AUTOMATIC CONCEPT MAP GENERATION AND SUMMARIZATION FROM TEXT

Shakeeb Sheikh[1],Ameya Bhingarkar[2], Sakshi Dhamne[3], Shriganesh Lokapure[4], Sourabh Mahajan[5]

Student, Department of Computer Engineering, Sinhgad College of Engineering, Vadgaon, Pune, Maharashtra, India[2345]

Assistant Professor, Department of Computer Engineering, Sinhgad College of Engineering, Vadgaon, Pune Maharashtra, India [1]

ABSTRACT

There is an enormous amount of textual information on the web and it is growing every single day. It is present in the form of web pages, blogs, articles, etc. and it needs to be structured so that it is easy for someone to get the context and the gist of the overall data in a concise form. It is a tedious task for humans to arrange the data manually and it also requires rigorous analysis of data. For that purpose, the data needs to be summarized. Nowadays, due to advancing technologies, understanding and interpreting concepts has become easier due to the visual representation of data. Concept maps are one of the ways to visually structure and represent data. Concept maps allow the users to group related information so that generation of a concept map becomes easier. This study proposes the method of a machine learning model for summarization and concept map generation without using the existing tools that are already present on the web. Various steps are performed to achieve the final output that is data summarization and concept map generation.

KEYWORDS

Concept map, Concept map mining, Text analysis, text summarization.

I. INTRODUCTION

Text summarization is the technique of extracting important information which gives us the overall idea of the entire document. It also shortens large pieces of data to give us a summary. Text summarization intends to create a coherent and conversational summary. It contains the main points outlined in the document. Text summarization helps us reduce the reading time. While searching for documents, the selection process is made easier. The effectiveness of indexing is also improved. Automatic summarization algorithms are less inclined than human summaries [5].

The visual organization and representation of knowledge are presented through a Concept Map. It is also called semantic maps [4]. It portrays concepts, ideas, and the relationships among them. It may be used by various professionals like instructional designers, engineers, writers, and others to organize and structure knowledge.

In a concept map, each word or phrase is connected to the other and is linked back to the original idea, word, or phrase. Through Concept maps, we can develop study skills and logical thinking by revealing connections. It helps the students to see how individual ideas form a larger whole idea. Concept maps are widely used in education, business, and management [5]. They help in recalling the memory, clarify and structure ideas. They help in developing higher-level thinking skills. They help in communicating complex ideas, arguments and help learners to understand learning objectives, concepts, and the relation between those concepts. The main figures of concept maps are concepts put in circles or boxes.

Concept map mining (CMM) refers to the automatic extraction of concept maps from the present documents. There are two sub-tasks to be performed in CMM: identification and summarization. Identification involves concept identification and relationship identification.

The figure given below explains the concept map mining process:

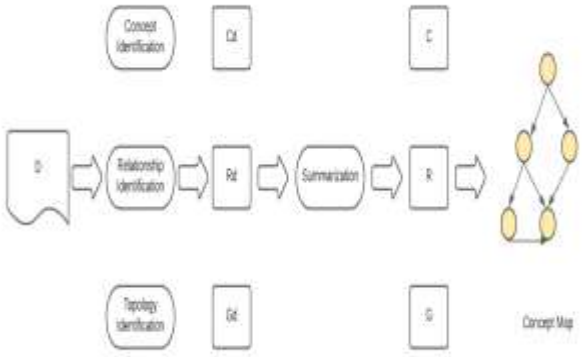

Figure 1. Concept map mining process

D represents the document from which the concept map is generated. Cd, Rd, and Gd refer to all the concepts, prepositions, and levels of generalization expressed in the document. C, R, and G refer to the summarized product of the document D. In the above diagram, concept extraction is the primary phase because it helps in identifying relations R and generalized document G. After the identification process, the identified document undergoes summarization which helps us to shorten data without missing important strings of the document. It delivers the summarized product of the document (C, R, and G). Lastly, the concept map is generated using data provided by C, R, and G [2].

Coreference resolution is a task of clustering data that refers to the same underlying texts. Coreference is the concept that

binds the same entities together. There are various approaches to coreference resolution which can be classified into various categories such as mention-pair, mention-ranking, and entity-based algorithms [1]. Some algorithms are based on neural network architectures. Coreference resolution is a versatile tool used for summarization, sentimental analysis, text understanding, information extraction, and many more. It has shown major development in the field of computation but it has not done much progress in the field of clinical free text. It requires linguistic pre-processing and rich language resources for identifying and resolving expressions automatically.

Apache Solr is an open-source platform that helps us to build search applications. Yonik Seely created Solr in 2004. It is non-relational data storage and data pre-processing technology. It can be used along with Hadoop as Hadoop can handle a huge amount of data and helps us to find the required information. Solr is scalable, fault-tolerant, search engine optimized to search large volumes of data, storage optimized. It has automated failover and recovers our data easily.

The remaining part of the document contains a brief idea about data pre-processing, concept extraction, relationship extraction, and visualizing data.

II. RELATED WORK

Anandika et.al. discuss how important Named Entity Recognition (NER) is in Natural Language Processing (NLP) and hence in the process of identifying entities and summarization. Different approaches for NER are mentioned by the author like Rule-Based NER, Machine learning based NER, and Hybrid NER. Rule-Based NER uses grammatical, syntactic, orthographic, and feature-based rules for the recognition of the named entities [1]. Machine Learning-based NER is better than Rule-Based NER because it is statistical. Different Machine Learning based models are mentioned which use probabilistic and statistical aspects for the classification of the entities. Comparative analysis of different approaches is also done.

1. Support Vector Machine (SVM)

SVM is one of the most demanded Supervised Learning algorithms, which is used for Classification as well as Regression problems. The main motive of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes which helps us to put the new data point in the correct category easily in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help us to create the hyperplane. These extreme cases are called support vectors. Hence, we can term this algorithm as a Support Vector Machine [7].

Applications of SVM:

- Face detection.
- Image classification
- Text categorization
- Bioinformatics
- Handwriting recognition
- Generalized predictive control
- Protein fold and remote homology detection

Types of SVM:

i. Linear SVM -

Linear SVM is used for linearly separable data. Thus, we can say that in this algorithm, the dataset can be segregated into two classes by using a single straight line. We term such data as linearly separable data, and this classifier is known as the Linear SVM classifier.[6]

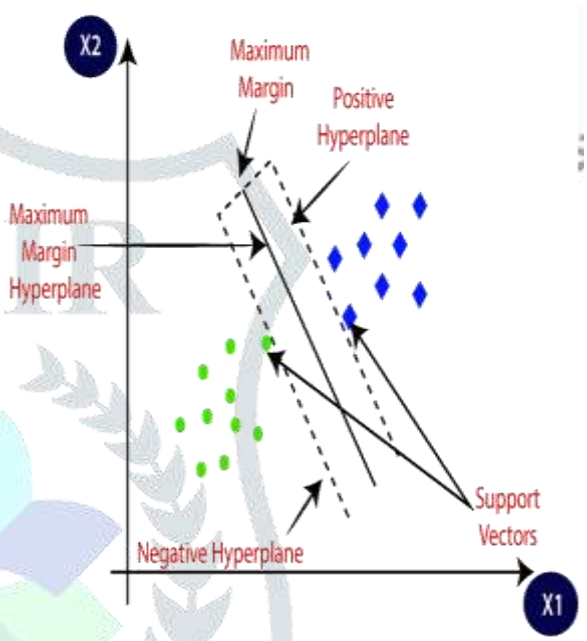

Fig.2. Linear SVM

ii. Non-Linear SVM -

Non-Linear SVM is used for non-linearly separated data. It means that if a dataset cannot be classified by using a straight line, then that type of data is termed non-linear data. The classifier used is called the Non-linear SVM classifier

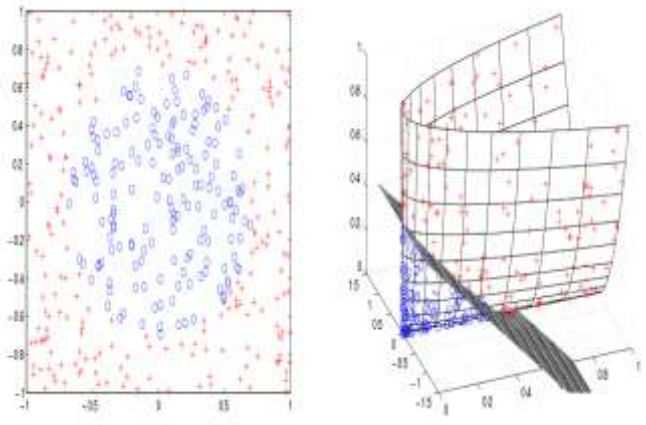

Fig.3. Non-Linear SVM

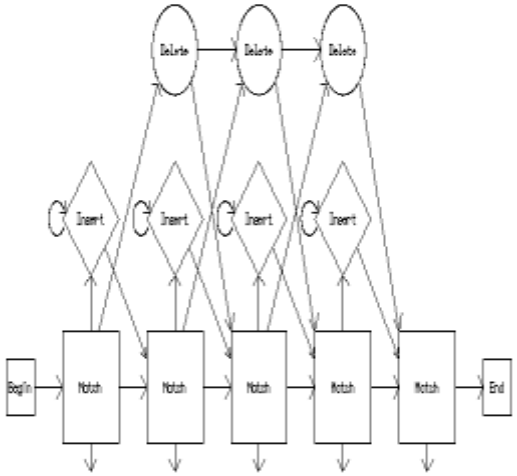

Fig.4. Profile HMM

2. Hidden Markov Model (HMM)

The Hidden Markov model (HMM) is a statistical model that was first put forward by Baum L.E. (Baum and Petrie, 1966). It uses a Markov process that contains hidden and unknown parameters. The observed parameters are used to identify the hidden parameters in the HMM. These parameters are then used for further analysis. The HMM is a type of Markov chain. HMM works on the principle that the observed events should have no one-to-one correspondence with states but should be linked to states through the probability distribution.

HMM is a doubly stochastic process. It includes a Markov chain as the basic stochastic process. Description of the state transitions and the stochastic processes that describe the statistical correspondence between the states and observed values is done. From the observers' perspective, only the observed value can be contemplated, while the states cannot [3].

Applications of HMM:

- Thermodynamics
- Statistical mechanics
- Physics, Chemistry, Economics
- Signal processing
- Information theory
- Pattern recognition
- Part of speech tagging

Types of HMM:

i. Profile HMM-

Profile HMMs are probabilistic models which encapsulate the evolutionary changes that have occurred in a set of related sequences. Profile-HMMs are HMMs with a specific architecture that is suitable for modeling sequence profiles. Profile-HMMs have a strictly linear left-to-right structure that does not contain any cycles which is not the case of general HMMs.

ii. Pair HMM-

The pair hidden Markov model (pair-HMM) is a variant of the basic HMM that is especially used for finding sequence alignments and evaluating the significance of the aligned symbols. The original HMM generates only a single sequence while a pair-HMM generates an aligned pair of sequences.

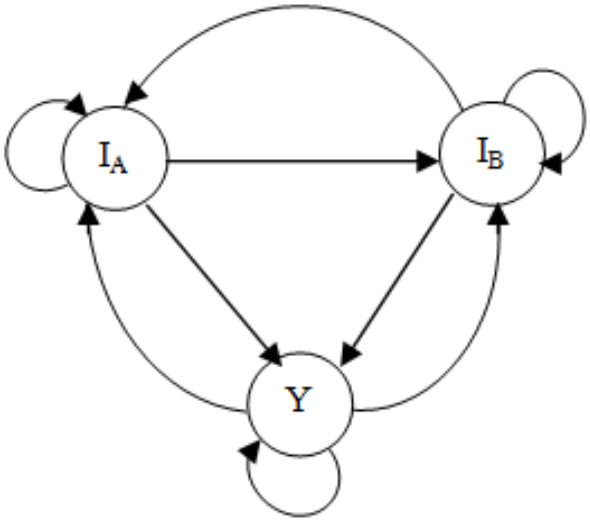

Fig.5. Pair HMM

iii. Context-sensitive HMM-

The main difference between a context-sensitive HMM and a traditional HMM is that a csHMM can use part of the past emissions to adjust the probabilities at certain future states. The utilization of such contextual information is very useful in describing long-range correlations between symbols. This

context-dependency expands the descriptive capability of the HMM considerably [3].

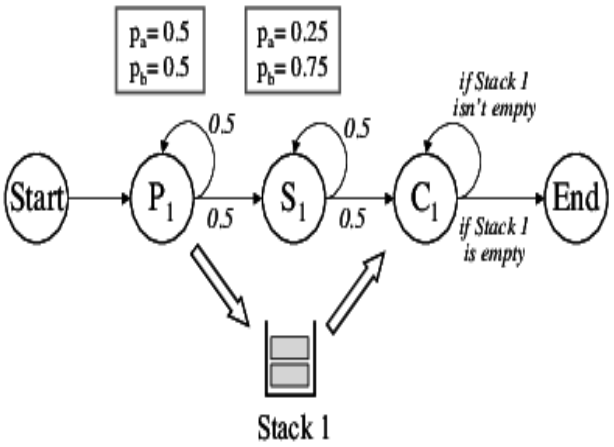

Fig.6. Context-sensitive HMM

III. PROPOSED SYSTEM

The web application should be developed to execute the whole process. The login page is created where a user gives the email address and password to log in or the sign-up option is also available if a new user comes. Log in is done using OTP to maintain privacy. There are 500 words as the input. The given text is first pre-processed for coreference resolution [1].

The sentences obtained from the pre-processing are fed to the tool used. The output of the tool is in the form <subject, predicate, object >. The application then does the concept map extraction. This system allows user to generate the concept map and do summarization together in a single application.

IV. FUTURE SCOPE

Mobile application development should also be considered for future requirements. The login page can store the data of the user and maintain the history for future references instead of doing the same process again and again for the same data. Text limitations can be removed further. The language barrier can be removed and allowance of other languages can be considered. The use of brackets and special symbols can also be allowed. The use of images/diagrams can also be included for the concept map generation and summarization process.

V. CONCLUSION

A lot of research has been done on other languages as compared to Indian languages; hence, it is difficult for Indian languages because of deficiency of annotated corpora, agglutinative nature, and multiple writing methodologies, demanding morphology, no capitalization conception, and many more other deficiencies. HMM does not give us considerable results to identify named entities for Indian languages. SVM is not suitable for large datasets or for the datasets having large noise (target classes are overlapping). In this paper, we discussed two methods that can be applied to NER having their pros and cons.

VI. REFERENCES

[1] Amrita Anadika, Smita Prava Mishra(2019),"A study on Machine Learning Approaches for Named Entity Recognition".

[2] N.R. Kasture, Neha Yargal, Neha Nityanand Singh, Neha Kulkarni and Vijay Mathur," A survey on Methods of Abstractive".

[3] SanjayShimpi and Vijay Patil,"Hidden Markov Model as Classifier: A survey".

[4] Addel Ahmed, Dr.Syed Saifur Rehman ,"DBpedia based Ontological Concepts Driven Information Extraction from Unstructured Text"

[5] Raghavendra Katagall, Rakesh Dadde, R. H. Goudar and Sreenivasa Rao,"Concept Mapping in Education and Semantic Knowledge Representation: an Illustrative Survey"

[6] Peng Sun, Suexhen Yang, Xiaobing Zhao and Zhijuan Wang,"An Overview of Named Entity Recognition"

[7] Hyeran Byun and Seong-Whan Lee,"Applications of Support Vector Machines for Pattern Recognition: A Survey".