International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

AMHARIC TEXT DOCUMENT SUMMARIZATION USING PARSER

Getahun Tadesse Mekuria[1], Getahun Tadesse Mekuria[2],
Department of Computer Science and Engineering,
Computer Science and Engineering ,
Pune, India.
getahun.mekuria@sitpune.edu.in,
aniket.jagtap@sitpune.edu.in

May 26, 2018

Abstract

Amharic document summarization is the way of producing currently important and relevant text document from the source text document. Document summarization could be single document or multiple document or combinations of the two text document summarizations. Based on the availability of multiple Amharic text document on different data center.The newly proposed system solve the problem of existing system by using different techniques of text document summarization approach. Summarization performed depending on its input text document. It might be single text document or multiple text documents as input and its final output. Text document summarization developed depending on generated out put such as abstractive or extractive type. It is the blue print task in natural language processing and a summary of a huge text document into a short and unique format from multiple sources. Multi-document summarization mainly target to condense the most important and relevant information from a collections of source documents to produce a short summary. We propose a system that addresses the existing system problem. The system implemented using Java programming language. For

matrices computation use java library called JAMA. PageRank algorithm used for finding out their sentence score and its weights of a sentence in the text document. Summarization of a source document generated using summarization rate of (35%,30%,25%) and by selecting summarization method. ***Key Words*:**-AATS,Amharic Text document summary,MDS, JAMA.

1 INTRODUCTION

Amharic is National language of Ethiopia and it has been a written language for a long time years a go.It has more than 100 million speakers inside the country Ethiopia and world wide.Like Canada, LosAngeless, Israel, Egpt and other. Amharic is the second most spoken Semitic language in the world next to Arabic. It has been the working language of the federal government of Ethiopia. Amharic language has its own Key Dialects like Gondar Amharic, Gojam Amharic, Wollo Amharic and also Shoa Amharic dialects(standard written and spoken Amharic)[1]. But there is no major variation among dialects to dialects, if so very easy to understand. Amharic text document summarization is the process of generating 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 candidate election, day to day 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 together, extracted exactly as it appears on the original document. Where as Abstractive text summarization was written to convey the main information and may reusability of features of phrases, clauses, sentences 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. The textual nature of the observed text document might be changed by new text document that describe the original text document in the abstract manner. Abstractive text document summarization techniques gives full meaning of the original document. 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 Page Rank 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 about Literature review or researches done in this area; section 4 brief minute discussions about proposed model, analysis of results, section 5 conclusion, and future work.

- Statement problem of M.Tech research

Currently, information flow increases, as well as technology growth in rapid speed and digital text document availability, increases on the net. To use such digital text/data for specified level of content from the huge text document was too difficult with out losing

its originality. So, automatic Amharic text document summarizer proposed to solve such problem.

- Objectives

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

- General Objective

Main goal of the proposed research are to find out an applications of domain-based approaches to automatic Amharic text document summarization using Parser.

- Specific Objective

To achieve the general objectives of the proposed study, the following list of specific objectives are identified: Conduct the literature review, use domain-based algorithms, develop a prototype of automatic Amharic text document summarization system, and evaluate the performance of the developed system. 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 tokenizations, 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. We are motivated to conduct such research by observing huge volume of information in the text document generated across the world which is written in Amharic. Documents like News, political candidate elections data, opinion mining and monitored data, intelligence analysis, web-based documents and the likes. In the current scenario technology and information emerged with high speed and it is huge and complex and to find out key information from the huge text 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, most research done by using extractive text summarization techniques even if abstractive text summarization used but it is nominal.

2 RELATED WORK

A. History of Automatic text document summarizations

Luhn (1958) start text document summarization within the most important work has been started using word frequency (15).He is the first person that produce summarized document in the earlier stage. Based on C Fang et al(2017)[2] developing a word-sentence supervised ranking model using extractive text document summarizations of a single document by applying page rank 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)[17] produce a summary of a text document from the huge source of document, extract a single document

by using deep auto-encoder and compute the feature space from its term frequency of input. It make Analysis local and global vocabulary. Prepare ensemble noise auto-Encoder (ENAE). ENAE is a stochastic improvement of auto-encoder (AE) and helps to add noise to the input text document and choice the most important key sentences from the ensemble noise runs. In many research extractive text document summarization used. Most of the time called Sentence Ranking. The system 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)[12] developed a text summarizer model using soft computing and general natural language processing process data through syntax lexical, semantic, programmatic analysis of the document[3]. Intelligent text summarization is one of the most challenging tasks in natural language processing (NLP).Natural language processing(NLP) used in the area like story telling, question answering, SVO rules, and part of speech tagging (POS Tag) Process data through Pos Tagger, NLP parser, Semantic Regression, Sentence Reduction(minimization). Sentence ambiguity Removal and grouping of sentence together. Text document summarizations techniques can be grouped into three, these are Extractive, Abstractive (Linguistics) and Hybrid (union of extractive and abstractive). Extractive Text document summarization Techniques: This technique summarized a text document based on text features. Text features can be keyword, title word, cue word, sentence positions (where the sentence located) and how much it long and how much it short [6, 8, and 16].

Abstractive Text document summarization Techniques: Abstractive text document summarization techniques helps to identify inter relationships of terms in the document through Part of speech tagging, grammar analysis, Extractions of the meaningful sentence and other [3]. Automatic text document summarization using soft computing represented using seven steps listed below [4]. Accept input text document from the user and identify its purpose and the user of these text documents. Ignore unwanted words like pronouns, and enhance its level. Clustering of text document according to its relevance. Verifies its structural error and removes its ambiguity using NLP parser, prepare a sentence in reduce format by using sentence similarity score, and leading title character score,

group a sentence by using subject, verb, and object (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. Madhuri [3], text document summarization categorized into:- Surface-based approach (statistics, graph-based),Semantic-based approach (Abstractive-Linguistic),Combined approach (Statistics and Semantics),General summarization techniques-clustering, learning, fuzzy logic. Cost of Graph-based summarization techniques is very high and it is a purely linguistics, implementation needs a long period of time executions. Sunitha C. et al (2016).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. Natural language processing play a great role in various areas of study in the field of computer science. Summarizations techniques is handy in various application area like online News article summary, product review summary, online e-mail summary, automated research summary, abstracted information summary for government officials, business organization summary with minimum human intervention. Document summarization is an effective task in natural language processing by concatenating text documents together to produce meaning full, short and precise key summary of a given text document [4].In most case document summarization takes place using extractive techniques even if abstractive document summarization techniques were done but it is nominal and symbolic due to the reason abstractive summarization techniques require dictionary look up table that means it is not working on the real or practical arena. Abstractive type of summarization techniques is very important but not yet developed for Amharic text document summarization. Summarization techniques are very useful for 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 sum-

mary. 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. Abstract text document summarization techniques Structure-based, at the beginning important terms, words, sentences, paragraph, collected together in the first predefined structured format to form the specified abstracted text document summary without missing its originality. The predefined structured format can be template based, background root information, tree-based structure, lead and body-based phrase structure approaches with similar sentence, phrases 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) fuzzy evolutionary cellular learning automata model used for text summarization using text feature [8]. Text features such as Word and Sentence feature. Word features such as keyword, pronoun, proper noun, cue word, and the likes. Sentence features are features characteristics of sentence like position (where the sentence is located beginning, middle, end) and sentence length (how it long or how it short). The collections of Fuzzy Logic, Swarm intelligence, Cellular Learning Automata give a better result. Fuzzy logic used for text document 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 its 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. Likewise, part of switch optimization similarity measure is a measure to discover hidden knowledge from a textual data store. Usually, text similarity is applicable likewise text summarization, text categorization or grouping of text or transformation , and machine translation. Learning automata bring out joint n-gram text within specified and short periods of time, artificial bee colony selective techniques to classify n-friend and optimize similarity measure, follow up of these PSO-GA helps to assign fair text feature by dividing text into two parts these are most important and less important. Finally used fuzzy logic system utilized to score sentence and produce summary [9].

Amharic Writing System

Amharic writing systems is a component of knowledge base systems. It has its own philosophical behaviour, and writing systems that support in synthesizing thoughts, ideas, and indeed via the use of symbols and other pictorial representations. Basically, writing is a process of recording, objectify, and organize activities and thoughts through images and graphs and a means to inscribe meanings that are expressed through sounds. Writing system helps record to previously and currently available data or informations and transmission from generation to generation without missing or losing of real content.

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

- 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.
- Tokenization- identify the word token from the given text document and breaking the sentence into words and extract a word from a sentence.
- Part of speech tagging (pos-tagging) here analysis output of tokenization or sequence of words and assign appropriate speech tag for each word.
- Named Entity Detection (NED): it helps to provide identifications of predefined categorizations of an object such as a person, organizations, locations, percentage etc[4].
- Relation detection: helps to identify the possible relation among chunked sentence. By providing: - co-occurrence of words, provide a link between pronouns, corresponding noun, 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 such as 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. Tools used for proposed system development

Basically, hardware and software tools are used for the development of the proposed text document summarization system. Java support platform independence and it is suitable for encoding Unicode. It has its own library that has a capability JAMA and other Library of JUNG. Both Java library used for the development of proposed automatic Amharic text summation system . Lucene used as information retrieval.Several summarizations types exist developed by using a number of programming language there but now we select java programming language eclipse/netbean.

D. Performance Evaluations criteria

Performance of a system can be evaluated by comparing the text document summarized manually using human expert and the system generated summary.Well, known performance evaluation metrics are precision, recall, and f-measure.

3 SYSTEM ARCHITECTURE AND ITS MINUTE DESCRIPTION

Proposed ystem designed for amharic text document, and by the help of this system, generate appropriate text summary from the huge source. Text document summarization can be performed by selecting its summarization rate as well as methods of summarization used for summary generation. The rate of text document can be 35,30,25 percent of the original text document. Most researchers conduct a text document summarization system by using extractive summarization approach. Abstractive type of summarization techniques also used, but it is too much rare and nominal. Abstractive techniques needs dictionary/Table Look up.

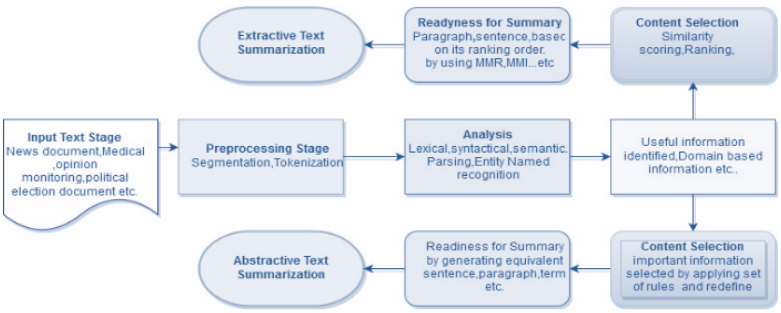

Figure 1 Architecture of the proposed system

From the above architecture of the system, we observe that input text document can be entered by the user. The system accepts the input text document from the user and immediately perform pre-processing task. Under pre-processing, a number of activities performed like tokenization, segmentations, normalizations and other tasks take place. Consequentively After pre processing the step the analysis of the pre-processed text document has been

parsed. 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.

4 RESULT ANALYSIS OF THE SYSTEM

LSA can be written as Latent Semantic Analysis. It use parsed word as input which are defined as unique characters string. First, it generate term document matrix which contain row words and column-paragraphs of a given document. Secondly, it generates document-document or paragraph-paragraph matrix and term-term matrix using the first term document matrix. Thirdly it computes singular value decompositions (SVD). SVD helps to identify patterns and relationships between terms and provide text document. LSA get its name because of SVD which is applied to document word matrices, document or word text graph that are semantically related to each other even if there are no common word.Words which have the same style are also in the same singular value decomposition. The result generated based on the selected method of summarization and corresponding to its selected summarization rate. Summarization rate can be 35%,30%,25%.

The performance of the system can be measured using precision, recall, and f-measure

$\mathbf{P} = \frac{TP}{TP+FP}$, $\mathbf{R} = \frac{TP}{TP+FN}$, $\mathbf{F} = \frac{2*P*R}{P+R}$,

where P= $precision, R = \ Recall, F = \ F - measure$

5 CONCLUSION

Summarization is the process of filtering and generating irreducible short and important text document from huge text document based on the users desire. Text can be summarized by using different summarization approaches or techniques. It may be static or dynamic approach. Based on its input such as single document or multiple documents. Summarizing a single document input is generating

summary by taking only a single document as input and produce a summary by selecting most important and high scoring paragraph, sentence etc. When the user wants to find out main point of a huge document without taking much time is too difficult so, text document summarization is an appropriate solution. To save time and effort use document summarization techniques. Document summarization techniques can be used based on summarization rate (35%,30%,25%) within the selected method(Topic,PageRank,HITS). Summarizations also made based on its usage such as generic type, topic-based and domain-based, and query oriented. These summarization models implemented using java programming language within its appropriate IDE called Eclipse/Netbean. Text document summarization implemented using java programming language due to its platform independence as well as code reusability. We use java powerful library for matrices computation called JAMA for information retrievals Lucerne used. Currently solve many problem faced like text summarization, information retrival,etc on the extractive summarization techniques and abstractive unique document summarization[22] 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 and video, audio.

References

[1] Changjian Fang a, Dejun Mu a, Zhenghong Denga, Zhiang Wub. Word-sentence co-ranking for automatic extractive text summarization, Expert Systems With Applications, 2017.

[2] [2] Fang, C., Mu, D., Deng, Z., Wu, Z. Word-sentence co-ranking for automatic extractive text summarization. Expert Systems with Applications, 2017.

[3] Geetha JK. Kannada Text Summarization Using Latent Semantic Analysis. IEEE, 2015.

[4] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J. Summarizing text documents: sentence selection and evaluation

metrics. 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[5] Hongjie Chen. Modelling Latent Topics and Temporal Distance for Story Segmentation of Broadcast News, 2016.

[6] Kai Li. Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis, JUNE 2015.

[7] [7]Kanapala, A., Pal, S., Pamula, R. Text summarization from legal documents: a survey. Artificial Intelligence Review, 2017.

[8] Patel, S. M., Dabhi, V. K., Prajapati, H. B. Extractive Based Automatic Text Summarization. JCP, 2017.

[9] R. Abbasi-ghalehtaki, et al., Fuzzy evolutionary cellular learning automata model for text summarization, Swarm and Evolutionary Computation, 2016.

[10] Shagan Sha, et al. Semantic Text summarization of Long Video. IEEE Winter conference on applications of computer vision, 2017.

[11] Sunitha, C., Jaya, A., Ganesh, A. A Study on Abstractive Summarization Techniques in Indian Languages. Procedia Computer Science, 2016.

[12] Tayal, M. A., Raghuwanshi, M. M., Malik, L. G. ATSC: Development of an approach based on soft computing for text summarization. Computer Speech Language, 2017.

[13] Vinay Kumar Jain et al. Extraction of emotions from multilingual text using intelligent text processing and computational linguistics, a journal of computational science, Elsevier, 2017.

[14] Yang, Y., Pedersen, J. O. A comparative study on feature selection in text categorization. Icml, 1997.

[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. Topic-based Amharic text summarization with probabilistic latent semantic analysis. In Proceedings of the International Conference on Management of Emergent Digital EcoSystems, 2012.

[17] Yousefi-Azar, M., Hamey, L. Text summarization using unsupervised deep learning. Expert Systems with Applications, 2017.

[18] Yogesh Kumar Meenaa, Dinesh Gopalanib. Domain Independent Framework for Automatic Text Summarization,2015.

[19] Mekuria, G. T., Jagtap, A. S. Automatic Amharic Text Summarization using NLP Parser.

[20] Tachbelie, M. Y. Morphology-based language modeling for Amharic, 2010.

[21] Hennig, L. Content modeling for automatic document summarization, 2011.