A Survey of Extraction of Important Texts for Effective Multi Document Summarization Using Improved Fuzzy Logic

Patel Birva[1], Shah Aakash[2]
[1, 2] Dept of CE
[1, 2] Silver Oak College Of Engineering & Technology, Gota, Ahmedabad,Gujarat, India

Abstract- *Text mining has different aspects to process sentences, words and texts. Many time it has become a lengthy and cumbersome process to understand and co-relate words and texts as part of sentences to generate a meaning which is in combination of these selected words.In this paper I have focused on advanced level text summarization where in extractive text summarization, important words and texts are selected based on certain important features which in turn extracting sentences containing it. The importance of some extractive features is more than the some other features, so they should have the balance weight in computations. Based on which a graph is generated which has same weight balanced node at each level.The purpose of this paper is to use novel method and WorldNet dictionary features such as relative words and synonyms to handle the issue of ambiguity and imprecise values with the traditional two value or multi-value logic.*

Keywords- text summarization, extractive method, graph creation,WordNet.

I. INTRODUCTION

In today's world, the daily hustle-bustle does not permit a human being to devote time for manually summarizing various lengthy documents.[1]

Text Summarization produces a shorter version of large text documents by selecting most relevant information. Text summarization systems are of two types: extractive and abstractive.

Text Summarization plays an important role in the area of text mining and natural language processing. As the information resources are increasing tremendously, readers are overloaded with loads of information[3]. Finding out the relevant data and manually summarizing it in short time is much more difficult, challenging and tedious task for a human being[4].

Text summarization can be done by three different methods: fuzzy logic based method,bushy path method, and wordnet synonyms method are used to generate summaries[3]. Wordnet ontology is also used to generate abstractive summary from extractive summary.

Multi-document summarization aims to produce a compressed version of numerous online text documents and preserves the salient information[1].

A particular challenge for multi-document summarization is that there is an inevitable overlap in the information stored in different 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[2].

This paper introduces a clustered genetic semantic graph approach for multi-document abstractive summarization. The semantic graph from the document set is constructed in such a way that the graph vertices represent the predicate argument structures The clustering algorithm is performed to eliminate redundancy in such a way that representative PAS with the highest salience score from each cluster is chosen, and fed to language generation to generate summary sentences.

This paper proposes an innovative graph-based text summarization model for generic single and multi-document summarization. The approach involves four unique processing stages: parsing sentences semantically using Semantic Role Labeling (SRL), grouping semantic arguments while matching semantic roles to Wikipedia concepts, constructing a weighted semantic graph for each document and linking its sentences (nodes) through the semantic relatedness of the Wikipedia concepts[6].

- The contributions of this paper are as follows[4].

1) 1)we avail SRL-based semantic representation of sentences to group similar arguments from each role-set and project them onto corresponding Wikipedia concepts.
2) we propose a weighted semantic document graph where each sentence is represented by the sub-nodes containing the concepts of its semantic arguments. The semantic relatedness between the Wikipedia concepts of the semantic arguments forms the edge-weights.
3) 3)The performance of our summarizer is empirically validated using the standard DUC2002 dataset.

- In each cluster, each sentence is assigned five different weights

1. Chronological weight of sentence (Document level)
2. Position weight of sentence (position of sentence in the document)
3. Sentence weight (based on term weight)
4. Aspect based weight (sentence containing aspect words)
5. Synonymy and Hyponym Weight.

- Then top ranked sentences having highest weight are extracted from each cluster and presented to user.

APPLICATIONS

- Tracking and Summarizing news articles on a daily-basis.
- Summarizing Chapters from various reference books.
- Speech Summarization

II. BACKGROUND

Most of the studies have focused on multi-document extractive summarization using techniques of sentence extraction, statistical analysis, discourse structures and various machine learning techniques. Different graph-based methods have also been investigated for multidocument extractive summarization. However, abstractive summarization is a challenging area for researchers. To date, a few research efforts have been done in this direction.

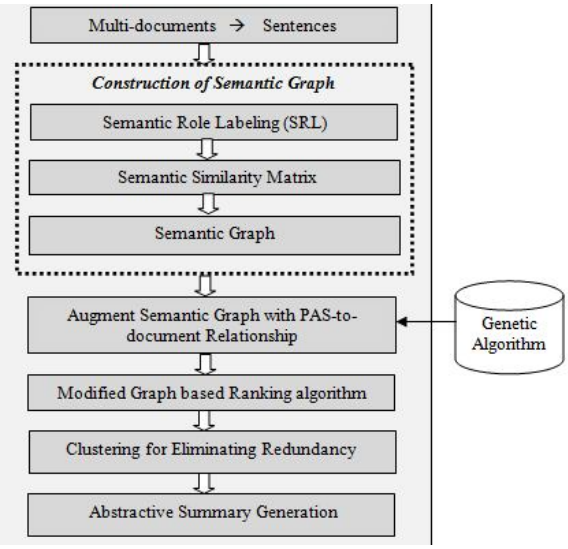
A particular challenge for multi-document summarization is that topically related documents usually contain overlapping information[3]. Thus, suitable summarization methods are required to merge similar information content across several documents. Specifically, the aforementioned graph based methods attempted for multi-document extractive summarization treat sentence as bag of words and did not
take into account the semantic structure of sentence and semantic relationships between sentences[8]. These methods determine sentence similarity by utilizing content similarity measure, which may not able to identify redundant sentences that are semantically equivalent. Thus, the final summary would contain redundant information. To our knowledge, semantic graph based approach has not been investigated for multi-document abstractive summarization (MDAS). Therefore, this study aims to introduce a clusteredgenetic semantic graph based approach for MDAS, which will automatically merge similar information across the documents, and employs language generation to generate abstractive summary[10]. The approach constructs semantic graph from the document text in such a way that the graph vertices represent the predicate argument structures (PASs) which are extracted automatically by employing semantic role labeling (SRL), and the edges of graph corresponds to semantic similarity weight determined from PAS-to-PAS semantic similarity, and PAS-to-document relationship. The salience (importance) score of graph vertices (PASs) is determined based on modified weighted graph basedranking algorithm, and finally the graph vertices (PASs) are sorted in reverse order based on salience scores[15]. We apply agglomerative hierarchical clustering to eliminate redundant PAS in such a manner that the most representative PAS (the one with the highest salience score) is chosen from each cluster. The representative PASs are then given to summary generation phase to produce summary sentences. Our
contributions are summarized as follows:

- Propose a clustered genetic semantic graph approach for multi-document abstractive summarization.
- Propose modified weighted graph based ranking algorithm to take into account the PAS-to-PAS semantic similarity and PAS-to-document relationship.
- Examine semantic similarity measure to detect redundancy by capturing semantically similar predicate argument structures (PASs).
- To evaluate the proposed semantic graph based approach with Pyramid and ROUGE evaluation measures on DUC 2002 multi-document summarization shared tasks.

Limited research studies have dealt with multi-document abstractive summarization. Two mainstream approaches are applied to multi-document abstractive summarization:linguistic and semantic based approaches.

Linguistic based approaches proposed for abstractive summarization employ tree based method and information item based method .All linguistic based approaches rely on syntactic representation of the source document, and therefore the general limitation of these approaches is the lack of semantic representation of source text. On other hand, different semantic based approaches have also beenintroduced for abstractive summarization such as template based methods and ontology based methods . The obvious drawback of template based methods is that linguistic patterns and extraction rules for template slots are manually created by humans, which is time consuming.

Moreover, these methods could not handle similar information across multiple documents. Moreover, theontology based methods heavily rely on domain expert to build domain ontology, which require more effort and time, and these methods are not applicable to other domains[11]. In recent years, different graph based approaches have been employed for multi-document extractive summarization. These methods use PageRank algorithm or its variants to rank sentences or passages. However, these approaches treat sentence as bag of words and did not consider the semantic structure of sentence i.e. predicate argument structure. Moreover, these approaches rely on content similarity measure and did not consider semanticrelationships between sentences while computing the salience score of sentences. These approaches may fail to detect redundant sentences that are semantically equivalent, and therefore the final summary would be inadequate. The only graph based approach introduced for abstractive summarization constructs semantic graph from manually built ontology[12]. This approach heavily relies on human expert and is limited to single document.



A. Semantic Role Labeling

The goal of this step is to extract predicate argument structure from each sentence in the document set. At first, we split the document set into sentences in such a way that each sentence is preceded by its corresponding document number and sentence position number. Since abstractive summarization requires deep semantic analysis, therefore we employ SENNA semantic role labeler [20] to parse each sentence and properly labels the semantic word phrases.

B. Semantic Similarity Matrix

The goal of this step is to construct a matrix of semantic similarity scores for each pair of predicate argument structure. In this step, similarity of the predicate argument structures (PASs) is computed pair wise based on acceptable comparisons of noun-noun, verb-verb, location-location and time-time[4].

C. Semantic Graph

The goal of this phase is to construct semantic graph from the semantic similarity matrix constructed in previous phase[3]. The undirected weighted semantic graph is constructed from similarity matrix (representing similarity scores of predicate argument structures (PASs))

D. Augment Semantic Graph with PAS-to-document Relationship

In order to reflect the local impact of document on predicate argument structures(PASs), this phase additionally augments the edge of semantic graph (representing semantic similarity weight between PASs) with PAS-to-document relationship. We assume that the PASs which appear early in the document and have close distance to the centroid of the document set will be considered as salient and will getmore chances to be selected for summary[13]. Thus, we express the correlation/relationship of PAS to document by three features discussed in our previous study and are given as follows: PAS semantic similarity to the document, position, and title features[6]. Since text features are sensitive the quality of summary i.e. not all features have same relevance with respect to summary.

E. Modified Weighted Graph based Ranking Algorithm

Previous graph based methods exploit relationships/associations between sentences based on content similarity and did not consider semantic relationships between [3]. These methods apply similar procedure like PageRank to

choose sentences based on number of "votes", received from their neighbouring sentences. To our knowledge, graph based ranking algorithm has not been considered for multi-document abstractive summarization. This study employs a modified weighted graph based ranking algorithm (MWGRA), which will take into account the edge weights in the vertices (PASs) ranking process (or importance analysis)[7]. The edge weight corresponds to PAS-to-PAS semantic similarity, and **PAS-to-document set** relationship.

F. Clustering for Eliminating Redundancy

Clustering of sentences for the purpose of removing redundancy is a common step in multi-document summarization[9]. Agglomerative hierarchical clustering is well-known technique in the hierarchical clustering method, which has been found useful in the range of applications

G. Abstractive Summary Generation

This phase takes the top scored predicate argument structures (PASs) from previous phase, employs SimpleNLG [4] and a simple heuristic rule implemented in it to generate summary sentences from PASs.

III. CONCLUSION

Our main goal in this paper is to find the best featured sentence that could impact to form the summary. For these, we have created cluster for each different weigh word from all documents[3]. Apart form text pre processing, we have introduced feature based selection that help in ranking the sentences and to get the best sentences from each vector.

The summary generated by the proposed algorithm follows the extraction method, when it finds the most unique sentences based on each word. it is containing that one sentence contains a formal person, place or thing.

We are dealing with conceptual synopsis in which we are thinking to combine sentences to make another single sentence examine their impact on summarization.

REFERENCE

[1] Jyoti Yadav & Dr. Yogesh Kumar Meena, "Use of fuzzy logic and wordnet for improving performance of extractive automatic text summarization" 2016,IEEE, 978-1-5090-2029-4

[2] Harsha Dave & Shree jaswal "Multiple text document summarization system using hybrid summarization technique" 2015,IEEE, 978-1-4673-6809-4

[3] Atif Khan, Naomie Salim & Haleem Farman "Clustered genetic semantic graph approach for multi-document abstractive summarization" 2016, IEEE(copy right), 978-1-4673-8753-8

[4] Muhidin Mohamed & Mourad Oussalah, " An iterative graph-based generic single and multi document summarization approach using semantic role labeling and wikipedia concepts" 2016, IEEE, 978-1-5090-2251-9

[5] Deepak Sahoo , Rakesh Balabantaray , Mridumoni Phukon & Saibali Saikia , *"Aspect Based Multi-Document Summarization"* 2016, IEEE (copyright), 978-1-5090-166

[6] Amol Tandel, Brijesh Modi, Priyasha Gupta, Shreya Wagle & Mrs. Sujata Khedkar, "Multi-document text summarization - A survey" 2016,IEEE

[7] A. Khan, N. Salim, and Y. J. Kumar, "A framework for multi-document abstractive summarization based on semantic role labelling," *Applied Soft Computing*, vol. 30, pp. 737-747, 2015.

[8] Y. J. Kumar, N. Salim, A. Abuobieda, and A. Tawfik, "Multi document summarization based on cross-document relation using voting technique," in *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*, 2013, pp. 609-614.

[9] D. Das and A. F. Martins, "A survey on automatic text summarization," *Literature Survey for the Language and Statistics II course at CMU*, vol. 4, pp. 192-195, 2007.

[10] H. P. Luhn, *The Automatic Creation of Literature Abstracts*, IBM Journal of Research and Development, 2(2), 159-165, 1958.

[11] P. B. Baxendale, *Machine-made Index for Technical Literature: An Experiment*, IBM J. Res. Dev., 2(4), 354-361, 1958.

[12] H. P. Edmundson, *New Methods in Automatic Extracting*, J. ACM, 16(2), 264-285, 1969.

[13] Pragnya Addala, Text Summarization A Literature Survey , https://www.scribd.com/doc/235008952/Text-Summarization-Literature-Surveyscribd,on Jul 24, 2014

[14] Amit.S.Zore1, Aarati Deshpande, Extractive Multi Document Summarizer Alogorithim, In (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5245-5248 ISSN:0975-9646

[15] Nenkova, Ani, Sameer Maskey, and Yang Liu. "Automatic summarization."Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011. Association for Computational Linguistics, 2011.