RESEARCH ARTICLE OPEN ACCESS

AUTOTEXT COMPACTOR BASED ON SENTENCE TOKENIZATION AND EXTRACTION

Komal Sarode, Ms. Suvarna Hajare

Department of Computer Science & Engineering

Jhulelal Institute of Technology, Nagpur

Abstract:

In today's fast-growing information age we have an abundance of text, especially on the web. New information is constantly being generated. Often due to time constraints we are not able to consume all the data available. It is therefore essential to be able to summarize the text so that it becomes easier to ingest, while maintaining the essence and understandability of the information. We aim to design an algorithm that can summarize a document by extracting key text and attempting to modify this extraction using a thesaurus. Our main goal is to reduce a given body of text to a fraction of its size, maintaining coherence and semantics. . In recent times, data is growing rapidly in every domain such as news, social media, banking, education, etc. There is a need of autotext compactor which will be capable to summarize the data especially textual data in original document without losing any critical purposes. Recent literature on automatic keyword extraction and tokenization are presented since text summarization process is highly depend on sentence extraction. This literature includes the discussion about different methodology used for sentence extraction and text summarization.

Keywords: **Abstractive summary, extractive summary, sentence Extraction, Natural language processing, Text Summarization**

I. INTRODUCTION

In the era of internet, plethora of online information are freely available for readers in the form of e-Newspapers, journal articles, technical reports, transcription dialogues etc. There are huge number of documents available in above digital media and extracting only relevant information from all these media is a tedious job for the individuals in stipulated time. There is a need for an automated system that can extract only relevant information from these data sources. To achieve this, one need to mine the text from the documents. Text mining is the process of extracting large quantities of text to derive high-quality information. Text mining deploys some of the techniques of natural language processing (NLP) such as parts-of-speech (POS) tagging, parsing, tokenization, etc., to perform the text analysis. It includes tasks like automatic keyword extraction and text summarization.

Automatic sentence extraction is the process of selecting words and phrases from the text document that can at best project the core sentiment of the document without any human intervention depending on the model .The target of automatic sentence extraction is the application of the powerand speed of current computation abilities to the problem of access and recovery, stressing upon information

organization without the added costs of human annotators.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses various methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. Actually, Natural language processing (NLP) is the ability of a computer to understand what a human is saying to it. NLP is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial intelligence (AI).

The development of NLP applications is challenging because computers traditionally require humans to speak to them in a programming language that is precise, unambiguous and highly structured or,

perhaps through a limited number of clear voice commands. Human speech, however, is not always precise. It is often ambiguous and the linguistic structure can depend on many complex variables, including slang and social context. Major tasks occours in nlp as follow

a. Chunking

Chunking can be defined as the process of dividing the sentence into a set of non-overlapping chunks. Chunks can be identified by observing the applications of stress on certain portions and also the pause/duration, followed by humans while reading a particular statement.

b. Parsing

The term is used to refer to the formal analysis by a computer of a sentence or other string of words into its constituents, resulting in a parse tree showing their syntactic relation to each other, which may also contain semantic and other information.

II. SENTENCE TOKENIZATION AND EXTRACTION OF TEXT

In summarizing document, people may perform a changeable order to ensure the summary document is smooth and coherence. This fact requires a new sentence reduction with the order of reduced sentence is different from the original. In addition to using sentence reduction for text summarization, the information of syntactic is not enough. The semantic information of original sentences should be incorporated with reduction process to enhance the accuracy of reduction process. This fact is also similar to the behavior of human in reduction sentence that they can understand the meaning of original sentences to ensure that important words is remained in reduced sentences.

To satisfy the new requirements mentioned above, we proposed a new sentence reduction based on decision tree model where semantic information is used to support reduction process. The decision tree model is also extended to cope with the changeable order between original sentences and reduced sentences.

It uses various methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

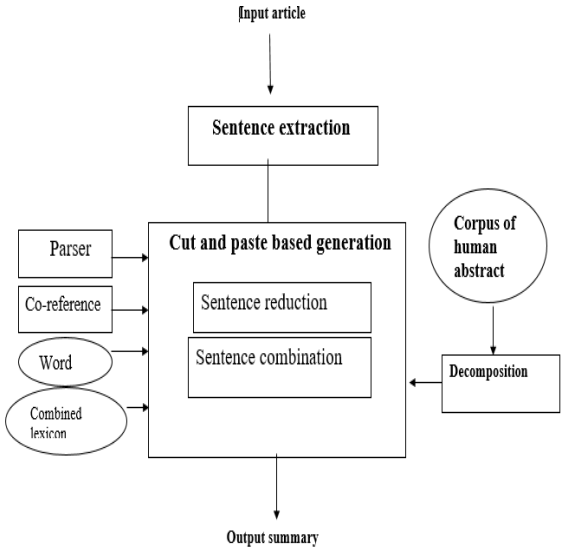

Figure 1: automatic sentence extraction on the basis of approaches used in existing literature

sentence extraction systems can be classified into four classes, namely, simple statistical approach, linguistics approach, machine learning approach, and hybrid approaches.

Simple Statistical Approach

These strategies are rough, simplistic and have a tendency to have no training sets. They concentrate on statistics got from non-linguistic features of the document, for example, the position of a word inside the document, the term frequency, and inverse document frequency. These insights are later used to build up a list of keywords. Cohen utilized n-gram statistical data to discover the keyword inside the document automatically. Other techniques in- side this class incorporate word frequency, term frequency (TF) or term frequency-inverse document frequency (TF-IDF) word co-occurrences and PAT-tree . The most essential of them is term frequency. In these strategies, the frequency of occurrence is the main criteria that choose whether a word is a keyword or not. It is extremely unrefined and tends to give very unseemly results. Similarly, word co-occurrence methods manage statistical information about the number of times a word has happened and the number of times it has happened with another word. This statistical information is then used to compute support and confidence of the words.

a. Linguistics Approach

This approach utilizes the linguistic features of the words for keyword detection and extraction in text documents. It incorporates the lexical analysis , syntactic analysis, discourse analysis , etc. The resources used for lexical analysis are an electronic dictionary, tree tagger, WordNet, n-grams, POS pattern, etc. Similarly, noun phrase (NP), chunks (Parsing) are used as resources for syntactic analysis.

b. Machine Learning Approach

Keyword extraction can also be seen as a learning problem. This approach requires manually annotated training data andtraining models. Hidden Markov model , support vector machine (SVM) , naive Bayes (NB) , bagging , etc. are commonly used training models in these approaches. In the second phase, the document whose keywords are to be extracted is given as inputs to the model, which then extracts the keywords that best fit the model's training. One of the most famous algorithms in this approach is the keyword extraction algorithm (KEA) . In this approach, the article is first converted into a graph where each word is treated as a node, and whenever two words appear in the same sentence, the nodes are connected with an edge for each time they appear together. Then the number of edges connecting the vertices are converted into scores and are clustered accordingly. The cluster heads are treated as keywords. Bayesian algorithms use the Bayes classifier to classify the word into two categories: keyword or not a keyword depending on how it is trained. GenEx is another tool in this approach.

c. Hybrid Approach

These approaches combine the above two methods or use heuristics, such as position, length, layout feature of the
words, HTML tags around the words, etc. . These algorithms are designed to take the best features from above mentioned approaches.

III. TEXT SUMMARIZATION PROCESS

Based on the literature, text summarization process can be characterized into five types, namely, based on the number of the document, based on summary usage, based on techniques, based on characteristics of summary as text and based on levels of linguistics process as shown in Figure 2.

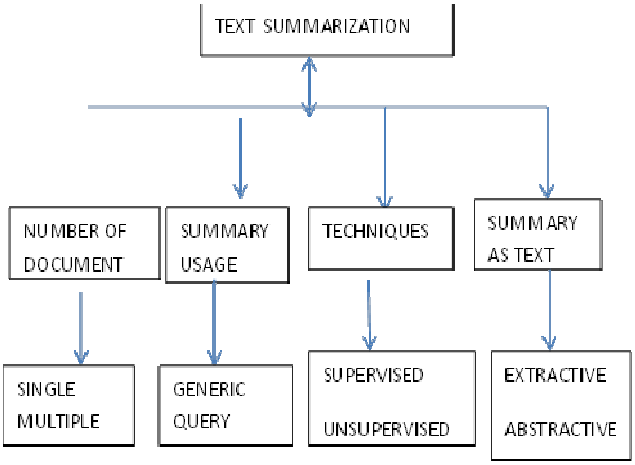

Figure 2: characterization of text summarization

a. **Single Document Text Summarization**

In single document text summarization, it takes a single document as an input to perform summarization and produce a single output document [. Thomas *et al*. [5] designed a system for automatic keyword extraction for text summarization in single document e-Newspaper article. Marcu*et al.* [35] developed a discourse-based summarizer that determines adequacy for summarizing texts for discourse-based methods in the domain of single news articles

b. **Multiple Document Text Summarization**

c. **Query-based Text Summarization**

In this summarization technique, a particular portion is utilized to extract the essential keyword from input document to make the summary of corresponding document . Fisher *et al*. developed a query-based summarization system that uses a log-linear model to classify each word in a sentence. It exploits the property of sentence ranking methods in which they consider neural query ranking and query-focused ranking. Dong *et al.* developed a query-based summarization that uses document ranking, time-sensitive queries and ranks recency sensitive queries as the features for text summarization.

d. **Extractive Text Summarization**

In this procedure, summarizer discovers more critical information (either words or sentences) from input document to make the summary of the corresponding document . In this process, it uses statistical and linguistic features of the sentences to decide the most relevant sentences in the given input document. Thomas *et al*. designed a hybrid model based extractive summarizer using machine learning and simple statistical method for keyword extraction from e-Newspaper article. Min *et al*.used freely available, open-source extractive summarization system, called SWING to summarize the text in multi-document. They used information which is common to document sets belonging to a common category as a feature and encapsulated the concept of category-specific importance (CSI). They showed that CSI is a valuable metric to aid sentence selection in extractive summarization tasks. Marcu*etal*.developed a discourse-based extractive summarizer that uses the rhetorical parsing algorithm to determine discourse structure of the text of given input, determine partial ordering on the elementary and parenthetical units of the text. Erkan*et al*.developed an extractive summarization environment. It consists of three steps: feature extractor, the feature vector, and reranker.

In multiple documents text summarization, it takes numerous documents as an input to perform summarization and deliver a single output document .Mirroshandel*et al*. presents two different algorithms towards temporal relation based keyword extraction and text summarization in multi-document. The first algorithm was a weakly supervised machine learning approach for classification of temporal relations between events and the second algorithm was expectation maximization (EM) based unsupervised learning approach for temporal relation extraction. Min *et al*.used the information which is common to document sets belonging to a common category to improve the quality of automatically extracted content in multi-document summaries.

.

e. **Abstractive Text Summarization**

In this procedure, a machine needs to comprehend the idea of all the input documents and then deliver summary with its particular sentences . It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. Brandow*et al*. developed an abstractive summarization system that analyses the statistical corpus and extracts the signature words from the corpus. Then it assigns the weight for all the signature words. Based on the extracted signature words, they assign the weight to the sentences and select few top weighted sentences as the summary. Daume*et al.* developed an abstractive summarization system that maps all the documents into database-like representation. Further, it classifies into four categories: a single person, single event, multiple event, and natural disaster. It generates a short headline using a set of predefined templates. It generates summaries by extracting sentences from the database.

IV. CONCLUSION

Text summarization is growing as sub – branch of NLP as the demand for compressive, meaningful, abstract of topic due to large amount of information available on net. Precise information helps to search more effectively and efficiently. Thus text summarization is need and used by business analyst, marketing executive, development, researchers, government organizations, students and teachers also. It is seen that executive requires summarization so that in a limited time required information can be processed. This paper takes into all about the details of both the extractive and abstractive approaches

along with the techniques used, its performance achieved, along with advantages and disadvantages of each approach. Text summarization has its importance in both commercial as well as research community. As abstractive summarization requires more learning and reasoning, it is bit complex then extractive approach but, abstractive summarization provides more meaningful and appropriate summary compare to extractive. Through the study it is also observed that very less work is done using abstractive methods on Indian languages, there is a lot of scope for exploring such methods for more appropriate summarization.

REFERENCES

1) Art of abstracting. ISI press, Philadelphia.
2) Cut and paste based summarization handout, Dept. CS, Colombia University.
3) ANSI1997.Guidelinesfor abstracts. Technical reportZ39.141997
4) [Aone et al., 1997] Aone, C., Okurowski, M., Gorlinsky, J., and Larsen, B. (1997). A scalable summarization system using robust nlp. In Proceedings of ACL/EACL'97 workshop on summarization, Madrid, Spain.
5) C. Zhang, "Automatic keyword extraction from documents using conditional random fields," Journal of Computational Information Systems, vol. 4 (3), 2008,
6) E. Hovy, C.-Y. Lin, "Automated text summarization and the summaristsystem.