INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Integrated Development of a Query-Based Document Summarization for Afaan Oromo Using Morphological Analysis

Jemal Abate Jilo[1]*, **Ashenafi Tulu Alemu**[1], **Fikirte Zemene Abera**[1], **Faizur Rashid**[2]

1 Lecturer, Department of Information Science, Haramaya University Haramaya, Ethiopia
2 Assistant Professor, Department of Computer Science, Haramaya University, Ethiopia

OPEN ACCESS

Received: 26.06.2021
Accepted: 28.09.2021
Published: 18.11.2021

Citation: Jilo JA, Alemu AT, Abera FZ, Rashid F (2021) An Integrated Development of a Query-Based Document Summarization for Afaan Oromo Using Morphological Analysis. Indian Journal of Science and Technology 14(38): 2946-2952. https://doi.org/10.17485/IJST/v14i38.1182

* **Corresponding author**.

abatejemal@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To develop document summarization for the Afaan Oromo language based on the query entered by the user(s). **Methods:** This study follows the design science analysis technique as a result of its considerations of thoughtful, intellectual, and ingenious activity throughout problem-solving and the creation of knowledge. The developed query-based framework has used the TF-IDF term weight methodology. Development tools such as HornMorpho are employed for morphological analysis; whereas, Natural Language Processing Toolkit is employed for the text process. The system has experimented on the various extraction rates of 10%, 20%, and 30%. The result's evaluated exploitation recall, precision, and F-measure for objective analysis; whereas, subjective analysis has been evaluated by language consultants. **Findings:** The results of the evaluations showed that the proposed system registered f-measure of 90%, 91% and 93% at a summary extraction rate of 10%, 20%, and 30% respectively. The informativeness and coherence of the proposed system also registered its best performance summary of 51.67%, 56.67 % and 54.17% average score on five scale measures at an extraction rate of 10%, 20%, and 30% respectively when both methods were used together. **Novelty:** By using a morphological analysis tool the performance of the system is improved from 80.67% to 91.3% F-measure when we compare it with the previous work even supposing there's still a requirement to conduct additional analysis to enhance the Afaan Oromo text summarization.

Keywords: Document; Summary; Natural Language Processing; Morphological Analysis; Text Ranking

1 Introduction

As a result of the huge amount of information available, a new technology that can process this information is required by users. Document summarizing may be a vital tool for addressing this issue. We can provide a brief version of a source text that contains useful information for consumers by employing document summarizing. Users will receive a concept of the complete content of the document from the summarized result and can decide whether or not to read the entire document[1]. According to Michael[2] the resulting summary is about the query asked in query-based document summarization. Document summary that is based on a query is known as query-based summarization. The task is to create a summary from the document that can deliver informative information relating to the user's information demands, given a user query. The text report is associate degree activity meant to make a transparent and simple outline having solely the key ideas of the documents by shortening long items of text. The creation, gathering, organization, storing, and spreading of data has been simplified with the advancement of data and communication technologies. Being capable to simply associate degreed shortly acknowledge this data in an organized, short, and precise approach offers the reader an outline of the ideas towards the contents of the scripts.

Afaan Oromo is one of the foremost wide spoken languages in a geographic area that accounts for regarding forty million speakers in the African nation, it's conjointly spoken in Ethiopian neighbouring countries like an African country, Somalia, Djibouti, and Egypt[3]. Currently, Afaan Oromo is that the official operating language of Oromia national regional state and also the instructional language for grammar school (1-8) students within the Oromia region. As a result, loads of works for official and/or personal functions use Afaan Oromo for social communication.

On the other hand, information users face a challenge in evaluating, filtering, and choosing information that meets their information needs[3,4]. Afaan Oromo text readers are littered with these issues as alternative under-resourced language readers within the world[4]. To repair these issues there's a necessity to possess a mechanism for Afaan Oromo which will evaluate, filter, and choose data in a very summarized type that meets the knowledge wants of users. Document summarization is the method of extracting the contents of the initial text in a very shorter type that gives helpful data to the user[5]. Few works have been done for Afaan Oromo text summarizer however they lack coherence. additionally, to the present, the developed summarizer doesn't take into account the requirements of users as a result of they are doing not settle for the user question. So, there's a necessity to eliminate the matter of coherence and permit users to urge the summarized document Turkish monetary unit their question. Therefore, the target of this study is to develop a query-based document report for Afaan Oromo. So that the user will get a summarized version of the document which might facilitate the reader to seek out the relevant data of the documents in a very summarized approach.

Different researchers attempt to develop a document summarization system for various languages. Manju[1] attempted to design, construct, and evaluate a sentence score method for Malayalam language multi document extractive summarization. The TextRank algorithm was used as a baseline for testing several text embeddings models, including fasttext and smooth inverse frequency embedding. Bharti, Babu and Jena[6] examined text summarization research from the perspectives of automatic keyword extraction, text databases, summarization process, summarization methodology, and evaluation matrices. The author suggests that text summarizing work be focused on languages with limited resources.

However, for Afaan Oromo only few works are done on document summarization. Tesema and Tamirat[7] used unsupervised machine learning to analyze the Afaan Oromo language structure and built an excellent file editing tool as a Microsoft Word plug-in to enable text entry and input techniques tool. Data for the training was gathered from official media, as well as cultural, historical, sports news, political, and economic papers from Afaan Oromo users. After that, N-gram methods (particularly Unigram, Bigram, Trigram, and Fourth Gram) were used to collect the trained data based on linguistic structure. As a result, Afaan Oromo is one of the limited resources (small dataset) for training, with Unigram, Bigram, and Trigram being the only options.

Open Oromo Text Summarizer (OOTS), developed by Dinegde and Tachbelie[4], is the first automatic text summarizer for Afaan Oromo news text. It is based on the Open Text Summarizer (OTS). OTS has been modified to support the Afaan Oromo language by altering the code. Term frequency and sentence position methods were utilized in this work, along with language-specific lexicons (synonyms and abbreviations) to assign weights to the sentences to be retrieved for the summary. Even if the study contributes to the development of natural language processing applications for the Afaan Oromo language, the summarizers' results are incoherent.

One of the work is done by Kabeta[8] in 2015 entitled "query-based automatic summarizer for Afaan Oromo text". He has enforced varied approaches to perform the various activities of the system like a vector space model (VSM) of information retrieval to compute the significant sentence score and rank the sentences within the document. He conjointly used the position methodology at the side of the vector area model in an endeavour to enhance the quality of the produced final summary. Experimental result shows that the system achieves high recall and low precision. However, for a query-based document summarization task, it is necessary to pursue high precision, without affecting the recall of the system.

Kabeta[8] shows that the unavailability of tools which help to obtain more clues in finding important sentences in the final summary and some final summaries contain unresolved references that may cause difficulties in understanding. For summarizing documents in various languages, such as English, there exist a number of natural language processing tools with advanced operational features. There has been an ongoing specialization on the characteristics in the development of these natural language processing systems. These systems are becoming increasingly sophisticated in order to meet the expectations of users for precise information. The use of word stemming was one of the attempts to make natural language processing technologies more effective in information retrieval[7].

In this study, Afaan Oromo HornMorpho was used as a tool to find important sentences in the final summary and some final summaries contain unresolved references that may cause difficulties in understanding. Starting from the previous research gap and recommendation, the design and develop a query-based document summarization for Afaan Oromo to provide a short version of a source text that provides informative information for users.

2 Methodology

2.1 Research design

In this study design science research is followed, as a result of it is a concern with the systematic creation of information concerning associated with design as an intentional, intellectual, and inventive activity for problem-solving.

2.2 Data Collection

To perform the experiments, the dataset set was collected from information sources that discuss diverse political, economic, and social issues. Oromia Broadcasting Network (OBN) is that the dataset supply for this study. OBN is that the radio and tv organization that broadcasts in Afaan Oromo. The OBN is chosen for dataset collection because of the easy availability of Afaan Oromo data the size of the corpus used for the experiments is concerning two hundred sentences, ready from the preceding online sources. A corpus containing about 433 sentences has been constructed.

2.3 Approaches and Tools

The developed query-based text summarization has been used the approaches to identify the sentence location within a paragraph. Based on this the average TF-IDF term weight and the relevance of the sentence for summarization has been calculated. Python is the software that is used to develop the prototype. Python is used because it is easy to work with, learn, and adaptable scripting language which makes it attractive for development[6]. HORN MORPHO 2.5 is a tool used for POS tagging activity. It is a program that analyses Amharic, AO, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given root or stem and a representation of the word's grammatical structure NLTK (Natural Language Processing Toolkit) is another tool used for text summarization[7].

2.4 The Proposed Architecture

The design was designed by using varied techniques and procedure tools. The figure below shows the projected Afaan Oromo query-based document summarization design.

At the pre-processing stage, texts are normalized to own the same format, and unnecessary and non-significant texts are removed and/or eliminated. At this stage, all text letters are normalized into small letters, and texts are tokenized into sentences. Unnecessary and non-significant texts like stop-words also are removed at the pre-processing stage. A list of stop-words adapted from Girma[9] and Nigussie[10] works are utilized in this study. At the second stage stemming from the word is finished, this can be the method of removing suffixes and prefixes from the word. HornMorpho may be a tool that's used to process the word stemming activity. The output from the stemming text is saved on a text file. Then the user is prompted to insert the question. The inserted query is additionally stemmed from the mistreatment of the HornMorpho tool. within the third stage, the weight of the sentence is calculated depending on the user query. Finally, the outline is generated for the ranked sentence.

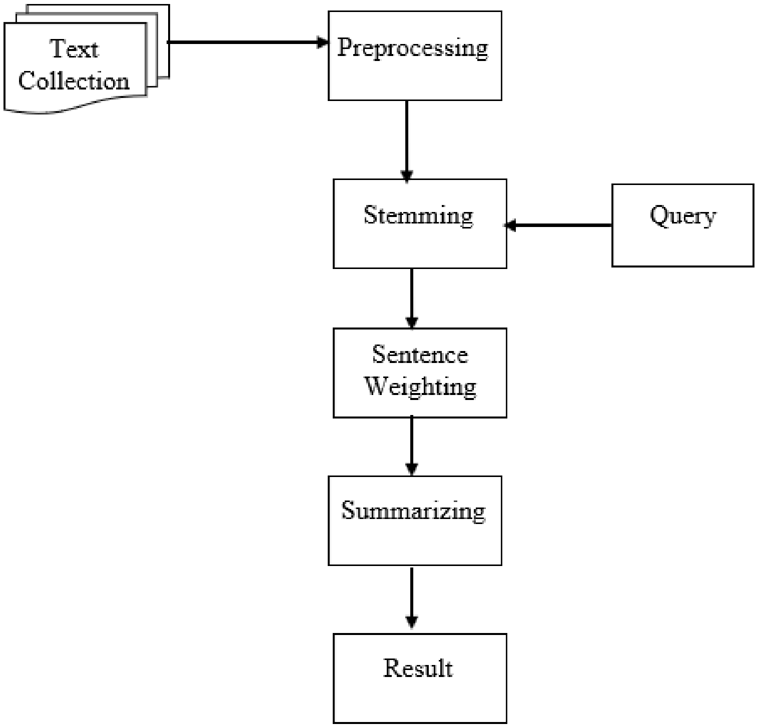
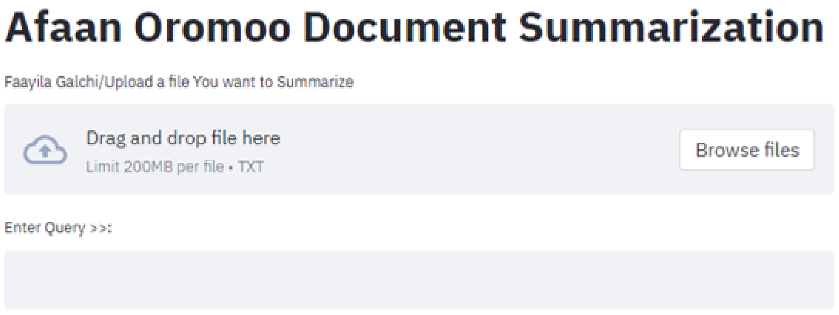


Fig 1. Afaan Oromo query-based document summarization architecture

3 Implementation and Performance Evaluation

3.1 Implementation

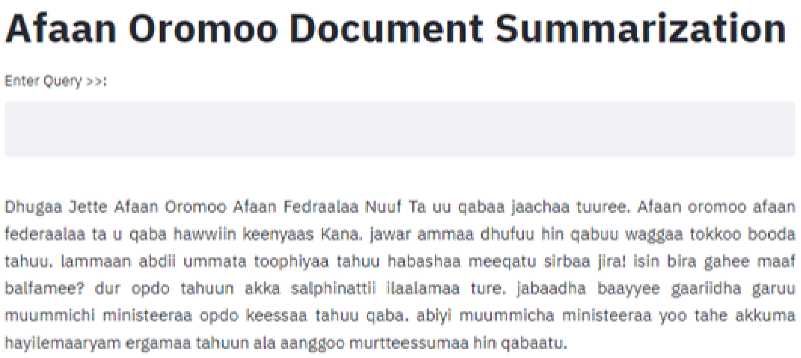
The experimentation was conducted by using three different summarization rate methods strategies, are 10%, 20%, and 30% of the document to be summarized Before experimenting, researchers have manually ready the outline of the text with the assistance of the Afaan Oromo language experts. Similar text is given to three different experts that facilitate the evaluation of the performance of the system. In Figure 2 below, the user interface of the developed system is shown. The user is expected to upload the file needed for summarization. After uploading the file, the query has to be entered to perform the summarization activity.


Afaan Oromoo Document Summarization

Faayila Galchi/Upload a file You want to Summarize

Enter Query >>:

Fig 2. Screenshot Interface


Afaan Oromoo Document Summarization

Dhugaa Jette Afaan Oromoo Afaan Fedraalaa Nuuf Ta uu qabaa jaachaa tuuree. Afaan oromoo afaan federaalaa ta u qaba hawwiin keenyaas Kana. jawar ammaa dhufuu hin qabuu waggaa tokkoo booda tahuu. lammaan abdii ummata toophiyaa tahuu habashaa meeqatu sirbaa jira! isin bira gahee maaf balfamee? dur opdo tahuun akka salphinattii ilaalamaa ture. jabaadha baayyee gaariidha garuu muummichi ministeeraa opdo keessaa tahuu qaba. abiyi muummicha ministeeraa yoo tahe akkuma hayilemaaryam ergamaa tahuun ala aanggoo murtteessumaa hin qabaatu.

Fig 3. Summarized Result

3.2 Performance Evaluation

The performance of the system is evaluated in two other ways that are objective evaluation and subjective evaluation. Recall, precision, and F-measure are used to evaluate the system objectively whereas informativeness and coherence are used to evaluate the system subjectively.

3.2.1 Objective Evaluation

After implementation, the performance of the system must be tested using objective analysis metrics. Therefore, the summarizers are evaluated using recall, precision, and F-measure. As a result, the experimental results of the Afaan Oromo text summarization are illustrated within the tables below severally for every extraction rate below.

Table 1. Performance result of Summarization on the rate of 10%

Doc.No	Summarized	Correctly Summarized	Measurement		
			Recall	Precision	F-Measure
1	9	8	0.9	0.89	0.89
2	8	8	0.8	1.00	0.90
3	9	7	0.9	0.78	0.84
4	9	9	0.9	1.00	0.95
Average			0.88	0.92	0.90

Table 2. Performance result of Summarization on the rateof 20%

Doc.No	Summarized	Correctly Summarized	Measurement		
			Recall	Precision	F-Measure
1	18	16	0.9	0.89	0.89
2	19	17	0.95	0.89	0.92
3	17	17	0.85	1.00	0.93
4	17	16	0.85	0.94	0.90
Average			0.89	0.93	0.91

3.2.2 Subjective Evaluation

For subjective analysis, the summarized result and also the documents square measure given to three language specialists. every language specialist named Ev#1 for 1st evaluator one, Ev#2 for the second evaluator a pair of and Ev#3 for the third evaluator. The specialists are needed to rate the result out of four (4) using the evaluation questionnaire.

The mean value of the rate of the evaluator is taken as the performance score of the system the subsequent tables show the subjective analysis results of the system supported the given summarization needs (10%, 20%, and 30%). The questions are

Table 3. Performance result of Summarization on the rate of 30%

Doc.No	Summarized	Correctly Summarized	Measurement		
			Recall	Precision	F-Measure
1	30	26	1	0.87	0.93
2	29	26	0.97	0.90	0.93
3	29	25	0.97	0.86	0.91
4	28	27	0.93	0.96	0.95
Average			0.97	0.90	0.93

named Qn#1, Qn#2, Qn#3, and Qn#4 for each questionnaire respectively. The following are the questions used for subjective evaluation.

- Qn#1→ Did the idea in the summary flows coherently
- Qn#2 → Did the summary represents all the important points from the document
- Qn#3 → Does the sentence in the summary is important
- Qn#4 → Is the information in the summary not redundant

Table 4. Result of Subjective Evaluation Rating Scores 10%

Questions	Rating Scores 10%			
	Ev#1	Ev#2	Ev#3	Mean
Qn#1	3	2	3	2.67
Qn#2	2	3	3	2.67
Qn#3	3	2	3	2.67
Qn#4	3	2	2	2.33

Table 5. Result of Subjective Evaluation Rating Scores 20%

Questions	Rating Scores 20%			
	Ev#1	Ev#2	Ev#3	Mean
Qn#1	3	3	4	3.33
Qn#2	3	2	2	2.33
Qn#3	2	3	2	2.33
Qn#4	4	3	3	3.33

Table 6. Result of Subjective Evaluation Rating Scores 30%

Questions	Rating Scores 30%			
	Ev#1	Ev#2	Ev#3	Mean
Qn#1	2.50	3	3	2.83
Qn#2	3	2.5	2	2.50
Qn#3	3.5	2	3	2.83
Qn#4	3	3	2	2.67

Based on the subjective evaluation of the experts the average performance of the system on informativeness and coherence of the summarized result is presented in the table below out of 100 percent.

4 Result and Discussion

The experimental result of objective evaluation for each rate of extraction has scored the higher performance of in recall than the precision. This means the system has performed well to extract the sentences that match the user query. Therefore, if the sentence contains a user query it is considered an important sentence for the summary.

Table 7. The overall performance of the system for subjectiveevaluation

S.No	Extraction Rate	The Overall Performance (100%)
1.	10%	51.67
2.	20%	56.67
3.	30%	54.17

For subjective evaluation, the scored result shows low system performance. This is because the system uses the ranked sentence without any modification on the structure of the sentence. Whereas, on the manually summarized text, it's observed that some modification has been made to make the flow of sentences on its phase. The system only considers the sentence that contains the query term due to this synonym and antonym words that have significance for the summary have been eliminated. As compared to the previous studies conducted on summarization for Afaan Oromo this study has scored better performance. However, there is still further work that needs to be done to solve this problem by incorporating the effects of synonym and antonym words in the text.

5 Conclusion and Recommendation

To eliminate the problem of consistency, time-consuming, long to search out the acceptable documents for Afaan Oromo readers there's a necessity to implement the text summarization scheme that allows users to get the summarized document as per their query. Therefore, this study has tried to implement a query-based text summarization for Afaan Oromo documents. So, a summarization system has been implemented which can create the document summaries depending on the query of the user.

Text is summarized using part-of-speech (POS) data indicating parts of speech for tokens in the text. The POS tagging has been used to find the groups of tokens with in the sentence. The POS data can be used to acquire summarized text by removing sentences that meet the removal criterion. Whereas this study incorporates a promising result to generate a summary of the document based on the user query, there's still more work to be done to improve the Afaan Oromo text summarization. The limitation of this study is, the semantic analysis of the sentence is not considered. Therefore, the subsequent recommendations are created for the long-run analysis direction:

- The summarized sentence can have syntactically correct structure but semantics related to the meaning of the sentences and the flow of ideas aren't considered in this study. So, more analysis work must be done on semantics analysis of the structure for the formation of paragraphs that have a semantically correct and descriptive flow of ideas on the summarized results of the system.
- Since it's tough to construct the paragraph semantically by using the algorithm used in this study. Therefore, the machine learning methodology combined with a rule-based approach could improve the performance of the Afaan Oromo text summarization task.

References

1) Manju K, Peter SD, Idicula S. A Framework for Generating Extractive Summary from Multiple Malayalam Documents. *Information*. 2021;12(1):41–41. Available from: https://dx.doi.org/10.3390/info12010041.
2) Michael JG. A Quick Introduction to Text Summarization in Machine Learning. 2018. Available from: https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f.
3) Kualo. Oromoo language, alphabet and pronunciation . 2010. Available from: http://www.omniglot.com/writing/oromo.htm.
4) Asthana A, Tiwari EV, Pandey E, Misra EA. A Novel Architecture for Agent Based Text Summarization. 2017.
5) DebeleDinegde G, Tachbelie MY. Afan Oromo News Text Summarizer. *International Journal of Computer Applications*. 2014;103(4):1–6. Available from: https://dx.doi.org/10.5120/18059-8990.
6) Naidu R, Bharti SK, Babu KS, Mohapatra RK. Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers. *Smart Computing and Informatics*. 2018;77:555–564.
7) Tesema W, Tamirat D. Investigating Afan Oromo Language Structure and Developing Effective File Editing Tool as Plug-in into Ms Word to Support Text Entry and Input Methods. . Available from: www.pubicon.in.
8) Kabeta AB. Query-based Automatic Summarizer for Afaan Oromo Text. 2015.
9) Girma D. Afan Oromoo news text summarizer. Addis Ababa University. 2012.
10) Nigussie E. Afaan Oromoo – Amharic Cross Lingual Information Retrieval: A corpus Based Approach. Addis Ababa University. 2013.