DEVELOPMENT OF A QUERY-BASED DOCUMENT SUMMARIZATION FOR AFAAN OROMOO

Jemal Abate **Email**: abatejemal@gmail.com
Department of Information Science
Haramaya University
Haramaya, Ethiopia

Ashenafi Tulu **Email**: asheetulu@gmail.com
Department of Information Science
Haramaya University
Haramaya, Ethiopia

Abstract

Aims: This study attempted to develop document summarization for Afaan Oromoo that can summarize based on the query entered by the user(s). So that user can get summarized results of the document which can help the reader to find the relevant information of the documents in a summarized way.

Method: This study follows the design science research method because it concerns thoughtful, intellectual, and inventive activity during problem-solving and creation of knowledge.

Methodology: The developed query-based summarization framework has implemented the average TF-IDF term weight has used methodology. Various development tools have been used such as; HornMorpho is used for morphological analysis whereas Natural Language Processing Toolkit is used for text processing.

Result: The system has experimented on the different extraction rates of 10%, 20%, and 30%. The result is evaluated using recall, precision, and F-measure for objective evaluation whereas subjective evaluation has been evaluated by language experts. The experimentation result has scored a good performance even though there is still a need to conduct further research to improve the Afaan Oromoo text summarization.

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

1. INTRODUCTION

According to (Michael, 2018) text summarization an activity intended to create a clear and easy summary having only the key ideas of the documents by shortening long pieces of text. The creation, gathering, organization, storing, and spreading of information has been simplified with the advancement of information and communication technologies. Being capable to easily and shortly recognize this information's in an organized, short, and precise way gives the reader an overview of the concepts towards the contents of the scripts.

Afaan Oromoo is one of the most widely spoken languages in east Africa which accounts for about 40 million speakers in Ethiopia, it is also spoken in Ethiopian neighboring countries like Kenya, Somalia, Djibouti, and Egypt (Kualo, 2010). Currently, Afaan Oromoo is the official working language of Oromia national regional state and the educational language for primary school (1-8) students in the Oromia region. As a result, a lot of works for official and/or personal purposes use Afaan Oromoo for interpersonal communication.

On the other hand, information users are facing a challenge in evaluating, filtering, and selecting information that meets their information needs (Dr.Amit, Er.Vagish, Er. M.C., & Er.Ankit, 2017). Afaan Oromoo text readers are also

suffering from these problems as other under-resourced language readers in the world (Girma & Martha, 2014). To fix these problems there is a need to have a mechanism for Afaan Oromoo that can evaluate, filter, and select information in a summarized form that meets the information needs of users. Text summarization is the process of extracting the contents of the original text in a shorter form that provides useful information to the user (Rasmita, Rakesh, & Anisha, 2015). Few works have been done for Afaan Oromoo text summarizer but they lack coherence. In addition to this, the developed summarizer does not consider the need of users because they do not accept the user query. So, that there is a need to eliminate the problem of coherence and allow users to get the summarized document asper their query. Therefore, the objective of this study is to develop a query-based document summarization for Afaan Oromoo. So that user can get a summarized version of the document which can help the reader to find the relevant information of the documents in a summarized way.

2. RELATED WORKS

The development of formulating a document summarization using a computer technique in the form of abstract or summary is referred to as Automatic Text Summarization. A lot of fruitful activities have been done in the past by different researchers on the development and implementation of document summarization. Pioneering work has been done by Basagic, Krupic, & Suzic (2009) from 1955 to 1979 by using simple extraction and linguistic approaches.

During the 1980s and 1990's Basagic, et al., (2009) shifted the interest of researchers to multi-document summarization and extraction of summary from multimedia documents and a series of knowledge-based text summarization systems by implementing artificial intelligence approaches.

Kifle & Martha (2017) introduces Graph-based Automatic Amharic Text Summarizer (GAATS), a generic and domain-independent graph-based model for automatic single-document summarization task, and shows how this model can successfully be used to generate extracts of high quality from Amharic texts. They have extended the two prominent graph-based link analysis algorithms: PageRank and HITS with two-sentence centrality measures: cumulative sum and discounted cumulative sum for exploiting the relation between sentences in a text and/or node in a graph, and shows the results of their experiments. The results demonstrated that extractive summaries of better quality can be generated when discounted cumulative sum paired with HITS. The results also revealed that our approach is domain-independent and more effective than reference summarization systems.

Different works have been done on text summarization for different languages, but for Afaan Oromoo only a few works have been done. One of the studies conducted by (Girma & Martha, 2014) has been used three methods for the development of Afaan Oromoo news text summarizers. These are: first they use term frequency and position methods without Afaan Oromoo stemmer and language-specific lexicons (synonyms and abbreviations); then they combine of term frequency and position methods with Afaan Oromoo stemmer and language-specific lexicons and finally with improved position method and term frequency as well as the stemmer and language-specific lexicons. The evaluation result of the experiment shows that the improved position method and term frequency along with stemmer has good performance. However, the summarizers result of this study's lack coherence so that more advanced method that would help avoid the problem of coherence has to be implemented.

The other work has been done by Asefa in 2015 entitled "query-based automatic summarizer for Afaan Oromoo text". He has implemented various approaches to perform the different activities of the system such as a vector space model (VSM) of information retrieval to compute the significant sentence score and rank the sentences in the document. He also used the position method along with the vector space model in an attempt to improve the quality produced final summary.

3. METHODOLOGY

3.1. Research design

In this study design science research is followed, because it`s a concern with the systematic creation of knowledge about and with design as an intentional, intellectual, and creative activity for problem-solving.

3.2. Data Collection

To perform the experiments, the dataset was collected from data sources that discuss various political, economic, and social issues. Oromia Broadcasting Network (OBN) is the data source for this study. OBN is the radio and television organization that broadcasts in Afaan Oromoo. The OBN is selected for data collection because of the easy availability of Afaan Oromoo data. The size of the corpus used for the experiments is about 200 sentences, prepared from the above-mentioned online sources.

3.3. Approaches and Tools

The implemented approach in this study are query-based along with feature set includes elements such as the location of a sentence within the document and its subsections and paragraphs, sign phrases, information on whether the sentence contains named entities, sentence length, average TF-IDF term weight, and data on whether the sentence contains a quotation or is inside a blockquote. Various development tools have been used in this study, among this python is the software that is used to develop the prototype. Python is easy to work with, learn, and adaptable scripting language which makes it attractive for the development (Masheet, 2011). HORN MORPHO 2.5 is a tool used for POS tagging activity. It is a program that analyzes 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 (Gasser, 2012) NLTK (Natural Language Processing Toolkit) is another tool used for text summarization.

4. THE PROPOSED ARCHITECTURE

The architecture was designed by using various techniques and computational tools. The figure below shows the proposed Afaan Oromoo query-based document summarization architecture.

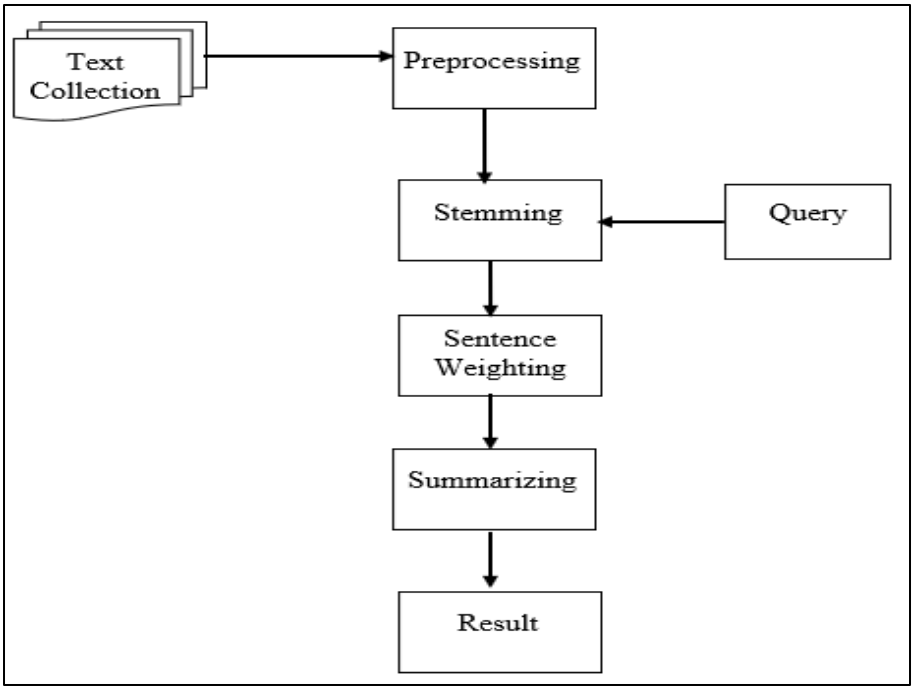

Figure 1 Afaan Oromoo query-based document summarization architecture

At the preprocessing stage, texts are normalized to have a similar format, and unnecessary and non-significant texts are removed and/or eliminated. At this stage, all text letters are normalized into lowercase letters, and texts are tokenized into sentences. Unnecessary and non-significant texts such as stop-words are also removed at the preprocessing stage. List of stop-words adopted from (Girma D., 2012), (Nigussie, 2013) and (Debela, 2010) works are used in this study.

At the second stage stemming from the word is done, this is the process of removing suffixes and prefixes from the word. HornMorpho is a tool that is used to process the word stemming activity. The result of HornMorpho stemming from text is saved on a text file. Then the user is prompted to insert the query. The inserted query is also stemmed from using the HornMorpho tool. In the third stage, the weight of the sentence is calculated depending on the user query. Finally, the summary is generated ranked for the ranked sentence.

5. EXPERIMENTATION AND PERFORMANCE EVALUATION

5.1. Experimentation

The experimentation was conducted by using three different summarization rate methods, these are 10%, 20%, and 30% of the document to be summarized. Before experimenting, researchers have manually prepared the summary of the text with the help of Afaan Oromoo language experts. The same text is given to three different experts that help to evaluate the performance of the system.

5.2. Performance Evaluation

The performance of the system is evaluated in two different ways which 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.

5.2.1. Objective Evaluation

After implementation, the performance of the system has to be tested using objective evaluation metrics. Therefore, the summarizers are evaluated using recall, precision, and F-measure. As a result, the experimental result of the Afaan Oromoo text summarization is illustrated tables below respectively for each extraction rate below.

Doc.No	Summarized	Correctly Summarized	The extraction rate of 10%		
			Recall	Precision	F-Measure
1	**3**	**2**	1.00	0.67	0.83
2	**2**	**2**	0.80	1.00	0.90
3	**4**	**3**	0.89	0.75	0.82
4	**2**	**1**	0.77	0.50	0.63
5	**3**	**2**	0.88	0.67	0.77
Average			**0.87**	**0.72**	**0.79**

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

Doc.No	Summarized	Correctly Summarized	The extraction rate of 20%		
			Recall	Precision	F-Measure
1	**5**	**4**	0.83	0.80	0.82
2	**4**	**3**	0.80	0.75	0.78
3	**9**	**7**	1.00	0.78	0.89
4	**5**	**3**	0.96	0.60	0.78
5	**6**	**4**	0.88	0.67	0.77
Average			**0.90**	**0.72**	**0.81**

Table 2 Performance result of Summarization on the rate of 20%

Doc.No	Summarized	Correctly Summarized	The extraction rate of 30%		
			Recall	Precision	F-Measure
1	**9**	**6**	1.00	0.67	0.83
2	**7**	**5**	0.93	0.71	0.82
3	**13**	**10**	0.96	0.77	0.87
4	**7**	**6**	0.90	0.86	0.88
5	**9**	**7**	0.88	0.78	0.83
Average			**0.94**	**0.76**	**0.85**

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

5.2.2. Subjective Evaluation

For subjective evaluation, the summarized result and the documents are given to 3 language experts. Each language experts are named as Ev#1 for first evaluator 1, Ev#2 for second evaluator 2, and Ev#3 for the third evaluator. The experts are required 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 following tables show the subjective evaluation result of the system based on the given summarization requirements (10%, 20%, and 30%). The questions are 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 are important
- Qn#4→ Is the information in the summary not redundant

Questions	Evaluation Result of Each Summarization Rates											
	Rating Scores 10%				Rating Scores 20%				Rating Scores 30%			
	Ev#1	Ev#2	Ev#3	Mean	Ev#1	Ev#2	Ev#3	Mean	Ev#1	Ev#2	Ev#3	Mean
Qn#1	3	2	3	2.67	3	3	4	3.33	2.50	3	3	2.83
Qn#2	2	3	3	2.67	3	2	2	2.33	3	2.5	2	2.50
Qn#3	3	2	3	2.67	2	3	2	2.33	3.5	2	3	2.83
Qn#4	3	2	2	2.33	4	3	3	3.33	3	3	2	2.67

Table 4 Result of Subjective Evaluation

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.

Extraction Rate	**The Overall Performance (100%)**
10%	51.67
20%	56.67
30%	54.17

Table 5 The overall performance of the system for subjective evaluation

6. RESULT AND DISCUSSION

The experimental result of objective evaluation for each rate of extraction has scored the high 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.

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 antinomy words that have significance for the summary have been eliminated. As a result, there is still further work is need to be done to solve this problem by incorporating the effects of synonym and antinomy words in the text.

7. CONCLUSION AND RECOMMENDATION

The world of documents containing text is huge and expanding every day in the World Wide Web. The majority of the data is in the form of natural language text. To eliminate the problem of consistency, time-consuming to

find the appropriate documents for Afaan Oromoo reader there is a need to implement the text summarization scheme that allows user to get the summarized document asper their query. Therefore, this study has attempted to implement a query-based text summarization for Afaan Oromoo documents.

Whereas this study, have a promising result to generate a summary based on the user query, there is still further work has to be done to improve the Afaan Oromoo text summarization. Therefore, the following recommendations have been made for the future research 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. Therefore, further research work has to be done on semantic analysis of the sentence structure for the formation of paragraphs that have a semantically correct and descriptive flow of ideas on the summarized result of the system.
- Since it is difficult to construct the paragraph semantically by using the algorithm used in this study it must implement another method. Therefore, the machine learning method combined with a rule-based approach may improve the performance of the Afaan Oromoo text summarization task.

REFERENCES

Basagic, R., Krupic, D., & Suzic, B. (2009). Automatic Text Summarization.The Graz University of Technology, Institute for Information Systems and Computer Media, Graz.

Debela, T. (2010). Designing a Stemmer for Afaan Oromoo Text: A Hybrid Approach. *MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia*, 100-101.

Dr.Amit, A., Er.Vagish, T., Er. M.C., P., & Er.Ankit, M. (2017). A Novel Architecture for Agent-Based Text Summarization. *Asian Journal of Applied Science and Technology (AJAST), 1*(5), 2.

Gasser, M. (2012). HORN MORPHO2.5 User's Guide. *Research group of human language technology and the democratization of information, 1*, 1.

Girma, D. (2012). Afan Oromoo news text summarizer. *MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia*.

Girma, D., & Martha, Y. (2014). Afan Oromoo News Text Summarizer. *International Journal of Computer Applications, 103*(04), 2.

Kifle, D., & Martha, Y. (2017). Graph-based Automatic Amharic Text Summarizer. *International Journal of Scientific & Engineering Research*, 2.

Kualo. (2010). *Oromoo language, alphabet, and pronunciation*. Retrieved October 14, 2017, from omniglot: http://www.omniglot.com/writing/oromo.htm

Masheet. (2011, April 7). *what is python*? Retrieved October 16, 2017, from javatpoint: http://www.javatpoint.com/what-is-python

Michael, J. G. (2018, September 19). *towards data science*. Retrieved from A Quick Introduction to Text Summarization in Machine Learning: http://www.towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

Nigussie, E. (2013). Afaan Oromoo – Amharic Cross-Lingual Information Retrieval: A corpus Based Approach. *MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia*.

Rasmita, R., Rakesh, C., & Anisha, B. (2015). Document Summarization Using Sentence Features. *International Journal of Information Retrieval Research, 2*(1), 3.

Tadesse, A., & Tomio, T. (2009). Development of an Amharic Text-to-Speech System Using Cepstral Method. *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages.* Athens.