Automatic Amharic Text Summarization using NLP Parser

Getahun Tadesse Mekuria[*1], Aniket S. Jagtap*2

Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune - 412115, Maharashtra, India

***Abstract-** The proposed system investigates the problem of building the domain based single and multiple document Amharic text summarization. Multi-document summarization is the main task in natural language processing and summarizing a huge text document into a short and precise format from multiple documents. Multi-document summarization targets to condense the most important information from a set of documents to produce a short summary. Multi-document summarization is also an integral tool for document understanding and well interpreting in the existing system of text summarization. But single text document summarization has been done from a single text document only. Text summarization can be done based on its input, purpose, and output. In the existing system, most research has been done on extractive single document summarization, but now we propose the new system that solves the existing problem by developing the combinations of extractive and abstractive summarization approach on a single as well as multiple document input from the user. To solve such existing problem by using Java programming language for their flexibility and it has a powerful library Java universal network graphic for text summary. PageRank algorithm plays a great role in finding out their sentence score and its weights of a sentence in the document. The proposed model summarizes only text document but in the future, develop text summarization model for all types of document including graph, image, picture, video and other form in addition to the text document.*

Keywords - *JUNG, Amharic Text summary, Abstractive, Extractive summarization, Domain-based summarization, MDS, AATS, JAMA.*

I. INTRODUCTION

Amharic which is the second most spoken Semitic language in the world after Arabic. It has been the working language of the federal government of Ethiopia.Amharic Text summarization is the process of producing automatic short and important summary from a huge document based on the user needs. Text summarization can be categorized based on its input (single document and multi-document), based on purpose (generic, domain or topic based, query-based, etc.), and based on its output (Extractive and abstractive approach). Most of the time researchers initiated when people convert one form of natural language to another language. Natural language processing (NLP) has different applications such as intelligence analysis, information retrieval, public opinion monitoring, political election, news recommendation and others. Text summarization techniques are two types: - Extractive and Abstractive text summarization. Extractive text summarization is produced by condensing or concatenating many sentences, paragraph…, extracted exactly as it appears on the original document. Where as Abstractive text summarization was written to convey the main information and may reuse phrases, clauses from the original document and replace those sentences, phrases, clauses by using general other sentences, phrases, or clauses by finding its dictionary equivalence to the input text document. Abstractive summarization gives full meaning of the original document. Text summarization has the objectives to the decent summary, to produce at least for a technical document, reduction ratio was not a primary concern, reductions did cleverly, and sentence grammar should be maintained accurately. Basically, text summarization follows word-sentence co-ranking approach on the word level. Co-ranking uses for sentence ranking in the graph-based model, which is implemented in PageRank algorithms, used in undirected weighted graph each node of a graph represented by sentence and the weight between sentences represented by words present in the sentence.

As described above many types of research have been done on Amharic text summarization especially on the extractive approach and the researchers suggested for the new coming researcher to come up with better solutions for the existing summarization problems of the text document due to its complexity. Still now Abstractive summarization techniques, not yet developed in the very effective manner, but the techniques of abstractive text document summary are very crucial and meaning full with the original text document, it generates a summary by replacing the existing word, phrase its similar meaning and contains more general information. Some of the research done Amharic text summarization like graph based automatic Amharic text summarizer by Mattias et al (2015)[14,15], Alemebante et al (2015) works on Amharic text prediction system [11,18].

The rest of the paper organized in the following manner. Section 2 statement of the research project problem, objectives, and state motivations, Section 3 discuss related works done in this area; section 4 brief minute discussions about proposed model, section 5 conclusions and future work.

II. Statement problem of the research project

The rapid growth of the information flows on the web and increase the users demand from time to time as well as it is too difficult to find out the most important and currently needed information from the huge text document require automatic text document summarization to save their time and human unbiased expert. The best solution that we are proposed now called automatic Amharic text document summarizations using NLP parser.

Objectives

General and specific objectives of a proposed research projects are stated below.

General Objective

General objectives of the proposed research project are to investigate the applications of domain-based approaches to automatic Amharic text document summarization using NLP Parser.

Specific Objective

To achieve the general objectives of the proposed study, the following list of specific objectives are identified:

- Conduct the literature review.
- Develop domain-based algorithms.
- Develop a prototype of automatic Amharic text document summarization system.
- Evaluate the performance of the developed system.
- Methodology and tools
- Literature review

Conduct on the domain-based approaches that should be used for Amharic text document summarizations. Focus will be given to domain-based modelling and text document summarization works that use these statistics or abstractive methods. Extractive summarization techniques summarize the input text document by filtering out its most important and available document after tokenization's, normalizations and removal of all stop words which are not important and order the sentence based on its similarity score. In the case of abstractive text document summarization, the system made summarizations by following its rules and dictionary meaning of each terms and generalize the input text document by paraphrasing it with equivalent terms. Abstractive types of summarization are similar to human summarization and it is to meaning full. Conduct latest literature review on the several of summarization. Domain-based text document summarizations in the graph-based modeling approach PLSA, LSA, LDA, google page ranking, develop domain-based algorithm, prototype of automatic Amharic text document summarization system. Evaluate the performance of the developed system. comparisons and analysis of the final results of the developed system[15,16].

Literature review to conducted on domain-based approaches that should be used for text document summarizations. Particular focus will be given to techniques of domain-based modelling and text document summarization works that use these statistical or abstractive methods.

Motivation

A huge volume of information in the text document generated across the world . Documents like Reporter news, political elections, opinion monitoring, intelligence analysis, web-based documents. In the current scenario technology and information become huge and complex and to find out very short and key information from the huge document is too difficult so, it needs a short and automatic summarization model to overcome such problem. Secondly, it is too difficult to get unbiased human expert for text summarization. Thirdly, the accuracy of the summarized text document is not yet gained and others. Fourthly, mostresearch done by using extractive text summarization techniques even if abstractive text summarization used but it is nominal.

III. RELATED WORK

A. *History of Automatic text document summarizations*

Text document summarization starts within the 1958 by Luhn(15), most important work has been started by using word frequency.

According to **C** Fang et al(2017)[2] develop a word-sentence supervised ranking model by using extractive text summarizations of a single document and apply page ranking algorithms of undirected weighted graph-based modeling techniques [1,2]. Among many similarities measuring techniques such as dice, cosine, Jaccard methods select Jaccard similarity techniques. In these techniques the undirected weighted graph, each node represented by the sentence and the weight of the graph represented by its edge. Term-frequency and inverse document frequency(TF-IDF)[1,2,4,7,15] used to compute the most important and convergence of the sentence.

M Yousefi-Azar et al(Expert Systems with Applications, 2017 – Elsevier)[7] summarized a text document of extractive query oriented single document by using deep auto-encoder and compute feature space from its term frequency input. Analysed local and global vocabulary. Prepare ensemble or combined noise auto-Encoder (ENAE). ENAE is a stochastic improvement of auto-encoder (AE) and that helps to add noise to the input text document and selects the most important top key sentences from the combined noise runs. In many research extractive text summarization, most of the time called Sentence Ranking. The model helps to summarize the data set of more than 40 email threads from 11 emails and below it.MA Tayal, MM Raghuvanshi, LG Malik (- Computer Speech & Language, 2017 – Elsevier) develop a text summarizer model by using soft computing and in the general natural language processing process data through lexical, syntax, semantic, programmatic analysis of the document[3]. Intelligent text summarization is one of the most challenging tasks in natural language processing (NLP).NLP used for story telling, question answering, SVO rules, and part of speech tagging (POS Tag) Process data through Pos Tagger, NLP parser, Semantic Regression, Sentence Reduction. Sentence ambiguity Removal and sentence combination. Text summarizations techniques grouped into three,these are -1. Extract 2. Abstractive (Linguistics) and 3.Hybrid (combinations of extractive and abstractive).

Extractive Text summarization Techniques: These types summarization techniques based on text features such as keyword, title word, cue word, sentence positions and length of sentence [1, 3.6, 8, and 16].

Abstractive summarization Techniques: It helps to identify relationships of terms the document via Part of speech tagging, grammar analysis, Extractions of the meaning full sentence and the likes [3].Generally, automatic text summarization using soft computing represent in the following seven steps [4].

- Take the input document from the user and differentiate the document purpose and the user of these text documents.
- Remove the unwanted words like pronouns, and enhance its level.
- Clustering of text document according to its relevance.
- By using NLP parser verifies its structural error and removes its ambiguity.
- Identify and prepare its sentence in reduce format by using similarity score, and leading title character score.
- Combinations of sentence done in these step by using subject, verb, andobject (SVO) rules representation and based on its similarity in first-order logic principle.
- In the final stage summary of the text document produced according to the required level of the summary percentage.

Based on the Madhuri [3] text document summarization classified into:-

- Surface-based approach (statistics, graph-based)
- Semantic-based approach (Abstractive-Linguistic)
- Combined approach (Statistics and Semantics).
- General summarization techniques-clustering, learning, fuzzy logic…

Graph-based summarization techniques is a purely linguistics and implementation cost is very high and it needs a long period of time executions.

Sunitha C. et al (2016). A Study on Abstractive Summarization Techniques in Indian Languages. *Procedia Computer Science*, *87*, 25-31. Natural language processing becomes a wide research with great advantages for the people, after starting interpreting one natural language in another language. Text document summarization becomes an effective task in NLP by concatenating of text together to produce meaning full and short key summary of a given text [4].in most case summarization takes place in extractive techniques even if abstractive summarization techniques were done but it is nominal and symbolic that means it is not working on the real or practical arena. Abstractive type of techniques are very important but not developed for Amharic text document summarization. Summarization techniques are very useful for in many applications such as- online email summarization, online new article summary, product review summary, automated research for business organizations summary with minimum human interventions. The problem of abstractive text summarization due to its complexity rises in the following manner such as how to select the most important part without losing its original text document meaning, secondly how to represent in a condensed manner and how to produce a reducible generated summary. Infact, an abstractive type of text summarization techniques can be divided into two broad categories such as 1. Structured based and 2.semantic based summarization techniques.

Structure-based abstractive text document summarization techniques, at the very beginning important terms, sentences, paragraph, collected in the first predefined structured format to obtain the required abstractive text document summary without losing its meaning. The predefined structured format can be template based, background information, tree-based structure, lead and body-based phrase structure approaches with similar sentence and terms are extracted from the original text document.As stated in the above Template based structure abstractive summarization techniques extract main part of the text document by using keyword and represent each-template format. Tree-based structured approach extracts the original text document by using the parser and organized or populated into a tree structure and it follows predicate tree structure. In the ontology-based technique, summarized text document of the original document preprocess text document to extract the important key term which is mapped to concepts and relations and the predefined ontology that will be converted in the meaningful abstractive summary. For the overall process to be performed the rules applied to every module to get the needed and meaningful text document which are representatives of the original text document. A number of modules or functions exist some of them such as preprocessing module, a categorization module, character analysis module, summary generation module and others depending on the text required to be done. At the end of the paper, abstractive summarization comparison has been done by using collected data and performance evaluation metrics such as precision, recall, and f-measure computed.

R. Abbasi-ghalehtaki et al (2016) used fuzzy evolutionary cellular learning automata model for text document summarization by using text feature [8]. Text features including Word feature and Sentence feature. Word features include keyword, pronoun, proper noun, cue word, and others. Sentence features such as sentence position, sentence length, the type of word the sentence contains the topic word or not and other. The combinations of Fuzzy Logic, Swarm intelligence, Cellular Learning Automata yield a better result. The fuzzy logic used for text summarization by applying First order logic .Genetic algorithm used as a tool for extracting a sentence from the text document summarization time. Particle swarm optimization used for feature selection problem in the text document summarization and it’s a good application in text clustering and text categorization. Term frequency (TF) and similarity measure also help for better summarization. Term frequency plays an important role in text summarization and it is an approach to identify very crucial sentence along with reductions of the sentence in information redundancy. Like the PSO similarity measure used as a measure to discover unseen knowledge from a textual database. Usually, short text similarity is applicable such as text summarization, text categorization, and machine translation. Cellular learning automata bring out joint n-gram within short periods of time, artificial bee colony used to classify n-friend and optimize similarity measure, next to these PSO-GA helps to allocate fair text feature by separating text into two parts these are most important and less important. Lastly fuzzy logic system utilized to score sentence and extract key point summary [9].

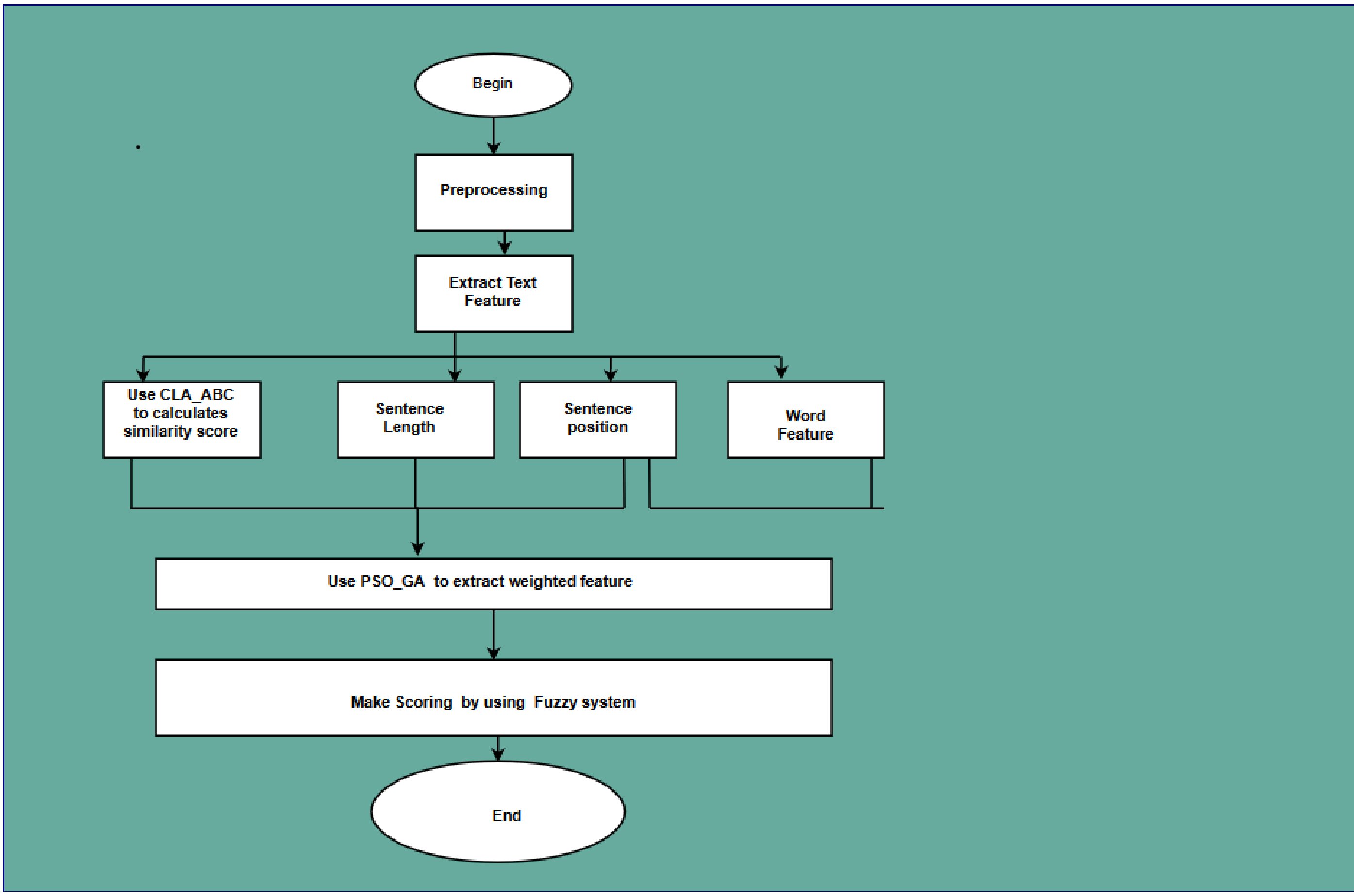
Begin
Preprocessing
Extract Text Feature
Use CLA_ABC to calculates similarity score
Sentence Length
Sentence position
Word Feature
Use PSO_GA to extract weighted feature
Make Scoring by using Fuzzy system
End

Figure 1. Represent Fuzzy Evolutionary Text summarization.

Patel et al. (2017) stated automatic text summary as the process of summarizing text content from a huge content of the original document. Based on Patel study, follow two summarization techniques such as Extractive and Abstractive approach. From the two approaches, Extractive summarization has two phases. These are -Preprocessing and processing phases [8]. Preprocessing is the initial step for generating structural representations of the text document. In preprocessing phase, the following actions performed. Some of the tasks performed under preprocessing such as

a) **Sentence Segmentation**-it is the process of dividing the text document into a sentence and converts a raw text document into a list of sentence strings.
b) **Tokenization-** identify the word token from the given text document and breaking the sentence into words and extract a word from a sentence.
c) **Part of speech tagging** (pos-tagging) here analysis output of tokenization or sequence of words and assign appropriate speech tag for each word.
d) **Named Entity Detection** (NED)

It is one of the major research areasespecially for text summarization; it helps to provide identifications of predefined categorizations of an object such as a person, organizations, locations, percentage etc. The system called Named Entity Recognition (NER) and provides entity detection dynamic processing used linguistic grammar based techniques and identify the statistical model to identify the well-known entity [4].

e) **Relation detection**

Relation detection helps to identify the possible relation among chunked sentence. By providing: -co-occurrence words, provide a link between pronouns, corresponding nouns,co-reference.

B. *Data collection*

The data used for evaluating the proposed text document summarization system model can be find out from any type of Amharic text document. It can be collected from Ethiopian Reporter News=>Sport news, marketing news, weather forecasting news, and others collected from the website that freely available. For the evaluations of the text document summary three unbiased human expert were required. Human expert summarize the Amharic text document manually.

C. Development tools used for proposed system

Different tools are used for the development of the proposed summarization system. Java programming language used for the development of the summarization model. Java support platform independence and it is suitable for encoding Unicode. It has its own library called JAMA and other Library of JUNG. Both library used for the development of proposed automatic text document summation model. Several summarizations are there but now we select java programming language and netbeans 8.3.1.

D. Performance Evaluations criteria

Performance of the proposed system can be evaluated by comparing the three text document summarized manually using human expert and the system generated summary. The well known evaluation metrics are precision P, recall R, and f-measure F.

IV. The proposed system Architecture and its Minute descriptions

The proposed system design to summarize both single as well as multiple input text document besed on the user needs. Most researchers conduct a text document summarization system by using extractive summarization approach.But now, we propose for both extractive and abstractive text document summarization approach. See the architecture of the proposed system below.

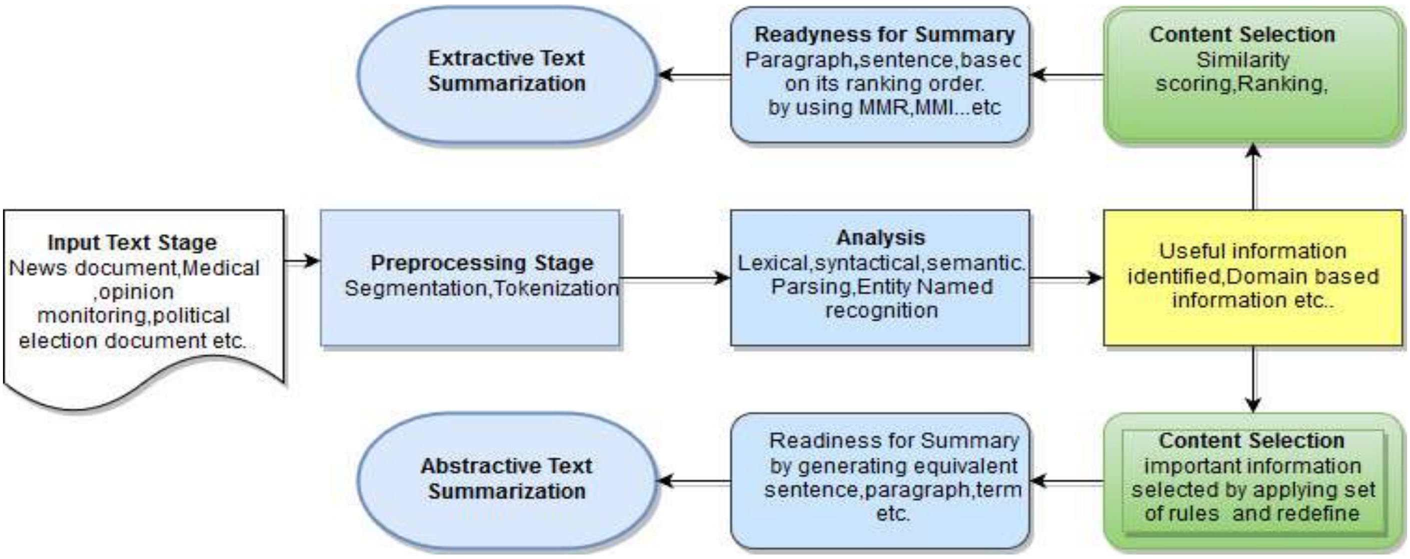
Extractive Text Summarization
Readyness for Summary Paragraph,sentence,basec on its ranking order. by using MMR,MMI...etc
Content Selection Similarity scoring,Ranking,
Input Text Stage News document,Medical ,opinion monitoring,political election document etc.
Preprocessing Stage Segmentation,Tokenization
Analysis Lexical,syntactical,semantic. Parsing,Entity Named recognition
Useful information identified,Domain based information etc..
Abstractive Text Summarization
Readiness for Summary by generating equivalent sentence,paragraph,term etc.
Content Selection important information selected by applying set of rules and redefine

Figure 2. Architecture of the proposed system

As we see in the above figure 2 the user give input text document (text document can be News, sport, business, weather forecasting, etc) to the system and the system accept the input text document proceed to pre-processing (sentence segmentations, tokenization's, normalizations etc.) takes place. After pre-processing the step the analysis of the pre-processed text document has been done (parsing=>lexical, syntactical and semantic, Named entity identification,). The analysis phase completed successfully the thematic information shall be identified and give options to select type of summarization techniques. After selecting type of summarizations the system generate the required summary to the end user.

V. Conclusion

Text document summarization is the process of generating a short and significant part of the text document from a huge text document based on the user's needs. Text document can be summarized depending on different summarization approaches or techniques. These are: Based on its input such as single document or multiple documents. Summarizations made from single document input is generating summary by taking only a single document as input and produce a summary by pick up most important and high scoring paragraph, sentence, clauses, phrases, and words. When the user wants to find out only main point from a document, it is too difficult to get important point within a short period of time without the text summarizer support. To save time and effort it is too important single document summarization techniques. Whereas multi-document summarization approaches help to summarize a text document form multiple input of text document of multiple source and generate a summary of by picking the most important and currently needed only based on its scoring values of the input text document of the input text document. Summarizations made based on its purpose such generic type, topic-based, domain-based, query oriented, and other. For the implementations of the text document summarization java programming language is preferred within its powerful library JAMA and JUNG that helps to summarize the text document.

The research conducted using natural language processing mainly focused on the extractive summarization techniques and unique document summarization but in the future the researcher conduct on an improvement of this and extend it into summarizations of all types of documents which is represented by images, large video, and any type of document that needs the user in the summarized form without fail, including diagrams, images.

Reference

1. Changjian Fang a, Dejun Mu a, Zhenghong Denga, Zhiang Wub,∗. Word-sentence co-ranking for automatic extractive text
summarization, *Expert Systems With Applications 72 (2017) 189–195*
2. Fang, C., Mu, D., Deng, Z., & Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, *72*, 189-195.
3. Geetha JK. Kannada Text Summarization Using Latent Semantic Analysis,2015 IEEE.
4. Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999, August). Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 121-128). ACM.
5. Hongjie Chen. Modelling Latent Topics and Temporal Distance for Story Segmentation of Broadcast News, *VOL. X, NO. X, MONTH YEAR, 2016.*
6. Kai Li. Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis, *VOL. 37, NO. 6, JUNE 2015.*
7. Kanapala, A., Pal, S., & Pamula, R. (2017). Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 1-32.
8. Patel, S. M., Dabhi, V. K., & Prajapati, H. B. (2017). Extractive Based Automatic Text Summarization. *JCP*, *12*(6), 550-563.
9. R. Abbasi-ghalehtaki, et al., Fuzzy evolutionary cellular learning automata model for text summarization, *Swarm and Evolutionary Computation (2016), http://dx.doi.org/10.1016/j.swevo.2016.03.004*.
10. Shagan Sha, et al. Semantic Text summarization of Long Video,*2017 IEEE Winter conference on applications of computer vision.*
11. Sunitha, C., Jaya, A., & Ganesh, A. (2016). A Study on Abstractive Summarization Techniques in Indian Languages. *Procedia Computer Science*, *87*, 25-31.
12. Tayal, M. A., Raghuwanshi, M. M., & Malik, L. G. (2017). ATSC: Development of an approach based on soft computing for text summarization. *Computer Speech & Language*, *41*, 214-235.
13. Vinay Kumar Jain et al. Extraction of emotions from multilingual text using intelligent text processing and computational linguistics, *a journal of computational science 21(2017)316-326 Elsevier*.
14. Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412-420).
15. Yirdaw, E. D. (2011). Topic-based Amharic Text Summarization. *Master's thesis, Faculty of Computer and Mathematical Science, Addis Ababa University*.
16. Yirdaw, E. D., & Ejigu, D. (2012, October). Topic-based Amharic text summarization with probabilistic latent semantic analysis. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (pp. 8-15). ACM.
17. Yousefi-Azar, M., & Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications*, *68*, 93-105.
18. Yogesh Kumar Meenaa.*, Dinesh Gopalanib. *Domain Independent Framework for Automatic Text Summarization, 2015.*