2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

A Survey on Abstractive Text Summarization

N. Moratanch
Research Scholar,
Department of Computer Science and Engineering,
CEG, Anna University,
Chennai, India
tancyanbil@gmail.com

Dr. S. Chitrakala
Associate Professor
Department of Computer Science and Engineering,
CEG, Anna University,
Chennai, India
au.chitras@gmail.com

***Abstract*—Text Summarization is the task of extracting salient information from the original text document. In this process, the extracted information is generated as a condensed report and presented as a concise summary to the user. It is very difficult for humans to understand and interpret the content of the text. In this paper, an exhaustive survey on abstractive text summarization methods has been presented. The two broad abstractive summarization methods are structured based approach and semantic based approach. This paper collectively summarizes and deciphers the various methodologies, challenges and issues of abstractive summarization. State of art benchmark datasets and their properties are being explored. This survey portrays that most of the abstractive summarization methods produces highly cohesive, coherent, less redundant summary and information rich.**

***Index Terms*—Text Summarization, structure Based Approach, semantic Based Approach, Sentence Fusion, Abstraction Scheme, Sentence Revision, Abstractive Summary**

I. INTRODUCTION

In recent times text summarization has gained its importance due to the data overflowing on the web. This information overload increases in great demand for more capable and dynamic text summarizers. It finds the importance because of its variety of applications like summaries of newspaper articles, book, magazine, stories on the same topic, event, scientific paper, weather forecast, stock market, News, resume, books, music, plays, film and speech. Due to its enormous growth, many topnotch universities like Aarhus University-Denmark, National Centre for Text Mining (NaCTeM)-Manchester University, etc. have been staunchly working for its improvement.

As the volume of information and published data on the World Wide Web is growing day by day, accessing and reading the required information in the shortest possible time are becoming constantly an open research issue. It is a tedious task to gather all the information and then give the output in a summarized form. Internet is a platform that fetches the information from databases. But still this information is massive to handle. So text summarization came into demand that condenses the document into shorter version by preserving the meaning and the content. A summary is thus helpful as it saves time as and retrieves massive documents data. Prior to this time, it was done by manual labour but now-a-days automation has brought forth many advantages.

Text summarization approaches can be typically split into two groups: extractive summarization and abstractive summarization. Extractive summarization takes out the important sentences or phrases from the original documents and groups them to produce a text summary without any modification in the original text. Normally the sentences are in sequence as in the original text document. Nevertheless, abstractive summarization performs summarization by understanding the original text with the help of linguistic method to understand and examine the text. The objective of abstractive summarization is to produce a generalized summary, which conveys information in a precise way that generally requires advanced language generation and compression techniques.

Abstractive summarization is an efficient form of summarization compared to extractive summarization as it retrieves information from multiple documents to create precise summary of information. This has gained its popularity due to the ability of developing new sentences to tell the important information from text documents. An abstractive summarizer displays the summarized information in a coherent form that is easily readable and grammatically correct. Readability or linguistic quality is an important catalyst for improving the quality of a summary.

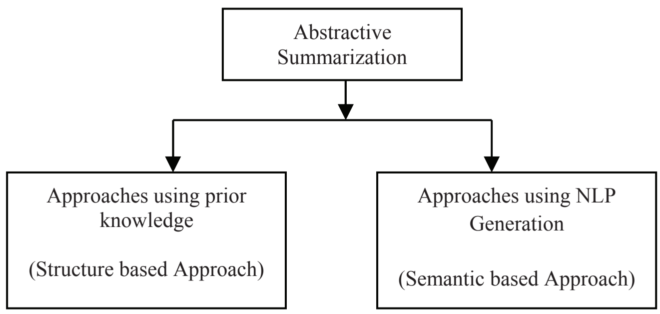


Fig. 1. Overview of Abstractive Summarization

This paper collectively summarizes the major methodologies adapted, issues found, research and future directions in text summarization. This paper is organized as follows. Section 2 depicts about the Structured based approach. Section 3 describes about Semantic Based approach. Section 4 depicts about Inferences made. Section 5 describes about Challenges and future research directions. Section 6 depicts about Experimental evaluation and finally this paper is concluded in Section 7.

II .STRUCTURE BASED APPROACH

Structured primarily based approach encodes most vital data from the document(s) through psychological feature schemas like templates, extraction rules and alternative structures like tree, ontology, lead and body, rule, graph based structure. Completely different ways that are used in this approach area unit are mentioned as follows.

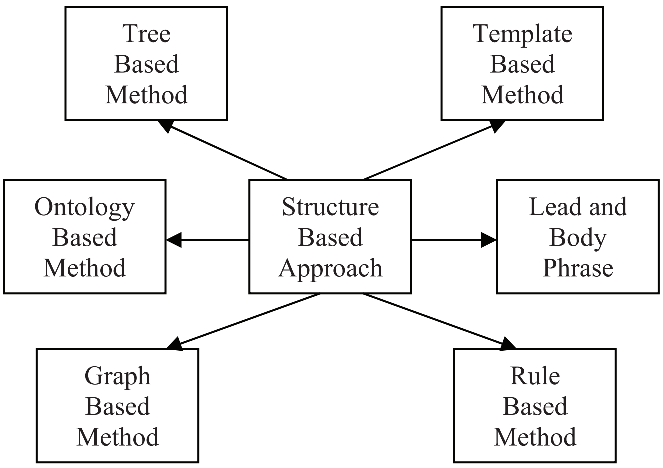


Fig. 2. Overview of Structure based approach

A. Tree based method

This technique utilizes a dependency tree that represents the text/contents of a document. Completely different algorithms are used for content choice for outline e.g. theme intersection algorithmic program or an algorithmic program that uses native alignment try across of parsed sentences. The technique uses either a language generator or associate degree algorithm for generation of outline. Connected literature victimization this methodology is as follows.

Regina Barzilay et al. [1] proposed a sentence fusion technique that identifies the common information phrases by using bottom-up local multi-sequence alignment. Sentence fusion is a technique used in multigene summarization system. In this approach, multiple documents are given as inputs and the central theme is identified by processing those inputs using theme selection and once the theme is finalized, they do ordering for the sentences and this is done by using clustering algorithm. Once the sentences are ordered, they are fused using sentence fusion and the corresponding statistical summary is generated.

B. Template based method

This technique uses a guide to represent a full document. Linguistic patterns or extraction rules area unit are matched to spot text snippets that may be mapped into guide slots. These text snippets are the area unit indicators of the outline content.

Sanda M. Harabagiu et al [2] proposed both single and multi-document summarization. They have adopted the techniques that were presented in GISTEXTER for producing both extracts and abstracts from the documents. GISTEXTER is a summarization system implemented for information extraction that targets the identification of topic-related information in the input document and translates it into database entries and later from these databases, the sentences are added to the summary based on user requests.

C. Ontology based method

Many researchers have created effort to use the ontology (knowledge base) to boost the method of summarization. Most documents on the online are domain connected which leads to same topic being discussed. Every domain has its own information structure which is highly represented by ontology. In the connected literature exploitation, this technique is mentioned as follows.

Lee et al [3] proposed the fuzzy ontology with its ideas introduced for Chinese news summarization to model uncertain information and thus will precisely describe the domain knowledge. In this approach, the domain ontology for news events is outlined by the domain experts followed by the Document preprocessing phase that produces the meaningful terms from the news corpus and also the Chinese news dictionary. For each of the fuzzy concept in the fuzzy ontology, the fuzzy inference phase generates the membership degrees. Various events of the domain ontology is associated with the collection of membership degrees for every fuzzy ideas.

D. Lead and body phrase method

This methodology relies on the operations of phrases (insertion and substitution) that have same syntactic head chunk within the lead and body sentences, so as to rewrite the lead sentence.

Tanaka et al. [4] proposed a method for summarizing broadcast news by analyzing the lead and body chunks of the sentence syntactically. The baseline of this idea is inferred from sentence fusion techniques. The summarization method involves in identifying common phrases in the lead and body chunks followed by insertion and substitution of phrases to generate a summary of news broadcast through sentence revision process. The initial step includes syntactically parser of the lead and body chunks which are followed by identifying trigger search pairs, followed by phrase alignment by using different similarity and alignment metrics. The final step involves insertion or substitution or both. The insertion step involves decision of insertion point, redundancy check and discourse coherence check to ensure coherency and elimination of redundancy. The substitution step ensures to increase information by substituting body phrase in the lead chunk.

E. Rule based method

In this technique, the documents to be summarized are depicted in terms of classes and listing of aspects. Content choice module selects the most effective candidate among those generated by data extraction rules to answer one or lot of aspects of a category. Finally, generation patterns are used for generation of outline sentences.

Pierre-Etienne et al. [5] suggested information extraction rules to find semantically related nouns and verbs. After extraction, content selection tries to avoid mixing candidates and sends the data to the generation. It is used for sentence structure and words in straight forward generation pattern. After generating, content guided summarization is performed.

Huong Thanh Le et al. [6] proposed an approach to abstractive text summarization based on discourse rules, syntactical constraints and word graph. The sentence reduction step is based on input sentences, keywords of the original text and syntactic constraints. Word graph is used only in the sentence combination stage. The method of generating a sentence from the essential fragment is split into finishing the start of a sentence and finishing the tip of a sentence. Sentence Combination is performed by observing and adhering to few syntactical cases.

Ansamma John et al. [7] proposed text Summarization based on feature score and random forest classification. The given input is pre-processed and then it computes the feature scores followed by training and cross validation of classifier and finally generating the summary of required size by maximal marginal relevance. The classification is a binary problem that determines which class the sentence belongs to either summary or non-summary class. The main task is to generate summary sentences from the summary class. The selected sentences are based on maximum relevance and minimum redundancy.

F. Graph based method

Many researchers use a graph data structure (called Opinosis-Graph) to represent language text. The novelty introduced in the system is that every node represents a word unit representing the structure of sentences for directed edges.

Dingding Wang et al. [8] suggested Multi-document summarization systems based on a variety of strategies like the centroid-based method, graph-based method, etc to evaluate different baseline combination methods like average score, average rank, borda count, median aggregation etc., for achieving a consensus summarizer to improve the performance of the summarization. A novel weighted consensus scheme is proposed to collect the results from individual summarization methods . Natural language generation (NLG) system is fed using linguistics illustration of document(s) in semantic based technique. This technique specializes in identifying noun phrases and verb phrases by linguistic data.

Kavita Ganesan et al. [9] proposed a novel graph-based summarization framework (Opinosis) that generates compact abstractive summaries of extremely redundant opinions. It has some distinctive properties that are crucial in generating abstractive summaries: Redundancy Capture, Gapped ssubsequence ccapture, Collapsible sstructures. The model generates an abstractive summary by exhaustively searching the Opinosis graph for appropriate sub-graphs encoding a valid sentence and high redundancy scores. The major components of the system are meaningful sentence and path scoring. Then a valid path is selected and it's marked with high redundancy score, collapsed paths and generation of summary. Then all the paths are ranked in the descending order of the scores and are eliminated duplicated paths by using Jaccard measure.

Elena Lloret et al. [10] focused on generating abstract summaries using word graph based method. The approach combines both extractive and abstractive information to generate abstracts. This method compresses and merges information based on word graph method thus generating abstracts. The words in the document form a set of vertices in the graph and the edge that represents the adjacent relationship between two words. A weighting function has been formulated to define the importance of the threshold using Page Rank value. The shortest path algorithm is used since it gives minimal length sentence with more information from relevant nodes in the graph. The important content is found using Compendium Text Summarization approach through two methods i) a set of sentences are given as input to the word graph and then it is forwarded to the compendium ii) selecting important content from the source document and then applying word graph method.

TABLE I

A COMPARATIVE STUDY ON STRUCTURED BASED APPROACH

Author/ Year	Techniques/ methods	Text representation	Content selection	Summary generation
Regina Barzilay, 1999	Tree Based	Dependency Tree	Theme intersection algorithm	Sentence fusion
Sanda M.Harabagiu, 2002	Template Based	Template having slots and fillers	Linguistic patterns or Extraction rules	IE based summarization algorithm
Lee and Jian, 2005	Ontology-Based	Fuzzy Ontology	Classifier	News Agent
Tanaka and Kinoshita, 2009	Lead and Body Phrase	Lead, body and Supplement structure	Revision Candidates (Maximum phrases of some head in lead and body sentences)	Insertion and Substitution operations on phrases
Ganest and Lapalme, 2012	Rule-Based	Categories and Aspects	Information Extraction rules	Generation Patterns
Elena Lloret, 2011	Graph based	Compresses and merge	Word Graph	Minimal length sentences compression

III .SEMANTIC BASED APPROACH

In semantic based technique, linguistics illustration of document(s) is employed to feed into natural language generation (NLG) system. This technique specialize in identifying noun phrases and verb phrases by processing linguistic data.

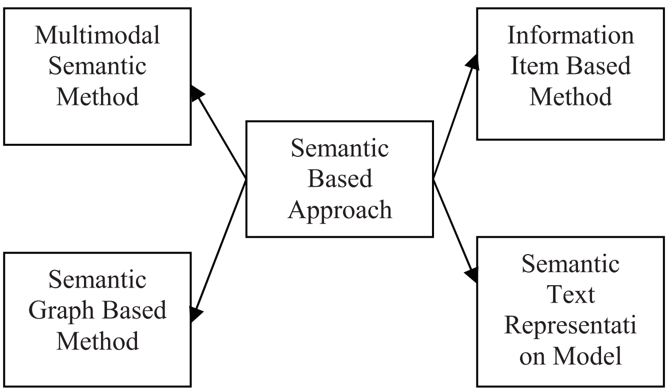


Fig.3. Overview of Semantic based approach

A. *Multimodal semantic model*

In this technique, a linguistics model, that captures concepts and relationship among ideas, is made to represent the contents like text and images that are used for multimodal documents. The important ideas are rated using some measures and eventually the chosen concepts are expressed as sentences to create summary.

Albert Gatt et al. [11] proposed a realization engine that aims in building large-scale data-to-text NLG systems, whose task is to summarize massive volumes of numeric and symbolic data. SimpleNLG provides interfaces to offer direct control over the way phrases are built and combined, inflectional morphological operations, and linearization. The major steps in constructing a syntactic structure and linearizing it as text with SimpleNLG are initializing the basic constituents that are required with the lexical items, combining constituents into larger structures, passing the resulting structure to the linearizer that traverses the constituent structure by applying the correct inflections and linear ordering depending on the features, and later the realized string is returned.

B. Information item based method

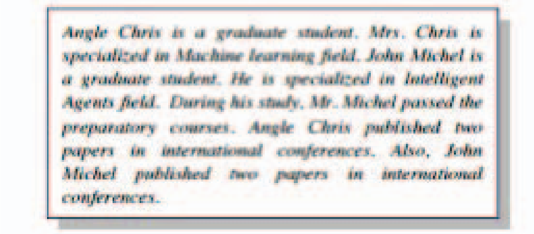
In this methodology, the information about the summary are generated from abstract representation of supply documents, instead of sentences from supply documents. The abstract illustration is Information Item that is the smallest part of coherent information in a text.

Pierre-Etienne Genest et al. [12] generated summarization by an information item (INIT) which is the smallest element of coherent information in a text or a sentence. The important goal is to identify all text entities, their attributes, predicates between them, and the predicate characteristics. During selection, the analysis of the source documents that leads to a list of INIT will proceed to select content for the summary. Frequency based models, such as those used for extractive summarization, could be applied to INIT selection instead of sentence selection. Most INIT do not give rise to full sentences, and there is a need of combining them into a sentence structure before being realized as a text. Local decisions are designed how to present the information to the reader and in what order during generation are now led by global decisions of the INIT selection step. The final summary generation is done by ranking of the generated sentences and a number of sentences that are intentionally in excess of size limit of the summary is first selected.

C. Semantic Graph Model

This technique aims to summarize a document by creating a linguistics graph known as rich semantic graph (RSG) for the initial document by reducing the generated linguistics graph and then generating the final abstractive outline from the reduced linguistics graph.

Ibrahim et al. [13] summarized a single document by creating a semantic graph called Rich Semantic Graph (RSG) from the original document. Then it reduces the generated semantic graph, and the final abstractive summary is produced from the reduced semantic graph. In Rich semantic sub-graphs generation module, for each input Word senses instantiation process instantiates a set of word concepts for both verb and noun senses based on the domain ontology. Concept validation processes are interconnected and validated to generate multiple rich semantic sub-graphs. The sentence concepts are inter-linked through the syntactic and semantic relationships generated in the pre-processing module. Sentence ranking process aims to threshold the highest ranked semantic sub-graphs for each sentence. During Rich semantic graph generation module, a set of heuristic rules are applied to the generated rich semantic graph to reduce it by merging, deleting or consolidating the graph nodes.


Angle Chris is a graduate student. Mrs. Chris is specialized in Machine learning field. John Michel is a graduate student. He is specialized in Intelligent Agents field. During his study, Mr. Michel passed the preparatory courses. Angle Chris published two papers in international conferences. Also, John Michel published two papers in international conferences.

Example Text

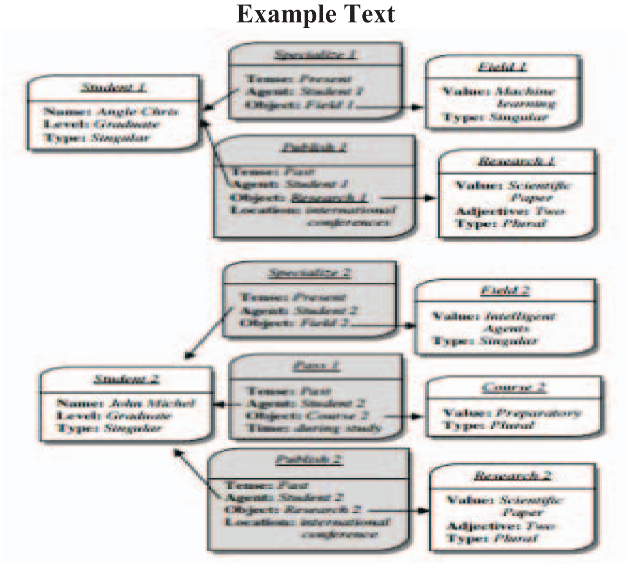


Fig. 4. Semantic graph text representation [13]

Fig 4.Semantic graph representation for the example text represents the nodes in the rich semantic graph represents the objects of the domain ontology classes for the input text nouns and verbs.

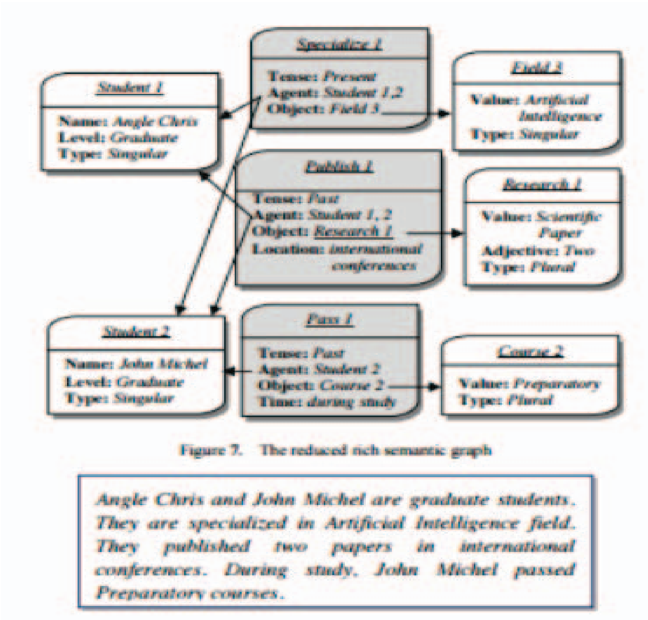


Fig. 5. Semantic graph reduction summary [13]

Fig.5.represents the reduced semantic graph and the summarized text obtained from the reduced graph. The reduced semantic graph is obtained by applying reduction heuristic rule. The generated abstractive summary contains only 50 percent of the original text document.

Lloret.E et al. [14] developed a model for generating ultra-concise concept-level summaries. The system initially converts the input document into its syntactic representation after lexical analysis. Summary generation is achieved using language generation tool with the lexical units as the input. This system lacks semantic representation of text and works on an assumption that all the sentences are anaphora resolved.

Nikita Munot et al. [15] performed summarization that the documents can be splitted into sentences and built by SPO triples to create a semantic graph called rich semantic graph(RSG). After performing RSG, it reduces the graph nodes to Subject Noun(SN), Main Noun(MN), and Object Noun(ON). After reducing Semantic graph, summary can be generated.

Jurij Leskovec et al. [16] proposed that the original document is represented as a semantic graph in the form of triples consisting of subject-predicate-object. The sub-structure of the graph is extracted to generate summaries. SVM classifier is used to identify a set of triplet that contribute to sentence extraction. A rich set of linguistic attributes are incorporated into the model to increase the performance of the proposed model. The set of triplets generated in the previous step are refined through a series of steps involving co-reference resolution, cross sentence pronoun resolution and sentence normalization and finally merge them to generate a semantic graph.

D. Semantic Text Representation Model

This technique aims to analyse input text using ssemantics of words rather than syntax/Structure of text.

Khan Atif et al. [17] proposed a framework for abstractive summarization of multiple documents in the form of semantic representation of supply documents. Content selection is done by ranking of the most significant predicate argument structures. Finally summary is generated using a language generation tool. But the system does not handle more detailed semantics in the summarization approach due to its assumption that the text to be handled are anaphora resolved and sense disambiguated.

Atif et al. [18] suggested Semantic role labelling to extract predicate argument structure from each sentence and the document set is split into sentences with its document number and position number. The position number is assigned by using SENNA semantic role labeller API. The similarity matrix is constructed from semantic Graph for Semantic similarity scores. After that, modified graph based ranking algorithm is used to determine predicate structure, semantic similarity and document set relationship. After predicate, MMR is used to reduce redundancy for summarization.

TABLE II

A COMPARATIVE STUDY ON SEMANTIC BASED APPROACH

Author/ Year	**Techniques/ Methods**	**Text Representation**	**Content Selection**	**Summary Generation**
Albert Gatt, 2009	Multimodal Semantic Based	Semantic model	Simple NLG	Generation technique: Synchronous tree
Pierre-Etienne Genest, 2011	Information item Based	Abstract representation: Information item(INIT)	Generated sentences can be ranked based on document frequency	NLG realize simple NLG
Ibrahim, 2012	Semantic Graph based	Rich Semantic graph	Calculation of each concepts and their sentences	Reduced Semantic graph
Atif, 2015	Reduced Semantic graph	Semantic Role Labelling	Semantic graph for similarity score	NLG realize simple NLG

IV .INFERENCES MADE

Quality of the summary is improved in structure based approach since it produces coherent , less redundant summary with higher coverage. The structure based method may have some grammatical issues since it does not take semantic representation of the document into consideration. The semantic based model provides better linguistic quality to the summary since it involves semantic representation of the text document capturing the semantic relations. The semantic method overcomes the issues of structure based that is it reduces redundancy in the summary , ensures better cohesion and also provides information rich content with better linguistic quality.

V .CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The major issues of abstractive summarization is there is no generalized framework , parsing and alignment of parse trees is difficult. Extracting the important sentences and sentence ordering as it has to be in the order as in the original source document for producing an efficient summary is an open issue. Compressions involving lexical substitutions, paraphrase and reformulation is difficult with abstractive summarization. The capability of the system is constrained by the richness of their representation and their way to produce such structure is the greatest challenge for abstractive summary. Still Information diffusion is not handled properly using abstractive text summarization.

VI. EXPERIMENTAL EVALUATION

Various benchmarking datasets are used for experimental evaluation of abstractive summarization. Document Understanding Conference(DUC) is the most common benchmarking dataset used for text summarization. There are number of datasets like DUC, TAC, DUC-2002,DUC-2004,2005, CNN, DUC-2006, DUC-2007,2008, TIPSTER, TREC. This dataset contains documents along with their summaries that are created manually, automatically and submitted summaries[19] [20] [21]

ROUGE toolkit is used to measure the summarization performance, which is widely applied by DUC. Several ROUGE methods are ROUGE-N, ROUGE-L, ROUGE-W and ROUGESU. Higher order ROUGE-N score (N>1) estimates the fluency of summaries.

$$\text{ROUGE-N} = \frac{\sum_{S \in ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ref} \sum_{gram_n \in S} Count(gram_n)}$$

VII. CONCLUSION

This survey has showcased various methods of abstractive summarization. Abstractive summarization methods produce highly cohesive, coherent, less redundant summary and information rich. The goal is to provide an extensive survey and comparison of different techniques and approaches of abstractive summarization. In this survey some of the challenges and future research directions are also highlighted. In summary, the literature in abstractive summarization depicts major progress in various aspects. However these works still have not addressed the various challenges of abstractive summarization to its full extent in terms of space and time complexity.

REFERENCES

[1] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," Comput. Linguist., vol. 31, no. 3, pp. 297–328, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1162/089120105774321091

[2] S. H. Finley and S. M. Harabagiu, "Generating single and multidocument summaries with gistexter," in In U. Hahn & D. Harman (Eds.), Proceedings of the workshop on automatic summarization, 2002, pp.30–38.

[3] C.-S. Lee, Z.-W. Jian, and L.-K. Huang, "A fuzzy ontology and its application to news summarization," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 35, no. 5, pp. 859–880, Oct 2005.

[4] T. K. T. K. Hideki Tanaka, Akinori Kinoshita and N. Kato, "Syntaxdriven sentence revision for broadcast news summarization," in NHK Science and Technology Research Labs. 1-10-11, Kinuta, Setagaya-ku, Tokyo, Japan, 2009.

[5] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 354–358. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390665.2390745

[6] H. T. Le and T. M. Le, "An approach to abstractive text summarization," in Soft Computing and Pattern Recognition (SoCPaR), 2013 International Conference of, Dec 2013, pp. 371–376.

[7] A. John and M. Wilscy, "Random forest classifier based multi-document summarization system," in Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in, Dec 2013, pp. 31–36.

[8] D. Wang and T. Li, "Weighted consensus multi-document summarization," Information Processing & Management, vol. 48, no. 3, pp. 513 – 523, 2012, soft Approaches to {IA} on the Web. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457311000732

[9] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in Proceedings of the 23rd International Conference on Computational Linguistics, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 340–348. [Online]. Available: http://dl.acm.org/citation.cfm?id=1873781.1873820

[10] E. Lloret, D. o. S. Manuel Palomar, and A. S. Computing Systems University of Alicante Apdo. de correos, 99 E-03080, "Analyzing the use of word graphs for abstractive text summarization," in The First International Conference on Advances in Information Mining and Management, 2011.

[11] A. Gatt and E. Reiter, "Simplenlg: A realisation engine for practical applications," in Proceedings of the 12th European Workshop on Natural Language Generation, ser. ENLG '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 90–93. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610195.1610208

[12] P.-E. Genest and G. Lapalme, "Framework for abstractive summarization using text-to-text generation," in Proceedings of the Workshop on Monolingual Text-To-Text Generation, ser. MTTG '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 64–73. [Online]. Available: http://dl.acm.org/citation.cfm?id=2107679.2107687

[13] [13] I. Moawad and M. Aref, "Semantic graph reduction approach for abstractive text summarization," in Computer Engineering Systems (ICCES), 2012 Seventh International Conference on, Nov 2012, pp. 132– 138.

[14] E. Lloret, E. Boldrini, T. Vodolazova, P. MartÃ nez-Barco, R. MuÃ´soz, and M. Palomar, "A novel concept-level approach for ultraconcise opinion summarization," Expert Systems with Applications, vol. 42, no. 20, pp. 7148 – 7156, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417415003541

[15] N. Munot and P.-N. P. I. Sharvari S. Govilkar Department of Computer Engineering, Mumbai University, "Conceptual framework for abstractive text summarization," International Journal on Natural Language Computing (IJNLC) Vol. 4, No.1, February 2015, 2015.

[16] J. Leskovec, M. Grobelnik, and N. Milic-Frayling, "Learning substructures of document semantic graphs for document summarization," in Proceedings of the KDD 2004 Workshop on Link Analysis and Group Detection (LinkKDD), Seattle, WA, USA, 2004.

[17] A. Khan, N. Salim, and Y. Jaya Kumar, "A framework for multidocument abstractive summarization based on semantic role labelling," Appl. Soft Comput., vol. 30, no. C, pp. 737–747, May 2015. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2015.01.070

[18] A. Khan, N. Salim, and Y. Kumar, "Genetic semantic graph approach for multi-document abstractive summarization," in Digital Information Processing and Communications (ICDIPC), 2015 Fifth International Conference on, Oct 2015, pp. 173–181.

[19] http://duc.nist.gov/data.html.

[20] http://trec.nist.gov/data.html.

[21] http://www.nist.gov/tac/data/forms/.