International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

A Survey on Various Methodologies of Automatic Text Summarization

Rahul Lahkar
Department of Information Technology
Gauhati University
Guwahati, Assam, India

Anup Kumar Barman
Department of Information Technology
Gauhati University
Guwahati, Assam, India

***Abstract*— Conversion of text-to-text, to generate summary has been a key research area now a days. Automatic text summarization reduces human effort in generating summary from text document(s) with the help of computer program. Various approaches, methods and systems have been suggested and developed so far till date. This survey focuses on some of the existing techniques of statistical document summarization as well as summarization using semantic approaches to deal with the improvements that can be done for Extractive Text Summarization.**

Keywords*— Extraction based summarization, Abstraction based summarization, pronominal resolution, POS tagging, Semantic Roll Labeling.

I. INTRODUCTION

When a text extracted or generated which is an important portion of an original text document(s), which conveys the information carried by the original text(s), can be called as a summary of that original text(s). When this is done automatically, i.e. with the help of a computer program then we call this as automatic text summarization. In short, a summary should preserve the essence of the document which helps in finding the relevant information quickly. *Radev et al, 2002* [1] proposed a definition of a summary as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". The definition implies that summaries may be produced from both single and multiple documents and it should be shorter enough with meaningful information conveyed from the original text(s). So depending on the input, automatic text summarization has been classified into two processes. If the system takes only one document as input and produce summary of that document then this process is called as Single Document Summarization whereas Multiple Document Summarization is when the input is multiple-documents of same type.

II. TYPES OF SUMMARIZATION

To achieve automatic summarization, there are two types of summarization techniques followed. These are Extraction based Text Summarization and Abstraction based Text Summarization.

A. Extraction Based Summarization

The main aim of Extractive summaries is to point out the most important regions (words, sentences, paragraphs etc.) from the input source document(s). Summaries produced by Extraction method contain several concatenated sentences taken exactly as they appear in the documents being summarized. For each sentence, in extractive summarization method a decision is made whether or not it will be extracted to be included in the summery. For example, Search engines typically use Extractive summary generation methods to generate summaries from web page(s). Various techniques have been proposed so far by using logical and mathematical formulations to score the regions and pick them out (having highest score) from the text to be in the summary. In short, sentence extraction can be imagined as; it works as a text filter, which allows only important sentences to pass through it. Most of the summarization research is on Extractive summarization as it is easier to implement. Luhn [2] 1958, Edmondson [3] 1969, Barzilay & Elhadad [11] 1997, Marcu [12] 1997, Summarist [13] 1998, FociSum [14] 1999, chen & lin [15] 2000, Copeck et al [16] 2002, Newsblaster [17] 2002, CATS [18] 2005 are some systems which use Extraction based methods to generate summary.

B. Abstraction Based Summarization

Human beings generally write abstractive summaries. After reading a text, Human beings have the intelligence to understand the topic and write a very short summary in their own way generating their own sentences without losing any important information. But for a machine, it is a challenging task to generate abstractive summaries. So, it can be said that the goal of abstraction based summarization is to generate a summery containing new sentences (like human beings do) which are grammatically correct, using advanced natural language generation techniques. To do so, it requires understanding the topic of the original text. Abstractive summary generation is relatively harder as it needs semantic understanding of the text to be fed into the Natural Language Generation system. Sentence Synthesis being the major problem here; gives rise to incoherence in the generated summary, as it is not a well-developed field yet. MultiGen [19] 1999, Cut & Paste [20] 2001, sumUM [21] 2008 are some systems which generate abstractive summaries.

III. APPROACHES TO SUMMARIZATION

To extract, or to generate a summary from a document(s), there are various methods or approaches that are being followed. Some methods use word frequency counts, sentence positions, extraction of headings, detection of cue phrases etc., without going deep into the meaning of the sentences of the document. There are some methods which use to identify the topic of the input text or find the meaning of the sentences, to generate summaries. This type

of summarization methodology gets close to the quality summary produced by human beings. Human beings understand the meaning of each sentence and this understanding helps in writing the quality abstractive summary, also very compact in size compared to the original text.

Methods which use the frequency counts of words, sentences, phrases, sentence positions etc. can be said as *Statistical Methods* and methods which try to find the meaning of the sentences to generate summary can be termed as *Semantic Methods.* Semantic approaches can go deep into the document to find the meaning as well as relations among the sentences. Finding relations among the sentences helps to reduce repeated information, this can help in generation of quality compact summary. Semantic based summarization uses the help of WordNet[9], Part-Of-Speech tagger, Named Entity tagger to achieve semantic understanding of the text.

A. Statistical Methods

The earliest work that was known to be done in 1958 by **Hans Peter Luhn**, was for single document summarization. Luhn, in his paper "*The automatic creation of literature abstracts*" [2] described his work as well as recharge done in text summarization in those days at IBM. In his work, he proposed that words having relatively high occurrences in the text are significant. That means he used statistical approach to count the occurrences of words and based on that he extracted sentences with higher word frequency and phrase frequency. Sentences with highest score extracted to present as the abstract of the document.

Baxendale [10] in the year of 1958 also performed some work on summarization at IBM using the concept of *sentence position*, the concept being-sentences of first paragraphs are more important than the sentences of last paragraphs. The concept of this positional feature can play a great role in extractive summarization.

H. P Edmondson in 1969 [3] used a corpus-based methodology and put more interest on sentence locations and *heading words* or titles of the document. According to him, position of a sentence, phrase can a play a significant role in finding out the important sentences as headings or sentences in the beginning as well as sentences in the last of a text can convey more information. Techniques used for summarization are *word frequency*, *cue phrases* (presence of some significant words like significant, certainly, important, hardly etc.*)*, *title and heading words* and *sentence location*.

Kupiec, Pedersen, and Chen in 1995 [4] built a statistical summarization system which extracts important sentences deploying the Bayesian classification algorithm for summarization. The trainable system uses some discrete feature sets to classify sentences as important and unimportant. The feature sets are: Sentence length cut-off feature, thematic word feature, Fixed-phrase feature, paragraph feature and uppercase word feature.

B. Semantics Based Methods

Divyanshu Bhartiyaa and Ashudeep Singh [5] developed a summarization system based on *semantic approach* with the use of WordNet. The approach begins with *Anaphora resolution* or *Pronominal resolution* followed by *Part-of-Speech tagging* and *Semantic Roll Labeling.*

Pronominal Resolution

Pronominal resolution is used to avoid referenced structures. *As for example, consider this sentence–*

John helped Mary[1]. She was happy for the help provided by him[2]. [5]

Here sentence 2 is more informative than 1. But with only sentence 2, generated summary will not make any sense and will be incomprehensible. However, by performing *pronominal resolution* to this sentence we can get-

John helped Mary. **Mary** was happy for the help provided by **John**.

The system takes a document as input and performs *pronominal resolution* to it. It helps to form chains in the document and resolves the pronouns with their respective subjects.

Part-of-Speech tagging

Part-of-Speech (POS) tagging is the task of identifying the part-of-speeches (Noun, Pronoun, Verb, Adverb, Adjective, etc.) present in sentences. In this system Stanford's implemented POS tagger is used to identify the grammar tagging. POS tagging forms the basis of Semantic Roll Labeling.

Semantic Roll Labeling

Semantic Roll Labeling (SRL) is a shallow semantic parsing technique in NLP by which we can detect predicates associated with a verb in a sentence. It helps finding the meaning of a sentence along with actions associated with the sentence. SRL is analogical to a function having certain parameters where each function can be considered as a verb corresponding to an action. Each action is associated with an agent and a theme, the parameters of the function can be considered as the agent and themes. Each verb is associated with modifiers like temporal, locational or an adverb.

If a sentence is represented by the following pattern [5], <Agent> <action> <theme> <modifiers>, then the sentence can be translated as F (arg1, arg2 …argN) where F *is the <action> and <arg1>…<argN> are* the <agent>, <theme> and <modifiers> respectively.

Example: [ARG0 John] helped [ARG1 Mary];

The system uses software SENNA to find SRL and PropBank Annotation to form frames from a sentence.

Example: *"Mr Bush met him privately, in the White House, on Thursday."* [5]

Here,

Relation: met

Arg0: Mr Bush

Arg1: him

ArgM-MNR: privately

ArgM-LOC: in the White House

ArgM-TMP: on Thursday.

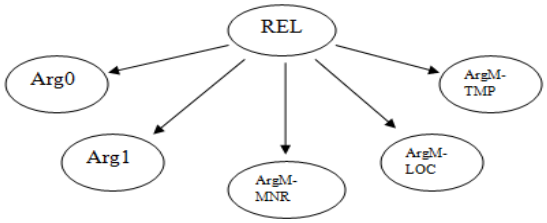

Figure1. By SRL, the structure of a frame [5]

After framing is done for all the sentences, it can be said that the document is now decomposed into some frames. After this, WordNet is used to find the hyponym synsets and hypernym synsets. This use of WordNet finds to capture the presence of repeated data or redundant data. The synsets created by WordNet is used to find textual entailment, i.e. finding the match between a frame's arguments' synsets and another frame's arguments' synsets. The matching of two frames is scored by assigning a matching score. After assigning the matching scores, frames are linked according to their matching score. A Graph is constructed with frames being the nodes of a graph and the edges being the matching between the two frames. This Graph connects all the related frames. They proposed one Segmentation algorithm to find segments in the Graph. From these segments the some frame are extracted as final frames to be connected to generate sentences. The sentences are formed by concatenating [Arg0] [verb] [Arg1] [Arg2]. Taking assumption that [Arg0] is the subject of [verb] and [Arg2] represents the object of [verb]. Sentences obtained thus are considered to be the summary of the original input document but this assumption always may be not a successful grammatical generation.

In the year 1997, **Regina Barzilay & Michael Elhadad** [6] proposed a methodology of constructing *lexical chains* for the purpose of text summarization. A lexical chain is a chain or sequence of words grouped together by finding any semantic cohesion amongst the words. They used the WordNet thesaurus, part-of-speech tagger, shallow parser and other derived segmentation algorithm to compute the lexical chains. The topic of the input text is identified by computing lexical chains. After construction of all possible lexical chains from the input text, chains are scored and best scoring chains are selected. From these high scoring chains the original whole sentences are extracted from the input text document to form the summary.

C. Hybrid Approaces

There are some methods which use statistical approach as well as they go for semantic approach also. These methods can be termed as hybrid methods.

Diana Trandabat [7] proposed a method to improve summarization done by extractive technique. The method uses both semantic analysis and statistical approaches to achieve the improvement. This method consists of three stages- named entity identification, sentence parsing and semantic roles extraction, extracting sentences containing specific semantic roles which have highest occurrences in the text.

Example:

Text fragment: *Hercules, of all of Zeus's illegitimate children seemed to be the focus of Hera's anger. She sent a two-headed serpent to attack him when he was just an infant.* [7]

Summary of this short fragment text using sentence elimination method hypothetically could be:

She sent a two-headed serpent to attack him. (which is incomprehensible, who is "She", who is "him") But, using anaphoric references the proposed method makes this more understandable. It identifies Hera as "She" and Hercules as "him" and thus presents an improvised summary of that text: Hera sent a two-headed serpent to attack Hercules.

After the anaphora resolution step, using statistical method the main character is identified. The main character considered as the Named entity which occurred most in the text. Sentences with most occurrences of the main character are extracted to be presented as summary

Abdullah Bawakid and Mourad Oussalah [8] developed a semantic summarization system to participate in TAC (Text Analysis Conference) 2008. The Query based system functions with mainly three steps; *Preprocessing*, *Extracting and Analyzing* and *Generation* of the summaries.

In *Preprocessing,* the input document is cleaned i.e. the unnecessary information and various tags (HTML, XML, etc.) are removed. Sentence splitting is done and key parts from the documents are extracted such as Headlines, Document IDs, publication dates etc. Sentence and word boundaries are detected and different features are extracted based on user query and using NE tagger (Locations, Persons, Organizations, etc.) and POS tagger for Part-of-Speech tag and co reference resolution.

In the Analyzing phase positioning of sentences, named entities, Title/Query are analyzed as all these things play a major role in conveying more information. To determine similarity between two sentences Sentence-Sentence Semantic similarity is measured and sentences are scored signifying their importance.

In *Generation*, the most significant (highest scoring) sentences are extracted and arranged in a chronological order for a better readability of the generated summary.

IV. CONCLUSION

The goal of extractive summaries is to find out the most important sentences from the text to be presented as a summary. But introducing semantic analysis to extraction improves its quality. Text summarization using semantic understanding of texts is more considerable, more comprehensible and more significant than summarization achieved by statistical approach only. So, the inference of this survey can be stated simply as- Using semantics and statistical approaches together, not only important sentences can be extracted but also it improves the quality of extracted sentences to make it more acceptable.

REFERENCES

[1] Dragomir R. Radev, Edward Hovy, Kathleen McKeown: *Introduction to special issue on summarization*, Computational Linguistics.
[2] Hans Peter Luhn: *The automatic creation of literature abstracts*, IBM Journal of Research Development.
[3] H. P Edmondson: *New Methods in Automatic Extracting*, Journal of the ACM.
[4] Julian Kupiec, Jan Pedersen and Francine Chen: *A Trainable Document Summarizer*, In Proceedings SIGIR '95.
[5] Divyanshu Bhartiya, Ashudeep Singh: A Semantic Approach to Summarization, course project, Department of Computer Science and Engineering, IIT Kanpur.
[6] Regina Barzilay, Michael Elhadad: *Using Lexical Chains for Text Summarization.* In Proceedings ISTS'97.
[7] Diana Trandabat: *Using semantic roles to improve summaries*, in Proceedings of the 13th European Workshop on Natural Language Generation ENLG2011.
[8] Abdullah Bawakid, Mourad Oussalah: *A Semantic Summarization System: University of Birmingham at TAC 2008*
[9] Miller, G. A. (1995). *Wordnet: a lexical database for English*, ACM.

[10] Baxendale, P. (1958). *Machine-made index for technical literature - an experiment*. IBM Journal of Research Development.
[11] Regina Barzilay, Michael Elhadad, *Using Lexical Chains for Text Summarization.* In Inderjeet Mani and Mark Marbury -editors. Advances in Automatic Text Summarization. MIT Press, 1999.
[12] Daniel Marcu, *Discourse Trees Are Good Indicators of Importance in Text*. In Inderjeet Mani and Mark Marbury -editors. Advances in Automatic Text Summarization. MIT Press, 1999.
[13] Eduard Hovy and Chin-Yew Lin, *Automated Text Summarization in SUMMARIST.* In Inderjeet Mani and Mark Marbury -editors. Advances in Automatic Text Summarization. MIT Press, 1999.
[14] Min-yen Kan, Kathleen Mckeown, *Information extraction and summarization: Domainn independence through focus types.* Technical report, Computer Science Department, Columbia University.
[15] Hsin-Hsi, C., Chuan-Jie, L. *A multi-lingual news summarizer*. In Proceedings of the 18th conference on Computational linguistics, 2000. Association for Computational Linguistics.
[16] Copeck, T., Szpakowicz, S., Japkowic, N. *Learning How Best to Summarize*. In Workshop on Text Summarization, Philadelphia, 2002
[17] Kathleen McKeown, Regina Barzilay et al, *Tracking and summarizing news on a daily basis with the Columbia's Newsblaster*. In Proceedings of the Human Language Technology (HLT) Conference. San Diego, CA, 2002.
[18] Farzindar, A., Rozon, F., Lapalme, G. *CATS A Topic-Oriented Multi-Document Summarization system at DUC 2005*. In the Document Understanding Workshop, Vancouver, B.C., Canada, 2005.
[19] Regina Barzilay, Kathleen McKeown, Michael Elhadad, *Information fusion in the context of multi-document summarization*. In Proceedings of ACL 1999.
[20] Hongyan Jing, *Cut-and-paste text summarization*. PhD thesis, 2001. Adviser- Kathleen R. Mckeown.
[21] Saggion, H, Lapalme, G. *Generating Indicative Informative Summaries with SumUM*. Computational Linguistics, 2002.