Journal of Engineering Sciences

Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

Text Summarization using Natural Language Processing

Syed Muqtadir Uddin Hussaini[1], Faraaz Mohd Khan[2], Faisal Khan[3], Dr. Abdul Subhan[4]

UG Scholar[1, 2, 3], Asst. Professor[4,]
Department of Computer Science & Engineering[1, 2, 3, 4],
Isl Engineering College, Hyderabad, India.

Abstract

In this project, Automatic text summarization is basically summarizing of the given paragraph using natural language processing and machine learning. There has been an explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods. The system works by assigning scores to sentences in the document to be summarized, and using the highest scoring sentences in the summary. Score values are based on features extracted from the sentence. A linear combination of feature scores is used. Almost all of the mappings from feature to score and the coefficient values in the linear combination are derived from a training corpus. Some anaphor resolution is performed. The system was submitted to the Document Understanding Conference for evaluation. In addition to basic summarization, some attempt is made to address the issue of targeting the text at the user. The intended user is considered to have little background knowledge or reading ability. The system helps by simplifying the individual words used in the summary and by drawing the pre-requisite background information from the web.

Introduction

In the modern Internet age, textual data is ever increasing. Need some way to condense this data while preserving the information and meaning. We need to summarize textual data for that. Text summarization is the process of automatically generating natural language summaries from an input document while retaining the important points. It would help in easy and fast retrieval of information. There are two prominent types of summarization algorithms.

• Extractive summarization systems form summaries by copying parts of the source text through some measure of importance and then combine those part/sentences together to render a summary. Importance of sentence is based on linguistic and statistical features.

• Abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text. Naturally abstractive approaches are harder. For perfect abstractive summary, the model has to first truly understand the document and then try to express that understanding in short possibly using new words and phrases. Much harder than extractive. Has complex capabilities like generalization, paraphrasing and incorporating real-world knowledge. Majority of the work has traditionally focused on extractive approaches due to the easy of defining hard-coded rules to select important sentences than generate new ones. Also, it promises grammatically correct and coherent summary. But they often don't summarize long and complex texts well as they are very restrictive.

Potential applications

Possible current uses of summarization:

1. People need to learn much from texts. But they tend to want to spend less time while doing this.

2. It aims to solve this problem by supplying them the summaries of the text from which they want to gain information.
3. Goals of this project are that these summaries will be as important as possible in the aspect of the texts' intention.
4. The user will be eligible to select the summary length.
5. Supplying the user, a smooth and clear interface.
6. Configuring a fast replying server system.

Objectives

The objective of the project is to understand the concepts of natural language processing and creating a tool for text summarization. The concern in automatic summarization is increasing broadly so the manual work is removed. The project concentrates creating a tool which automatically summarizes the document.

Scope

The project is wide in scope | all of the limitations stated below may seem to contradict that, but they are the only restrictions applied. This project looks at single document summarization - the area of multi document summarization is not covered. Also, the summaries produced are largely extracts of the document being summarized, rather than newly generated abstracts. The parameters used are optimal for news articles, although that can be changed easily.

Methodologies

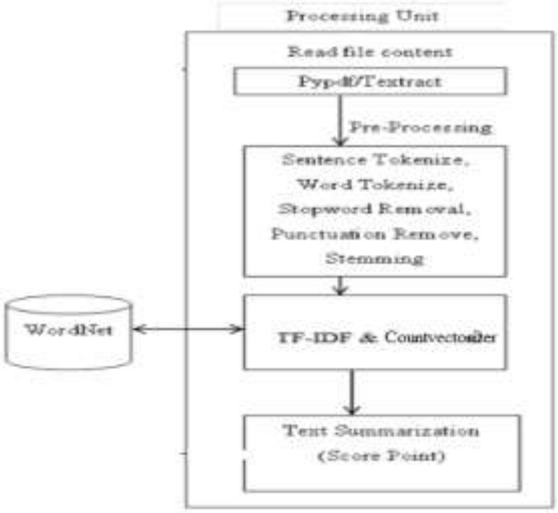
For obtaining automatic text summarization, there are basically two major techniques i.e.-Abstraction based Text Summarization and Extraction based Text Summarization.

Extraction Based Extraction

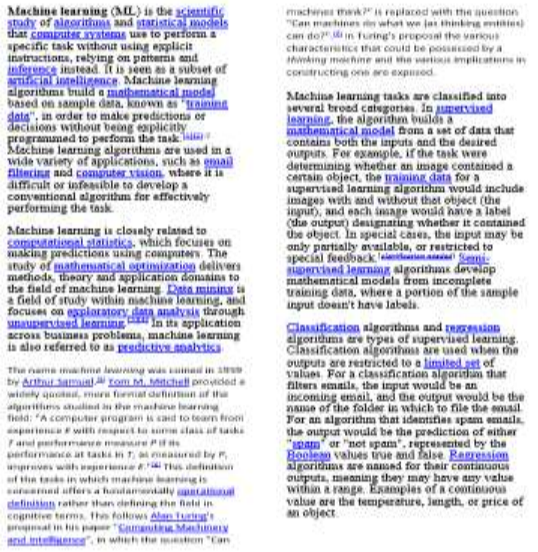
The Extractive summaries are used to highlight the words which are relevant, from input source document. Summaries help in generating concatenated sentences taken as per the appearance. Decision is made based on every sentence if that particular sentence will be included in the summary or not. For example, Search engines typically use Extractive summary generation methods to generate summaries from web page. Many types of logical and mathematical formulations have been used to create summary. The regions are scored and the words containing highest score are taken into the consideration. In extraction only important sentences are selected. This approach is easier to implement. There are three main obstacles for extractive approach. The first thing is ranking problem which includes ranking of the word. The second one selection problem that includes the selection of subset of particular units of ranks and the third one is coherence that is to know to select various units from understandable summary. There are many algorithms which are used to solve ranking problem. The two obstacles i.e. - selection and coherence are further solved to improve diversity and helps in minimizing the redundancy and pickup the lines which are important. Each sentence is scored and arranged in decreasing order according to the score. It is not trivial problem which helps in selecting the subsets of sentences for coherent summary. It helps in reduction of redundancy. When the list is put in ordered manner than the first sentence is the most important sentence which helps in forming the summary. The sentence having the highest similarity is selected in next step is picked from the top half of the list. The process has to be repeated until the limit is reached and relevant summary is generated.

People by and large utilize abstractive outlines. In the wake of perusing content, Individuals comprehend the point and compose a short outline in their own particular manner creating their very own sentences without losing any essential data. In any case, it is troublesome for machine to make abstractive synopses. Along these lines, it very well may be said that the objective of reflection based outline is to make a synopsis utilizing regular dialect preparing procedure which is utilized to make new sentences that are syntactically right. Abstractive rundown age is difficult than extractive technique as it needs a semantic comprehension of the content to be encouraged into the Common Dialect framework. Sentence Combination being the significant issue here offers ascend to irregularity in the produced outline, as it's anything but an all around created field yet. Abstractive arrangement to grouping models is by and large prepared on titles and captions. The comparative methodology is embraced with archive setting which helps in scaling. Further every one of the sentences is revamped in the request amid the inference. Document synopsis can be changed over to regulated or semi-administered learning issue. In directed learning methodologies, indications or signs, for example, key-phrases, point words, boycott words, are utilized to recognize the sentences as positive or negative classes or the sentences are physically labelled. At that point the parallel more tasteful can be prepared for getting the scores or synopsis of each sentence. Anyway they are not effective in removing archive explicit summaries. If the report level data isn't given then these methodologies give same expectation independent of the record. Giving archive setting in the models diminishes this issue.

System Architecture :



Results:



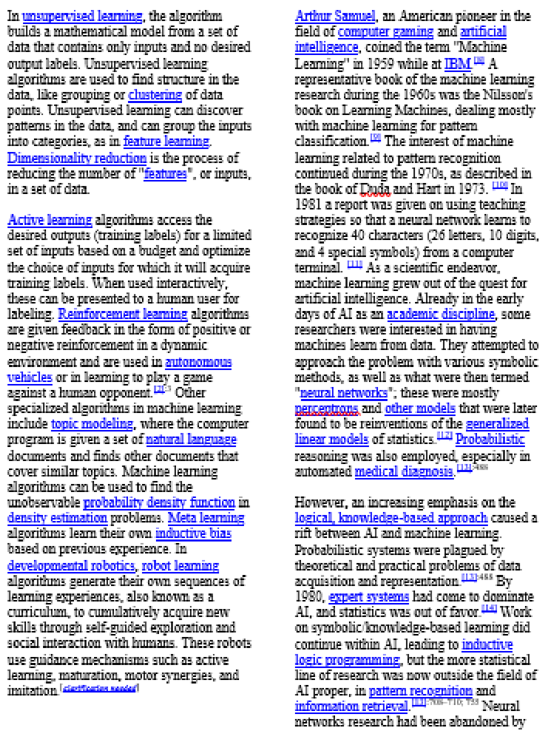
Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2] Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning.[3][4] In its application across business problems, machine learning is also referred to as predictive analytics.

The name machine learning was coined in 1959 by Arthur Samuel.[5] Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."[6] This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turing's proposal in his paper "Computing Machinery and Intelligence", in which the question "Can machines think?" is replaced with the question "Can machines do what we (as thinking entities) can do?".[7] In Turing's proposal the various characteristics that could be possessed by a *thinking machine* and the various implications in constructing one are exposed.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback.[clarification needed] Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.



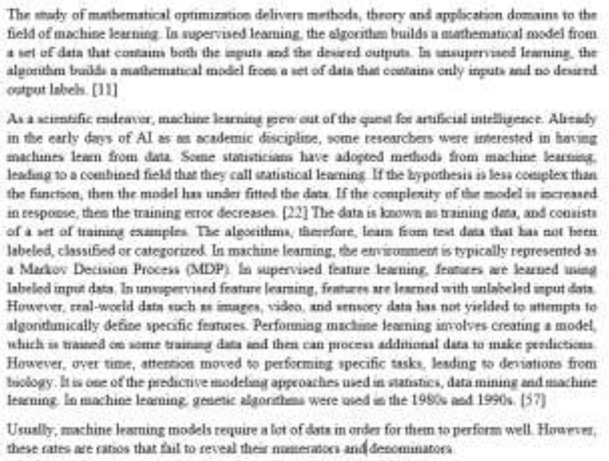
In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

Active learning algorithms access the desired outputs (training labels) for a limited set of inputs based on a budget and optimize the choice of inputs for which it will acquire training labels. When used interactively, these can be presented to a human user for labeling. Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment and are used in autonomous vehicles or in learning to play a game against a human opponent.[2]:3 Other specialized algorithms in machine learning include topic modeling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems. Meta learning algorithms learn their own inductive bias based on previous experience. In developmental robotics, robot learning algorithms generate their own sequences of learning experiences, also known as a curriculum, to cumulatively acquire new skills through self-guided exploration and social interaction with humans. These robots use guidance mechanisms such as active learning, maturation, motor synergies, and imitation.[clarification needed]

Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM.[8] A representative book of the machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification.[9] The interest of machine learning related to pattern recognition continued during the 1970s, as described in the book of Duda and Hart in 1973. [10] In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal. [11] As a scientific endeavor, machine learning grew out of the quest for artificial intelligence. Already in the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics.[12] Probabilistic reasoning was also employed, especially in automated medical diagnosis.[13]:488

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation.[13]:488 By 1980, expert systems had come to dominate AI, and statistics was out of favor.[14] Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming, but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval.[13]:708–710; 755 Neural networks research had been abandoned by

Summarized Text

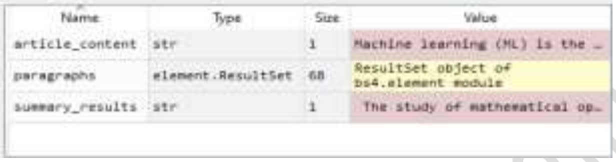


The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels. [11]

As a scientific endeavor, machine learning grew out of the quest for artificial intelligence. Already in the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning. If the hypothesis is less complex than the function, then the model has under fitted the data. If the complexity of the model is increased in response, then the training error decreases. [22] The data is known as training data, and consists of a set of training examples. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. In machine learning, the environment is typically represented as a Markov Decision Process (MDP). In supervised feature learning, features are learned using labeled input data. In unsupervised feature learning, features are learned with unlabeled input data. However, real-world data such as images, video, and sensory data has not yielded to attempts to algorithmically define specific features. Performing machine learning involves creating a model, which is trained on some training data and then can process additional data to make predictions. However, over time, attention moved to performing specific tasks, leading to deviations from biology. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. In machine learning, genetic algorithms were used in the 1980s and 1990s. [57]

Usually, machine learning models require a lot of data in order for them to perform well. However, these rates are ratios that fail to reveal their numerators and denominators

Collected Information as paragraph so, total 68 paragraphs have been collected from URL web page and extracted the main content as summary.



Name	Type	Size	Value
article_content	str	1	Machine learning (ML) is the …
paragraphs	element.ResultSet	68	ResultSet object of bs4.element module
summary_results	str	1	The study of mathematical op…

Conclusion

Text summarization is one of the major problems in the field of Natural Language Processing. Methods such as Deep Understanding, Sentence Extraction, Paragraph Extraction, Machine Learning, and even some which employ all these methods along with Traditional NLP Techniques(Semantic Analysis, etc.). As such, keeping these accomplishments in mind, there is still ample amount of research left in the domain of Text Summarization, as a meaningful summary is still difficult to attain in all domains and languages.

As with time internet is growing at a very fast rate and with it data and information is also increasing. it will going to be difficult for human to summarize large amount of data. Thus there is a need of automatic text summarization because of this huge amount of data. Until now, we have read multiple papers regarding text summarization, natural language processing. There are multiple automatic text summarizers with great capabilities and giving good results. We have learned all the basics of Extractive and Abstractive Method of automatic text summarization and tried to implement extractive one. We have made a basic automatic text summarizer using nltk library using python and it is working on small documents. We have used extractive approach to do text summarization.

References

[1] Ahmad T. Al-Taani. “,Automatic Text Summarization Approaches” International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017)

[2] Neelima Bhatia, ArunimaJaiswal, “Automatic Text Summarization: Single and Multiple Summarizations ”, International Journal of Computer Applications

[3] Mehdi Allahyari, SeyedaminPouriyeh, Mehdi Assefi, SaeidSafaei, Elizabeth D. Trippe, Juan B. Gutierrez, KrysKochut, “ Text Summarization Techniques: A Brief Survey”, (IJACSA) International Journal of Advanced Computer Science and Applications

[4]Pankaj Gupta, Ritu Tiwari and NirmalRobert,”Sentiment Analysis and Text Summarization of Online Reviews: A Survey”InternationalConzatiference on Communication and Signal Processing,August 2013

[5]Vishalgupta,Gurpreet Singh Lehal,”A Survey of Text Summarization Extractive Techniques.”JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[6]Jiwei Tan, XiaojunWan,Jianguo Xiao Institute of Computer Science and Technology,Peking University “Abstractive document summarization with a GraphBased attentional neural model. ”

[7]SeonggiRyang, Graduate school of Information science and technology, University of Tokyo Takeshi Abekawa, National institute of informatics “Framework of automatic text summarization using Reinforcement learning” 48

[8]Tianshi, YaserKeneshloo, Narenramakrishnan, Chandan K. Reddy, Senior member, IEEE " Neural Abstractive text summarization with sequence-to -sequence models"

[9] Josef Steinberger, KarelJežek, "Using latent Semantic analysis In Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, UniverzitníCZ-306 14 Plzeň.

[10] Sumitha C., Dr. A. Jaya, Amal Ganesh, "A study on Abstract Summarization Techniques in Indian Languages", Elsevier Proceeding of Computer Science, No. 87, pp.25-31, 2016.

[11] Dipanjan Das, Andre F.T. Martins, "A Survey on Automatic Text Summarization",Language Technologies Institute, Carnegie Mellon University, November 2007.