International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 3 No. 3 July 2013, pp. 857-865

http://www.issr-journals.org/ijias/

Literature Review of Automatic Single Document Text Summarization Using NLP

Md. Majharul Haque[1], Suraiya Pervin[1], and Zerina Begum[2]

[1]Department of Computer Science & Engineering,
University of Dhaka,
Dhaka, Bangladesh

[2]Institute of Information Technology,
University of Dhaka,
Dhaka, Bangladesh

ABSTRACT: In the time of overloaded online information, automatic text summarization is especially demanded for salient information retrieval from huge amount electronic text. For the blessing of World Wide Web, the mass of data is now enormous in its volume. Researchers realized this fact from various aspects and tried to generate an automatic abstract of the gigantic body of data from the commencement of the last half century. Numerous ways are there for characterizing different approaches to passage recapitulation: extractive and abstractive from single or compound document, objective of content abridgement, characteristic of text summarization, level of processing from superficial to profound and sort of article's content. A significant précis is very much helpful in our day to day life which can save valuable time. The investigation was at first commenced naively on single document abstraction. In this paper, automatic single document text summarization task is addressed and different methodologies of various researchers are discussed from the very beginning of this research to this modern age. This literature review intends to observe the trends of abstraction procedure using natural language processing. Also some promising approaches are indicated and particular concentration is dedicated for the categorization of diversified methods from raw level to similar like human professionals, so that in future one can get precious direction for further analysis.

KEYWORDS: Information retrieval, World Wide Web, electronic text, automatic abstract, human professionals.

1 INTRODUCTION

Outcome of the information retrieval becomes necessary for user to find out concrete information for the abstraction because of the stridently escalation of data on the web. Internet is widely used by people to come across information using proficient information retrieval (IR) tools, such as Google, Yahoo, AltaVista, etc., where findings are abundant. In most of the cases, users feel bore with the very tedious and time consuming job to reveal the main gist of the outcome of the IR. Academics and researchers are very much benefitted by using automatic text summarization system as a tool to lessen the amount of time spent manually extracting the chief thoughts from large documents. In addition to the above reason, automatic text summarization also provides its users with numerous benefits as well as:

(i) Increase efficiency of other researches to choose documents/information from search engines' output, which usually contain an excess amount of replicated information.

(ii) Solve the limitation of information presentation on small communication devices such as PDA and mobile phone etc., which is able to display abridged version of the full document.

(iii) The running time of machine for translation is significantly reduced if a short version of text is given.

Corresponding Author: Md. Majharul Haque (mazharul_13@yahoo.com)

American research libraries spawned the initial interest in automatic text shortening during sixties [1]. A considerable body of research over the last sixty years has explored different levels of analysis of text, to help determine what information in the text is salient for a given summarization task [1]. Radev et al [2] in 2002 defined a summary as a text that is engendered from one or more texts, which conveys essential information of the original text(s), and that is no longer than half of the original text(s) and usually notably less than that. Simply a summary text is a derivative of a source text condensed by selection and/or generalization on important content [3]. A large document is entered into the computer and a recapitulated content is returned, which is a non redundant extract from the original passage. Automatic text summarization can be classified into single document text summarization and multiple documents text summarization [4] as in figure 1. This paper focused on single document text summarization as single document was the target from the commencement of such research on automatic abstraction.

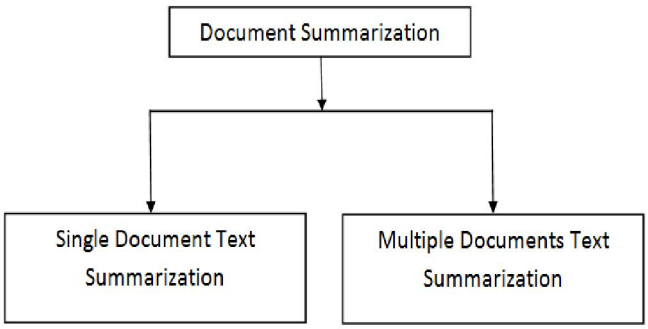

Fig. 1. Classification of document summarization

Various methods that utilize passage categorization, such as neural networks, semantic graphs, decision trees, regression models, fuzzy logic and swarm intelligence, etc. are involved on the study on finding crucial portion of text. The objective of this paper is to present a comprehensive literature review on automatic single document text summarization using natural language processing and investigate the movement of document abridgement.

The rest of the paper is organized as follows. Section 2 briefly describes about Natural Language Processing. Section 3 presents a comprehensive literature review about different procedures of automatic single document text summarization. At a glance comparison among the various techniques is depicted in section 4. Section 5 turns conclusion with a brief about this paper.

2 NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a field of computer science, artificial intelligence and linguistics as all those specified arena brought it into play. Generally it deals with the interactions between machines and human languages that accomplish task on analyzing, understanding and generating the language, which human use naturally in order to interact with computers in both oral and written contexts using natural human languages instead of computer languages [5]. It is an interdisciplinary area based on versatile arena of study including computer engineering, which provides methods for model illustration, algorithm devise and accomplishment; linguistics, which categorizes linguistic forms and practices; mathematics, which provides formal models and methods; psychology, which studies models and theories of human behavior; statistics, which offers procedures for predicting measures based on sample records; and biology, which travels around the underlying architecture of linguistic processes in the human brain [6].

3 REVIEW ON AUTOMATIC SINGLE DOCUMENT TEXT SUMMARIZATION

In the very beginning of the research in the arena of launching artificial intelligence to generate abridged version of a large document, disclosed the paradigms for extracting salient features.

Automatic text summarization process model can be divided into three steps [7] as in figure 2: (1) in the preprocessing step source text interpretation to source text representation, (2) source representation transform to summary text representation with an algorithm and (3) in the final step, summary text generation from summary representation.

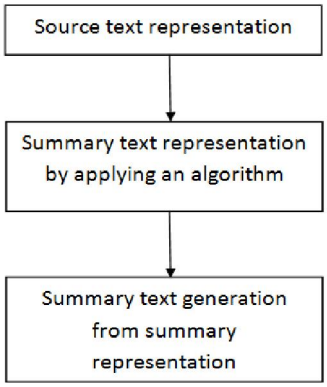

Fig. 2. Process flow of automatic text summarization

Numerous types of research work have been accomplished by various researchers where we can be familiar with multiple types of way of summary generation from single document text. In the following part of this section, methodologies of several researchers that incorporated this topic are depicted in brief, from pioneering works to the era of modern science where the thoughts of similar like human professionals' abstraction techniques are being explored.

3.1 PIONEERING WORKS

The voice over automatic i.e. computerized abstraction initiated around sixty years ago. As accomplishment of automatic passage summarizer was often cited in the oldest publication in 1958 by H. P. Luhn [8]. This method is based on the word frequency and clearly emphasized that the most frequent words represent the most important concepts of the text. In the next step, frequencies are used to score and decide on sentences to be extracted. In this paper, application of machine is emphasized and expressed that because of the absence of the variations of human capabilities and orientation, auto-abstracts have a high degree of reliability, consistency, and stability, as they are the product of statistical analysis of the author's own words. This paper is worthy of being appreciated as it is almost the earliest paper in this arena of automatic text abstraction [9]. Besides this, the proposed method mostly works on scientific article and journal paper.

P. B. Baxendale [10] in 1958 investigated machine techniques for reducing scientific credentials to their important sharp indices. Human scanning patterns were tried to be simulated here for selecting topic sentences special phrases. It was declared in this paper that sentence' position and containing certain cue-words (i.e., words like 'crucial' or 'pertinent' etc.) or the word exist in the heading are special indicator for being in the salient parts of the document.

G. J. Rath et al [11] in 1961 in their research scrutinized about four types of lexical indicators of content to determine which one is the best for detecting relevant document from repository and to answer a set of question. After their experiment, it was claimed that utilizing complete text is better than only abstract for answering question. But for distinguishing relevant document only abstract is enough. They also proclaimed that purely statistical method of producing extracts was suspected of being inadequate, and hence other methods were sought.

H. P. Edmundson [12] in 1969 accomplished a notable progress after ten years of the beginning of the research on text recapitulation. He described three additional methods with the standard keyword method, disregarding the very high frequency common words to determine the sentences' weight. Those are:

(i) Cue Method: The hypothesis of this technique is that the presence or absence of certain cue words will compute the significance of a sentence.

(ii) Title Method: The weight of a sentence is calculated as a sum of all the content words materializing in the title, headings and sub-headings of a text.

(iii) Location Method: Here relevance is assumed on the basis of location, expressed that sentences taking place in initial position of paragraphs have a higher probability of being pertinent.

The result was very fruitful and assumed that by using a combination of these three schemes the best correlation between the automatic and human-made abstracts can be achieved. This paper emphasized on indicative abstracts rather than on the production of informative abstracts. So, if a user doesn't know about the document to be summarized earlier, can't get summary through this proposed methodology.

3.2 ALGEBRAIC METHODS

Julian Kupiec et al [13] in 1995 explored an innovative algebraic method. Using naïve-bayes classifier, this classification procedure sorts out each sentence as worthy of being in the abstract or not. A trained classifier is generated at first, where the authors used a corpus of 188 couples of complete documents/summaries. The distinguishing features used uppercase words, length of sentence, structure of phrase, position in paragraph besides word frequency.

ChinYew Lin et al [14] in 1997 in their paper addressed the difficulty of identifying probable topics of texts by their location in the corpus, and scored sentences by its position in given text. They considered techniques of tailoring the position method towards optimality over a variety and how it can be estimated for efficiency. The idea that texts generally follow a predictable discourse structure, and that the sentences of greater topic centrality tend to occur in certain specifiable locations (e.g. title, abstracts etc), arises the position method for topic recognition. Here also predicted that discourse structure significantly varies over domains, for which position method is a bit tough. Given a text T and the list of topic keywords, each sentence of the text is labeled with the paragraph number and the sentence number. With the number of 2097 documents in this method, the result illustrated that the first and the last 50 positions fully cover majority of the text.

Eduard Hovy et al [15] in 1999 attempted to create a robust automated text summarization system and named it SUMMARIST. Instead of irregular term counting, SUMMARIST combines symbolic world knowledge (embodied in WordNet, dictionaries, and similar resources) with strong NLP processing (using IR and statistical method) to resolute concept relevance. Their procedure based on the following equation:

Summarization = topic identification + interpretation + summary generation

S. P. Yong et al [16] in 2005 introduced an automatic text summarization system that incorporates learning ability by combining a statistical approach, keywords extraction and neural network with unsupervised learning. It was claimed that their proposed technique can extract 83.03% of significant contents. The procedure is made up of three steps, as follows:

(i) Text pre-processing subsystem: two pre-processing methods are applied, one is stop words removal to remove words like “the”, “a”, “can” and “will” etc. and another one is stemming to convert each word to its stem by eliminating suffixes and prefixes.

(ii) Keywords extraction subsystem: from a group of sample heterogeneous text documents, main features of each word are determined by computing term frequency tf(w, s) as the number of times that the word w appears in the sentences s, and the inverse sentence frequency isf(w) is calculated as the number of sentences in which the word w occurs. From the generated tf-isf(w, s) matrix, most frequent terms are listed as keyword in the text to be summarized.

(iii) Summary production subsystem: in the final step the system chooses sentences, which contain the keywords, as part of the summary. It is suggested to run through another round of stop words checking procedure before selecting sentences for being ensured that there is no stop word is working as keyword.

3.3 PROCEDURES BASED ON TEXT CORRELATION

Previously discussed methods used statistical probability for choosing a sentence for generating abridged version of text. But those panoramas are fully dumb about the cohesion of sentences with each other. So, the relations between concepts in a text can’t be captured using extractive methods. If a sentence is extracted which contains link to any previous context then the summary will be difficult to understand [17]. In those circumstances various research work done for exploring the cohesive properties to comprise relations among expressions of text.

M. A. K. Halliday et al [18] in 1976 performed the first research to explore the degree of subjectivity of two aspects of the lexical cohesion perceived by readers of text: the word cluster (lexical chains) that are formed and the lexical semantics relations that are perceived between the words. Building blocks of lexical cohesion, cohesive harmony, and the concept of patterns of lexical affinity are the lexical semantic relations. The linguistic study was emphasized here and tried to forms inter sentence groups of related words, and the original outlook of them was very wide and universal; but there had to be a recognizable relation between two words as the only decisive factor.

Ruqaiya Hasan [19] in 1984 extended the investigation to include the concept of cohesive harmony, which adds lexico-grammatical structure to word groups by first dividing them into two types: (i) identity-of-reference chains that combine reference and lexical cohesion, (ii) similarity chains that used only classical relations, and linking these chains with grammatical intra-sentence relations.

Rhetorical Structure Theory (RST) of text summarization can also be included into the groups of methods based on text correlation as the text coherence is attributed principally to the presence of rhetorical relation. RST was developed during 1980s by researchers in natural language generation, which models the discourse structure of a text by means of a hierarchical tree diagram [20].

William C. Mann et al [21] in 1988 established a new definitional foundation of RST and also scrutinized three claims: the predominance of nucleus/satellite structural patterns, the functional basis of hierarchy, and the communicative role of text structure. Various kinds of consequences of RST have also been reviewed in this paper. A quantity of allegorical relations, which hook up together text units, forms a hierarchical structure called RST tree. The relations tie collectively a nucleus as central to the author's objective and a satellite as subsidiary parts. Texts in terminal nodes of RST tree are supposed to be encoded and non-terminal nodes represent contiguous text spans, whose sibling spans are joined by discourse relations. At last the nucleus and some tightly connected terminal nodes are selected as the principal theme of passage.

Jane Morris et al [22] in 1991 utilized cohesion chains and Regina Barzilay et al [23] in 1997 exploited lexical chains to characterize contents of a document. A perception in the document is symbolized by a sequence of associated words in these representations. The sequence is a list of words that confines a segment of the consistent structure of the document and is generated as independent of the grammatical formation of the text. The procedure usually launch from a set of words in the heading of the document to construct lexical chains, adjoining of words that have similar meaning or are related to the title. WordNet thesaurus has been used for determining cohesive relations between terms (i.e., repetition, synonymy, antonymy, hypernymy, and homonymy). In the next step, scores are calculated for sentences considered to be important by the previous step and those scores will be used to generate the final summary.

Branimir Boguraev et al [24] in 1997 proposed a novel technique for saliency-based content characterization of text document. A procedure was offered, which uses a salience feature as the basis for a "ranking by importance" of an unstructured referent set, and ultimately for topic stamp identification. Co-reference resolution system is used here, which is a process of determining whether two expressions in natural language refer to the identical entity. Then a threshold value is calculated and where the occurrences of frequently mentioned objects overcome the value is included into the summary.

Li Chengcheng [25] in 2010 presented an effective method using RST for successful automatic text summarization, which is based on natural language generation. The rhetoric structure of the text is extracted with a compound that relates the sentences in this theory. Here the summarization is accomplished using the nodes i.e. nucleus, those are given weights based on script based analysis. Then the entire text is divided into individual sentences based on commas, quotes, semicolons and punctuation marks exist in the sentences. This is then done into a graph, deletes the unimportant sentences and then summarizes the entire text. Principal scheme of this procedure is as follows: (i) analyzing the candidate sentence, (ii) discover the rhetoric relations and (iii) forming the essential part of sentence constructive for ultimate recapitulation.

3.4 METHODS CLOSE TO HUMAN ABSTRACTION CONCEPT

Artificial Intelligence is much matured in this modern age, as the application of this is noticeable in various regions such as robotics, machine learning and knowledge based system etc. Automated summary is also the grace of Artificial Intelligence through natural language processing. But, still there is a qualitative dissimilarity between synopsis generated by existing automated summarizers and the abstracts written by human. Extraction is a common technique utilized by most of the summarizers, which may be ambiguous or incoherent to the original abstract. In spite of the existence of some shortcomings, a number of methods have started to emerge lately with either sentence compressing capability or re-producing technique.

Hongyan Jing [26] in 2000 presented a novel sentence lessening method, which eliminate extraneous phrases from the extracted sentences that are chosen for abstraction purpose. A parse tree is generated at first in this procedure and then grammar is checked to decide which terms of a sentence must not be removed to keep its' structure accurate. Then it finds the sentences that are closely related to the main topic by using corpus evidence, consisting of sentences compressed by human and original sources, delete unnecessary sub-tree from parse tree for producing the final outline.

Kevin Knight et al [27] in 2000 proposed an innovative method by incorporating the procedure of regenerating sentence for coherent abstract, rather than just extracting from original source. A training corpus is used here, which is a collection of newspaper articles paired with human written abstracts. Their first goal is to rewrite a compressed version of given input parse tree. The rewriting process starts with an empty Stack and Input List that contains the sequence of words subsumed by the given large parse tree. Four types of operations are incorporated here: SHIFT operations move the first word from the input list into the stack; REDUCE operations pop the top syntactic trees, combine them and again push them to the stack; DROP operations are used to remove from the input list that has already been compressed; ASSIGN TYPE operations change label of trees at the top of the stack. The procedure ends when the input list is vacant and the stack contains only one tree.

After that an in-order traversal of the leaves of this tree generates the compressed form of the sentences those are given as input.

K. McKeown et al [28] in 2000 stated that there is a significant difference between the summaries produced by current automatic system and the human professionals. Because automatic summarizer cannot always recognize key topics of an article and automatic procedure has no robust text generation method. This paper presented a "Cut and Paste" strategy for text summarization that derived from an analysis of human written abstract. "Cut and Paste" method not just extracts sentences but smooth the extracted sentences by editing them which mainly cutting phrases and pasting them together in a novel ways. It was stated that generated summary bears a resemblance to the human summarization process more than extraction does. Six operations were defined to transform chosen sentences from an article into the corresponding summary sentences in its human written abstracts: (i) sentence reduction, (ii) sentence combination, (iii) syntactic transformation, (iv) lexical paraphrasing, (v) generalization and specification, and (vi) reordering. An evaluation was done for this procedure with 50 human-written abstracts, consisting of 305 sentences in total and claimed that 93.8% of sentences were correctly decomposed.

4 COMPARISON AMONG THE TECHNIQUES

At a glance comparison among the discussed techniques of single document text summarization has been shown in table 1:

Table 1. Comparison Among the Techniques of Single Document Text Summarization

#	Researcher(s), Year, Reference	Category	Technique	Basis of procedure
1	H. P. Luhn, 1958, [8]	Pioneering works	Word frequency	Frequent words represent the most important concepts of the text.
2	P. B. Baxendale, 1958, [10]	Pioneering works	Position in text	Sentence position and containing certain cue-words or the word exist in the heading are special indicator for being the salient parts of the document.
3	G. J. Rath, 1961, [11]	Pioneering works	Lexical indicator	Lexical indicators of content could be utilized best by subjects to determine relevant from irrelevant documents.
4	H. P. Edmundson, 1969, [12]	Pioneering works	Cue words and heading	The sentences that contain specific cue words, words those are exist in title and heading or sub-heading, are significant to be selected for summary.
5	Julian Kupiec, 1995, [13]	Algebraic method	Naïve-Bayes classifier	Uppercase words, length of sentence, structure of phrase, position in paragraph with word frequency are the distinguishable features of sentence.
6	ChinYew Lin, 1997, [14]	Algebraic method	Position in text	Texts generally follow a predictable discourse structure, and that the sentences of greater topic centrality tend to occur in certain specifiable locations (e.g. title, abstracts, etc).
7	Eduard Hovy, 1999, [15]	Algebraic method	Symbolic world knowledge	Instead of irregular term counting, this method combines symbolic world knowledge (embodied in WordNet, dictionaries, and similar resources) with strong NLP processing (using IR and statistical method) to resolute concept relevance.
8	S. P. Yong, 2005, [16]	Algebraic method	Neural network	Using neural based system this method generate summary using three sub system such as: Summarization = Text pre-processing sub-system + Keywords extraction sub-system + Summary production sub-system.
9	M. A. K. Halliday, 1976, [18]	Text correlation	Lexical cohesion	Building blocks of lexical cohesion, cohesive harmony, and the concept of patterns of lexical affinity are the lexical semantic relations.
10	Ruqaiya Hasan, 1984, [19]	Text correlation	Lexical cohesion	Identity-of-reference chains and similarity chains, that linking with grammatical intra-sentence relations.
11	William C. Mann, 1988, [21]	Text correlation	Rhetorical Structure Theory(RST)	Texts in terminal nodes of RST tree are supposed to be encoded and non-terminal nodes represent contiguous text spans.
12	Jane Morris, 1991, [22]	Text correlation	Cohesion chains	A perception in the document is symbolized by a sequence of associated words in the representation.
13	Regina Barzilay, 1997, [23]	Text correlation	Lexical chains	The procedure usually launch from a set of words in the heading of the document to construct lexical chains, adjoining of words that have similar meaning or are related to the title.
14	Branimir Boguraev, 1997, [24]	Text correlation	Saliency-based content characterization	This method uses a salience feature as the basis for a "ranking by importance" of an unstructured referent set.
15	Li Chengcheng, 2010, [25]	Text correlation	Rhetorical Structure Theory(RST)	Principal scheme of this procedure is as follows: (i) analyzing the candidate sentence, (ii) discover the rhetoric relations and (iii) forming the essential part of sentence constructive for ultimate recapitulation.
16	Hongyan Jing, 2000, [26]	Human abstraction concept	Lessening method	Find the sentences that are closely related to the main topic and eliminate extraneous phrases from the extracted sentences that are chosen for abstraction purpose.
17	Kevin Knight, 2000, [27]	Human abstraction concept	Regenerating sentence	Rewrite a compressed version of given input parse tree. The rewriting process starts with an empty Stack and Input List that contains the sequence of words subsumed by the given large parse tree.
18	Hongyan Jing, 2000, [28]	Human abstraction concept	Cut and paste	Smooth the extracted sentences by editing them which mainly cutting phrases and pasting them together after reduction, combination, transformation, specification, reordering, etc.

5 CONCLUSION

In this paper the concepts of single document text summarization that categorize different approaches in this arena have been reviewed. This literature review emphasizes on extractive approaches for text summarization. Deriving the classification of the automatic text abstraction procedure has also been attempted. Recent trend in summarization system that comes from novice procedure to resemble with human written summary has been scrutinized here. We believe that the study of document summarization is a productive region for further research, both by linguists performing text analysis and by computational linguists trying to create techniques to produce summaries conforming to one or more of the characteristics listed above. Around 18 papers have been briefly discussed and various key topics from other historical publication relevant with text abstraction have been analyzed here from 1958 to 2010. There exist some other procedures similar with those briefed in this paper, the discussion of which has not been included here as it will be a large corpus. Nowadays, in the age of computer and internet, information is found from various sources about a single topic, so the research of multi-document text abstraction is a burning issue. After all it is expected that any researchers can get help from this literature review for better understanding of different types of summary generation techniques even for multi-document text abridgement. It will also assist for better perception of the diversified sorts of abstraction procedure, which will help to construction of new formula and systems that significantly serve the various principle of summarization in broad-spectrum.

REFERENCES

[1] Martin Hassel, "Evaluation of Automatic Text Summarization - A practical implementation," Licentiate thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden, 2004. [Online] Available: http://www.csc.kth.se/~xmartin/papers/licthesis_xmartin_notrims.pdf (May 21, 2013)

[2] Dragomir R. Radev, Eduard Hovy and Kathleen McKeown, "Introduction to the special issue on summarization," *Journal of Computational Linguistics*, vol. 28, no. 4, pp. 399-408, December 2002.

[3] Abdelwahab Hamou-Lhadj and Timothy Lethbridge, "Summarizing the Content of Large Traces to Facilitate the Understanding of the Behaviour of a Software System," Proceedings of the 14th IEEE International Conference on Program Comprehension, pp. 181-190, 2006.

[4] S. Suneetha, "Automatic Text Summarization: The Current State of the art," *International Journal of Science and Advanced Technology (ISSN 2221-8386)*, vol. 1, no. 9, pp. 283-293, November 2011.

[5] Eugene Charniak and Drew McDermott, Introduction to artificial intelligence, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1985.

[6] Bill Z. Manaris, Natural Language Processing: A Human-Computer Interaction Perspective, University of Southwestern Louisiana, Academic Press, New York, vol. 47, pp. 1-66, 1998.

[7] Karen Sparck Jones, Automatic summarizing: factors and directions, In: Inderjeet Mani and Mark T. Maybury (Eds.), Advances in Automatic Text Summarization, Cambridge MA: MIT Press, pp. 1-12, 1999.

[8] Hans P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159-165, 1958.

[9] Dipanjan Das and André F. T. Martins, "A survey on Automatic Text Summarization," Language Technologies Institute, Carnegie Mellon University, 2007.
[Online] Available: http://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf (May 21, 2013)

[10] P. B. Baxendale, "Machine-made Index for Technical Literature -An Experiment," *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354-361, October 1958.

[11] G. J. Rath, A. Resnick and T. R. Savage, "Comparisons of four types of lexical indicators of content," *Journal of the American Society for Information Science and Technology*, vol. 12, no. 2, pp. 126-130, April 1961.

[12] H. P. Edmundson, "New Methods in Automatic Extracting," *Journal of the Association for Computing Machinery*, vol. 16, no. 2, pp. 264-285, April 1969.

[13] Julian Kupiec, Jan Pedersen and Francine Chen, "A Trainable Document Summarizer," Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 68-73, 1995.

[14] Chin-Yew Lin and Eduard Hovy, "Identifying Topics by Position," In Proceedings of the Fifth conference on Applied natural language processing, San Francisco, pp. 283-290, 1997.

[15] Eduard Hovy and Chin-Yew Lin, Automated Text Summarization in SUMMARIST, In: Inderjeet Mani and Mark T. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, chapter 8, pp. 18-24, 1999.

[16] S. P. Yong, A. I. Z. Abidin and Y. Y. Chen, "A Neural Based Text Summarization System," 6th International Conference of Data Mining, pp. 45-50, 2005.

[17] Karel Jezek and Josef Steinberger, "Automatic Text Summarization (The state of the art 2007 and new challenges)," Znalosti, pp. 1-12, 2008.
[18] M. A. K. Halliday and Ruqaiya Hasan, Cohesion in English, Longman, London, 1976.
[19] Ruqaiya Hasan, Coherence and Cohesive Harmony, In: Flood James (Ed.), Understanding Reading Comprehension: Cognition, Language and the Structure of Prose. Newark, Delaware: International Reading Association, pp. 181-219, 1984.
[20] William C. Mann and Sandra A. Thompson, Relational Propositions in Discourse, Defense Technical Information Center, Information Sciences Institute, Marina del Rey, 1983.
[Online] Available: http://www.tandfonline.com/doi/abs/10.1080/01638538609544632?journalCode=hdsp20#preview
[21] William C. Mann, Sandra A. Thompson, "Rhetorical Structure Theory: Toward a functional theory of text organization," Text, vol. 8, pp. 243-281, 1988.
[Online] Available: http://www.cis.upenn.edu/~nenkova/Courses/cis700-2/rst.pdf (May 21, 2013)
[22] Jane Morris and Graeme Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text," *Journal of Computational Linguistics*, vol. 17, no. 1, pp. 21-48, March 1991.
[23] Regina Barzilay and Michael Elhadad, "Using Lexical Chains for Text Summarization," In Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, Association of Computational Linguistics, pp. 10-17, 1997.
[24] Branimir Boguraev and Christopher Kennedy, "Salience-based Content Characterization of Text Documents," In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.
[25] Li Chengcheng, "Automatic Text Summarization Based On Rhetorical Structure Theory," International Conference on Computer Application and System Modeling (ICCASM), vol. 13, pp. 595-598, October 2010.
[26] Hongyan Jing, "Sentence Reduction for Automatic Text Summarization," In Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, USA, pp. 310-315, 2000.
[27] Kevin Knight and Daniel Marcu, "Statistics-Based Summarization Step One: Sentence Compression," In Proceeding of the 17th National Conference of the American Association for Artificial Intelligence, pp. 703-710, 2000.
[28] Hongyan Jing and Kathleen R. McKeown, "Cut and Paste Based Text Summarization," In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, USA, pp. 178-185, 2000.