International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

NLP: Text Summarization By Frequency And Sentence Position Methods

N.Kannaiya Raja, Naol Bakala, S. Suresh,

Abstract: ***In today's fast-growing online information age we have an abundance of text, especially on the web. New information is constantly being generated. Often due to time constraints we are not able to consume all the data available. It is therefore essential to be able to summarize the text so that it becomes easier to ingest, while maintaining the essence and understandability of the information. The summarizer basically uses the combinations of term frequency and sentence position methods with language specific lexicons in order to identify the most important sentence for extractive summary. We aim to design an algorithm that can summarize a document by their performance both objectively and subjectively in Afan Oromo Language. The performance of the summarizers was measured based on subjective as well as objective evaluation methods. The techniques used in this paper are term frequency and sentence position methods with language specific lexicons to assign weights to the sentences to be extracted for the summary.***

Keywords : Natural Language Processing (NLP), Text Summarization (TS).

I. INTRODUCTION

In the world of information, the increasing availability of online information has necessitated intensive research in the area of automatic text summarization within the field of Natural Language Processing (NLP). In this first growing information age, Text summarization has become an important and timely tool for assisting and interpreting text information. With the huge availability of text document in the internet, it gives more information than is needed. It is very difficult for human beings to manually summarize large documents of text. So searching for relevant documents through an overwhelming number of documents available in the web is a very difficult task. In order to solve the above two problems, the automatic text summarization is very much necessary. Before going to the Text summarization, first we, have to know that what a summary is. A summary as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is shorter than that of the original text(s). The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. A summary [2] can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An Abstractive summarization [9][10] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. In this paper we focus on novel techniques which are based on extractive text summarization methods.

Manuscript published on 30 September 2019
* Correspondence Author
Dr.N.Kannaiya Raja*, M.E., Phd.,,Professor, Department of Computer Science ,Ambo University, Ambo, Ethiopia.
Mr. Naol Bakala, M.Sc., Head/Department of Computer Science, Ambo University, Ethiopia.
Mr. S. Suresh, M.E., Asst. Professor, Asst Professor/ Department of Computer Science & Engineering, C. Abdul Hakeem College of Engineering & Technology, Vellore, Tamil Nadu, India.

II. STATEMENT OF THE PROBLEM

These days, documents in paper and electronic format are growing dramatically. As a result, the users (readers) are facing information overload problem with vast quantities of text. In almost all languages in the world, texts in any domain are written in detail and readers are forced to see unwanted detail without being interested in it unless it is summarized to save the readers' time. Afan Oromo text readers are not exceptional to suffer from this problem.

There are many domain areas that produce large content of textual information which needs summarization to save the time of readers. Some of the textual information are large volumes of legal judgments which is very essential if they are used by the experts (for timely justice) and by law students for their study, newspaper texts and online news articles produced by media agencies, criminal investigation document produced by polices at different level, reports from government offices, etc

There are a number of media agencies and presses releasing news in electronic and non-digital format. There are a number of newspapers publishers that produce news articles. Some of such sources of newspaper are: *Barriisa*, *Kallacha Oromiya* and *Oromiya*. *Bariisa* is a weekly newspaper, whereas the rest two come out once in two weeks.

There are also radio broadcasts in Afan Oromo by Ethiopian Radio and Radio Fana for 14 and 30 hours weekly, respectively. Moreover, Oromia Radio and Television Organization found in Adama releases daily news through radio and television broadcast and on its official website.
On the other hand, magazines, judiciary documents and office reports also constitute some portion of the documents produced in the language.

1. RESEARCH METHODOLOGY

Primarily, literatures related to automatic text summarization have been reviewed. As the study conducted on Afan Oromo news text summarization, the nature of the language and the structure of the documents to be summarized for testing were investigated. To carry out this task, books, journal articles, and relevant websites are consulted.

1.1. CORPUS PREPARATION

A corpus to evaluate the summarizer (Afan Oromo news articles) was selected and prepared as there is no previous research and corpora in Afan Oromo for evaluating summarizer. The prepared corpus consists of 8 news items from Oromia Radio and Television Organization (ORTO) as well as Voice of America (VOA) and Afan Oromo official websites written on different topics .While selecting from news archives, longer articles (at least one page or more than 200 words) are considered due to the fact that as the text itself gets shorter summarizing it becomes unnecessary. The average length of news items, in the corpus, is approximately 300 words or 12 sentences.

1.2. SUMMARY GENERATION

For the purpose of manual summary generation, the corpus was provided to the human subjects together with the corresponding guideline. The four available experts ranked the sentences based on their ability of providing salient information for the reference summary. For a sentence, an average rank was calculated as the sum of its four ranks divided by four. The sentences have then been ordered according to their average rank. Finally, reference summaries were produced from the top ranking sentences at 10 %, 20%, 30 % and 40% of the original text's word length (compression rate) of randomly selected test sets.

1.3. SUMMARIZATION TECHNIQUE AND TOOLS USED

Most research on summary generation techniques still relies on extraction of important sentences from the original document to form a summary (Kaili and Pilleriin, 2005). There are several ways in which one can characterize different approaches to text summarization. The technique proposed for this study is extraction technique for single news text. Using extraction technique most important sentences from the document are extracted and displayed to the reader. To create a summary by this technique there is no need of rewriting the document by making linguistics analysis. To extract important sentence from a text to be summarized, sentence can be weighted based on cue phrases it contains, location of the sentence, sentence containing most frequent words in the document. Then sentences with the highest weight obtained by efficient combination of extraction features will be selected and a summary is written. This work is based upon the Open Text Summarizer (OTS) (Rotem, 2001), an open source tool for summarizing texts. The program reads a text and decides which sentences are important and which are not. It ships with Ubuntu, Fedora and other Linux distributions. OTS supports many (more than 25) languages which are configured in XML3 files. OTS incorporates natural language processing (NLP) techniques via an English language lexicon with synonyms and cue terms as well as rules for stemming. These are used in combination with a statistical word frequency based method for sentence scoring. Therefore, the source code available in python has been used and the XML file has been configured with Afan Oromo rule of stemming, stop list, synonyms and abbreviations such that it can support Afan Oromo news text summarization. The summarizer prototype is therefore customized from the existing OTS. Moreover, the researcher developed and integrated a tool for objective evaluation (compute standard recall and precision) with the summarizer.

1.4. EVALUATION TECHNIQUE

After configuring and developing the prototype text summarizer based on OTS, two forms of summaries prepared (system summary and reference summary) are used to evaluate the performance of the system. The evaluation process was conducted using an intrinsic method. It comprised of both subjective (qualitative) and objective (quantitative) evaluation methods. For both measures the four human subjects (expert journalists) are involved. Subjective evaluation was used to measure the linguistic quality, in formativeness and coherence of the automatically generated summaries.
The linguistic quality is basically aimed to measure the readability and fluency of the summary. We adopted subjective summarization techniques used by Greek text summarizer (Pachantouris, 2004). On the other hand, objective evaluation was basically used to measure the summarizer's performance in identification and extraction of salient sentences. This performance is measured by the standard recall and precision measures. Given an input text, human's (reference) summary and summarizer's extract, it measures how close the extracts are to the reference summary.

1.5. ALGORITHM

1. *Read a text in .txt or .rtf format and split it into individual tokens.*
2. *Remove the stop words to filter the text.*
3. *Assign a weight value to each individual terms. The weight is calculated as:*

wt = frequency of the term / Total Number of terms in the documents

4. *Add a boost factor to that terms which are appear in bold, italic, underlined or any combination of these. The boost Factor can be calculated as:*

*b = frequency of the special effect term * s_value / Total number of special effect term in the document*

5. Rank the individual sentences according to their weight value as : $wt_s = \sum_{i=1}^{n}(wt)/n$

where wt_s= weight of the sentence wt_1,wt_2,wt_3……wt_n are the weights of individual terms in the sentence

n= total number of terms in that sentence.

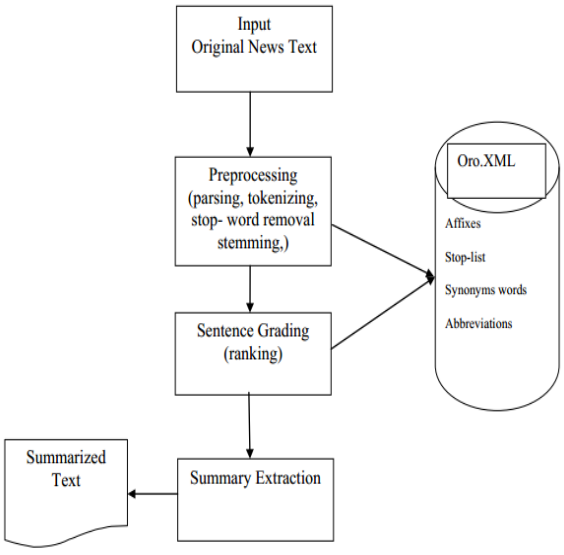
6. Finally, extract the higher ranked sentences including the first sentence of the first paragraph of the input text in order to find the required summary. The number of sentences extracted is based on the user requirement i.e. the percentages of summary the user give as input.
2. Implementation Of Afan Oromo News Text Summarizer.

2.1. ARCHITECTURE OF TS

We named our customized summarizer Open Oromo Text Summarizer (TS), the version based upon TS which summarizes Afan Oromo news texts. It is open because we planned to make it open to the public to serve as a framework that can be used for other Latin based Ethiopian languages.

2.2. PREPROCESSING

As is in other TS systems, preprocessing step includes tokenizing, stop-word removal, stemming and parsing (breaking the input document in to a collection of sentences). For stop word removal, we have used the Afan Oromo stop-word compiled from different literature in addition to the stop-word list prepared by Debela (2010).
Furthermore, using stemmer, a word is split into its stem and affix after stop-word removal. Affixes striped can be replaced by another affix or replaced by white space as per the rule it matches with. The design of a stemmer is language specific, and requires some significant linguistic expertise in the language. A typical simple stemmer algorithm involves removing suffixes using a list of frequent suffixes, while a more complex one would use morphological knowledge to derive a stem from the words. Since Afan Oromo is a highly inflectional language, stemming is necessary while computing frequency of a term.



2.3. Sentence Ranking

After an input document is formatted and stemmed, the document is broken into a collection of sentences and the sentences are ranked based on two important features: term frequency (TF) and sentence position.
TF is frequency of keyword appearance in an article. This method is the earliest known method to be used for automatic text summarization since research began in this area. It is based on the idea that the most relevant sentences are those containing the largest number of the most frequent words in the document (stop-words excluded) (Luhn, 1958). With the *tf(term frequency) method*, the importance value (score) of a sentence *s* (IVs) is given by: $IV=\sum tf$ Where, IV is Importance Value based on term frequency *tf*, is Term frequency On the other hand, positional value **(**score**)** of a sentence *s* is computed in such a way that the first sentence of a document gets the highest score and the last sentence gets the lowest score in news domain as the original TS uses constant multiplicative factor of term frequency score calculated.
The positional value for the sentence *s* is computed using the following formula by combining two parameters for sentence ranking. Therefore, the total importance value (score) of a given sentence *s* (TIVs) $TIVs=TIV * c$ Where, c is constant multiplicative factor. The value of c is 2 for first statement of first paragraph, 1.6 for first sentences of all other paragraphs. All other sentences are weighed only by their term frequency score. TIVs, is total score of importance value of a sentence based on term frequency and position value.

2.4. Summary Generation

A summary is produced after ranking the sentences based on their scores and selecting N-top ranked sentences, where the value of *N* is set by the user. To increase the readability of the summary, the sentences in the summary are reordered based on their appearances in the original text; for example, the sentence which occurs first in the original text will appear first in the summary.

III.EXPERIMENTATION RESULTS.

```
In [22]: # Removing Square Brackets and Extra Spaces
         article_text = re.sub(r'\[[0-9]*\]', ' ', article_text)
         article_text = re.sub(r'\s+', ' ', article_text)

In [23]: # Removing special characters and digits
         formatted_article_text = re.sub('[^a-zA-Z]', ' ', article_text )
         formatted_article_text = re.sub(r'\s+', ' ', formatted_article_text)

In [25]: import nltk
         sentence_list = nltk.sent_tokenize(article_text)

In [26]: stopwords = nltk.corpus.stopwords.words('english')

         word_frequencies = {}
         for word in nltk.word_tokenize(formatted_article_text):
             if word not in stopwords:
                 if word not in word_frequencies.keys():
                     word_frequencies[word] = 1
                 else:
                     word_frequencies[word] += 1

In [29]: import heapq
         summary_sentences = heapq.nlargest(7, sentence_scores, key=sentence_scores.get)

         summary = ' '.join(summary_sentences)
         print(summary)

         Boorlis aanga'oota Tarmaalifi kan Faransaayin wal anguuf daandii kana hunda erga imalanii booda, gocha akkasii raawwachuusaanii
         qaanidha,'' jechuun dubbachuseanii Pan gabaaseera. Tuffiifi of tulummaa namtichaa kan mullisuudha,'' jedhan Kaampel.Gabaasaa s
         ijaasaa 'Sky News; Tom Raayliner ammoo taatee kanarratti kallattii birraa kenneera, ''gochiichi qaana qoosaa waliin taasisaa tur
         antitti,'' jechuun. Maakroon ''finjaalli kun bakka mala irraa kaawutan ta'un tajaajileera'' joo jedhaman ture ture, Mr Boriis
         mila ol kaasuun finjaalicha irra kaawutan. Mr Boriis gocha sana daiigaa musaeraaf kan raawutan qoosaa Maakroon itti himaa tu
         ran irraa ka'uun akka ta'ee nagaaleen ni mullisu. Yeroo sanatti eegaa osoo pireidaantii Faransaayi Imaanu'el Maakiroon waliin
         taa'anii haasa'anu, Boorlis milasaantii ol kaasuun kan filingaala giingoo tokko yaabbattan. Durii taanaan intarneetiin waamuma hu
         nda affarsa'' jechuun ture gaacexaan Faransaay Le Parshiyaan jedhamu tokko kan barreesse. Gocha kanaanis Boorlis waan aanga'aan
         a biyya tokkootiif himaadaallee raawwatantiiru jedhamuun qeeqameeliru.Dhuggaatti kan ta'e maal ture garuu?
```

IV. CONCLUSIONS

In this paper we have done one new thing which has taken into consideration in online text summarization. The font based feature i.e. bold, italic, underlined and all the combination of these are considered to be more important when calculating the weight for ranking the sentences of the document. Since the summarization follows the extraction method, when it extract the important sentences it might happen that one sentence contains a proper noun and the next sentence contains a pronoun as a reference of the proper noun. In that case, if the summary considers the second sentence without considering the first one, then it does not give its proper meaning. It is a big issue in automatic text summarization. We are working to resolve this type of anaphoric problems in text summarization.

REFERENCE

1. Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, In proceedings of Seventh IEEE/ACIS International Conference on
2. Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.
3. Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg, 2005.
4. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159-165.
5. Baxendale, P. (1958). Machine-made index for technical literature - an experiment. IBM Journal of Research Development, 2(4):354-361.
6. Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2):264-285.
7. H. P. Edmundson.,” New methods in automatic extracting”, Journal of the ACM, 16(2):264-285, April 1969.
8. J. Kupiec, J. Pedersen, and F. Chen, “A trainable document summarizer”, In Proceedings of the 18th ACMSIGIR Conference, pages 68-73, 1995.
9. Ronald Brandow, Karl Mitze, and Lisa F. Rau. “Automatic condensation of electronic publications by sentence selection. Information Processing and Management”, 31(5):675-685,1995.
10. Vishal Gupta, G.S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009.

AUTHORS PROFILE

Dr.N.Kannaiya Raja, M.E., Phd., working as Professor in Department of Computer Science, Ambo University, Ambo, Ethiopia. His Major research interest is Natural Language Processing, Pattern Matching and data science.

Mr.Naol Bakala, M.Sc is working as Head in Computer Science Department in Ambo University, Ethiopia. His major research interest lies in Natural Language Processing

S. Suresh, M.E., currently working as Asst. Professor in C.Abdul Hakeem College of Engineering and Technology, Vellore, Tamilnadu, India. His research interest is Cognitive Radio Networks, Data Mining and Natural Language Processing.