I.J. Modern Education and Computer Science, **2019, 11, 31-38**
Published Online November 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2019.11.05

Design and Implementation of IR System for Tigrigna Textual Documents

Teklay Birhane[1]
Department of Information Science, Mekelle University, Mekelle, Ethiopia
Email: teklaybirhane12@yahoo.com

Birhanu Hailu[2]
Department of Information Science, Mekelle University, Mekelle, Ethiopia
Email: brhanuhaylu@gmail.com

Received: 10 October 2019; Accepted: 28 October 2019; Published: 08 November 2019

Abstract—Nowadays, various amount of information's are available on the internet. To search relevant documents from the internet development of information retrieval system or search engines is necessary. Therefore, this paper deals with development of Information Retrieval system for Tigrigna textual documents. It helps to find relevant documents from the internet, which are stored in Tigrigna language for the Tigrigna language users to satisfy their information need. The system includes two sub systems those are indexing and searching part. The indexing part is the process of organizing filtered Tigrigna documents using keywords extracted from the entire Tigrigna collection or corpus. It is an offline process carried out by the producers or authors world to speed up searching of information from the entire document as per users query. Searching is the process of scanning documents to find relevant documents that matches to the users query or information need. It is an online process mostly carried out by the users or readers world. Vector space model techniques was applied to implement this system. Vector space model is the most core information retrieval technique used to calculate similarity measure between the query and the documents finally it ranks the most relevant documents to the given query according their similarity score in descending order. According to this, the retrieval system was tested and the experimental results of the system in Tigrinya documents returned an encouraging and promising result. The system has registered, 70% precision and 84% Recall.

Index Terms—Corpus, Indexing, Information Retrieval, Searching, Tigrigna Language, Vector Space Model.

I. INTRODUCTION

Information Retrieval (IR) is one of the major branches of Information Science discipline [2, 8]. The trend in information storage and retrieval can be traced back to 2000 BC when people of Sumerians chose special place to store clay tablets with cuneiform inscription [2, 7]. After they understand their work is efficient on use of information, they developed special categorization system that identifies every tablets and its content. Information retrieval is defined as finding of relevant documents from unstructured way of large collection that satisfies user's information need [9]. An information retrieval system is system that stores and manages information on documents and enables users to find the information they need. It returns documents that contain answer to users question rather than explicit answer to their information need. Most of the time-retrieved documents satisfy user's information needs. Documents, which satisfy user's information needs, are called relevant documents, whereas documents, which are not satisfying user's information need, are called irrelevant documents. In fact, there is no perfect information retrieval system, which retrieves all relevant documents [10]. Information Retrieval has two main subsystems, Indexing and Searching. Indexing is an offline process of representing and organizing large document collection using indexing structure such as Inverted file, sequential files and signature file to save storage memory space and speed up searching time. Searching is the process of relating index terms to query terms and return relevant documents to users query. Both indexing and searching are interrelated and dependent on each other for enhancing effectiveness and efficiency of IR system [9]. Efficiency is about optimizing computing resource such as the needed storage space and time complexity, while effectiveness concerned with relevancy of document retrieved that satisfies the users information need.

There are more than 80 spoken languages in Ethiopia. Tigrigna (ትግርኛ) is one of the local languages and it is a member of Semitic language of the Afro-Asiatic language family [14,15]. It is spoken in Tigray-Ethiopia as well as in Eritrea. Currently this language has more than 10 million speakers worldwide. Tigrigna is the official language of Tigray regional state of Ethiopia and academic language for primary school of the regional state. Tigrigna literature and myths are delivered as a field of study in many universities in Ethiopia [12]. Nowadays magazines, journals, newspapers, online education, books, entertainment Medias, videos, pictures

and tutors in Tigrigna language are available in electronic format both online and offline sources. There is huge amount of information being released with this language, since it is the language of education and research, language of administration and political welfares, language of religious activities and social interaction [11, 13]. As a result, the Tigrigna documents are increasing in size from time to time. This shows that there are large collections of Tigrigna document available in the web. Due to that, it is necessary to implement and design an IR system for Tigrigna language.

II. OBJECTIVES

A. General Objective

The main objective of this study was to design and implement IR system for Tigrigna language documents.

B. Specific Objective

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

- ✓ Conduct the literature review.
- ✓ Understand and explore basics of Tigrinya language and perform text operation
- ✓ Develop a prototype of Tigrigna document retrieval system.
- ✓ Evaluate the performance of the developed system.

III. RELATED LITERATURE

3.1 Overview of Tigrigna Language

1) Alphabets

Alphabets are sets of letters arranged in fixed orders of the language they used to write. They are also called phonemes, which contain consonants and vowels. There are different alphabetical representations in the world. The most alphabets representation is Latin or Roman alphabets, which have been adapted by numerous languages. The Ethiopic writing systems have also their own writing systems. Tigrigna have their own alphabets (ፊደል) and it is used for writing different documents of Tigrigna languages. They have thirty-five (35) base symbols with seven orders, which represent seven vowels for each base symbol [1].

2) Punctuation Marks

Identifying punctuation marks is vital to know word demarcation for natural language processing [1]. According [1] the punctuation are word separator mark (:) is used in the old literature to separate one word from other words. In the current literature, it is rarely used only in churches bible authors are used it. As, a result a single space is used to separate words instead of this punctuation marks. The end of sentence mark (::) is used to shows when an idea is finished. The sentence connector mark (፤) is used to connect two sentences or compound sentences in to one sentence. The list separator mark (፥ , ፣) is used to list things, separate parts of a sentence, and indicate a pause in a sentence or question. Like the other punctuation marks, the beginning of the list mark (፡-) is used at the beginning of the lists.in addition to those listed the above dot (.) used for acronym and abbreviation like slash (/) example ዓ.ም or ዓ/ም (which means Ethiopian calendar) , ቅ.ል.ክ or ቅ/ል/ክ (which means BC) and ድ.ል.ክ or ድ/ል/ክ (which means AD).

3) Tigrigna Morphology

Tigrigna is a morphological rich and highly inflected language by its nature. It has the root and pattern morphological system. The root system is a sequence of consonants and it represents the basic form of word formation. Tigrigna affixes are prefixing, suffixing and infixes to form inflectional and derivational word forms. Nouns in Tigrigna language have different sound for gender, number and other cases. For example, ሰብ (seb)-single person, ሰባት (sebat) - peoples, ሰብኣይ (seb'ay) – male person, ሰበይቲ (sebeyti) - female person. Adjectives of Tigrigna language are also inflected for gender and number. For example, ቀይሕ (keyh), ቀያሕቲ (keyahti) meaning 'red', 'reds' respectively. Verbs in Tigrigna Language show different morpheme syntactic features based on the arrangement of consonant (C) - vowel (V) patterns. For example, the root ስበር-'sbr' /to break/ of pattern (CCC) has forms such as ሰበረ-'sebere' (CVCVCV) in Active, ተሰበረ-'te-sebre'(te-CVCCV) in Passive.

3.2 IR Models

IR model is the mechanism of predicting and explain the need of the user given query to retrieve relevance documents from the entire collection. IR models serves as blueprint to develop applicable IR system therefore, we applied in our study. In addition to that, they guide the matching process to retrieve a ranked list of relevant document for a given query [3]. They are broadly categorized in to three main categories, which are Boolean, vector space, and probabilistic model.

1) Boolean Model

The Boolean retrieval model is a form for information retrieval in which any created query is expressed in a Boolean expression terms structure, that is in which terms are combined with the operators AND, OR, and NOT. The Boolean model views each document as a set of words and both the documents to be searched and the user's query are conceived as sets of terms [4].

2) Vector space model (VSM)

In vector space model, both the document and the query are represented in vector form. We have chosen vector space model for our study since it is a term weighting scheme, and the retrieved documents could be sorted according to their relevancy degree. Another significant feature for using VSM technique is the ability

to get a relevance feedback from the users of the system. Users can judge whether the retrieved document is relevant to their need/query or not. The Vector Space Model (VSM) or term vector model is an algebraic model used for Information Filtering, Information Retrieval, relevancy rankings and indexing. It represents natural language documents in a formal manner by the use of vectors in a multi-dimensional space, which has only positive axis intercepts. Nowadays Vector space model is most popular model in Information Retrieval system since it use non binary weighting technique it gives partial match or ranking the retrieval relevant documents [5]. There are four techniques used in vector space model those are Inner Product, Cosine similarity, Dice Similarity Jaccard Similarity [4]. The weight of terms in the documents or in the queries assigned by using term frequency (*tf*) and inverse document frequency (*tf*idf*) scheme which are the most successful and widely used automatic generation of weights [3].

3) Probabilistic Model

In which the relevance of a document for a given query could be estimated by using the probability of finding relevant information and the probability of finding a non-relevant information [4].

3.3 Related Works

Many researchers has conducted a research on developing Information Retrieval Systems for different languages using different techniques and models. Their common focus is on improving the search mechanisms used in IR systems in order to satisfy the user defined query as most as the system can using their native language. The main objective of developing an IR system is to understand the contents of documents. So "The more the system able to understand the contents of the documents the more effective will be the retrieval outcomes." [4].

Gezehagn Gutema and Bilal Ahmed [2, 4] had done research on development of information retrieval system for Afaan Oromo and Arabic languages respectively. They used vector space model to design and develop the IR system for retrieval of relevant document from the unstructured corpus. The system registered an encouraging result for Arabic language, but it is only limited with text format it is better to improve the system performance with covering pdf and html formats. The system for afaan Oromo registered an average 0.575(57.5%) precision and 0.6264(62.64%) recall. the stemmer function is not functional in this study it needs further improvement to register encouraging performance.

Amanuel Hippra [3] has designed a probabilistic based IR system for Amharic language and the system registered on the average 73% F-measure without controlling the synonyms and polysemious terms. Since Amharic language is morphologically rich in nature the system performance is highly affect by the polysemy and synonym terms it is good to refine the performance by using VSM.

Tsegay Semere [1] has developed dictionary-based approach Tigrigna to Amharic Cross language information retrieval system using probabilistic model. The author was tried to improve and increase the effectiveness of the IR system that has been built to work with Tigrigna and Amharic documents. Since probabilistic model uses binary matching technique, still there are hidden relevant documents that are not retrieved because it does not compute the degree of similarity between a query and each document. The system registered an average recall of 84% and 93% and an average precision of 75% and 64%, To provide better matching between the given query and corpus for better retrieval of relevant documents that satsfies user informatio need it is better to use Vector space model to enhance effectiveness of the IR system.

The review of related litretures motivates the authors to conduct study focusing on building vector space model based IR system that is valid over Tigrigna Language textual documents since nothing is done for Tigrigna Language on the previous studies.

IV. Tigrigna IR System Design And Architecture

Fig. 1. Depicts the model followed in the development of Tigrigna Text Retrieval system generation. In the model, the main tasks that are done for preprocessing of Tigrigna text are Normalization, Tokenization, Stop words removal, Stemming, and similarity measure.

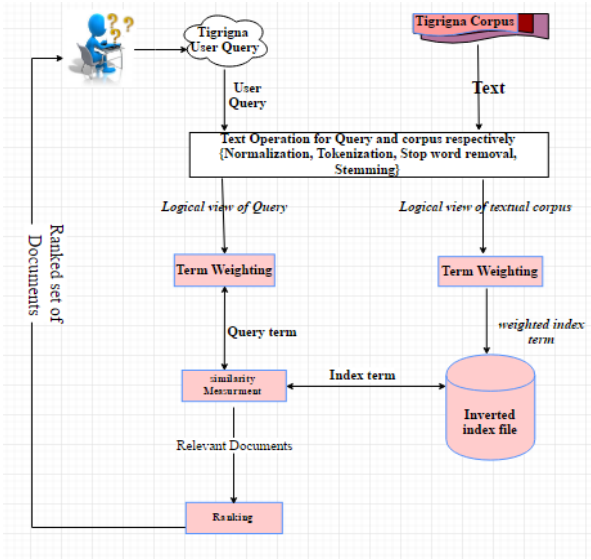

Fig. 1. General Architecture of Tigrigna Text Retrieval System.

A. Indexing Tigrigna Document and Query

In information retrieval, searching is possible or efficient when the database is small. However, in large databases searching will take much more time and space unless indexing structure is used to organize documents. Therefore, constructing and maintaining indexing on large database is necessary [1]. An Inverted Index is an index data structure storing a mapping from content, such

as words or numbers, to its document locations and is generally used to allow fast full text searches. Since there was no standard corpus for Tigrigna, we have constructed a large database corpus collection for Tigrigna language from different sources. Mainly to conduct this study Tigrinya documents are prepared from DWT information center (ድወት), mekalh Tigray Newspaper (ጋዜጣ መቓልሕ ትግራይ) and Tigrigna textbook for grade 10 and 12.

Text operation is performed to search relevant document from the Tigrigna corpus. Preprocessing is an important part of text processing. In the preprocessing stage file formats, character sets, and variant forms was converted, so that all text, regardless of its source, is in the same format. Text operation is language dependent. According to this point text operation for Tigrigna, language starts with normalization, Tokenization, removal of stop word and stemming. Those are also called as the steps followed by the researchers for Tigrigna language IR system design and development.

1) Normalization

Different alphabetical characters in Tigrigna writing system with the same sound are available and they have equivalent meaning to be standard those different symbols are converted in to one common representative form. Those are characters ሠ (se) and ሰ (se), ፀ (tse) and ጸ (tse), ሀ (he) and ኀ (he) should converted in to one common sound sense and word which are written in short/acronyms/ should also converted in to formal representation of words before tokenization. For example, መምህር - መምህር (teacher), ዶ/ር - ዶክተር (Doctor), ኣ/ክርስትያን - ኣብያተክርስትያን (churchs), ሕ/ሰብ - ሕብረተሰብ (society) etc. Because if we tokenize them before normalize, words that are written in the form of short/acronyms can lose their meaning and assumed as content bearing term.

```
def Normalization(N):
    h1=["ሀ","ሁ","ሂ","ሃ","ሄ","ህ","ሆ"]
    h2=["ኀ","ኁ","ኂ","ኃ","ኄ","ኅ","ኆ"]
    s1=["ሰ","ሱ","ሲ","ሳ","ሴ","ስ","ሶ"]
    s2=["ሠ","ሡ","ሢ","ሣ","ሤ","ሥ","ሦ"]
    ts1=["ፀ","ፁ","ፂ","ፃ","ፄ","ፅ","ፆ"]
    ts2=["ጸ","ጹ","ጺ","ጻ","ጼ","ጽ","ጾ"]
    d1=["እዩ","ኣሎ","እዮም","እውን","እምበር","ማሕበረሰብ","ዶክተር"]
    d2=[" 'ዩ","'ሎ","'ዮም","'ውን","'ምበር","ማ/ሰብ","ዶ/ር"]
    a1=["መምህር","ቤት ትምህርቲ","ኣብያተክርስትያን","ሕብረተሰብ","ሃይለስላሴ","ወይዘሮ'
    a2=["መ/ምህር","ቤ/ትምህርቲ","ኣ/ክርስትያን","ሕ/ሰብ","ሃ/ስላሴ","ወ/ሮ","ዝ/ዩ"]
    for i in range(len(h1)):
        N=N.replace(h2[i],h1[i])
        N=N.replace(s2[i],s1[i])
        N=N.replace(ts2[i],ts1[i])
        N=N.replace(d2[i],d1[i])
        N=N.replace(a2[i],a1[i])
    return N
```

Fig. 2. Python code for Normalization

2) Tokenization

In Tigrigna, language words are taken as tokens after all punctuation marks, control characters, numbers and special characters are removed from the text before the data is processed. All punctuation marks are converted to space and space is used as a word demarcation sometimes the (:) also used as word demarcation. Therefore, if a sequence of characters is followed by space or Punctuation (:), that sequence is identified as a word. A consecutive sequence of valid characters was recognized as a word in the tokenization process.

```
def Tokonization(token):
    p=['[',']','{','}',',',',','?','።','፦','“','..','”','፣','!','፤','፥','.','-','፡፡',' ? ']
    token=re.sub('[??o=፡?.??():“,፦<>-``/”??"??!‘’' '\ufeff]','',token)
    for i in p:
        token=token.replace(i,'')
        token= re.sub('[\d+]','', token)
        token=token.rstrip()
    return token
```

Fig.3. Python code for Tokenization

3) Stop word removal

Tigrigna language has domain independent stop words including prepositions, conjunctions, and articles. Generally, stop words are words, which serve as no purpose or meaning less for natural language processing (NLP) applications, but are used very frequently in composing documents, and these stop word lists are developed for two main reasons: These words may damage the stemming performance because stemming stop words have no any advantage in Natural Language Processing. Secondly, removing stop words helps to reduce the size of the file. The most frequently ranked stop word in Tigrigna language are Ab (ኣብ), Eyu (እዩ), kab (ካብ),Ewun (እውን), slezi (ስለዚ) and Nay (ናይ) like in English (the, a and an).

```
def Stopword(sw):
    stop=open("stopwords.txt",encoding="utf-8")
    stop=stop.read()
    stop=stop.replace('\ufeff','')
    stop=stop.split('\n')
    for i in stop:
        sw=sw.replace(i,'')
    return sw
```

Fig. 4. Python code for Stop word removal

4) Stemming

Like another languages Tigrigna has a rich of morphemes like prefix, suffix and infixes. Stemming is the process of converting the same word with different affixes into one-root form. It is one of the preprocessing technique made on automatic Tigrigna thesaurus construction. Since there is no available stemmer for Tigrigna language we have developed a simple stemmer that can remove common Tigrigna affixes (prefixes and suffixes) based on the concept of the algorithm developed by Yonas [6].

5) Term Weighting and Similarity Measurement

The fifth step in IR architecture design is term weighting and similarity measurement .Term weighting is highly related to recall and precision. In vector space model, term weighting based on single term statistics throughout the entire document. In order to calculate term weighting we need term frequency, document frequency and length normalization. Product of these three main factors is result of term weighting.

Weight (D_{ij}) = *tf*idf {where idf =log ((N/df), 10)}*
Tf= Term frequency of the term in each document
Df= document frequency which contains the term
Idf= inverse document frequency
N=total number of documents

6) Ranked Document

The final, result of the whole IR system process is to retrieve relevant document according to the scale of their relevancy. That is calculated in the similarity measurement. Since we have used vector space Model the IR model retrieve partial matching of similar documents to our query. The set ranked relevant document is obtained through the system in decreasing/descending order based on degree of similarity. High similarity score become first ranked relevant document. Which means the document with high similarity score have matched more with our query used to search relevant document from the given corpus collection. Always the order of ranking documents according their relevancy of similarity is done with reverse true this means in descending order, decreasing order or non- increasing order.

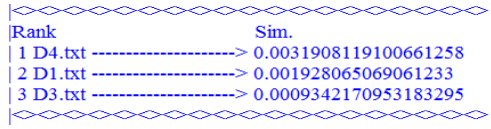

Fig. 5. Sample output of retrieved relevant document from the system

V. Experimental Results

A. Document Selection

Since there is no available standard documented corpus collection for Tigrigna Language like the TREC for English language, the authors develop their own corpus collections for this study. They took documents from different sources such as Wurayna Newspaper (ውራይና ጋዜጣ), DWT information center (ድወት ቴቪ), mekalh Tigray Gazeta (ጋዜጣ መቐለሕ ትግራይ) and Tigrigna textbook for grade 10 and 12. A total 30 Tigrigna documents were used as a document corpus to test the developed prototype IR system.

B. Query Selection

The authors were prepared a total six (6) queries samples to test for the selected sample Tigrigna documents. The preparation was done based on relevancy for the constructed Tigrigna corpus documents. After query and document selection, preprocessing is held the users can start asking and interact with system using the designed prototype.

The main task of the prototype is to integrate users with the system and search query terms and Tigrigna documents in the matrix to make comparison between documents and a given query. Then calculate the weight of each query terms based on the similarity measure notion implemented by vector space model, it calculates the score of each document and rank the retrieved relevant documents in descending order starting from the most relevant documents to the least.

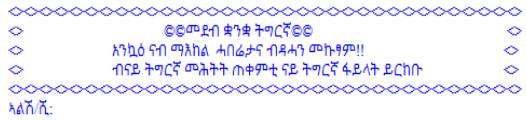

Fig. 6. The main page of the Designed IR system prototype

Meeting the design requirements is one factor that we used to evaluate our IR system. The following screen shots and explanations explain whether our system meets its design requirements. A user can enter a word as a query and must get the same or similar result. The following Figures are shown sample of screen shot experimental results generate by the developed system after the queries such as Query1, Query2, Query3 and Query4 are submitted respectively by the user. For instance, Fig.7 shows the result for the query ቁንቋ ትግርኛ እንታይ ማለት እዩ (means what it means by Tigrigna language). After the query is submitted to the system, it creates a list of vector by applying text operation. Finally, by taking the content bearing words like ቁንቋ and ትግርኛ from query 1 it calculates related terms from the indexed entire corpus collection by using similarity measure and rank out the most relevant documents to the query. The same process is followed for the other listed queries. Therefore, we can clearly see that our developed IR system indeed meets its design requirements.

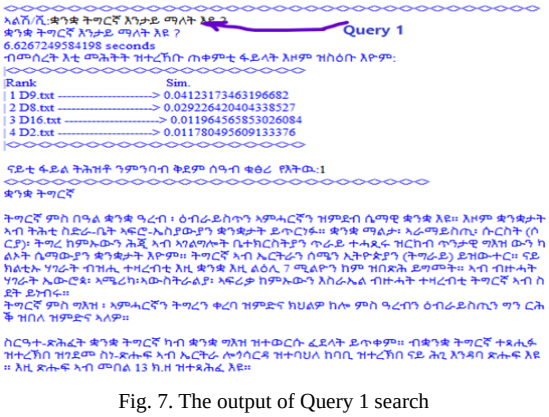

Fig. 7. The output of Query 1 search

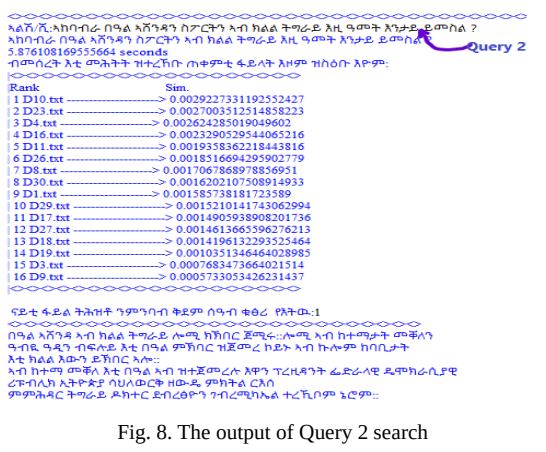

Fig. 8. The output of Query 2 search

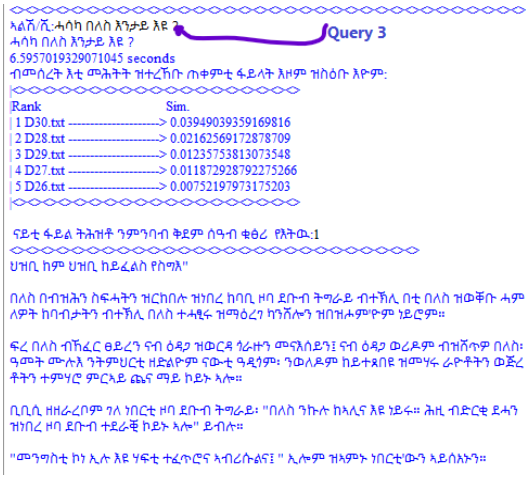

Fig. 9. The output of Query 3 search

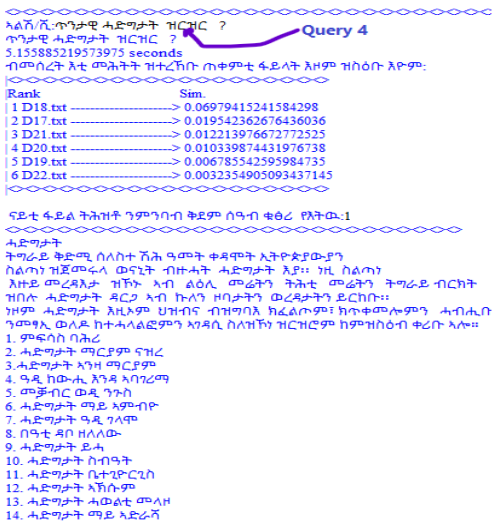

Fig. 10. The output of Query 4 search

VI. System Performance Evaluation

IR system should be evaluated its performance and accuracy after designed and developed. Evaluation of IR system involves two things effectiveness and efficiency [2]. It is important to evaluate both effectives and efficiency of the system. Efficiency of IR system is depends on time and space complexity so our system is finally low efficiency because it consumes more time and space to create the inverted file(vocabulary and posting file). We have used suffix tree to build the stemmed dictionary, takes more time and used word level inverted file to create the word list and posting file then it needs more space requirement. However, it is effective for searching time because it uses binary search and saves time for the users. The IR system effectiveness was evaluated in various way of evaluation metrics. The most common evaluation methods are precision and recall [2]. In addition to those evaluation metrics, F-Measure and E-measure are included in this study to measure the performance of the system.

Precision is the ratio of the number of relevant documents a search retrieves, by the total number of documents retrieved. In other word it is the fraction of the documents retrieved that are relevant to the user's information need. It evaluates the capability of the IR system to retrieve top-ranked documents that are most relevant to the user need query, and it is defined to be the percentage of the retrieved documents that are truly relevant to the users query.

Recall is the ratio of number of relevant documents retrieved, by the total number of existing relevant documents that should have been retrieved. It is the fraction of the total retrieved relevant documents per total successfully retrieved documents to the given query by the system. It evaluates capability of IR system to get all the relevant documents in the database. It is defined as the percentage of the documents that are relevant to the user query.

$$\text{Re}\,call = \frac{|\{\text{Re}\,levant\} \cap \{\text{Re}\,trieved\}|}{|\{\text{Re}\,levant\}|} \quad (1)$$

$$\text{Pr}\,ecision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (2)$$

F-Measure is performance measure that takes in to account or make balance both Precision and Recall. In addition, it is called as harmonic mean.

$$F = \frac{2PR}{P+R} \quad (3)$$

(F= F-Measure, P=Precision, R=Recall)

E-Measure is also another performance measurement of IR system, which allows the user to specify the importance of Precision and Recall.

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} \quad (4)$$

(E-Measure, P=Precision, R=Recall, β is constant value which is given by yourself).

Value of β controls trade-off: Where β = 1: Equal weight for precision and recall (E=F). β > 1: Weight recall more. It emphasizes recall. β < 1: Weight precision more. It emphasizes precision. Depending on the above listed evaluation techniques the system evaluation is done in the table below.

The table in the below shows a list of query samples used on system evaluation process to test whether the developed IR system is efficient and effective. The authors was selected six (6) sample queries to evaluate the system performance.

The results obtained from the system for the above queries are processes using the formulas in equation (1), (2), (3) and (4). For instance, the result for query 1 was calculated as follows: The total number of relevant documents assumed by the user for query 1 from the entire corpus are 4 and automatically 4 documents are retrieved by the system when the query submitted by the user the same is true all the retrieved documents are relevant.

*Using Equation (1):

Recall=4/4=1, it is100% effective.

*Using Equation (2):

Precision =4/4=1, it is 100% efficient.

*Using Equation (3):

*F-measure = {2(1*1)/ (1+1)} =2/2=1*

*Using Equation (4):

$$E-mesure=\frac{(1+1^2)1*1}{1^2*1+1}=2/2=1 \text{ where } \beta=1$$

The same process is also held for the remaining five query terms in the above table to calculate their performance. Finally, average precision, recall, F-measure and E-measure was calculated.

As it is shown in Table 1, the obtained result is 0.70(70%) an average precision, 0.84(84%) an average recall, 0.75(75%) F-Measure, and 0.75(0.75%) E-Measure. From the above result our Precision is high it evaluates that all relevant document are retrieved and recall was registered high percentage because of the stemmer algorithm is not well functional some irrelevant documents are retrieved. As we know stemming, enhance the value of Recall because when we are stemming the Words or terms are merges together, which means it increases the number of matches between the sample document and the query so additional irrelevant results are retrieved by the system, which looks like to the stemmed term and affects Precision. The system retrieves 70% relevant documents for these given six (6) queries. Which indicates that our system performance is good. In other hand, F-Measure and E-Measure scores the same result 0.75(75%) because the value of β is one. As we have discussed in the above if the value of β is equal to one (β=1) then E-Measure becomes the same as F-Measure. Therefore, our system realizes the truth in the experimental result.

Table 1. Performance measurement of the developed IR system

NO	List of Queries	Relevant	Retrieved	Relevant Retrieved	Recall	Precision	F-Measure	E-Measure β=1
1	ቋንቋ ትግርኛ እንታይ ማለት እዩ ?	4	4	4	1	1	1	1
2	በዓል ደቂ ኣንስትዮ ኣብ ክልል ትግራይ እንታይ ይመስል?	6	16	4	0.67	0.25	0.36	0.36
3	ኣከባብራ በዓል ኣሸንዳን ስፖርትን ኣብ ክልል ትግራይ እዚ ዓመት እንታይ ይመስል?	12	16	10	0.83	0.63	0.72	0.72
4	ሓሳካ በለስ እንታይ እዩ ?	5	5	4	0.8	0.8	0.8	0.8
5	ጥንታዊ ሓድግታት ዘርዝር?	4	6	3	0.75	0.5	0.6	0.6
6	ረብሓታት ክንክን ሃፍቲ ተፈጥሮ እንታይ እንታይ እዮም?	6	6	6	1	1	1	1
	Average				0.84	0.70	0.75	0.75

VII. Conclusion And Recommendations

A. Conclusion

We presented a study on Tigrigna language conducted with the aim of designing and implementing IR system for Tigrigna textual documents using vector space model. Since the World Wide Web has become a vital means of facilitating global communication and a huge repository of information in the form of text, audio, video, and image. To provide this wealth of information accessible to the user be needs inventions of search engines and information retrieval systems. Tigrigna language is one of the languages that have a representation in the global information space and increasing number of users we are struggling to implement complete language specific search engine like another global languages.

B. Recommendations

This is the beginning level of designing Information Retrieval System for Tigrigna language. The system has registered a promising result. Even if the system performance showed an encouraging result, there were expected works to be done in the future to improve the system activities. Therefore, the following are recommended as future research directions.

- Further investigation on construction of standard corpora for Tigrigna language is needed like TREC for English Language.
- The stemmer will highly enhanced in the future.
- The size of the collection used in this research was small. Larger collections must be set up and used in order to refine retrieval results. The larger the collection size, the finer the results.

REFERENCES

[1] T.Semere,"Probabilistic Tigrigna-Amharic Cross Language Information Retrieval (CLIR) ",Msc Thesis, School of Information Science, Addis Ababa University,2013.

[2] G. Gezehagn," Afaan Oromo Text Retrieval System ",Msc Thesis ,School of Information Science ,Addis Ababa University,2012.

[3] Hirpa, "Probabilistic Information Retrieval for Amharic Language", Msc Thesis, School of Information Science, Addis Ababa University, 2012.

[4] Ahmad," Applying Vector Space Model (VSM) Techniques in Information Retrieval For Arabic Language", N.D.

[5] Polyvyanyy, D. Kuropka," A Quantitative Evaluation of the Enhanced Topic-Based Vector Space Model ", Hasso-Plattner-Institut Für Softwaresystemtechnik and Der Universität Potsdam, 2007.

[6] Y. Fisseha, "Development of Stemming Algorithm for Tigrigna Text", Msc Thesis, School Of Information Science, Addis Ababa University, 2011.

[7] R. Baeza-Yates, Information Retrieval: Data Structure & Algorithms, 1st Ed. Waterloo: University of Waterloo, 2004, Pp. 1-630.

[8] P. Ingwersen, Information Retrieval Interaction, 1st Ed. London: Taylor Graham Publishing, 2002.

[9] D. Manning, P. Raghavan, and H. Schutze, An Introduction to Information Retrieval, Online Edition, Cambridge: Cambridge up, 2009.

[10] Hiemstra, Information Retrieval Models, Wiley Online. New York: John Wiley & Sons, Ltd, 2009.

[11] H. E. Wolff, sematic-Cushitic Languages, Encyclopedia Britannica. Encyclopedia Britannica, Inc., 2012.

[12] Kula Kekeba, V. Varma, and P. Pingali, Evaluation of Oromo-English Cross-Language Information Retrieval, Journal of International Joint Conference on Artificial Intelligence (IJCAI)-2007, Vol. IIIT/TR/20, June, 2008.

[13] S. Heinz, "Efficient Single-Pass Index Construction for Text Databases", Journal of The American Society, Vol. 54, No. 8, Pp. 713-729, 2003.

[14] Y. K. Tedla, "Nagaoka Tigrinya Corpus : Design and Development of Part-of-speech Tagged Corpus," Nagaoka University of Technology, pp. 1–4, 2016.

[15] H. Kidu, "A Mobile Based Tigrigna Language Learning Tool, University of Gondar" pp. 50–53, 2017.

Authors' Profiles

Teklay Birhane is earned his BSc. Degree in Information Science from Mekelle University, Ethiopia in 2017. He is staff member in Department of Information Science under College of Natural and Computational Science in Mekelle University. Currently he is joined to Haramaya University in Ethiopia to attend his MSc. Program in Department of Information Science. His research interests are in the areas of artificial intelligence, mobile application development (android system), natural language processing, knowledge management, data mining and machine learning.

Brhanu Hailu is earned his BSc. Degree in Information Science from Mekelle University, Ethiopia in 2017. He is staff member in Department of Information Science under College of Natural and Computational Science in Mekelle University. Currently he is joined to Haramaya University in Ethiopia to attend his MSc. Program in Department of Information Science. His research interests are in the areas of artificial intelligence, information retrieval natural language processing, knowledge management, data mining, Archival system and document management system.

How to cite this paper: Teklay Birhane, Birhanu Hailu, "Design and Implementation of IR System for Tigrigna Textual Documents ", International Journal of Modern Education and Computer Science (IJMECS), Vol.11, No.11, pp. 31-38, 2019.DOI: 10.5815/ijmecs.2019.11.05

