Comparative Analysis of Text Summarisation Techniques

Sakshi Bhalla 1Roma Verma2 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 ANALYSIS 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 naïve 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>
<thead>
<tr>
<th>Author/Year</th>
<th>Existing Work</th>
<th>Category</th>
<th>Techniques</th>
<th>Journal/Proceedings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osborne, 2002</td>
<td>Maximum Entropy</td>
<td>Neural network</td>
<td>ACL 2002 Workshop on Automatic Summarization</td>
<td></td>
</tr>
<tr>
<td>Lin, 2004</td>
<td>Similarity of Sentences</td>
<td>Neural network</td>
<td>ACL2004 Workshop</td>
<td></td>
</tr>
<tr>
<td>Nenkova, 2005</td>
<td>Proper ranking of sentences</td>
<td>Neural network</td>
<td>AAAI 2005</td>
<td></td>
</tr>
<tr>
<td>Yong, 2005</td>
<td>Neural Network</td>
<td>Neural network</td>
<td>International Conference on Data Mining</td>
<td></td>
</tr>
<tr>
<td>Svore, 2007</td>
<td>Neural Network algorithm (RankNet)</td>
<td>Neural network</td>
<td>EMNLP-CoNLL</td>
<td></td>
</tr>
<tr>
<td>Aone, 1990</td>
<td>Inverse Term Frequency &amp; NLP technique</td>
<td>NLP(Natural Language Processing)</td>
<td>Advances in Automatic Text Summarization</td>
<td></td>
</tr>
<tr>
<td>Barzilay, 1997</td>
<td>Deep NLP</td>
<td>NLP(Natural Language Processing)</td>
<td>ISTS 1997</td>
<td></td>
</tr>
<tr>
<td>McKeown, 1997</td>
<td>Lexical Chains</td>
<td>NLP(Natural Language Processing)</td>
<td>AAAI</td>
<td></td>
</tr>
<tr>
<td>Marcu, 1998</td>
<td>Rhetorical Structure Theory (RST)</td>
<td>NLP(Natural Language Processing)</td>
<td>6th Workshop on Very Large Corpora</td>
<td></td>
</tr>
<tr>
<td>Carbonell &amp; Goldstein, 1998</td>
<td>Maximal Marginal Relevance</td>
<td>NLP(Natural Language Processing)</td>
<td>SIGIR 1998</td>
<td></td>
</tr>
<tr>
<td>Zhan, 2007</td>
<td>Info Extraction of salient topics from online reviews</td>
<td>SAA</td>
<td>IEEE International Conference on Computer Science and Information Technology</td>
<td></td>
</tr>
<tr>
<td>Verma, 2007</td>
<td>Ontology Knowledge (e.g. WordNet &amp; UMLS) in</td>
<td>SAA</td>
<td>Document Understanding Conference DUC 2007</td>
<td></td>
</tr>
<tr>
<td>Bawakid, 2008</td>
<td>Semantic Analysis (sentence location, named entities,</td>
<td>SAA</td>
<td>1st Text Analysis Conference (TAC) 2008</td>
<td></td>
</tr>
<tr>
<td>Author, Year</td>
<td>Title</td>
<td>Conference/Journal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, 2009</td>
<td>Query-based Words Extraction &amp; New Sentence Ranking Formula</td>
<td>ICCPOL 2009, LNAI 5459, Springer-Verlag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TroelsAndreasen, 2009</td>
<td>Conceptual Clustering &amp; Semantic Similarity Measure</td>
<td>Springer-Verlag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamid Khosravi, 2008</td>
<td>Fuzzy Logic</td>
<td>Springer-Verlag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohammed SalemBinwahan, 2009</td>
<td>SI (Swarm Intelligence ) Fuzzy Swarm Based Text Summarization</td>
<td>Journal of Computer Science</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VI. COMPARISON IN CHRONOLOGICAL ORDER

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

<table>
<thead>
<tr>
<th>RESEARCH PAPER NAME</th>
<th>YEAR</th>
<th>ADVANTAGES</th>
<th>DISADVANTAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Challenges Of Automatic Summarization</td>
<td>2000</td>
<td>The Generic summary type used in this paper gives 90% length reduction 60% time reduction and 0% accuracy loss.</td>
<td>User focus summary type explain before provides only 77% length reduction 50% time reduction and 5% accuracy loss.</td>
</tr>
<tr>
<td>From Text To Speech Summarization</td>
<td>2004</td>
<td>This paper approach was able to give the summarization of spoken language and meetings with an accuracy loss of 63% only</td>
<td>The approach used in this paper was not able to provide the analysis of integrating text and speech.</td>
</tr>
<tr>
<td>Comprehensive Method For Text Detection And Summarization</td>
<td>2005</td>
<td>This paper provides video text detection and localization with a detection accuracy of 90.8%.</td>
<td>The method in (10) was providing only 67.3% detection accuracy</td>
</tr>
<tr>
<td>Automatic Text Summarization Using Fuzzy Logic</td>
<td>2006</td>
<td>This paper proves that fuzzy logic optimized with evolutionary algorithm gives the best result with an accuracy loss of .831 %.</td>
<td>Use of ga-gp does not improve the precision of Microsoft word summarizer with an average of only 291%.</td>
</tr>
<tr>
<td>The Research Of Data Mining Based On Neural Networks</td>
<td>2011</td>
<td>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</td>
<td>Algorithm efficiency enhancement is needed here.</td>
</tr>
<tr>
<td>An Approach For Text Summarization Using Deep Learning Algorithm</td>
<td>2012</td>
<td>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</td>
<td>Less features were considered and more hidden layers can be added to RBM algorithm for better results</td>
</tr>
<tr>
<td>Semantic Graph Reduction Approach For Text Summarization</td>
<td>2012</td>
<td>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%.</td>
<td>This semantic graph reduction was not working properly with different sizes of the document</td>
</tr>
<tr>
<td>Text Summarization Of Turkish Texts Using Latent Semantic Analysis</td>
<td>2013</td>
<td>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-t f-measure score. The results show that the cross method is better than all other approaches.</td>
<td>The modified tf-idf approach lacks performance because it removes some of the sentences/words from the input matrix, assuming that they cause noise.</td>
</tr>
<tr>
<td>Text Summarization Using Neural Networks And Rhetorical Structure Theory</td>
<td>2015</td>
<td>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</td>
<td>Precise summarization, more in-depth understanding of the sentence is required</td>
</tr>
<tr>
<td>Statistical And Analytical Study Of Guided Abstractive Text Summarization</td>
<td>2016</td>
<td>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)</td>
<td>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</td>
</tr>
</tbody>
</table>

[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