies the structure of knowledge. Without ontologies there cannot be a vocabulary of sentences for representing knowledge. Ontology is a conceptual model that can be used in multiple applications. It shared common understanding of the structure of information. Ontology is used to share the vocabulary (for human and agents)[6]. The domain ontology[6] is used to specify the concept and corresponding concept of that particular document. Information-retrieval systems, digital libraries, integration of heterogeneous information sources, and Internet search engines need domain ontologies to organize information and enhance the search processes. Domain ontology helps in developing object-oriented design. It also helps in building large knowledge systems e.g. in many areas of Artificial Intelligence, Ontology [5] [6] play the role of concept dictionary in natural-language understanding. WorldNet[7]can be used in ontology to define word meaning and models. WordNet tries to focus on the word meanings instead of word forms, though inflection morphology is also considered. WordNet consists of three types, one for nouns, one for verbs and a third for adjectives and adverbs. WordNet consists of a set of synonyms "synsets". A synset denotes a concept or a sense of a group of terms. Synsets provide different semantic relationships such as synonymy (similar) and antonymy (opposite), hypernoymy (super concept), hyponymy (sub concept) (also called Is-A hierarchy / taxonomy), meronymy (part-of) and holonymy (has-a). The structure of paper is as follows: Section 2 introduces the previous work i.e. literature survey done by different author on this summarization system. Section 3 is the proposed work where the methods to generate abstractive summary is explained in detail. And lastly the conclusion of paper is outlined in section 4.

## II. LITERATURE SURVEY

This section describes the various state of art technique for summarization. Summarization is the technique which is used to compress the lengthy data in short and concise form. Various methods have been proposed to achieve the extractive summary and abstractive summaries, most of them are based on scoring and ranking the sentences. Here these techniques are explained as follows.

- **Automatic Text Summarization:** The aim of automatic text summarization [2] is to condense the source text by extracting its most important content that meets a user's or application needs. Generally a summary is less than or equal to 50% of the original text.

  **Text Summarization Features:** In order to identify key sentences for summary, a list features as discussed below, can be used to for selection of key sentences.

  **Location:** It gives the important sentences according to their locations.

  **Cue Method:** Words that have positive or negative effect on the respective sentence weight to indicate significance or key idea such as cues: "in summary", "in conclusion", "the paper describes", "significantly".

  **Title/Headline word:** It states that the word belongs to headline or title is important sentence ion the summarizations i.e. they must be include in the summary.

  **Sentence length:** The length of sentences matters a lot. Short sentences are not include in summary where as long sentences are also not useful in summary.

  **Proper noun:** Sentences having proper nouns are considered important for document summary. Examples of proper nouns are: Fred, New York, Mars, and Coca Cola etc.

- **Yago Ontology:**Yago based summarizer [1] is a new technique which is used to generate document summaries. These summaries are in the form of extractive. There are basically three steps to generate summary, they are as follows.

  **Entity recognition:** This step analyzes the input document to identifying the most relevant concepts and their corresponding context of use. For this purpose a Yago knowledge base is used to map the words that occur in the document sentences.

  **Sentence Score & Ranking:** To generate a summary only the relevant and important sentences which are syntactically and semantically correct has to be considered in account. Sentence ranking and scoring method will be used to find out the important sentence from the original document.

  **Sentence Selection:** Sentence selection process is used to select the top ranked sentences from the document to generate the extractive summary.

- **Summarization in Disaster Management:**It is already known that, somewhere some unwanted thing happened like earthquake, hurricanes, or other natural resources causes' immense physical, social and hypothetical loss which was not recovered easily? So to recover at least important document will be helpful in future situation. Hence summarization technique is helpful in disaster management[4]. Here two different directions to generate the summary is used they are generic summarization and query-focused summarization.

  **Generic summarization:** For generic summarization in disaster management domain, the main task in general is to distill the most important overall information from a set of documents related to the disaster. The standard K-Means method is used to cluster the sentences of a document collection. The generic summary can be generated using sentence ranking, sentence scoring and sentence selection strategy.

  **Query- Focused summarization:** Query-focused summarization aims at generating a short summary based on a given document set and a given query. The generated summary gives the summary according to query.

- **FoDoSu:**Many approaches to multi-document summarization have used probability-based methods and machine learning techniques to simultaneously summarize multiple documents haring a common topic. However, these techniques fail to semantically analyze proper nouns and newly-coined words because most depend on an out-of-date dictionary or thesaurus. To overcome these drawbacks, a new technique is propose for multi-document summarization called FoDoSu [8], or Folksonomy-based Multi-Document Summarization, that employs the tag clusters used by Flickr, a Folksonomy system, for detecting key sentences from multiple documents. It first creates a word frequency table for analyzing the semantics and contributions of words using the HITS algorithm. Then, by exploiting tag clusters, it can analyze the semantic relationships between words in the word frequency table. Finally, creates an extractive summary of multiple documents[7] by analyzing the importance of each word and its semantic relatedness to others.

- **Text Summarization Systems :**There are different approaches for text summarization system[9]scoring and selecting sentences

  **Statistical approaches:** In Statistical methods, sentence selection is done based on word frequency. There are several methods for determining the key sentences such as, The Title Method, The Location Method, The Aggregation Similarity Method, The Frequency Method, TF- Based Query Method, and Latent Semantic Analysis.

  **Linguistic approaches:** Linguistic approaches are based on considering the connections between words and trying to find the main concept by analyzing the words.

  **Graph Theory:** Graph theory is used to represent the structure of text in the form of nodes and edges. Where nodes are sentences and edges represent the connection between the sentences.

## III. PROPOSED WORK

### Objective

Main objective is to make computer understand the document and how to make it generate the summary. Summary should match with human generated summary. The proposed system should accept the documents in .txt, pdf, html form and percentage of summary or key words from the user and generates its summary. System displays the summary as highlighted text in the original document. The proposed system generates summary based on both statistical and linguistic features of the document. In statistical we go for word frequency calculation. The idea here is that most important words are those that occur frequently in the document apart from stop words like is, a, an, the and so on. Plural resolution and abbreviations are considered. Linguistic analysis follows the concept of graph theory. As shown in Fig 2 the general model is explained in detail.

**Extractive Summary:**To generate extractive summary there are steps like Sentence Tokenization, Sentence Extraction, Sentence Scoring, Sentence Ranking. They are explaining in detail as follows.

Linguistic Analysis: It extracts sentences from the input documents, and then performs tokenization [10].**Tokenization:** Tokenization [10]is the process of breaking a sentence into word, phrases, symbols etc. [10]

**Redundancy Detection:** That the removal of sentences is based on the evolution of the event. Stemming is the process which is used for the purpose of redundancy. Stemming is defined as use the word as base word. Examples are connections, connective, consider being base word as connect.

**Sentence Representation:** Sentence representation includes calculating the frequency of relevant sentences. They are

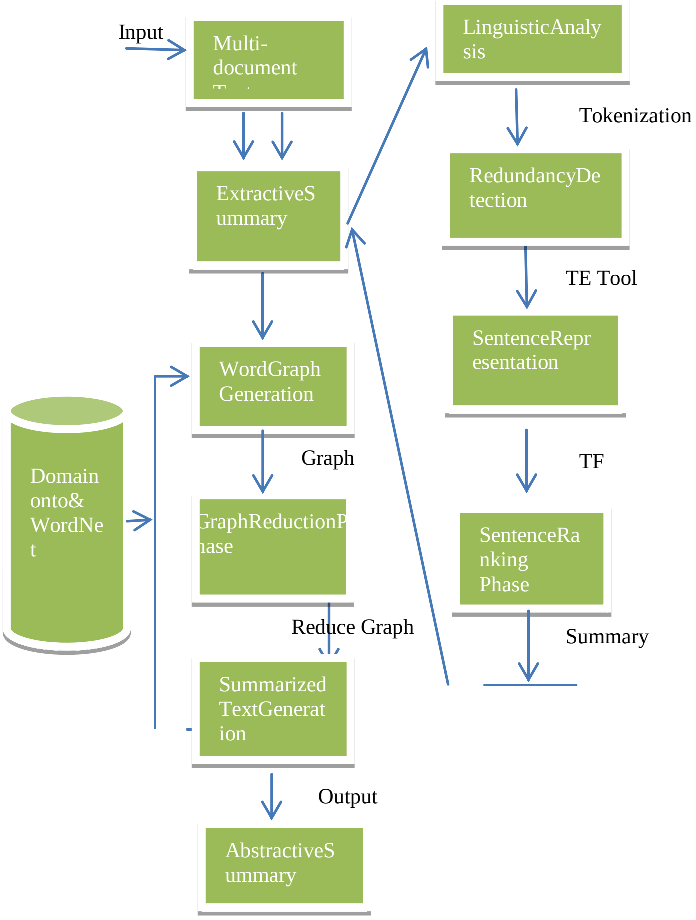

Fig.2 Multi-document Summarization technique

**Term Frequency Model:** In this model, each entry of a sentence vector denotes the term weight

$$tf_{i,j} = \frac{n_{i,j}}{\sum k \, nk,_{j}} \qquad (1)$$

Where $n_{i,j}$ is the number of occurrences of term $t_i$ in sentences$s_j$ ,$\sum k \, ni,_j$ is the sum of number of occurrencesof all the terms in sentence $s_j$

**Sentence Ranking:**Sentences are ranked according to their relevance and the highest ones are selected and extracted to generate a summary.

$$\max n = \log_2 \frac{N}{n_i} \qquad (2)$$

Where,

Maxn= maximum frequency of any term in the document
N= no of sentences in the document
$n_i$ = total no of occurrences of term i in the document

**Abstractive Summary:** To create abstractive summary word graph generation, graph reduction phase, and summarized text generation steps should be followed,

**Word Graph Generation**: This phase aims to reduce the original document to more reduced graph [11]. In this phase, a set of heuristic rules are applied to reduce it by merging, deleting, or consolidating the graph nodes. These rules exploit the WordNet semantic relations: hypernymy, holonymy, and entailment.

**Summarized Text Generation:** Summarized Text Generation [12][13]is the last step which is used to generate abstractive summary by replacing the word with the synonym from WorldNet Ontology. Here the word net Ontology is used.

**Lexicalization Process:**For each verb/noun object, its synonyms are selected by accessing the WordNet ontology to generate the target content.

$$W= (E+ (1-\frac{NR}{RT}) +(\frac{NGS}{TG}))/3)*10(3)$$

Where*E* is theExistence probability of the synonym in the document,*NR* is the synonym WordNet rank, *RT* is the total value of all synonym ranks, *NGS* is the WordNet group by similarity for synonym and *TG* is the total number of groups by similarity for all synonyms.The summary we get as output is an abstractive summary, where the abstractive summary is in human readable format.

**Aggregation Process:**In aggregation process system will collect all sentences having maximum weight and forms a paragraph.

**Paragraph Realization:**The summarized paragraph or sentences includes punctuation and corrected grammatically.

**Paragraph Selection:** Sentences with higher score are included in the final summary in the same order of their occurrence as in the original text document to retain their semantic meaning. Sentences are included in the final summary based on the following rules:

**Rule 1:** Sentences are selected to be included in the summary according to their highest to lowest rank score value.

**Rule 2:** If more than one sentence in the same paragraph shows same rank score then, sentence appearing earlier in the paragraph is given preference over the sentence appearing later to generate fixed length summary

## VII. CONCLUSION

Summarization is new system which help user to get a quick overview of whole document in fraction of second. Here a new system is proposed which give summary in abstractive form. Abstractive summaries form new sentences which are also called as compression and fusion to generate a correct summary. This generated summary can be human readable format. The sentences having highest score are select and using WordNet ontology the synonym are replace to generate a abstractive summary. These abstractive summaries are bit complicated than extractive summaries but are moreuseful as well.

## REFERENCES

[1] Elena Baralis , Luca Cagliero , Saima Jabeen, Alessandro Fiori , Sajid Shah, “Multi-document summarization based on the Yago ontology,” *Expert Systems with Applications* 40 (2013) ;6976–6984,Elsevier,http://dx.doi.org/10.1016/j.eswa.2013.06.047

[2] AtifKhan, Naomie Salim,“A REVIEW ON ABSTRACTIVE SUMMARIZATION METHODS,”*Journal of Theoretical and Applied Information Technology* 10th January 2014; Vol. 59 No.1

[3] Archana AB, Sunitha C , “An Overview on Document Summarization Techniques,” *International Journal on Advanced Computer Theory and Engineering (IJACTE)* 2013;Volume-1, Issue-2; 2319 – 2526,

[4] Lei Li and Tao Li, “An Empirical Study of Ontology-Based Multi-Document Summarization in Disaster Management,”*IEEE Transactions on Systems, Man, & cybernetics: Systems*, vol. 44, no.2, February

[5] http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[6] B. Chandrasekaran and John R. Josephson, V. Richard Benjamins, “What Are Ontologies, and Why Do We Need Them?,” *IEEE INTELLIGENT SYSTEMS*; 1999

[7] Feiyu Lin ,Kurt Sandkuhl, “A Survey of Exploiting WordNet in Ontology Matching,” *International Federation for Information Processing*, Volume 276; Artificial Intelligence and Practice II; Max Bramer; (Boston: Springer), 2008 pp. 341–350.

[8] Jee-Uk Heu, Iqbal Qasim, Dong-Ho Lee, “FoDoSu : Multi-document summarization exploiting semantic analysis based on social Folksonomy,”*Information Processing and Management* ;2014 Article in press . Elsevier http://dx.doi.org/10.1016/j.ipm.2014.06.003

[9] Qinglin Guo,Ming Zhang, “Multi-documents Automatic Abstracting based on text clustering and semantic analysis,”2009 ;*Elsevier Knowledge-Based Systems* 22 (2009) 482–485

[10] Tokenization:http://en.wikipedia.org/wiki/Tokenization

[11] Katja Filippova, “Multi-Sentence Compression: Finding Shortest Paths in Word Graphs,”*23rd International Conference on Computational Linguistics* August 2010; (Coling 2010),Beijing

[12] Elena Lloret, María Teresa Romá-Ferri, “COMPENDIUM: A text summarization system for generating abstracts of research papers,” *Data & Knowledge Engineering* 88 ;2013 164–175

[13] Ibrahim F. Moawad, Mostafa Aref, “ Semantic Graph Reduction Approach for Abstractive Text Summarization,”IEEE 2012; 978-1-4673-2961-3/12/$31.00