Comparative Analysis of Text Summarisation Techniques

Sakshi Bhalla [1]Roma Verma[2] Kusum Madaan [3]
[1,2, 3, 4] Dept.Of Cse , Hmr Institute of Technology & Management
[5] Assistant Professor, Dept.Of Cse, Hmr Institute of Technology & Management

***Abstract*-In the today's busy world everyone just want a ready to serve things and the same can also be observed in case of computer era, so to provide the empowerment in the field of computer the concept of TEXT SUMMARIZATION gains the utmost attention. With the abundance of text material available on the Internet, text summarization has become an important and timely tool for assisting and interpreting text information .The main methodology behind this paper is to provide the wider view of the text summarization, like what in actually the text summarization, which techniques are used till now in this field, various application of this text-summary. This paper surveys the techniques used in previous years till now for text summarization and compares on the basis of their performance to produce better results.**

Index Terms*: *NLP-NATURAL LANGUAGE PROCESSING, SAA-SEMANTIC ANANLYSIS APPROACH, SI-SWARM INTELLIGENCE

1. INTRODUCTION

Automatic text summarization is defined as a specific process of minimizing a text file into a compact and summarized pattern using a specific computer program or code so as to context, gather and highlight its most important and main points [1].The coherent summary should be one which can easily take into account of variables such as syntax, length of the text and writing style. However the main methodology of this summarization is to find a representative subset of the text that itself or best define the entire data, this summarization can include document summarization, image collection summarization and video summarization. Document summarization, basically rely to automatically create a *representative summary* or *abstract* of the entire document, by finding the most *informative and impressive* sentences. Similarly, in case of image summarization the system finds the most representative and important images whereas, in consumer videos as well one would want to remove the boredom or repetitive scenes, and extract out a much precise and concise version of the video. This is also fruitful say for surveillance videos, where one might want to extract only the useful events in the recorded video, since most part of the video may be uninteresting with nothing going on. But now-a-days as the problem of information overload is growing, and as the amount of data increases, the interest in automatic summarization is also increasing. However this paper focuses on text summarization only, generally, there are two approaches already defined to achieve automatic summarization such as: *extraction* and *abstraction*. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques for eg (we are making the use of back propagation technique) to create a summary that is closer to what a human being might generate. Such type of a summary might contain words not explicitly present in the original document. Researchers also conforms with the fact that use of abstractive methods is an increasingly important and active technique in the area of text summarization, however due to complexity constraints and due to various other problems, developers to date has focused primarily on extractive methods mostly.

Text mining refers to the process of deriving high quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

Typical text mining tasks include text categorization, clustering, concept extraction, sentiment analysis, document summarization, etc.

II. RELATED WORK

Most early work on single-document summarization focused on technical documents. Perhaps the most cited paper on summarization is that of (Luhn, 1958), that describes research done at IBM in the 1950s. In his work, Luhn proposed that the frequency of a particular word in an article provides an useful measure of its significance. There are several key ideas put forward in this paper that have assumed importance in later work on summarization. As a first step, words were stemmed to their root forms, and stop words were deleted. Luhn then compiled a list of content words sorted by decreasing frequency, the index providing a significance measure of the word. On a sentence level, a significance factor was derived that tells the number of occurrences of significant words within a sentence, and the linear distance between them due to the intervention of non-significant words. All sentences are ranked in order of their significance factor, and the top ranking sentences are finally selected to form the auto-abstract.

Related work (Baxendale, 1958), also done at IBM and published in the same journal, provides early insight on a particular feature helpful in finding salient parts of

documents i.e. the sentence position. Towards this goal, the author examined 200 paragraphs to find that in 85% of the paragraphs the topic sentence came as the first one and in 7% of the time it was the last sentence. Thus, a naive but fairly accurate way to select a topic sentence would be to choose one of these two. This positional feature has since been used in many complex machine learning based systems.

Edmundson (1969) describes a system that produces document extracts. His primary contribution was the development of a typical structure for an extractive summarization experiment. At first, the author developed a protocol for creating manual extracts that was applied in a set of 400 technical documents. The two features of word frequency and positional importance were incorporated from the previous two works. Two other features were used the presence of cue words (presence of words like significant, or hardly), and the skeleton of the document (whether the sentence is a title or heading). Weights were attached to each of these features manually to score each sentence. During evaluation, it was found that about 44% of the auto-extracts matched the manual extracts. A brief summary of existing research works is shown in TABLE I.

III. TEXT SUMMARIZATION USING NEURAL NETWORKS

With the abundance of text material available on the Internet, text summarization has become an important and timely tool for assisting and interpreting text information. The Internet provides more information than is usually needed. Therefore, Summarization is a useful tool for selecting relevant texts, and for extracting the key points of each text. A summarization tool for news articles would be extremely useful for almost everyone, since for given news topic or event, there are a large number of available articles from the various news agencies and newspapers. Because news articles have a highly structured document form, important ideas can be obtained from the text simply by selecting sentences based on their attributes and locations in the article.

Generally, there are two approaches for text summarization: extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. First clean the text file by removing full stop, common words (conjunction, verb, adverb, preposition etc.). Then calculate the frequency of each word and select top words which have maximum frequency. This technique retrieves important sentence emphasize on high information richness in the sentence as well as high Information retrieval. These related maximum sentence generated scores are clustered to generate the summary of the document.

IV. TYPES OF SUMMARIZATION

There are 3 types of summarization through which we can summarize any text, file, video, etc. Three types of summarization are extraction based, abstraction based and aided summarization. [1]

4.1 Extraction-based summarization

In this type of summarization, the automatic system captures and finds objects and its instances from the whole collection of the data, without changing the objects and their instances themselves. Examples of extraction based summarization are key phrase extraction, where the goal is to select individual words or phrases to "tag" a document, and document summarization, where the goal is to select whole sentences (without modifying them) to create a short paragraph summary. Similarly, in image collection summarization, the system extracts images from the collection without modifying the images themselves.

4.2 Abstraction-based summarization

Extraction techniques simply copies the information that is most important by the system to the summary (for example, key clauses, sentences or paragraphs), while abstraction involves paraphrasing sections of the source document. Generally, abstraction can condense a text more strongly than extraction, but the programs that can do this are harder to develop as they require use of natural language technology, which itself is a growing field.

While some work has been done in abstractive summarization (creating an abstract synopsis like that of a human), the majority of summarization systems are extractive (selecting a subset of sentences to place in a summary).

4.3 Aided summarization

Machine Learning techniques from closely related fields such as information retrieval or text have been successfully adapted to help automatic summarization.

Apart from Fully Automated Summarizers (FAS), there are systems that aid users with the task of summarization (MAHS, Machine Aided Human Summarization), for example by highlighting candidate passages to be included in the summary, and there are systems that depend on post-processing by a human (HAMS = Human Aided Machine Summarization)

V. TECHNIQUES THROUGH WHICH INFORMATIVENESS OF SUMMARY IS EVALUATED

5.1 Intrinsic and extrinsic evaluation:-

Intrinsic evaluation evaluated the summarization system within itself whereas an extrinsic evaluation system evaluates and tests the system on the basis of its effects on other tasks.

5.2 Inter – textual and Intra – textual :-

Intra textual evaluated the outcome of the system whereas Inter textual compares the outcomes of several summarization system.

VI. METHODOLOGIES USED

Table 1 shows the review of technologies used so far in various papers and we have categorised them on the basis of approach required for text summarization.[1][2][3][4]

TABLE 1

Author/ Year	Existing Work		
	Category	Techniques	Journal/ Proceedings
Osborne, 2002	Neural network	Maximum Entropy	ACL 2002 Workshop on Automatic Summarization
Lin, 2004		Similarity of Sentences	ACL2004 Workshop
Nenkova, 2005		Proper ranking of sentences	AAAI 2005
Yong, 2005		Neural Network	International Conference on Data Mining
Svore, 2007		Neural Network algorithm (RankNet)	EMNLP-CoNLL
Aone, 1990	NLP(Natural Language Processing)	Inverse Term Frequency & NLP technique	Advances in Automatic Text Summarization
Barzilay, 1997		Deep NLP	ISTS 1997
McKeown, 1997		Lexical Chains	AAAI
Marcu, 1998		Rhetorical Structure Theory (RST)	6th Workshop on Very Large Corpora
Carbonell & Goldstein, 1998		Maximal Marginal Relevance	SIGIR 1998
Daume&Marcu, 2002, 2004		Log Probability & Rhetoric Structure Tree	ACL 2002, DUC 2004
KaustubhPatil, 2007	SAA	Graph Theory, Latent Semantic Analysis (LSA), Node Centrality	International Journal on Computer Science and Information Systems (IADIS)
Zhan, 2007		Info Extraction of salient topics from online reviews	IEEE International Conference on Computer Science and Information Technology
Verma, 2007		Ontology Knowledge (e.g. WordNet& UMLS) in	Document Understanding Conference DUC 2007
Bawakid, 2008		Semantic Analysis (sentence location, named entities,	1st Text Analysis Conference (TAC) 2008

Liu, 2009		Query-based Words Extraction & New Sentence Ranking Formula	ICCPOL 2009, LNAI 5459, Spinger-Verlag
TroelsAndreasen, 2009		Conceptual Clustering & Semantic Similarity Measure	Springer-Verlag
Hamid Khosravi, 2008	Fuzzy Logic	Optimizing Text Summarization Based on Fuzzy Logic	Springer-Verlag
Mohammed SalemBinwahlan, 2009	SI (Swarm Intelligence)	Fuzzy Swarm Based Text Summarization	Journal of Computer Science

VI. COMPARISON IN CHRONOLOGICAL ORDER

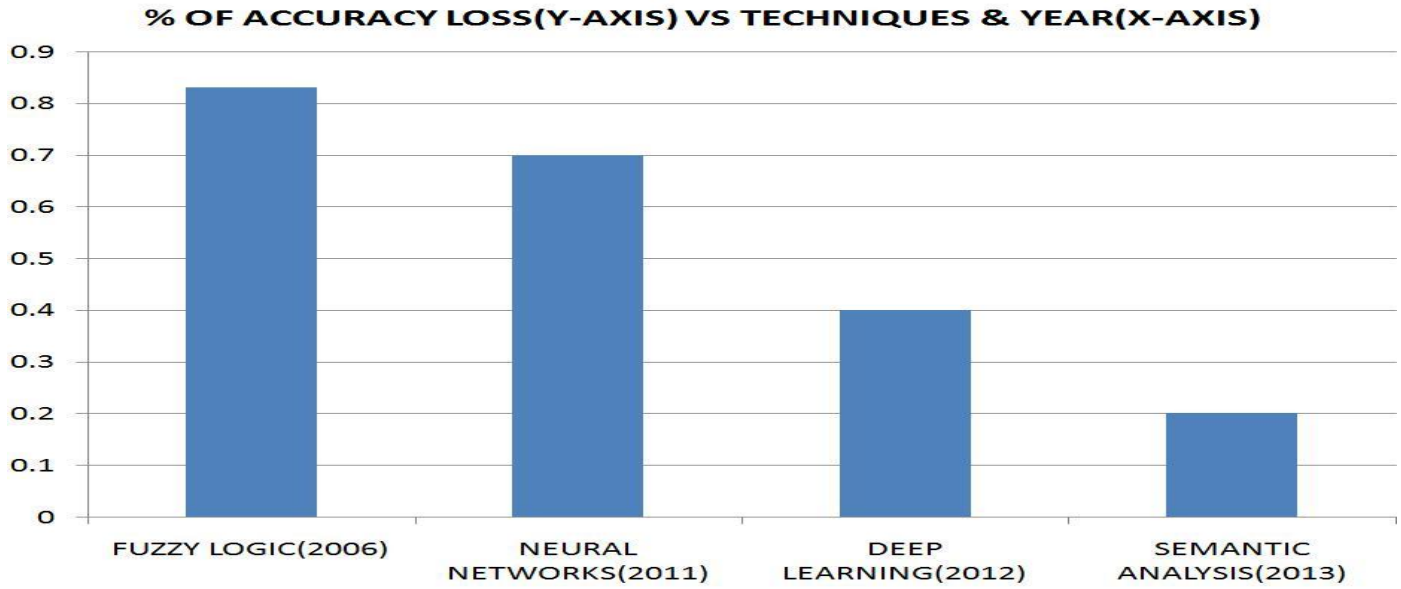
Final comparison has been made on the basis of accuracy and performance yielded by various techniques used so far.

RESEARCH PAPER NAME	YEAR	ADVANTAGES	DISADVANTAGES
Challenges Of Automatic Summarization	2000	The Generic summary type used in this paper gives 90%length reduction 60%time reduction and 0% accuracy loss.	User focus summary type explain before provides only 77% length reduction 50% time reduction and 5% accuracy loss.
From Text To Speech Summarization	2004	This paper approach was able to give the summarization of spoken language and meetings with an accuracy loss of .63% only	The approach used in this paper was not able to provide the analysis of integrating text and speech.
Comprehensive Method For Text Detection And Summarization.	2005	This paper provides video text detection and localization with a detection accuracy of 90.8%.	The method in (10) was providing only 67.3% detection accuracy
Automatic Text Summarization Using Fuzzy Logic	2006	This paper proves that fuzzy logic optimized with evolutionary algorithm gives the best result with an accuracy loss of .831 %.	Use of ga-gp does not improve the precision of Microsoft word summarizer with an average of only .291%.
The Research Of Data Mining Based On Neural Networks	2011	Network pruning algorithm and rule extraction algorithms have presented and improved, it makes the data mining based on neural networks more and more to favor for the majority of users and it has handle large amount of data	Algorithm efficiency enhancement is needed here. the rule extraction algorithm's computation complexity is a important limiting factor of this paper
An Approach For Text Summarization Using Deep Learning Algorithm	2012	Here proposed approach was based on deep learning algorithm i.e. RBM algorithm is used for better efficiency. The performance judging parameters f -measure has got value 0.49, 0.469,0.520 for 3 different document	Less features were considered and more hidden layers can be added to RBM algorithm for better results
Semantic Graph Reduction Approach For Text Summarization	2012	The approach used in this paper that is creating a semantic graph called rich semantic graph produced a good summarization by minimizing the original text to 50%.	This semantic graph reduction was not working properly with different sizes of the document
Text Summarization Of Turkish Texts Using Latent Semantic Analysis	2013	Latent Semantic Analysis are explained and two new approaches, namely cross and topic, are introduced. The comparison of these approaches is done using the rouge-l f-measure score. The results show that the cross method is better than all other approaches.	The modified tf-idf approach lacks performance because it removes some of the sentences/words from the input matrix, assuming that they cause noise.
Text Summarization Using Neural Networks And Rhetorical Structure Theory	2015	The numerical data feature was introduced . Numerical data feature, which will help to select highly ranked summary sentences. and rhetorical structure theory provides a combination of features that useful in several kinds of discourse studies	Precise summarization, more in-depth understanding of the sentence is required
Statistical And Analytical Study Of Guided Abstractive Text Summarization	2016	Accuracy was better than extractive summarization. The challenges in Indian languages are handled at each stage by writing i.e. rules and creating generic templates.(f score-0.815, precision-0.8642, recall- 0.7973, accuracy- 0.7217)	Template-based models generate flatness and monotony in the summary of the paragraph is generated. this can be resolved using wordnet12 (freely available lexical database) or simple nlg13 (java api) may be suggested to facilitate the generation

[6][7][8][9]

VII. GRAPH

A graph has been designed to show lost in accuracy while text summarization using various different techniques over these years. The comparison has been made in accordance to the above table and performance measures given in different tables. It can be observed that deep learning and semantic analysis helps to improve accuracy more than other techniques & semantic analysis is the recent one.



VIII CONCLUSION

It has been seen that there were many summarization techniques which has been developed till now. We have summarized almost all the summarization techniques in this paper out of which abstractive text summarization techniques are recently used. Abstractive text summarization techniques are better than extractive text summarization techniques in terms of their accuracy and performance.

REFERENCES

[1] KhosrowKaikhah, "Automatic Text Summarization with Neural Networks", in Proceedings of second international Conference on intelligent systems,IEEE, 40-44, Texas, USA, June 2004, link: https://digital.library.txstate.edu/bitstream/handle/10877/3819/fullte xt.pd f.

[2] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto, "Automated Summarization Evaluation with Basic Elements", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006, link: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.7845& rep =rep1&type=pdf.

[3] Josef Steinberger and KarelJezek, "Evaluation Measures for Text Summarization", In Computing and Informatics / Computers and Artificial Intelligence - CAI , vol. 28, no. 2, pp.251- 275, 2009, link:http://www.cai.sk/ojs/index.php/cai/article/download/37/24.

[4] Alkesh Patel, TanveerSiddiqui and U. S. Tiwary, "A language independent approach to multilingual summarization", InConference RIAO2007, Pittsburgh PA, U.S.A., 2007, link:http://users.cis.fiu.edu/~lli003/Sum/RIAO/2007/1.pdf

[5] Wikipedia – Artificial Neural Networks, link:http://en.wikipedia.org/wiki/Artificial_neural_networkS.

[6] https://www.google.co.in/url?sa=t&source=web&rct=j&url=http://w ww. ijarcce.com/upload/2015/june-15/IJARCCE%252012.pdf&ved=0ahUKEwitpPGdiZrTAhVGpZQ KHR h5D8kQFgghMAA&usg=AFQjCNHycqNWbR-

[7] https://pdfs.semanticscholar.org/54ba/ f0170bccaa30ba3ab97d23503d1f8 YiqqmWmookzpznr4KuA – text summarization using neural networks and rhetorical structure. 1a0854b.pdf - research on data mining using neural networks

[8] text summarization using deep belief networks – IIIT Bangalore link - https://www.google.co.in/url?sa=t&source=web&rct=j&url=http://ci sl.ii itb.ac.in/wp-content/uploads/2015/12/abinaya_thesis.pdf&ved=0ahUKEwi2qargi prT AhUJwI8KHcQDAggQFgghMAA&usg=AFQjCNHy8hhM0lQOR5 ZD KoCp7LOH_KBJNg

[9] https://www.google.co.in/url?sa=t&source=web&rct=j&url=http://ci tese erx.ist.psu.edu/viewdoc/download%3Fdoi%3D10.1.1.81.2807%26re p% 3Drep1%26type%3Dpdf&ved=0ahUKEwjHkIWri5rTAhWLOY8K HQJ WDS4QFgghMAA&usg=AFQjCNGn2g6RnzbokWz7p23KQLdiku CPG g – the challenges of automatic summarization – computer.

[10] https://www.google.co.in/url?sa=t&source=web&rct=j&url=htt p:// ieeexplore.ieee.org/document/7604928/&ved=0ahUKEwjPgKSJi5rT Ah UIQo8KHdReCQcQFggmMAI&usg=AFQjCNE2996GLNVpKT8d NpB j0tisXbOv9g – automatic text summarization using fuzzy interface