AUTOMATIC TEXT SUMMARIZER

Report Summary

ATS- An Overview

Text summarization is the method to reduce the source text into a compact variant, preserving its knowledge and the actual meaning[1]. The research paper thoroughly investigate the automatic text summarization (ATS). This paper outlines extractive and abstractive text summarization technologies and provides a deep taxonomy of the ATS domain[1].

Text summarization is imperative in today's fast paced world. People hardly have time to read the entire volumes of text being produced in numerous domains. Therefore a handy alternative in form of a quick access software like an automatic text summarizer can help people to extract the desired information they want from any text.

LITERATURE REVIEW OF EXISTING ATS SURVEY

A study on the collection of scholarly research articles dated between 1998 and 2021 is made . The PRISMA approach is followed in streamlining the articles and their contents into one paper.

The article performs a systematic review of the automatic text summarization, including the fundamental theories and evolutions. The survey includes the investigation of the existing dataset, feature extraction, text summarization approaches, text summarization algorithms, performance measurement, evaluation matrices, and challenges.The article compiles ATS architectures based on current methods, datasets, feature extraction, and summarization approaches. Moreover, this study explains the constraints and limitations of such methods. Subsequently, the study ends by distinguishing the current difficulties and challenges of ATS architectures, along with future research directions[1].

The next slide contains a tabulation[1] of the exploration and findings done by various researchers and scholars in the field of ATS, their purpose and limitations. Diverse methods are used to formulate the module of implementing this technique.

Existing ATS methods, their scope and limitations

Reference	Year	Main Purpose	Limitations
[10]	2009	Critical ways to summarize the texts and provides a taxonomy of the methods of summarization.	Extractive, abstract, NLP with Machine learning and Deep learning are missed.
[11]	2014	A hybrid approach can do both the extractive and abstractive methods efficiently.	Avoided complex processing like NLP approaches.
[12]	2014	Reviewed works between (2000-2013) year and proposed a hybrid statistical method.	This paper doesn't include cognitive aspects, including visual-ization techniques and evaluations of the impact.
[13]	2016	Describes two definite summarization techniques, which are abstractive, extractive.	Introduces techniques and methods only.
[14]	2017	A study based on automated keyword extraction and text summarization.	They do not briefly review every approach they included; they missed some feature extraction model.
[15]	2017	Topic Representation, frequency-driven, graph-based, and the effectiveness and Limitations.	The recent approaches are not surveyed.
[16]	2017	Processes of extractive methods and multilingual text summa-rization are discussed.	A precise classification and idea about feature scores and extraction is missing.
[17]	2020	Methods, processes, primary structure, strategies, datasets, measurements of ATS.	A detailed classification and description of feature extraction are missing.
[18]	2020	To handle multi- documents for summarization based on recent research work and comparison.	Does not represent any brief discussion of any topic.

Natural Language Processing

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can[2].

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment[2].

The algorithms of NLP has made text extraction, classification and modulation very comprehensive.

NLP techniques like Stop Words Filtering, Stemming, Tokenization, Lemmatization and Bag of Words is very helpful in generating a fitting summary for human speech and text.

Structure of ATS

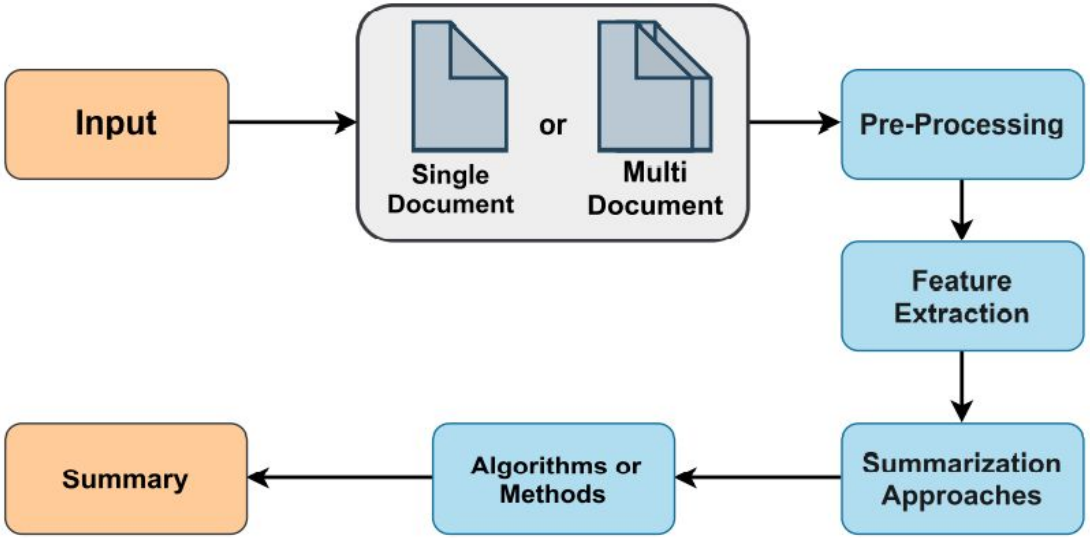


Fig1.A basic structure of an ATS[1].

The input is categorized in two ways-Single Documents that are dealt by SDS and multi documents that are dealt by MDS

Moreover, summarization can be done in two ways - abstractive and extractive.

PRE-PROCESSING

In this stage basic redundancy like the filtering of stop words, conjunctions, stemming and lemmatization, which is reducing a word or a token to its root form(example running becomes run) is done.

FEATURE EXTRACTION

Each sentence is subjected and tested against some characteristics. The sentences that score above a threshold are selected further for summarization process.

SUMMARIZATION AND ALGORITHM

The extracted text is then processed via number of optimum algos and modules to create a fitting model best resembling the original meaning of the text. Choosing the correct approach is imperative to create an apt summarizer.

DATASETS USED IN AUTOMATIC TEXT SUMMARIZATION

Datasets used in ATS to efficiently train our model are assembles of newspaper articles, emails, novels, and other forms of texts.

The datasets are used in their specific domains. Like a collection of emails will be used as a dataset to generate summary of emails as they tend to be written in the same pattern.

Dataset Name	Number of Documents / Sentences	Language	Used in	URL
New Taiwan Weekly	2738 sentences	English	[41]–[44]	http://www.newtaiwan.com.tw
Annotated English Gigaword	9,876,086 documents	English	[16], [45]–[50]	https://catalog.ldc.upenn.edu/LDC2011T07
DUC 2001	600 sentences	English	[51]–[56]	https://www-nlpir.nist.gov/projects/duc/data.html
DUC 2002	600 sentences	English	[57]–[62]	https://www-nlpir.nist.gov/projects/duc/data.html
DUC 2003	600 sentences	English	[51], [55], [58], [63]–[67]	https://www-nlpir.nist.gov/projects/duc/data.html

Fig 2.The table presented the popular datasets used in the text summarization domain[1]

Pre Processing Techniques

Parts Of Speech (POS) Tagging: The words belonging to specific categories like nouns, verbs are grouped together

Stop Word Filtering: Stop Words like ‘A, an ,the ‘ add no sense to the text in context of real meaning or information being conveyed. Thus they are filtered out.

Stemming: Words are reduced to their base forms to make processing easier. Like running is reduced to its root-”run”.

Named Entity Recognition (NER): Words reflecting the names of known entities like cities, persons, companies are categorized.

Tokenization: Individual sentences are broken down or tabulated in terms of independent entities called tokens

Capitalization: All text is converted into a single case preferably the lower case to make computation easy.

Slang and Abbreviation: Anomalies like slang and abbreviations are removed or filtered.

Noise Removal: Punctuation and special characters are removed to make classification easier

Spelling Correction: Spelling correction is optional but can be preferably done to extract the correct meaning of words and the document.

FEATURE EXTRACTION IN ATS

It is important to sort and select the most important sentences from the text. Thus a criteria consisting of several required features is applied to every sentence and each sentence is scored on that . The research paper presents a list of the most prevalent features for calculating the score of sentence.

- Term Frequency (TF)
- Term Frequency-Inverse Sentence Frequency (TF-ISF)
- Position Feature
- Length Feature
- Sentence–Sentence Similarity
- Title Feature (Tif):
- Phrasal Information (PI)
- Title Similarity (TS)
- Sentence Position (SP)
- Thematic Word (TW):
- Numerical Data

EXTRACTIVE TEXT SUMMARIZATION

The extractive text summarization method aims to identify important words and sentences in a text material and use them effectively to create a summary[1]. Its flowchart is:

FlowChart 1

Split the text into sentences and create an Indicator representation and topic representation for each sentence. Score the sentences- Topic Representation scores on the weightage of topic words in the sentence while Indicator Representation scores on Feature weightage. Select the highest scored sentences to form the summary.

SUPERVISED and UNSUPERVISED LEARNING

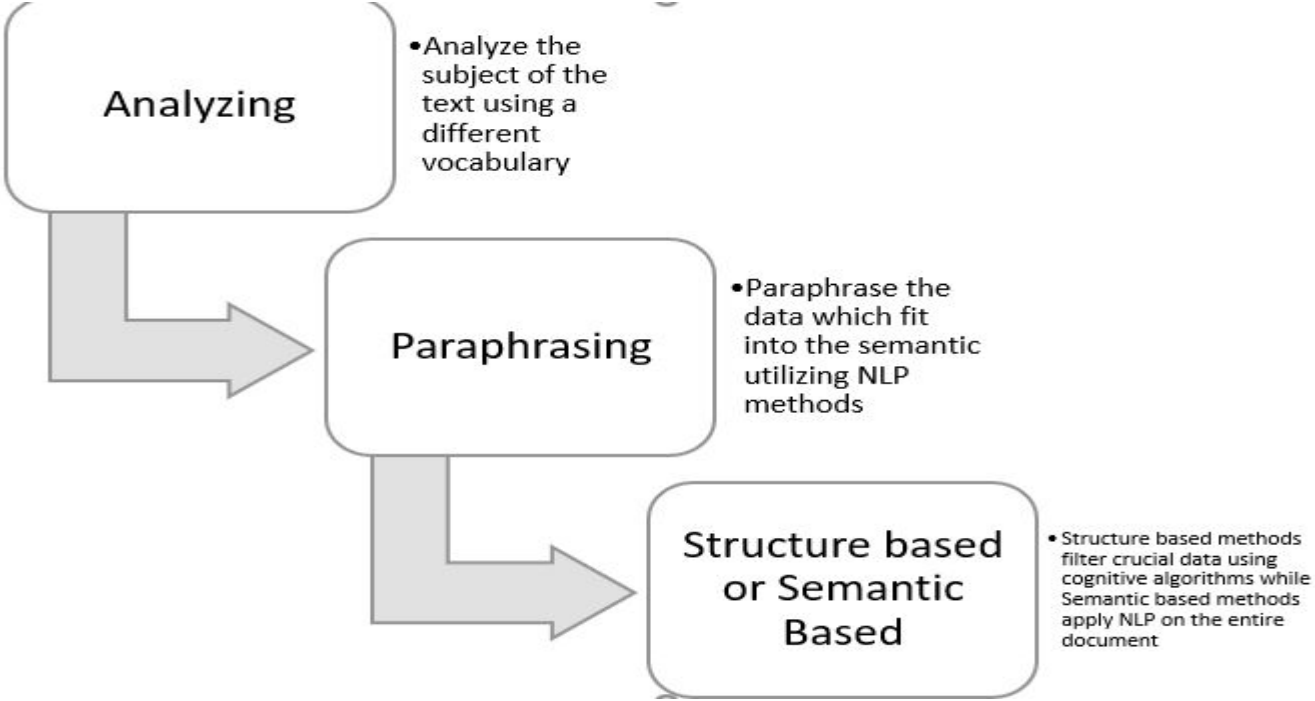
Supervised Learning-It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox[3]

Machine learning and neural network algorithms of these methods require a classified dataset for training, where summarized and non-summarized texts are available with labels[1].

Unsupervised Learning-is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision[4].The summarization process can be performed without any help, such as selecting the introductory sentences of the document from the user[1].

ABSTRACTIVE TEXT SUMMARIZATION

The abstractive process identifies key sections and the main ideas of a text document by paraphrasing them[1]. Its main steps are



Unsupervised Learning Methods

FUZZY LOGIC BASED METHOD-It selects the most important sentences from the document , requires a redundancy remover to achieve better results.

CONCEPT-BASED METHOD-It formulates an extracts concepts to score the sentences and then includes them in summary

LATENT SEMANTIC ANALYSIS-Latent semantic analysis (LSA) is an algebraic-statistical method for extracting hidden semantic structures of sentences and phrases[1].It can pick out the similar words appearing in the text document.

References

[1].M. F. MRIDHA 1 , (Senior Member, IEEE), AKLIMA AKTER LIMA 1 , KAMRUDDIN NUR 2 , (Senior Member, IEEE), SUJOY CHANDRA DAS 1 , MAHMUD HASAN3 , AND MUHAMMAD MOHSIN KABIR 1 .”A Survey of Automatic Text Summarization: Progress, Process and Challenges”,Received October 27, 2021, accepted November 18, 2021, date of publication November 22, 2021, date of current version November 30, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3129786

[2].Studied Online at https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can. [Accessed in May]

[3].Studied Online at https://www.ibm.com/cloud/learn/supervised-learning [Accessed in May]

[4].Studied Online At https://www.javatpoint.com/unsupervised-machine-learning {Accessed In May]