A Study of Different Text Summarization Techniques Based on Input

Nomi Baruah[1], Shikhar Kr. Sarma[2], Surajit Borkotokey[3]
[1]*(Dept. of CSE, Dibrugarh University, India)*
[2]*(Dept. of IT, Gauhati University, India)*
[3]*(Dept. of Mathematics, Dibrugarh University, India)*
Corresponding Author: Nomi Baruah

Abstract: *Automatic text summarization is the technique of distilling the most important information from a text or a set of text. The amount of online information services, social media and other digital format documents is increasing gradually which has necessitated intensive research in the area of automatic text summarization. Text summarization is the most challenging task for information retrieval. It requires integration of techniques, implementation of hybrid schemes, including NLP, data mining and analytics, as well as optimization and linguistic theories. Various researchers from the NLP community have been addressing it from various perspective in different domains and using different paradigms. Some summarization techniques perform well in identifying and summarizing single documents but their precision degrades sharply with multi-documents. As automatic summarization is becoming an important way to find relevant information precisely in large text in short time. This paper discusses various text summarization techniques both in the areas of single-document and multi-document summarization giving emphasis to approaches used for automatic summarization, methodologies, application in various domains and performance behavior. This has attempted to document and analyse the established techniques for text summarization. Analysis has been done on the methodologies, the schemes, and the approaches used in various research outcomes and proposals. The content of the present study, in the form of a critical analysis on text summarization techniques is a concise, comparative, and foundation block for further research works in the domain.*

Keywords: *Multi-Document, Single-Document, Text Summarization.*

Date of Submission: 15-07-2019 Date of acceptance: 31-07-2019

I. Introduction

A summary is defined as a text that is produced from one or more text[1].It contains a significant portion of information in the original text(s) and should be no longer than half of the original text(s).A summariser is a system that produces a condensed representation of its input's for user consumption[2].The research in automatic summarization started in late fifties when interest was shown in the production of abstract of technical documents by automatic summarization[3].But the interest in this area declined until Artificial Intelligence started showing interest in this topic[4].The interest in summarization restarted in the nineties with the organization of a number of relevant scientific events[5].However, the peak interest in summarization started from the year 2000 with the development of evaluation programs such as the Document Understanding Conference(DUC)[6] and Text Analysis Conference(TAC)[7].

Early research on summarization is proposed on scientific documents for extracting salient features on text. Luhn [3] extracted salient features in a text using features like word and phrase frequency. Baxendale [8] extracted salient features in a text using features like position in the text. Edmunson [9] extracted salient features in a text using key phrases. Apart from scientific documents, text summarization is done in various domains such as news articles, legal documents, tourism, etc. Summaries are extracted based on two approaches: Abstractive Approach and Extractive Approach [10]. Abstractive summarization is an elusive technological capability in which textual summaries of content are generated de novo [11]. Extractive summarization systems create summaries using representative sentences chosen from the input [12]. Text summarization is of two types: Single-document Text Summarization and Multi-document Text Summarization [13].

II. Text Summarization Techniques

Early research in summarization concentrated on summarization of single-document. There doesn't exist any standard length for the generated summary. Various language dependent and language independent are proposed based on single-document summarization technique. For a language independent summarization system, it should be portable to new languages and domain [14].Single document summarizers may be generic and query based [15].Generic summarisers are suitable for long documents containing a variety of topics. Query based summarisers use a user-query to summarize the document around this user query. Sentences are matched with the user-query using similarity measures. The sentences which are closer to the query are selected to be in the summary. A multi-document summary is defined as a brief description of the essential contents of a set of related documents. Multi-document summarization is useful in two types of situation: when the user is faced with a collection of dissimilar documents and wishes to access the information landscape contained in the collection. And when there is a collection of topically related documents which are extracted from a larger more diverse collection as a result of a query or a topically cohesive cluster[15].Section 2.1 to 2.5 falls under single-document summarization while Section 2.6 to 2.15 can be applied to single-document as well as to multi-document summarization.

2.1 Naive Bayes Method

Naive Bayes method was proposed by Kupiec et. al.[16] in 1995.The Naive Bayes classifier was used for learning from the data(corpus of document/summary pairs).It is a method derived from Edmundson[17] .Apart from the features in [17], it includes sentence length and uppercase words. The Naive Bayes classification calculates the probability of a sentence s with k features like F_1 ,F_2 ,..............,F_k .

$$P(s \in S|F_1 ,F_2 ,..............,F_k) = \frac{\prod_{i=1}^{n} P(F_i \ |s \in S).P(s \in S)}{\prod_{i=1}^{n} P(F_i)}$$

The sentences are scored and ranked for summary selection based on the above equation .The n top sentences were extracted based on the scores they have acquired. The system was evaluated based on a corpus of technical documents with manual abstract. Each sentence in the manual abstract analyzed its match with the actual document sentences and created a mapping. Evaluation of the auto-extracts is done against this mapping. System which uses position and cue features along with the sentence length sentence feature performs best.

Aone et. al.[18] proposed another naive-bayes classifier named DimSum with richer features. Features such as term-frequency (tf) and inverse document frequency (idf) were used to derive words which indicate key concepts in a document. The idf was computed from a large corpus of the same domain as the required documents. Two noun word collocations which were statistically derived were used as units for counting along with single words. Each entity was assumed as a single token with the help of a named-entity tagger. Some shallow discourse analysis were employed like reference to same entities in the text, maintaining cohesion. At a very shallow level, the references were resolved by connecting name aliases within a document .The synonyms and morphological variants were using Wordnet[19] while considering lexical terms. Corpora is used in the experiments, which were from newswire and some of which belonged to the TREC evaluations.

2.2 Position Method

Lin and Hovy[20] proposed a technique for Identifying Topics by Position. The method describes an automated training and evaluation of an Optimal Position Policy. It locates the likely positions of topic-bearing sentences based on genre-specific regularities of discourse structure. The method has two steps. Sentence Position Yields and the Optimal Position Policy; and Additional Measures and Checks. The optimal position for topic occurrence is determined as follows. A text T and a list of topic keywords t_i of T. Each sentence of T is labelled with its ordinal paragraph (P_m) and sentence number (S_n).All closed-class words are removed from the texts. Morphological restructuring and anaphoric resolution is not performed. A choice is made between the topic keywords and the abstracts accompanying each text in the corpus for determining the optimal position. Keywords and abstracts are considered as they appear in the original texts. A topic keyword has a fixed boundary, so it is easier to rank sentences than using an abstract. The yield of each sentence is computed by the number of different topic keywords contained in the appropriate sentence in each context and averaging over all texts. Formula sensitive is used for degree of overlap. The formula is based on Fibonacci function which monotonically increases with longer matched substrings and is normalized to produce a score of 1 for a complete phrase match. The hit function H measures the similarity between topic keyword t_i and a window w_{ij} for each sentence in the text. The H_s scores are computed from the beginning of a sentence to the end and added them together to get the total score H_s for the whole sentence. The H_s score for each of the sentence in the text is calculated. After obtaining all the H_s scores, all the sentences are sorted according to their paragraph and sentence numbers. The average H_{avg} score is computed for each paragraph and sentence number. Finally, the paragraph and sentence position is sorted by decreasing yield H_{avg} score. Additional measures and checks were performed inorder to prevent spurious or wrong rules. The yield of each sentence position is determined in the corpus empirically. It is measured against the topic keywords. The sentence position is ranked by their average yield to produce the Optimal Position Policy (OPP) for topic positions for the genre.

2.3 Hidden Markov Method

Conroy and O' Leary proposed a model which extracts sentences from a document using Hidden Markov Model(HMM)[21].The HMM handles positional dependence, dependence of features and Markovity[22]. The proposed HMM for text summarization consists of 2s+1 states with s summary states and s+1 non-summary states. The model is consisted of three parts: p the initial state distribution, M the Markov transition matrix and B the collection of multi-variant normal distributions associated with each state. Let $\alpha_t(i)$ be the probability in the sequence $\{O_1, O_2, \ldots\ldots O_t\}$ and is in state i (1 ≤ i ≤ N). $\alpha_t(i)$ is computed recursively. Assume $\alpha_t(i)$=p(i). α_t = $D_{O_t} M^T \alpha_{t-1}$ for t = 2 ...T.T is the number of sentences in the document. D_{O_t} = I - diag $\{b_1(o_1), b_1(o_1)$...... , $b_{2s+1}(o_{2s+1})\}$,I is the identity matrix ,b(.) is the cumulative density function, argument $o_i = (O_t - \mu_i)^T \xi^{-1} (O_t - \mu_i)$, μ_i is the mean for ith state. The probability of entire observation is w ≡ Pr(O) = $\sum_{i=1}^{2s+1} \alpha_T$(i).Two sentence extracting methods were used for HMM. In the first approach, sentences with the maximum posterior probability were chosen. From a summary of length k, sentences with k largest values of g_t were chosen. In the second approach, QR decomposition is used to remove any redundant sentence that might be included by the HMM maximum posterior probability method.

2.4 Log Linear Model

Osborne states that existing approaches to summarization have always assumed feature independence [23].It is possible to integrate together various sources of knowledge which are believed to be useful for the task with maximum entropy (log-linear) model. The model works incrementally and does not always need to process the entire document before assigning classification. Assume c be a label, s the item interested in labeling, f_i is the i^{th} feature, and λ_i the corresponding feature weight. The conditional log linear model [23] is stated as follows.

$$P(c|s)=\frac{1}{Z(s)}\exp\left(\sum_i \lambda_i f_i(c,s)\right) \qquad 3$$

where Z(s)=$\sum_c exp \sum_i \lambda_i f_i(c,s)$.A discrimination between the sentences to be extracted or not is made in the model. It is found that when sentences are extracted from technical papers, the score of precision levels are high and recall is very low. The disadvantage in the model is that features which predicted whether a sentence to be extracted tends to be very specific and occur infrequently.

2.5 Neural Networks

A task of creating 100 word summary of a single news article is issued in DUC 2001-2002.However,it is seen that the evaluations of the best performing systems could not outperform the baseline with statistical significance. Nenkova analyzed this extremely strong baseline and corresponds to the selection of the first n sentences of a newswire article [24]. Svore et al. proposed an algorithm which is based on neural networks and the use of third party datasets [25]. It tackles the problem of extractive summarization, outperforming the baseline with statistical significance.

A dataset containing 1365 documents is gathered from CNN.com. It consisted of the title, timestamp, three or four human generated story highlights and the article texts. Three machine highlights were created. Two metrics were used to evaluate the system. The first metric concatenates the three system generated highlights, concatenates three human generated highlights and compares the two concatenated highlights. The second metric considers the ordering and compares the sentences on an individual level.

Svore et al. trained a model for proper ranking of sentences in a test document [25]. The authors trained the model based on the labels and features for each sentence of an article. The sentences are ranked using RankNet [26].It is a pair-based neural network algorithm designed to rank a set of inputs which uses the gradient descent method for training.ROUGE-1 is used to score the similarity of a human written highlight and a sentence in the document.

2.6 Deep Natural Language Analysis Method

Barzilay and Elhadad proposed a work which uses linguistic analysis for performing the task of summarization [27]. The work is divided into the following steps. They are segmentation of the text, identification of lexical chains and use of strong lexical chains to identify the sentences worthy of extraction. The work is a mediocre between deep semantic structure of the text and word statistics of the documents.

Lexical chains are used as a source of representation for summarization and are present in the word level and word sequences. The words which are semantically related and word sequences are identified in the document. Several chains were extracted that represents the document. Wordnet [19] is used to find out the lexical chains. Three generic steps were applied. They are selection of a set of candidate words, finding an appropriate chain for each candidate word depending on a relatedness criterion among members of the chains. If the appropriate chain is found, the word is inserted and updated accordingly. Wordnet distance is used to measure the relatedness. The set of candidates is found using simple nouns and noun compounds. Finally, strong lexical chains were used to create the summaries. The score of the chains are calculated by their length and homogeneity. A few heuristics were used to select the significant sentences.

Ono et. al. proposed a computational model of discourse for Japanese expository writings[28]. They elaborated a practical procedure for extracting the discourse rhetorical structure. A binary tree is generated to represent the relations between chunks of sentences. The structure is generated by the following steps. They are sentence analysis, rhetorical relation extraction, segmentation, candidate generation and preference judgement. The evaluation is done based on relative importance of rhetorical relations. The evaluation is done based on sentence coverage. The dataset is of 30 editorial articles of a Japanese newspaper. The most important key sentence is selected from a set of key sentences from the articles which is judged by human subjects. It is found that the key sentence coverage was nearly 51% and the most important key sentence coverage was 74%.The results were found to be encouraging. Marcu proposed a unique approach towards summarization [29]. The approach uses discourse based heuristics with traditional features. The discourse theory is the Rhetorical Structure Theory(RST) that holds between two non-overlapping pieces of text spans i.e. the nucleus and the satellite. The difference between nuclei and satellite can be seen from empirical observation. The purpose of the writer is more expressed by the nucleus compared to the satellite. It is found that the nucleus of a rhetorical relation is independent of the satellite but the satellite is not independent of the nucleus.

2.7 Centroid-Based Summarization

Radev et. al. proposed MEAD- a centroid-based multi-document summarizer[31]. The centroid can be used to classify relevant documents and to identify salient sentences in a cluster. The documents which are relative to each other are grouped together into clusters. A weighted vector of TF*IDF is represented for each document. A centroid is generated by CIDR using only the first document in the cluster. The TF*IDF values are compared with the centroid with the process of each new document. The following formula is used to check whether a new document be included in the cluster or not based on the similarity measure sim(D,C) to be within a threshold.

$$\text{sim}(D,C) = \frac{\sum_k (d_k * c_k * idf(k))}{\sqrt{\sum_k (d_k)^2}\sqrt{\sum_k (c_k)^2}}$$

A small corpus of newsgroup is prepared consisting of a total of 558 sentences in 27 documents which is organized in 6 clusters, all organized by CIDR. The factors considered for selection of clusters are coverage of as many news sources as possible, coverage of TDT and non-TDT data, coverage of different types of news and diversity in cluster sizes.

2.8 Multilingual Multi-Document Summarization

Evans et. al. proposed Similarity-based Multilingual Multi-Document Summarization[32]. It summarizes machine translated documents using text similarity to related English documents. Sentences are identified to extract from the translated text to built the summary and replaces the machine translated sentences from the summary with similar sentences from a related English text. The purpose is to match the content of non-English documents with the content of English documents and thereby improving the grammatical and comprehensibility of the text. The approach is applicable for documents on the same topic. The English text is simplified through sentence simplification software [33].The long sentences are broken into two separate sentences by removing embedded relative clauses. This results in a more fine-grained matching between the Arabic and English sentences. The sentence simplification is examined with syntactic and syntactic with pronoun resolution. While evaluating with both the types, it is found that syntactic simplification performs 3% better on ROUGE scores than simplification with pronoun resolution. The similarity between the translated and relevant text is calculated using a tool named Simfinder[34]. The summarizer runs in multiple configuration system.

2.9 Graph Search and Matching

Mani and Bloedorn proposed Multi-document Summarization by Graph Search and Matching[35]. It summarizes the similarity and dissimilarity in a pair of related documents using a graph representation for text. The nodes in the graph are denoted by words, phrase and proper names in the document. For a given pair of documents to be summarized, a spreading activation technique is used to discover nodes in each document which are semantically related to the topic. The graphs which are activated from each document are matched to yield a graph corresponding to the similarity and dissimilarity in a pair. A sentence and paragraph tagger is used in this experiment which contains a very extensive regular-expression-based sentence boundary disambiguator. The Alembic part-of-speech tagger is invoked on the text [36]. SRA's NetOwlis used to extract names and relationships between names from the document [37].TF*IDF metric is used to extract salient words and phrases form the document in which a reference corpus is used. The phrase extraction method finds candidate phrases using several patterns defined over part-of-speech tags. The nodes which are extracted from the document and which is equivalent to topic terms are treated as entry points in the graph. Intrinsic evaluation is performed on the summaries which are generated by FSD-graphs with and without spreading activation.

2.10Topic Driven Summarization and MMR

Carbonell and Goldstein proposed The Use of MMR, Diversity-Based Re-ranking for Reordering Documents and Producing Summaries [38]. It combines query-relevance with information-novelty in the context of text retrieval and summarization. The MMR criterion reduces redundancy. The linear combination which is a measure of relevance and novelty independency is called marginal relevance. The MMR passage selection works better for longer documents and is useful in extraction of passage from multiple documents on the same topic.

2.11Abstraction and Information Fusion

SUMMONS is a multi-document summarization system that reads a database which is built by a template-based message understanding system. The architecture of SUMMONS consisted of a content planner that selects the information to be included in the summary through the combination of the input templates and a linguistic generator that selects the right words to express the information in grammatical and coherent text. The linguistic generator was devised by adapting existing language generation tools named FUF/SURGE system. Content planning is made through summary operators i.e. change of perspective, contradiction, refinement, etc. some of which require resolving conflicts. Finally the linguistic generator gathers all the combined information and uses connective phrases to synthesize a summary.

The framework seems promising for narrow domain but creates problem for broader domains. McKeown et al. [39] and Barzilay et al. [40] proposed an improved framework where the input is a set of related documents in raw text. Themes were identified i.e. sets of similar text units. These themes were formulated as a clustering problem. Inorder to compute a similarity measure between text units, these are mapped to vectors of features. It includes single words weighted by their TF-IDF scores, noun phrases, proper nouns, synsets from the Wordnet database and a database of semantic classes of verbs. A vector is computed for each pair of paragraphs which represented matches on the different features. Decision rules which were learned from data are used to classify each pair of text units either as similar or dissimilar. It places the most related paragraphs in the same theme.

After the identification of themes, the system enters into second stage i.e. information fusion. The main objective is to decide the sentences of a theme that is included in the summary. An algorithm is proposed that compares and intersects predicate argument structures of the phrases within each theme. It determines which are repeated enough to be included in the summary.

2.12Lexical Chain Method

Barzilay & Elhadad proposed Using Lexical Chains for Text Summarization [41]. They proposed a technique which produces a summary of an original text without requiring its full semantic interpretation but instead relying on a model of the topic progression in the text derived from the lexical chains. A new algorithm was proposed to compute lexical chains in a text as a result merging several robust knowledge sources such as WordNet thesaurus, a part-of-speech tagger, shallow parser for the identification of nominal groups and a segmentation algorithm. The summarization is defined in four steps: the original text is segmented, lexical chains are constructed, strong chains are identified and significant sentences are extracted.

2.13Latent Semantic Analysis

Gong & Liu proposed Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis [42]. They proposed two generic text summarization methods that create text summaries by ranking and extracting sentences from the original documents. Two methods are used.IR methods uses to rank sentence relevances while the LSA technique uses to identify semantically important sentences for summary creations. Sentences which are highly ranked and different from each other are selected to create a summary with a wider coverage of the documents main content and less redundancy. The performance is evaluated on the two summarization methods by comparing their summarization outputs with the manual summaries generated by three human evaluators.

2.14Lex Rank

Erkan & Radev proposed LexRank: GraphBased Lexical Centrality as Salience in Text Summarization [43]. The importance of the sentence is computed based on the concept of eigenvector centrality in a graph representation of sentences. A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of the sentences. It is found that the approach is quite insensitive to the noise in the data that may result from an imperfect topical clustering of documents. Generic extractive text summarization is produced based on multi-document.

2.15Graph-Based Summarization

Parveen and Strube proposed a graph based method for extractive single-document summarization which considers importance, non-redundancy and local coherence simultaneously. The input documents are represented by means of a bipartite graph consisting of sentence and entity nodes. The sentences are ranked based on the importance by applying a graph-based ranking algorithm to the graph and ensure non-redundancy and local coherence of the summary by means of an optimization step. The graph-based summarization technique achieves state-of-the-art results on DUC 2002 data.

2.16Hybrid Approach

Yang, Bu and Xia proposed Automatic Summarization for Chinese Text Using Affinity Propagation Clustering and Latent Semantic Analysis. The new approach is a hybrid approach which is based on Affinity Propagation (AP) and Latent Semantic Analysis (LSA).AP is a new clustering algorithm which takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars.LSA is a technique in vectorial semantics of analyzing relationships between a set of sentences.

III. Analysis of Text Summarization Techniques

Table Name

Techniques	Abstractive/ Extractive	Approaches	Domain Specific	Performance Behaviour
Naive Bayes Method [16][17][18]	Extractive	Statistical, Knowledge-based	Technical Documents	• Summaries which are 25% of the size of the average document,84% of the sentences are selected by the professionals. • Improvement of about 74% observed for smaller summary size
Position Method[20]		Knowledge-based	Newspaper Text	• 30% of the topic keyword are not mentioned directly in the text • Only 50% of the topic keywords are present in the title • The title including two most rewarding sentences contains 60% of the topic keywords
Hidden Markov Model[21][22]	Extractive	Knowledge-based, Computational	Associated Press, Financial Times, Los Angeles Times, Washington Post, Wall Street Journal, Philadelphia Inquirer, Federal Registry, Congressional Record, Short Stories	• F1 score range from 51 to 58.
Log Linear Model[23]	Extractive	Statistical	Technical Documents	
Neural Networks[24][25][26]	Extractive	Computational, Knowledge-based, Machine Learning	News Articles	• Out performs the standard baseline in the ROUGE-1 measure on over 70% of document set
Deep Natural Language Analysis Method[27][28][29]	Extractive	Linguistic, Computational, Knowledge-based	Japanese Editorial Articles, newspaper, technical papers	• A maximum of 74% of the most important sentences of the original text
Centroid-Based Summarization[31]		Computational, Knowledge-based, Statistical	News articles	
Multilingual Multi-Document Summarization [32][33][34]	_____	Machine Learning, Knowledge-based	News articles	• 68% of the sentence replacements improves summary
Graph Search and Matching		Computational, Knowledge-based		• F=32.36, $p < 0.05$, using analysis of variance F-test

[35][36][37]				
Topic-Driven Summarization and MMR[38]		Computational,	News Stories	• 70% accuracy on informative summaries
Abstraction and Information Fusion[39][40]	Extractive	Machine Learning, Statistical, Linguistic, Computational, Knowledge-Based	News Articles	• Recovers 39.7% of the similar pairs of paragraphs with 60% precision • Overall accuracy over both similar and dissimilar pairs is 97%
Lexical Chain Method[41]	Extractive	Linguistic, Computational	News Articles	• 2 summaries were constructed for a document: one at 10% length other at 20% length • 61% precision and 67% recall for 10% length of the document • 47% precision and 64% recall for 20% length of the document
Latent Semantic Analysis(LSA)[42]	Extractive	Computational	News stories	• The performance w.r.t. first summarizer are R= 0.52,P= 0.59,F=0.55 • The performance w.r.t. second summarizer are R= 0.53,P= 0.61,F=0.57
LexRank[43]	Extractive	Computational	News	• ROUGE 1 score is 0.3883 • 95% Confidence Interval is [0.3626,0.4139]
Graph Based summarization[44]	Extractive	Computational	Scientific Articles	• R-SU4 score is 0.121 w.r.t. to editors' summaries • R-SU4 score is 0.200 w.r.t. to authors' abstract
Hybrid Approach(LSA+Clustering)[45]	_______	Computational	News, Articles from web	• The performance are R=0.665,P=0.489,F_measure=0.548

IV. Conclusion

The increase in online information has brought a challenge and a need to develop efficient summarization systems. The summarization started with technical documents but its application is mostly found in news articles. It is found that most of the techniques had used extractive summarization to generate the summary.

This survey emphasizes on the approaches used to generate the summary. It is found that in case of single-document summarization most of the techniques used are statistical, knowledge-based and computational alongwith machine learning and linguistic whereas in case of multi-document summarization most of the techniques used are knowledge-based and computational alongwith statistical, machine learning and linguistic. Evaluation of the techniques are done with various measures.

References

[1] D. R. Radev, E. Hovy and K. McKeown, Introduction to the Special Issue on Summarization, *Computational Linguistics*, 28(4), 2002,399–408.

[2] I. Mani, *Automatic Summarization.* J. Benjamins Publ. Co. Amsterdam Philadelphia, 2001.

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

[4] G. DeJong, *An Overview of the FRUMP System.* In: W. Lehnert, M. Ringle(eds.) Strategies for Natural Language Processing, 149-176(Lawrence Erlbaum Associates,Publishers,1982).

[5] K. S. Jones and B. Endres-Niggemeyer,Automatic Summarizing, *Information Processing & Management*,31(5),1995,625-630.

[6] P. Over, H. Dang and D. Harman, DUC in Context, *Information Processing & Management*, 43, 2007, 1506-1520.

[7] K. Owczarzak and H. Dang,Overview of the TAC 2010 Summarization Track, *In: TAC 2010*, 2010,NIST.

[8] P. Baxendale, Machine-Made Index for Technical Literature - An Experiment, IBM Journal of Research Development,2(4), 1958, 354-361.

[9] H. P. Edmunson, New methods in automatic extracting, *Journal of the ACM*, 16(2), 1969, 264-285.

[10] C. Y. Lin and E. Hovy, Identify Topics by Position, *In Proc. of the 5th Conference on Applied NLP*, 1997.

[11] F. Liu, J. Flanigan, S. Thomson, N. Sadeh and N. A. Smith, Toward Abstractive Summarization Using Semantic Representations ,*Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 1077-1086,Denver,Colarado,(2015).©2015 Association for Computational Linguistics.

[12] M. Kageback, O. Mogren, N. Tahmasebi and D. Dubhashi,Extractive Summarization using Continuous Vector Space Models, *In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, 31–39, Gothenburg, Sweden, April 26-30 2014. © 2014 Association for Computational Linguistics.

[13] S. Suneetha, Automatic Text Summarization: The Current State of the Art, *International Journal of Science and Advanced Technology,* 1(9), 2011.

[14] R. Mihalcea, Language Independent Extractive Summarization, *In Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions, ACLdemo'05*,2005,49-52, Stroudsburg, PA, USA. Association for Computational Linguistics.

[15] H. Zha,Generic Summarization and Keyphrase Extraction using Mutual Reinforcement Principle and Sentence Clustering, *In Proceedings of the 25th Annual International ACM SIGR Conference on Research and Development in Information Retrieval*,SIGIR'02,2002,113-120,New York, NY,USA,ACM.ISBN 1-58113-561-0

[16] J. Kupiec, J. Pedersen and F. Chen, A trainable Document Summarizer, *In Proceedings SIGIR '95*, 1995, 68-73, New York, NY, USA.

[17] H. P. Edmundson, New Methods in Automatic Extracting, *Journal of the ACM*, 16(2), 1969,264-285.

[18] C. Aone, M. E. Okurowski, J. Gorlinsky and B. Larsen, *A Trainable Summarizer with Knowledge acquired from robust NLP Techniques*, In Mani, I. and Maybury, M. T., editors(Advances in Automatic Text Summarization,71-80, MIT Press1999).

[19] G. A. Miller,Wordnet: A Lexical Database for English, *Commun. ACM*, 38(11), 1995, 39-41.

[20] C. Y. Lin and E. Hovy,Identifying Topics by Position, *In: ANLC '97 Proceedings of the fifth conference on Applied Natural Language Processing* ,1997, 283-290.doi:10.3115/974557.974599

[21] J. M. Conroy and D. P. O' Leary, Text Summarization via Hidden Markov Models, *In: Proceedings of SIGIR '01*, 2001, 406-407, New York, NY, USA.

[22] L. R. Rabinder, A Tutorials on Hidden Markov Models and selected applications in Speech Recognition, *In: Proceedings of IEEE*, 1989,77(2), 257-286 .

[23] M. Osborne, Using Maximum Entropy for Sentence Extraction, *In: Proceedings of the ACL'02 Workshop on Automatic Summarization*, 2002, 1-8, Morristown, NJ, USA.

[24] A. Nenkova, Automatic Text Summarization of Newswire: Lessons learned from the Document Understanding Conference, *In: Proceedings of AAAI* 2005, Pittsburgh, USA.

[25] K. Svore, L. Vanderwende and C. Burges, Enhancing Single-Document Summarization by combining RankNet and third-party sources, *In: Proceedings of the EMNLP-CoNLL*, 2007,448-457.

[26] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton and G. Hullender, Learning to rank using gradient descent, *In: ICML '05Proceedings of the 22nd international conference on Machine learning*, 2005, 89-96, New York, NY, USA, ACM.

[27] R. Barzilay and M. Elhadad, Using lexical chains for text summarization, *In: Proceedings ISTS'97*, 1997.

[28] K. Ono, K. Sumita, and S. Miike, Abstract Generation Based on Rhetorical Structure Extraction, *In: Proceedings of Coling '94*, 1994, 344-348, Morristown, NJ, USA.

[29] D. Marcu,Improving Summarization through Rhetorical Parsing Tuning, *In: Proceedings of The Sixth Workshop on Very Large Corpora,* 1998, 206-215, Montreal, Canada.

[30] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, Multi-Document Summarization by Sentence Extraction, *In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, 4, 2000, 40-48.

[31] D. R. Radev, H. Jing, M. Stys and D. Tam, Centroid-based summarization of multiple documents, *Information Processing and Management*, 40,2004,919-938.

[32] D. K. Evans, K. McKeown, and J. L. Klavans, Similarity-based Multilingual Multi-Document Summarization,2005.https:// www.semanticscholar.org/paper/Similarity-based-Multilingual-MultiDocument-Evans-McKeown/90892be51eaf1ae17287 def7ba6e483fe3 5fc88e

[33] A. Siddharthan, Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs, *In: Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, 2002, 60–65, Philadelphia, USA.

[34] Hatzivassiloglou, J. L. Klavans, M. Holcombe, R. Barzilay, M. Y. Kan and K.R. McKeown,Simfinder: A flexible clustering tool for summarization, *In: NAACL'01 Automatic Summarization Workshop*, 2002.

[35] Mani and E. Bloedorn, Multi-document Summarization by Graph Search and Matching, *In: AAAI/IAAI,* 1997, 622-628.

[36] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson and M. Vilain, "MITRE: Description of the ALEMBIC System Used for MUC-6", *In: proceedings of the Sixth Message Understanding Conference (MUC-6)*, 1995, Columbia, Maryland.

[37] G. Krupka, SRA: Description of SRA System as Used for MUC-6, *In: Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 1995, Columbia, Maryland.

[38] J. Carbonell and J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, *In Proceedings of SIGIR '98*, 1998, 335-336, New York, NY, USA.

[39] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay and E. Eskin,Towards Multi-Document Summarization by reformulation: Progress and prospects, *In: AAAI/IAAI,* 1999, 453-460.

[40] R. Barzilay, K. McKeown, and M. Elhadad, Information fusion in the context of Multi-Document Summarization, *In: Proceedings of ACL '99*, 1999.

[41] R. Barzilay, and M. Elhadad, Using Lexical Chains for Text Summarization. www.aclweb.org/anthology/W97-0703,[accessed on 25.04.2018]

[42] Y. Gong, and X. Liu, Generic Text Summarization using Relevance Measure and Latent Semantic Analysis, *In: Proceedings of ACM SIGIR*,2001, New Orleans, USA.

[43] G. Erkan and D. Radev,LexRank: Graph-based Lexical Centrality as salience in Text Summarization, *Journal of Artificial Intelligence Research*, 22, 2004, 457–479.

[44] D. Parveen and M. Strube, Integrating importance, non-redundancy and coherence in graph-based extractive summarization, *In: Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, AAAI Press. 1298–1304.

[45] R. Yang, Z. Bu and Z. Xia, Automatic Summarization for Chinese Text Using Affinity Propagation Clustering and Latent Semantic Analysis, *In: Web Information Systems and Mining*, 2000, 543- 550.

Do ThiBich Ngoc. " A Study of Different Text Summarization Techniques Based on Input" IOSR Journal of Computer Engineering (IOSR-JCE) 21.4 (2019): 11-18.