ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

Somali Language Information Retrieval Using Query Expansion

Abdisalam Mahamed Badel

A Thesis Submitted to The Department of Computer Science in Partial Fulfilment for The Degree of Masters of Science in Computer Science

Addis Ababa, Ethiopia

June 2020

Addis Ababa University

College of Natural Sciences

Abdisalam Mahamed Badel

Advisor: Yaregal Assabie (PhD)

This is to certify that the thesis prepared by Abdisalam Mahamed Badel, titled: *Somali Language Information Retrieval Using Query Expansion* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining committee:

	Name	Signature	Date
Advisor:	Yaregal Assabie (PhD)	_____	_____
Examiner:	Ayalew Belay (PhD)	_____	_____
Examiner:	Solomon Gizaw(PhD)	_____	_____

Abstract

In these days the number of documents available online on the Internet is growing at a very rapid speed and has become quite large over the past few years. To search such documents various types of search engines or Information Retrieval systems has been performed for numerous languages. Based on the language the information seeker is using, Information Retrieval allows to find relevant information with respect to the information need of the user. During retrieval many words can be represented by using more than one representation and this can create a confusion in the retrieval process. Modern Information Retrieval systems employ techniques to deal with such confusion in retrieval process.

This research deals with Somali Language Information Retrieval system. Keyword based Information Retrieval systems confirm lower performance compared to the systems with query expansion. To solve such problem, we have implemented the research with query expansion technique using thesaurus. Thesaurus expands original query terms by finding synonyms of terms from manually constructed list of alphabetically ordered words.

A prototype has been developed to test the performance of the system. The prototype is implemented using java programming language with open source library called lucene. A testing collection of about 2007 documents of three different domains collected from various sites used. The experimental result of this study is a two phase, the first phase of the experiment is to test the system without query expansion. The second phase is testing the system with query expansion. Our experiment shows a hopeful result, and notable improvement by recording 53% and 72% of precision and recall respectively. After query expansion technique is applied the system recorded 64% and 84% of precision and recall respectively with unranked set of retrieval.

Keywords: Information Retrieval, Somali Language, Thesaurus, Query Expansion.

Dedication

This work is dedicated to my family.

Acknowledgement

First and foremost, praises and thanks to Allah, the Almighty, for giving me healthy throughout my research work to complete the research successfully.

I would like to express my profound and sincere gratitude to my research advisor, Dr. Yaregal Assabie for giving me the opportunity to do research and providing invaluable guidance throughout this research. His vision, dynamism, sincerity and motivation have profoundly inspired me. He has taught me the methodology to carry out the research and to present the research work as clearly as possible. It was a great honor and privilege to work and study under his guidance. I am very grateful for what he has offered me. I would also like to thank him for his friendship, kindness and compassion during the research work. I am very grateful to my parents for their love, caring prayers, and sacrifices for educating and preparing me for my future. I thank to my professors and staffs in the department of computer science for training me, through the years we being together. I would like to say thanks to my friends and research colleagues. for their constant encouragement. I would also like to say thank you to my classmates in the department of computer science for their support during my research work. I would like to thank to Mr. abdulahi abdi and Mr. dahir, for helping me, linguistically. Finally, I would like to thank to all the people who have gave me a support, to complete this research work directly or indirectly.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Acronyms/Abbreviations

ACM	Association of computing machinery
BM25	Best Match 25
CLIR	Cross-language Information Retrieval
DOCS	Documents
F	F-measure
GUI	Graphic User Interface
IDF	Inverse Document Frequency
IDF-TF	Inverse Document Term Frequency
IR	Information Retrieval
NLP	Natural Language processing
P	Precision
POS	Parts of Speech
Q	Query
QE	Query Expansion
REL	Relevance
R	Recall
RSJ	Robertson Spark Jones
TF	Term Frequency
TREC	Text Retrieval Conference
VSM	Vector Space Model

Chapter One: Introduction

1.1 Background

Shakespeare defined the seven ages of man starting from infancy and leading to senility [1]. The history of Information Retrieval matches such a life. Its idea started to be widely known in 1945, although people have started using it. Yet it has first been researched by vannervar Bush [1]. He has established a fast access to the content of the world's libraries and has shown the importance of Information Retrieval. Because of the change in technology [2] the need for information has massively increased and surpassed the one in the past.

In recent days, it is becoming undeniable that retrieving information is important in our lives [2]. Because our daily live activities depend on information, and the largest source of such information is the Internet, the www. The www contains different information of different sizes and contents [2]. There is a huge increase in the number of data available online in the world wide web [3]. Thus, such tremendous and fast growth of information [2]. In the Internet makes information seekers not to get the relevant information in satisfactory way.

Information Retrieval is getting items (typically documents) of unstructured nature (typically text) which pleases user's need of information from bulky or huge data [3]. It is interacted by many people every day, when searching on their emails, desktop and the web etc[4].

Information Retrieval is a technology that handles the retrieval of unstructured data usually textual based documents in response to a query [2]. It finds data from large collection of sources by searching and the searching is based on mainly metadata. IR systems were primarily initiated to assist manage the large scientific literature which has been developed since 1940's [5]. Information Retrieval systems retrieve and present relevant documents or information to users as per their query or information need [5].

In the development of Information Retrieval system two main subsystems are important and the IR system contains these two subsystems which are indexing and searching [6]. Giving files an index is an offline process of managing and arranging large collection of documents by using indexing methods such as signature files, inverted files and sequential files [6]. Indexing is used to minimize the memory consumption and increase the speed of searching.

The other subsystem of the IR system is searching and it is way of relating index terms to the asked query terms and then giving users the relevant documents. Indexing and searching are two IR subsystems which [6] are dependent on each other.

In addition to the subsystems that IR has, it has also retrieval models such as: The Boolean model, vector space model and probabilistic model [7]. The Boolean model is usually considered as set theory whereas the vector space and the probabilistic models are said to be statistical models as they use statistical approaches.

It sometimes becomes hard for IR systems retrieve relevant documents [8]. Such problem can be caused by word mismatch and information overload. To solve the problem researchers, use query expansion approaches to add meaningful words to the original query [8]. One of such query expansion techniques is the usage of thesaurus. The thesaurus lists relationships between words and finds synonyms for the query terms [8]. Until the start of Information Retrieval, many systems have been developed to support user's information need in search engines [9]. However, users of minority languages such as Somali still feel problems with searching in the search engines [9].

1.2 Motivation

Somali language is the official language of the Somali region (state) [10] furthermore Somali language is spoken by ethnic Somalis in Kenya and Djibouti. In addition to this Somali language is the official and working language of Somalia [10, 12]. Apart from the people of the above-mentioned countries the Somali language is spoken by the Somali diasporas. The language has had a written form only since 1972 [12].

The total population speaking Somali language in the Somali region (state) has been estimated 5.3 million, according to the Census of the Ethiopian Central Statistical Agency in 2013 [13]. Furthermore, the number of Somali speakers in Somalia has been estimated to be between 13 to 25 million [14]. Although there is no exact estimation because of security-related problems in the past.

The Somali state educational bureau, as well as the regional offices, print different documents online. Documents are also produced by the government of great Somalia and Somalis in Kenya and Djibouti. For this reason, documents available online increase, so that

searching big amount of information manually becomes tedious and nearly impossible. Therefore, IR can answer those mentioned problems by offering a system that the user can easily find the information he/she needs from the digital data. This motivated us to work Somali language Information Retrieval.

1.3 Statement of the problem

Since Information Retrieval came into existence, and Internet users have started using the IR system. Many IR studies have been conducted on different languages according to the review of the related work that the researchers have conducted, these languages could be internationally or locally spoken.

Information Retrieval researches have been conducted on English [50, 51, 52, 53, 54]. It is undeniable that Information Retrieval studies have been mostly conducted in English. As it has been one of the languages preferred to carry out researches in the previous times. IR is conducted on the German language [55]. Researches have also been conducted on Arabic IR [56, 57]. There were hopeful attempts to conduct Information Retrieval researches for different Ethiopian languages such as Amharic [59, 60, 61], Afaan Oromo [5, 62], Tigrinya [9, 63]. etc. Elizabeth Boschee *et al* [64] proposed a low-resource cross-lingual domain-focused Information Retrieval for effective rapid document triage. In this research Somali language is taken as one of the low-resourced languages. Authors in this paper express that English speakers can access what is written in Somali or another language. Homed Bonab *et al* [65] conducted CLIR translation resource scarcity using high-resource languages. This research has been conducted by using parallel corpora translation. In both of these studies queries written in one language are translated into another language mostly into English.

Knowing that every language has its own characteristics and sentence structure. We propose that it is better to have Information Retrieval on the native Somali language, then promote it into CLIR. Because the so far conducted CLIR researches cannot help much Somali speakers access stored information rather it helps English speakers. It is good to do research that studies Somali language properties and uses its own corpus instead of translating the query. Thus, the objective of this research is to study Somali Language Information Retrieval using thesaurus as query expansion.

1.4 Objectives

General Objective

The general objective of this research is to design Somali Language Information Retrieval using thesaurus as query expansion.

Specific objectives

The general objective of the study can't be achieved without doing specific tasks so, in order to attain the general objective, the following specific objectives have been identified.

- To review literature on Information Retrieval and query expansion.
- To identify the characteristics of Somali language.
- To design a general architecture for Somali Language Information Retrieval using thesaurus as query expansion.
- To collect a corpus of Somali text.
- To develop a prototype.
- To evaluate the performance of the system.

1.5 **Methods**

Through the study the following will be the methodologies which will be used by the researcher.

- **Literature review**

Review on the related studies of the research will be done, to understand the way to implement Somali language information retrieval. Books and other previous researches will be examined. The grammatical formation of the Somali language will be reviewed as it will be an important component.

- **Data collection**

Data collection will be an important task during the study, there will be data collection related to the structure and formation of Somali language. We will collect sample textual

data in the language. After textual data is collected then, this sample data will be used as a testing purpose. This collected data will also be used as an experiment.

- **Prototype development**

In this part of the study, the researcher will develop a prototype to implement Somali Language Information Retrieval system. The prototype mostly will be a user interface in which the user can provide his/her information need. To develop the prototype java programming and different modules of it will be used in addition NetBeans will be used as development editor.

- **Evaluation**

To check the effectiveness and efficiency of the system there will be evaluation methods which will be conducted. Evaluation methods like recall, f-measure and precision will be used for evaluation.

1.6 Scope and Limitation

The scope of this study will be designing and development of Somali Language Information Retrieval system that retrieves Somali text. The study will stem and normalize Somali words and then index them to make the retrieving process of the system faster. The limitation of this study is that, some tasks in NLP are not fully implemented and need enhancement such as: Morphological analysis, Parts of speech tagging and stemmer.

1.7 Application of Result

Successful completion of this study will help Somali speakers retrieve relevant information of their interests, without loss of time. This study will also help people learning Somali retrieve their relevant data when searching online. The Somali higher education will benefit from this study as they store books and learning materials online. Students those who study in Somali will get a benefit from this study as it will help them search and retrieve knowledge written in Somali and access online data. Regional offices will benefit from this research after complete as it will allow searching with less time of the relevant data.

1.8 Thesis Organization

The rest of this study is organized as follows: Chapter Two presents literature review. Chapter Three describes the related papers of different languages and possible techniques used by the researchers. Chapter Four presents the general architecture of the proposed Somali language IR system and algorithms used to implement the components in the architecture. Chapter Five explains the implementation and evaluation of the proposed system, data analysis and interpretation of the designed system. Finally, Chapter Six concludes the thesis with conclusion, contributions and future work of the research.

Chapter Two: Literature Review

2.1 Introduction

In this chapter, we will focus on models and techniques for Information Retrieval systems. The chapter contains four sections. In section one of this chapter, we will give detailed explanations of Information Retrieval and its subsystems. The second section of the chapter will present classical models of Information Retrieval. The third section will summarize the concept of query expansion and the last section will present details of Somali language. It's grammatical and sentence structures.

2.2 Information Retrieval

After the invention of the web in 1990, people started to store their documents and share ideas on the web [15]. Based on the storage and knowledge sharing on the web. The web became a universal repository of human knowledge and culture which has increased the level of knowledge share. However, the web has solved storage problems and facilitated the sharing of knowledge yet it brought problems. Which are related to the way people retrieve texts, one of the problems it has brought is finding a single document from the huge documents on the web [15]. For this reason, emphasis has been given to study the field of Information Retrieval.

According to [16] the objective of Information Retrieval is to query and manage free (unstructured) data by using modern computer tools so that it can quickly find the fields which the users need. The incompetence of the Information Retrieval system is caused by the inexactness of the user's query and keywords selected by the user [17]. A short and un-clear query makes the user unable to get his/her information need.

Information Retrieval is usually concerned about finding data that is mostly in an unstructured form [18]. IR systems consist of three main subsystems, such as Indexing, Processing, and Searching, which each has its importance for IR [18]. These components have different components inside, such as stop word removal, stemming, normalization and tokenization [19]. The next lines of text give brief explanations of IR system components:

Indexing: is an offline process used to select symbols that are important during searching, this process extracts documents [18, 19]. After the analyses and processing of documents, potential terms are indexed and an inverted file is created. The inverted index is a type of data structure that maps from the contents such as, numbers or words to their locations in documents [19]. It is one of the most popular data structures and allows a fast search of text. The [19] Inverted indexes are classified into two main ones: record level and word level inverted index. Record level inverted index usually contains the list of references to a document and word-level contains the posting of the words in addition to the references.

In collecting text documents to convert into an inverted index, follows few steps including [19]: tokenization, stop word removal, stemming.

Tokenization: is a method for splitting punctuations, special characters, and white spaces from the document. A document to collect may contain full characters such as full stops and quotations. Which might need removal for processing. In this case, tokenization is applied.

stop word-removal: these words are the frequently occurring words in a document. They don't have the power to distinguish themselves from one document to another document. Some common English stop words are a, the, an, so, this, for, etc. In text process removing stop, words save memory space of up to 25-30%.

Stemming: is a technique used usually to find the root form of a word [19]. It is a method used by many IR systems and sometimes one of the core techniques in text processing. Stemming is used by most search engines. To stem a word there are various algorithms for different languages. As languages have different inflectional forms the algorithms for them might be of different. In natural language the words are mostly inflected morphologically for this reason stemming is an important technique for text processing in information retrieval and similar systems.

2.3 Information Retrieval Models

Information Retrieval consists of diversified models for implementation however, there are well known and mostly used models [20]. these are written in the next sections in detail:

2.3.1 Boolean Model

According to Venkat N. Gudivada *et al.* [20] Boolean model is one of the simplest and earliest models of Information Retrieval. The documents in this model are either relevant or non-relevant to the user's query. Thus, the model does not consider relevance, it is based on set theory and Boolean logic. The terms in this model are shown as Boolean variables, the values of terms may be true or false, or in a binary format as 1 or 0.

The Boolean query is any valid Boolean expression that can be formed from dictionary terms combined with the three Boolean operators: AND, OR and NOT [20]. These operators are used to express the Boolean expressions disjunctive normal form (d1^d2^d3) V (d4^d5) or conjunctive normal form (d1Vd2Vd3) V (d4^d5).

2.3.2 Probabilistic Models

Information Retrieval has certain probabilistic models, which use the probability ranking principles such models include:

probability model

Maron and Kuhns familiarized ranking with the use of probability of relevance and this idea was turned into a principle by Stephen Robertson by formulating the probability ranking principle [21]. The main task that makes the probability retrieval model different than other models is that of ranking documents in decreasing order [21]. Stephen Robertson and Karen spark-jones proposed to rank documents with P(R|D). That means the probability of relevance R given the documents content description d.

The probability model ranks documents in contrast to the vector space model, the ranking of the documents is based on the relevance of the documents to the query [21]. The documents and queries are both assumed to be observations of the variables D and Q. The relevance is, in addition, assumed to be binary random variable R. In this sense the relevance can be

either R (relevance) or –R (non-relevance), the relevance of a query to a document is formulated as p(r|d, q).

Using the Bayes theorem and algebraic manipulation, p(r|d, q) can be expressed as:

$$P(R|D,Q) = log \frac{P(R|D,r)}{P(R|D,\bar{r})} \quad (1)$$

Okapi BM25 model

Waweru Mwangi *et al.* [22] present that the model Okapi Best Match 25 is a non-binary model that was developed as part of the TREC conference. This model is a probabilistic model and it depends on probability theory. It uses the average length of each document separately using tuning parameters. The parameters it uses are k1 and b to normalize the saturation of the formula. Stephen Robertson and Hugo Zaragoza [22] argue that BM25 is closely related to the notion of estimating probability of reference for each pair of documents. It ranks documents in relative to a given query, in descending order of the probability of relevance. For scoring function BM25 uses term-weighting and document score. They state that to use BM25 as ranking algorithm in Information Retrieval systems we have to choose values for b and k1 parameters. In addition to this RSJ needs to be instantiated [23]. This model is based on a bag-of-words retrieval function which ranks the documents based on the query terms appearing in each document [27, 22]. The formula of the scoring function used by BM25 is:

$$score(D,Q) = \sum_{t \epsilon q, f_t, d>0} \log(1 + \frac{N - df_t + 0.5}{df_t + 0.5}). \frac{f_{t,d}}{f_{t,d+k1.(1-b\frac{l(d)}{avgdl}}} \quad (2)$$

Where k1 and b are parameters and defined as k1 = 1.5, and k1 used to control term-frequency saturation. As it is obvious that low values may result in quicker saturation while high values result in slower saturation and b = 0.75 and it used to control field-length normalization of a document. l(d) is the length of the document in words, avgdl is the average document length in the text collection. Q is a query 0.5 is a normalization factor, N is a total document in the collection, ft is term frequency, df is document frequency. BM25 [22] calculates idf to ensure whether a term is rare or not across all documents uses the formula:

$$\mathrm{IDF}(\mathrm{qi}, \mathrm{D}) = \log \frac{(N - n(qi) + 0.5)}{n(qi) + 0.5} \quad (3)$$

Where n(qi) is the number of documents which contain term qi, f(qi,D) term frequency in the document D and N is total number of documents.

2.3.3 Vector Space Model

The vector space model (VSM) is a model that characterizes documents usually text and queries as vectors in multidimensional space. Its dimensions are the terms used to build an index to characterize the documents [24, 25]. In time of using large document collections, the vector space model is mostly used because of its simplicity and efficiency. During implementation, this model follows three phases: document indexing phase where meaningful words or terms are extracted from document text, the weighting of indexed terms for the purpose of increasing the relevance of the document to the query of the user. And thirdly ranking the document based on a similarity measure. Figure 1 shows an example of a vector space model with two documents, query and three terms [24]:

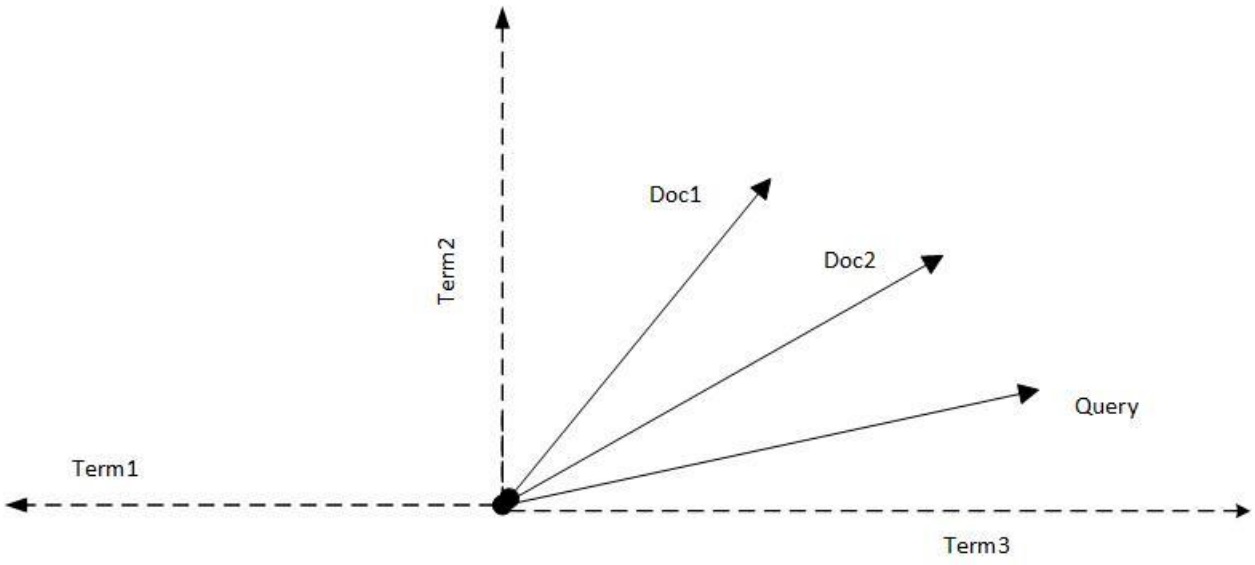

Figure 2.1: A Vector Space Model for two documents, 3 terms, and a query [24].

According to Singh *et al.* [26] the vector space model leads its users to documents that might be more similar and might be too significance by calculating the distance or angle measure between the query and terms or document. This model is based on the assumption that the

meaning of a document can be understood from the document's constituent terms. Documents in the vector space model are represented as vectors where document d is represented as terms d = (t1, t2, … tn), such that (1<= i <= t) is a non-negative value denoting the single or multiple occurrences of term i in document d [26].

In the vector space model, each term in a document represents a dimension in the space [22]. In addition to the terms the queries are also represented as vectors where Q= (t1, t2,…….tn) and term ti(ti>=1) is non-negative denoting the number of occurrences of term t in the query[26].

To know how similar two documents are in the vector space model [26]. There are number of techniques used such as inner product, cosine measure (similarity), dice coefficient and Jaccard coefficient. However, the cosine coefficient which measures the angle between document vector and query vector is the mostly used similarity measure.

Singh *et al*. [26] stated that cosine measure computes the angle between the vectors in a high-dimensional space. The authors add that for two vectors d and d' the cosine similarity between them is given by (D*D')/(|D|*|D'|) [24]. From the equation, D*D is the vector product of D and D', computed by multiplying the equivalent frequencies.

The vector space model uses term weighting [26]. This concept tells that not all words in a given document are meaningful, a word may or might be relevant to a document or it might not be relevant to it.

2.4 Term Weighting Methods

As the authors in [28, 7] state there are three factors for term weighting in information retrieval. The first factor of these factors is the term frequency (tf) in a document that represents the content of the document. The second factor is the inverse document frequency(idf) which has been intended to increase a term's distinguishing ability to select all relevant documents from irrelevant documents. These two factors are generally merged using multiplication operation and so enhance the performance of the Information Retrieval systems. The last and third factor is said to be the cosine similarity that equalizes the length of the documents [28, 5]. The term frequency and inverse document frequency, as well as cosine similarity of documents, is defined as follows:

The term frequency (tf) is the number of times in which the specific term appears in a document. Importantly the measurement of the term frequency in a document helps to know how important a term is in a particular document. The term frequency is found using the formula:

$$tf_{td} = f_{td} / \max(f_{td}) \quad (4)$$

From equation 4, the frequency of term t in a document d is found by dividing tf_{td} with $\max(f_{td})$. Where tf_{td} is the term frequency of term t in a document d and $\max(f_{td})$ is the maximum of the term t in a document d.

The inverse document frequency (idf) is the measure to know whether the term is common or rare in all documents. Its formula is taken by logarithm of total number of documents divided by the number of documents containing the term t:

$$\text{idf} = \log_2(N/df_t) \quad (5)$$

In equation 5, N is the total number of documents and df_t is the number of documents in which the term t appears.

The frequency of term t and the inverse document of the term is usually the dot product of the two and is found using:

$$tf * idf = tf_{td} * \log_2(N/df_t) \quad (6)$$

2.5 Similarity Measurement

According to [29] measuring the similarity between two documents, has a lot of measures that are obtainable for the calculation of inter-document relations. Retrieval systems have different similarity measuring techniques such as cosine, jacard, inner product and dice coefficient.

Cosine similarity: During calculating the similarity, the choice of a particular type of measure may change the result produced by the calculation process [29]. In Information Retrieval systems a similarity matrix is needed to cluster documents. The author in [29] argues that for every two documents Di and Dj which are in the same document collection. The similarity sim(Di, Dj) will have a value which will be the same in all queries which the

user sends to the Information Retrieval system. In IR the similarity between two documents is calculated as:

$$Sim(Di, Dj) = \frac{\sum_{k=1}^{n} dik.djk}{\sqrt{\sum_{k=1}^{n} d^2ik.\sum_{k=1}^{n} d^2jk}} \quad (7)$$

In equation 7, Di and Dj are two different documents, whereas (dik and djk) are the terms in the document collection, in which the similarity is decided by comparing them.

Jaccard Similarity Measure: This type of document measurement, measures document similarity by using two values which are normally between 0 and 1 [30]. The 0 shows that documents are not similar, while 1 shows similarity between the documents. In Jaccard similarity there is always probability of similarity between the documents and this is between the 0 and 1 [30]. The similarity of documents in Jaccard measurement is calculated using the formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (8)$$

Where A and B are the two documents which their similarity is to be measured.

Dot product: Dot product is a measurement technique applied for document measurement [3]. It's the document is vector (d) of a document which derived, from the document d, with one component in the vector for each dictionary term. Dot product can be assumed that the components are computed using tf-idf weighting methods. The dot product is also the set of documents in a collection which may be regarded as a set of vectors in a vector space in this case there is one axis for each term in the vector.

2.6 Query Expansion

Query expansion (QE) is a transitional phase between the indexing and matching process in Information Retrieval systems [31]. Users in the Information Retrieval system are responsible for submitting asking meaningful queries in to the system. Sometimes users give a week or strong queries to the search engines, the reason is that the users don't know about the information resources they are searching for [31].

According to [31] users encounter two main problems when they are using search engines these problems are word mismatching and submitting short queries. Query expansion methods are vital and mostly used for enhancing the performance of textual Information Retrieval systems [32]. These methods help IR systems to overcome the problems of word mismatch because of IR emphases on finding documents, whose contents match a user query from a huge document collection.

Fang w u *et al.* [33] describe that QE methods can be broadly classified into three main ones: interactive QE in this approach the user is presented to a list of terms suggested by the system after a user submits his/her query to the system. This method requires a user to have knowledge of the information he/she wants. Furthermore, a good result is displayed to users when using this approach. The second approach is a semantic dictionary; this method uses semantic dictionaries such as the WorldNet. This approach uses related words from the WorldNet, such words might give a useless result. The third approach of QE is based on document set which is further classified into automatic global analysis, automatic local analysis, and user relevance feedback. However, as stated in [34] query expansion techniques can be classified as:

Query Expansion Using Wikipedia: The author in [34] states that Wikipedia is the largest encyclopedia that is obtainable on the web without payment. The author claims that the contents of Wikipedia are well organized and correct and many studies advise the use of Wikipedia as a source for query expansion. In Wikipedia based query expansion, the base query is run against a Wikipedia collection of documents and each category is assigned a weight proportional to the number of top-ranked documents assigned to it.

Query Logs Based Expansion: With the growth in usage of web search engines, it is easy to gather and use user's query logs. The query logs are conserved by each search engine in order to analyze the behavior of the user while interacting with the search engine. In [34] search engines analyze the user's preferences and add corresponding terms to the query. This method can fail when the user tries to search something that is not related to what searched so far by the user, in this case, the system does not get suggestion words for the user.

2.6.1 External Resource-Based Query Expansion

In this approach, the user's query is expanded with the help of some external resource, such as WorldNet (WorldNet based query expansion), lexical dictionaries or thesaurus (thesaurus-based query expansion). These dictionaries are built manually and they contain mappings of the terms to their relevant terms. For that reason, they concern about looking up the terms in the resources and adding them to the query of the user. These techniques work on lexically and semantically arranged databases. Detailed narrations of Thesaurus and WordNet from the literature is given in the next lines of text.

WordNet

According to [35] WordNet is the outcome of group of synsets that contain words which have the same meaning, its Words are grouped and interconnected. WordNet is a lexical database which is helpful for the knowledge base in natural language processing and natural language [36]. Unlike thesaurus, WordNet groups its words semantically, it's like a network of words associated with their senses [36]. Words in WordNet mostly contain four part of speech categories [35]. Some of the part of speech categories in which it contains are: adverbs, adjectives, nouns and verbs. In addition to the arrangement of words based on their semantical relationships WordNet adds definitions and possible examples to the synonym words [36, 34]. For that reason, WordNet is combination of thesaurus and traditional dictionary.

Thesaurus

Thesaurus is a semantically similar group of words which contain term to term mappings of alphabetically ordered vocabularies [35]. It orders synonym words in an alphabetical order; thesaurus is different from dictionary in that dictionary gives definitions of the words while thesaurus doesn't. A term in a thesaurus can be either a word or phrase. As stated in [37] Thesaurus generally defines three types of term relationships. These types of term relationships are defined as broader term (BT), narrower term (NT) or hierarchical relationships, related terms (RT) or associative relationship and synonyms (SYN) or equivalence relationships. Increase in digital media [37] makes thesaurus a valuable knowledge for Information Retrieval as it solves the gap between the indexer and the

searcher. Constructing thesaurus [38] might be mainly of two types which are either manual or automatic:

Automatic thesaurus: - To generate thesaurus using this type of thesaurus construction researchers always use statistical methods. The significant of words is assessed and selected based on statistical techniques. In this method of thesaurus construction there is no need for language expert instead everything is done automatically.

Manual thesaurus: - In manual thesaurus construction, one should first identify the subject area, in which the thesaurus is to be constructed. This technique needs experts in the subject area, as it is to verify the validity of the relationships of the terms collected. To construct manual thesaurus different sources are used such as: books, websites, dictionaries and existing thesauri.

2.6.2 Relevance Feedback Based Expansion

In this method the initial query is executed on the collection of documents and the top k documents are retrieved from the documents and ranked [34]. Then to enhance the retrieval performance the ranked documents are used, in relevance feedback-based expansions. The top retrieved documents are considered as relevant and the most occurring words of these documents are used to be the expansion terms. Relevance feedback-based expansions fail when the algorithm of the initial search engines is poor, some of the relevance feedback-based expansion models are:

Explicit feedback from a user: In [34] it is argued that this is an interactive technique in which the early retrieved documents are offered to the user and the user is requested to select the relevant documents. This model is not user-friendly as it bothers the user by asking to modify his/her query repeatedly instead the system does independently the task for the user. Developers use this kind of model when testing the search engines for the purpose of interacting with the system.

Implicit Feedback: In this model, the user's feedback is deduced by the system. The feedback can be concluded from the user's behavior such as the pages which user opens for reading or pages in which user visits when the results are displayed back to the user.

Pseudo Relevance Feedback (PRF): In the pseudo relevance feedback-based models, the initial query is sent and the top k results are gotten. Then the important terms usually based on co-occurrence from the retrieved documents are extracted and added to the original query of the user. The expanded query is re-sent to retrieve the final collection of documents that are then exposed to the user. As presented in [34] the pseudo relevance feedback is good for search engines built for low-resource languages.

2.7 Linguistic Properties of Somali Language

2.7.1 Somali Language and its Dialects

Somali language had its first writing system in the year 1972 [39]. In this year the Somali Latin alphabet has been formally introduced. The writing system and alphabets used by the language was introduced by Osman Yusuf (pronounced in Somali as Cusman Yusuf). This man is credited with the invention of the spelling system, for the invention of the language's writing system. There were an estimated eighteen numbers of nominees, for new orthographies, from the eighteen competing writers' shire Osman Yusuf's script has been chosen.

To this end the main purpose behind the invention of the Somali Latin alphabet was to encourage literateness [39]. As Somalis at that time used Arabic language to write, while speaking Somali [39]. Furthermore, Somali is easier for strangers compared to Amharic, to write because of its Latin alphabets. In [39] Somali language is written from left-to-right similar to English and other languages that use the Latin scripts.

According to Sunita Shah [40] Somali uses all but three letters (p, v, and z) of the English alphabets. Fifteen sounds of Somali consonants such as (b,d,f,g,h,j,k,l,m,n,s,sh,t,w, and y) are very much like their English counterparts [40]. In addition to this, the names of the letters are based on Arabic letter names. According to [40] Somali language has six consonants (c, dh, kh, q, r, x) that don't match any one of the English alphabets. People who don't speaker Somali feel difficulties in representing Somali alphabets that are represented by the letters c, q, r, and x, since these letters are pronounced quite differently in Somali [40].

The authors in [40, 41] state that in Somali, the consonants r, d, g, l, m, n, and b can be doubled to indicate a sound which is pronounced with much more force than its single equivalent. Thus, because Somali language has some doubled words, Somalis often pronounce the doubled consonants in English words such as "bigger," "middle," "merry," "simmer," and "nibble" with more strength than they would be pronounced by a native speaker of English. In addition, shah [40] state that Somali language has 21 consonants and five vowels.

According to Tosco [42] Somali language uses Latin scripts called **Osmania** which has been named after its inventor Osman Yusuf Kenadid. The language has its own grammatical structure called **naxwe (nahwe).** As stated in [43] Somali language uses different punctuations such as, **Joogsi (.)** (full stop), **Hamse (‘)** (quotes), **Hakad (,)**(comma), **Calaamad su.aal** (?) (question mark) and **calaamatu layaab**(!) or exclamation mark**.**

In Lisa Peters *et al*. [44, 45] the dialects in Somali language are three major ones: Northern (the most common and the basis for standard Somali including Somali Region spoken dialect), Benadir (Indian coast) and May (southern Somalia). As argued by [44, 45] the Somali language spoken by Somalis in the Somali region is the most esteemed in part because it’s the dialect mostly used by Somali poets who are highly respected. Tables one and two show the consonants of Somali language and vowels respectively.

Table 2.1 Somali consonants

Capital Consonants of Somali																			
B	T	J	X	KH	D	R	S	SH	DH	C	G	F	Q	K	L	M	N	W	H	Y
Small Consonants of Somali																				
b	t	j	x	Kh	D	r	s	sh	dh	c	g	F	q	k	l	m	n	w	H	y

Table 2.2 vowels in Somali

Capital Somali vowels				
A	E	I	O	U
Small Somali vowels				
A	E	i	O	U
Long vowels				
AA,aa	EE,ee	II,ii	OO,oo	UU,uu

2.7.2 Grammatical Features of Somali Language

Somali language has its own grammatical features, usually called **naxwe** (**nahwe**) [45, 46]. It marks the grammatical structure of person, gender and number, which differentiates one language from the other [45]. The grammatical features of Somali language are described as follows:

Nouns

As stated in [47] Somali language is morphologically an agglutinative language. This means grammatical formation of the language is determined morphologically by the way of attaching affixes (prefixes or suffixes) to word roots and stems. Diana Briton Putman and Mohamood Cabdi Noor [46] state that Somali nouns are highly inflected than are nouns in English. Somali is agglutinative language so that the nouns in English are inflected only for number, for that case they have different forms for singular and plural. In Somali in addition to the inflection in number, the nouns are also inflected based on gender (masculine or feminine) and case (nominative, genitive, and vocative). Somali language is different than English language because the differences in gender and number are marked by grammatical tone Table 2.3 shows:

Table 2.3 Somali nouns

Somali noun	English noun	Case
Inan	Boy	Gender
Inan	Girl	Inflected by tone gender
Dameer	Donkey/jack	Gender
Dameer	Jenny	Inflected by tone
Dibi	Ox	Number
Dibi	Oxen	Infected by number
Muuse	Moses	Vocative
Muuse	Hey muuse	Inflected by tone Vocative
Wiilka	The boy	Nominative
Wiilasha	The boys	Inflected

Nouns in Table 2.3, are differentiated based on their grammatical tone or pronunciation. Example of such in the Table 2.3 is inan and inan.

Adjectives

As the authors in [46, 48] state adjectives in the Somali language are formed by adding **aa** or **san** to a verb or noun. Thus, the word **xun** (Bad) becomes xumaa (badness) in the adjective, and the word **walwal (worry)** becomes **walwalsan (worried)**. Adjectives in the Somali language often occur with a short form of the verb to be suffixed to them, an example of such will be, **yar** (**small**) becomes **yaraa (he was small)**. [46] states that Somali speakers of English tend to add **aa** to English adjectives. Thus, this causes confusion between Somalis and British English speakers as Somalis say something similar to "small-ah" instead of saying "smaller" which makes British speakers understood it as "smaller". Table 2.4 shows common adjectives in Somali:

Table 2.4 Somali Adjectives

Somali Adjectives	English Adjectives
Quruxsan	Beautiful
Wacan	Better
Gaaban	Shortness
Yaraa	Smaller
Dheeraa	Toller
Naxsan	Fearful
Walwalsan	Worried
Waalan	Mad
Yaxyaxsan	Ashamed/embarrassed

Prepositions

In Somali there are four mainly used prepositions, those prepositions are used to mean differently. Somali prepositions come before the verb instead of the noun [40, 42]. Table 2.5 shows the prepositions in Somali language and their different ways of usage:

Table 2.5 Somali Prepositions

Somali prepositions	English prepositions
Ka	from, away from, out of and about
Ku	in, into, on, at and with
U	to, towards, for and on behalf of
La	with, together with, in the company of

Verbs

Somali language has verbs called **ficil** and these verbs in the language come at the end of the sentences [40]. As a result, speakers of the language make mistakes when speaking English by putting the verb at the end of the sentences like in the Somali language.

According to [40] Somali does not have a passive voice and instead of the passive voice Somalis use the indefinite pronoun **la,** as **goormaa la dhisay** (when has it been built), **miyaa la keenay**(has it been brought). Some common Somali verbs are shown in table 2.6:

Table 2.6 Somali Verbs

Somali Verb	English Verb	Somali Verb	English Verb
Dilay	Killed	Kacay	stood
Galay	Entered	Baraa	Teaches
Keenay	Brought	Cunay	eat
Seexday	Slept	Sameyaa	Makes
Dhisay	Built	Xidhay	Dressed
Dumay	Collapsed	Qor	Write
Qoray	Wrote	Akhriyey	Read

Conjunctions

According to J. W. C. KIRK [48] Somali language has conjunctions called **xidhiidhiye,** similar to one in English language. The most commonly used **xidhiidhiye** in Somali language is shown in Table 2.7:

Table 2.7 Somali Conjunctions

Somali conjunction	English conjunction
Ama	Or
Laakiin	But
Iyo	And
Sida daraadeed	So

Pronouns

Ettien Koffi *et al*. [49] Argue that Somali pronouns can either occur in their full forms as in column 2 of Table 2.8 or in their clitic forms as in column 3 of Table 2.8. The forms in column 3 are called clitics because these pronouns have to attach themselves to other lexical items to be fully realized. Somali pronouns don't need to occur in the same sentence, but both can and do occur, the pronoun "Anigaa" may precede the verb, however, it can also be omitted. Somali has a morpheme "waa" that occurs together with pronouns.

Table 2.8 Somali pronouns

Persons and Numbers	Full Forms2	Clitic Forms	English Equivalent
1st Person singular	Aniga	Aan	I
2nd Person singular	Adiga	Aad	You
3rd Person singular masculine	Isaga	Uu	He
3rd Person singular feminine	Iyada	Ay	She
3rd Impersonal		La	One
1st Person plural inclusive	Annaga	Aynu	We
1st Person plural exclusive		aannu(aan)	Not Applicable
2nd Person plural	Idinka	Aydin	You(plural)
3rd Person plural	Iyaga	Ay	They

Table 2.9 Somali Adverbs and Interrogative Pronouns

Somali	English	Somali	English	Somali	English
Goorma	When	Qunyar	slowly	Barbar	Alongside
Maxay	what	Gortas	then	Maalinwalba	Everyday
Abab	Why	Kolkol	sometimes	Xagan	This way
Sidee	How	Dhakso	Quickly	dabadeed	Afterwards
Kuma	Who	Gudaha	Inside	labbagoor	Twice

2.7.3 Somali Sentence Construction

Grammatically the order of a Somali sentence is Subject, Object, Verb [48]. What makes it different than other languages is its agglutinative behavior. It is stated that adverbial clauses, especially expressions of Time, come first. In Somali, if the subject of an adverbial or conditional clause is the subject of the principal sentence, it is placed first [48]. The verb always requires the simplest form of the personal pronoun to immediately precede it, whether the true subject is expressed or not [48]. Complicated sentences are entirely avoided [48]. Some Somali sentences are as follows:

- **Weedh Fudud** (**Simple Sentence**): these sentences are usually built in a subject (faacul) and a verb (ficil). Examples of such sentence are: **xagee tagaysaa**(where are you going), **ma maqashaa** (do you hear), **wax buu baranayaa** (He is learning), **ma ilawday** (did you forget). **Iyadu way karisay** (she has cooked). **Gaadhi buu kaxeeyey** (He drove a car).
    - **Weedh Dhafan** (**Compound Sentence**): a sentence consists of at least two independent clauses joined by coordinating conjunctions such as **u,ka, ku, la, da, kii, kaa**. Examples of such sentences are: **ninka hal kaa fadhiya ayaan uyeedhi** (I will call the man sitting over there), **maxaad u maqli wayday wixii la guyidhi** (why don't you listen to what you are told). **Waxaa rabaa inaad ka jawaabto su aashayda** (I need you to answer my question).

Chapter Three: Related Work

3.1 Introduction

This chapter summarizes a brief investigation of Information Retrieval systems of both international and national (Ethiopian) languages. The first four sections present international papers on Information Retrieval. The second section of this chapter presents with Information Retrieval systems on local languages such as Amharic, Afaan Oromo, and Tigrinya. The third section deals with cross-lingual. The last section of the chapter briefly summarizes the reviewed papers together.

3.2 English Language Information Retrieval Systems

Sanjeev K Sunny and Mallikarjun Angadi [50] have carried a study on Potential Roles and Applications of Thesauri in Digital Information Retrieval Systems. The authors claim that the objective they have conducted this study was to show the potential roles and applications in which the thesauri has on digital IR systems. The data set which used as testing purposes was 77 relevant publications. From the 77 relevant primary data set 75 relevant documents have been identified. Thesauri has an impact on information retrieval systems. The authors present that it would be better to analysis the computing methods of automatic key-phrase. The usage of thesauri in this study has improved the consistency of metadata and query reformulation.

Rong Yan and Guanglai Gao [51] conducted Pseudo-Based Relevance Analysis for Information Retrieval. The authors have proposed a unique Pseudo Relevance Feedback (PRF), by diversifying the feedback documents. They have developed an abstract pseudo document to represent the value of each document feedback, to cover the diverse aspects of feedback. The challenge met by the authors is that PRF lies in how to get the reliability relevant content for the user query. In this paper, the authors have used the Lemur Toolkit for both indexing and retrieval. They have conducted their data sets on two standard languages such as OHSUMED, AP (the Associated Press) and Chinese data set including XINHUA (Simplified Chinese). The authors have taken three numbers for the data sets 60, 70, and 100 for OHSUMED, XINHUA, and AP respectively.

WordNet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain [52] is a study conducted by M. Uma Devi and G. Meera Gandhi. These two authors have an accomplished Information Retrieval system using WordNet and anthology. The authors have selected specific domain specially, sports domain. The authors claim that they have expanded their queries to find and add more meaningful terms to the original query. Stanford parser has been used as a tool by the authors to parse the ontological data and data extracted from the web. The authors have persisted that synonyms may cause errors by not finding the relevant documents for this reason ontology is needed. An experiment has been conducted by the authors and they claim that the approach shown better results. Accuracy of the system is found to be about 87.1% whereas the increase in precision is about 40% and 20 to 30% increase in recall.

Bhavadharani M *et al.* [53] proposed performance analysis of ranking models in information retrieval. The authors have discussed about various IR models for ranking documents. They have analyzed the accuracy of each algorithm in the category selected, and they took the most efficient one. After the analysis, they say that the performance of every algorithm in the study can be improved. They model analyzed models such as vector space model and query expansion techniques such as pseudo relevance feedback. The data set they have used is CACM which is open source and freely available in the Internet. The data set contained about 22000 text of documents with journal abstracts. The authors claim that there where challenges in the study such as word mismatching, short query from the users and the performance of the search engines. Compared to the other techniques used in the study the pseudo relevance feedback technique became more accurate and performed well.

Rui Zhang, Caitlin Westerfield, *et al.* [54] conducted research on Improving low-resource Cross-lingual Document Retrieval by Re-ranking with Deep Bilingual Representations. According to the paper presented by the researchers, the authors suggest an approach for low-resources cross-lingual document retrieval performance by using deep bilingual query-document representation. The researchers have used query likelihood retrieval and tried to include the score as extra feature. They said that their approach after evaluation outperforms than other translation-based query likelihood retrieval and monolingual deep relevance ranking approach. For experimental purpose they have used the indri system which uses the query likelihood with dirichlet smoothing.

3.3 German Language Information Retrieval System

Stemming and Decomposing for German Text Retrieval [55] by Martin Braschler, Brbel Ripplinger, *et al*. The authors of this paper claim that this study made major contributions that transcends its focus on German, by exploring complete spectrum of approaches ranging from language-dependent to detailed linguistic methods. The main findings were that stemming was important even when using a simple approach, the authors said that the splitting of the compounding words importantly boosted performance. These authors have based their findings on analysis using methods they said were reliable, they have exhibited that stemming was important for German text retrieval in any case. With their system they have compared with system that was without stemming and they got performance enhancement and improved their precision performance to 23% for short and up to 11% for long queries. As they got for recall 12% for short and 4% long queries which was better performed then the previous systems.

3.4 Arabic Language Information Retrieval Systems

Arabic Information Retrieval [56] is a study and probably survey conducted by Kareem Darwish and Walid Magdy. The authors of this survey described that their study covers the following seven points: general properties of the Arabic language, some of the aspects of Arabic that affect retrieval, Arabic processing necessary for effective Arabic retrieval, Arabic retrieval in public IR evaluations, specialized retrieval problems, namely Arabic-English CLIR, Arabic Document Image Retrieval, Arabic Social Search, Arabic Web Search, Question Answering, Image retrieval, and Arabic Speech Search; Arabic IR and NLP resources, and open IR problems that require further attention. The authors say that previous researchers carried out in the area of Information Retrieval in the Arabic language were not enough to solve problems of the language related to retrieval. Because of their study was a survey they have only studying the characteristics of the language instead of using evaluation methods and carrying experiments.

Ibrahim Moawad, Waseem Alromima and Rania Elgohary [57] developed Bi-Gram Term Collocations-based Query Expansion Approach for Improving Arabic Information Retrieval. In their paper they have proposed a language-independent semantic-based Information

Retrieval approach. The authors conducted query expansion approach to expand user's query using bi-gram term collocations. On the paper that these authors proposed describes that it contributes two contributions, first one is mining bi-gram collocations from text corpus, and the second is indexing way has been constructed to minimize the cost and effort of stemming process. For experimentation the researchers have used the holy Quran for testing, as the quran uses the Arabic scripts. The authors say that their new system outperforms then the stem-based method in terms of recall and precision.

3.5 Turkish Information Retrieval System

Kutlu Emre Yılmaz *et al.* [58] carried study on Turkish text retrieval experiments using lemur toolkit. The authors have used lemur toolkit and developed an automated indexing and retrieval experiments on TREC-like test collection for Turkish text. In the paper three models have been studied and compared later, the authors claimed that all retrieval models benefit from language specific preprocessing in terms of retrieval quality. 45,000 documents and 72 da-hoc queries and relevance judgments. The authors have found optimum parameters and have done language specific improvements for all three retrieval models and the tool used for experiments in this research were lemur toolkit. The authors said that result of 2000 without stemming and 1000 with stemming being found. The challenge in the study was that there were no Turkish test collections on Information Retrieval (IR).

3.6 Amharic Information Retrieval Systems

Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique [59] conducted by Abey Bruck and Tulu Tilahun. This study has been carried to improve the recall of previous work, by using statistical co-occurrence technique. The authors of this research have used query expansion in addition to the technique. They claim that the main attitude of using query expansion is to give user relevant documents those please their information need. Statistical o-occurrence considers frequently appearing terms in the query term regardless of their position. The authors claim that this technique outperformed then the previous ones by improving recall 6% and f-measure 2%.

Amharic text retrieval: an experiment using latent Semantic indexing (LSI) with singular value decomposition [60] is a study undertook by Tewodros Hailemeskel Gebermariam. In

this research the potential of latent semantic approach in Amharic text retrieval was explored. 206 Amharic documents and 25 queries were used by the author to test the technique. Automatic indexing of the documents resulted in 9256 unique terms which were not in the stop list used for the research. Finally, they have claimed that the performance of the LSI approach was compared with the standard vector space and they said that precision of the LSI was above then that of the standard vector space.

Amharic-English Information Retrieval [61] proposed by Atelach Alemu Argaw and Lars Asker. These authors have accomplished Amharic-English cross-lingual information retrieval experiments using adhoc bilingual. The authors said that their queries have been supported by morphological analysis and parts of speech tagger. They used different machine-readable dictionaries for translation purposes. The terms those were not found on their experiments were maintained by the fuzzy matching, as the authors claim they have used four different experiments including fuzzy matching and term weighting. To this end the authors concluded that words boosted gave 10 times worse result then the words those have not been boosted.

3.7 Afaan Oromo Information Retrieval Systems

Afaan Oromo Text Retrieval System [5] proposed and carried out by Gezehagn Gutema Eggi, was afaan Oromo text retrieval system. As mentioned by the author the main aim of this study was to implement and develop text retrieval in afaan Oromo with the use of modern approach of Information Retrieval. For retrieval purpose of the documents vector space model has been used. The researcher claims that this model is widely used one, the corpus used in this study consisted of different news articles and to test the system 9 queries were used. The challenge in this study was handling the synonymy and polysemy in the language, the author says that if the thesaurus is used that the problem of the synonymy and polysemy will be solved. The experiment shown average result in both recall and precision where the recall was 62.64% and precision was 57.5%.

Applications of Information Retrieval for Afaan Oromo text based on Semantic based Indexing [62] Berhanu Anbase have conducted an afaan Oromo text retrieval system applications based on semantic indexing. The data set which the author has used was about 70 documents of afaan Oromo text from different news channels for texting purpose. The

total number of queries used by the author o test the system were 9 in number. The author claims that it is challenging to develop an afaan Oromo text retrieval system without a standard afaan Oromo corpus. As challenges. The author insists that with the use of the relevance feedback the performance of the system being measured in terms of recall and precision. The score of the prototype was an average in this case with a rate of 67% precision and 63% recall. Afaan Oromo text retrieval system work well with a good stemmer and standard corpus.

3.8 Tigrinya Information Retrieval Systems

Indexing Tigrinya language documents [9] is a study conducted by Omer Osman Ibrahim and Yoshiki Mikami, these two authors have reported and developed an indexing system for the Tigrinya language. During the search in the search engines for the language, they claim that indexing facilitates information retrieval. An original and unique analyzer has been come up with, the analyzer developed makes efficient to search Tigrinya language. This study consisted of two main components: analyzer and indexer. The indexer in this study first tokenizes the words and then removes the stop words after this the text is normalized and then stemmed for processing. For implementation, Lucene library has been used by the authors and in the study 1176 stops have been used which were constructed from their corpus.

Teklay birhane and birhanu hailu [63] conducted a study on Design and Implementation of IR System for Tigrigna Textual Documents. These authors have presented that thy have used vector space model for the implementation of this research. Their study shows that there wasn't any Tigrinya document corpus prepared before them, like the one in English called TREC. For this reason, the authors have collected documents themselves. The documents were collected from different sites which Tigrinya language is published as well as the text books for grade 10,11 and 12. The total documents collected and used as testing purpose was a bout total of 30 documents.

The queries they have used were about 6 queries which makes their system less queries. The performance of their system was good and showed a recall of 84% and precision of 70%. The challenges in the system were like corpus preparation and luck of stemmer.

3.9 Cross-Lingual Information Retrieval Systems

Elizabeth Boschee *et al.* Proposed a simulation system in [64] which they named SARAL: A Low-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage. The authors here included Somali language as one of the low resourced languages. In this paper, monolingual English speakers effectively access what is written in another language. The authors have demonstrated an end-to-end CLIR, and summarization system that combines searcher and traditional IR techniques and then applies them to text and speech for the documents of low-resourced languages. They have used two approaches term-level matching and shared embedding architecture for effective retrieval. In this system, serious challenges can arise when the content of information is in a language that the searcher can't understand. The authors said that their system summarized results of 87% true positive and 45% false negative.

Simulating CLIR Translation Resource Scarcity using High-resource Languages [65] proposed by Hamed Bonab *et al*. Is a contrastive study framework that uses high-resource languages to simulate low-resource Languages. In their framework, the authors have focused on corpora, on parallel translation that they aimed for the factors that impact CLIR Performance to be understood better. In this paper, the authors have sampled a high-resource language to become an artificial low-resource language. They have formulated that the problem in this paper was an NP-hard problem. For this reason, they have proposed two greedy algorithms with polynomial complexities, with the use of four high-resourced languages (Italian, French, Finnish, and German). And then they compared their technique with other alternate methods, to simulate two low-resource languages (Somali and Swahili). Their result in the experiment suggested that language families are important for the problem. they simulate Swahili with Finnish and Somali with German, achieving 97% and 98% on the similarity percentage in terms of CLIR performance, respectively.

Abdillahi Nimaan *et al.* [66] proposed an approach to access African oral corpora by combining automatic speech recognition and information retrieval. The authors in this paper presented a hybrid language model including words and sub-words. They lastly presented a new strategy to combine sub-words and words to enhance information retrieval results. Because there has not been textual data matching the speech corpora, the task became very

challenging for speech recognition, the data set used was 20 minutes of radio broadcasting and 20k lexical words. The result obtained was 46.4%.

3.10 Summary

As a summary of the review of the related work, in which the researcher has explored. Several studies have been undertaken in different languages using different techniques. From the review IR studies have been conducted in English [50, 51, 52, 53], German language [55], Arabic [56, 57], Turkish [58], Amharic [59, 60, 61], Afaan Oromo [5, 62], Tigrinya language [9, 63] etc. However, Information Retrieval is language-dependent. Thus, algorithm developed for one language may not work for another language. The reason is that languages have different characteristics and sentence structures in which one language is different from another. Besides IR being language-dependent, the increase of the documents on the internet of various languages makes important for the development of Information Retrieval in different languages. Searching a document manually is very tedious and time taking.

According to the existing literature, there is only two attempts to conduct Somali language IR. These studies are Somali-English and Somali-Swahili query translation simulation systems and papers. In these two studies, parallel corpora and term matching approaches has been used respectively. Translation mechanisms have been used by translating query written in one language into another language. Such translation systems may not work and produce good results with the language properties. As language has its own characteristics, which need to be addressed separately on its own. Furthermore, a query translated using parallel corpora cannot be an effective for Somali speakers' information need. For all these reasons, cross-language systems can't answer the need for Information Retrieval in Somali language. Thus, in this research, we aim to design an Information Retrieval using thesaurus for Somali language. It is obvious that keyword-based IR systems don't give high performance, so in this study we use thesaurus-based query expansion. According to the review thesauri are semantically related vocabularies which help users find the synonym of query terms. This technique finds synonyms from arranged thesauri and uses them as expansion terms.

Chapter Four: Design of Somali Language Information Retrieval Using Thesaurus as Query Expansion

4.1 Introduction

In this study an attempt is made to conduct Somali Language Information Retrieval system using controlled vocabulary (thesauri) as a query expansion technique. In this chapter the architecture of IR system for Somali language is described. The first part of the chapter gives details of the proposed architecture of the system. The second part explains the components of the architecture, algorithms and techniques used to attain the required standard of the study.

4.2 Architecture of Somali IR System

The main objective of our proposed system is to design the architecture of an IR system for Somali language. Using manually constructed thesaurus for helping users formulate more meaningful queries and find increased relevant documents. The proposed architecture shows the main functionality of the system. Our proposed architecture of the system consists of three main modules. These three modules are responsible for the operations which the system does in terms of functionality. The architecture shows that the system reads collections of documents from Somali text corpus. The documents follow steps carried out by the document preprocessing module. The steps which the document follows include: stop word removal, normalization, stemming and indexing. The initial query accepted by the system also follows the same steps followed by the document. As depicted in Figure 4.1 in addition to the document preprocessing the architecture shows also that the similarity between documents is measured through the matching component. The three key modules, of Somali Language Information Retrieval system as described in the architecture are: query processing module, result processing module and docs preprocessing module. Details of what each module in the architecture does and components of every module is discussed in detail.

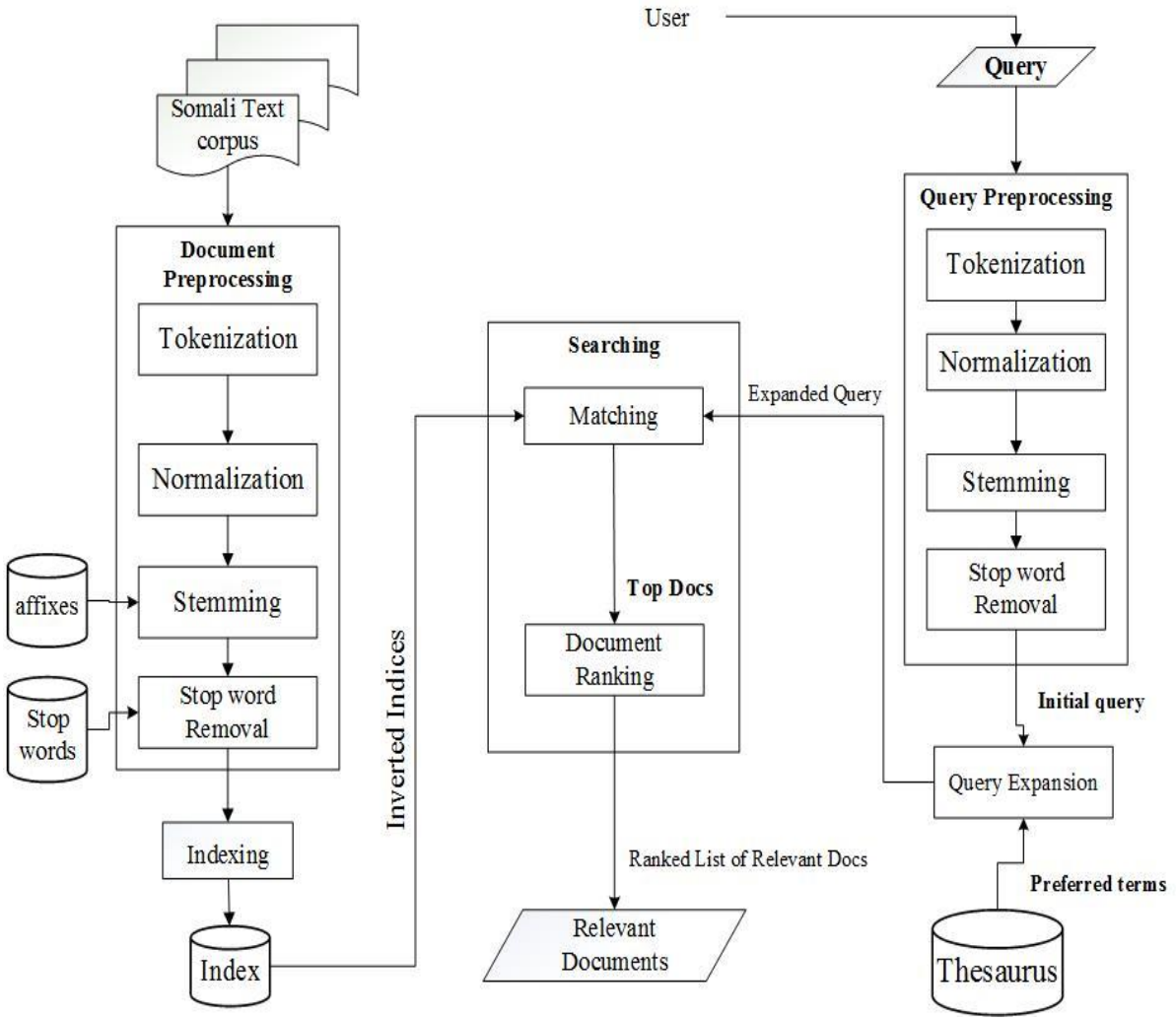

Figure 4.1 Proposed Architecture of Information Retrieval System for Somali language using thesaurus

In this architecture, we show that the stemming part is unique for Somali language. For example, stemming Somali language is different than other Ethiopian languages, as it is morphologically agglutinative language. Two morphemes make a meaning when combined. So a morpheme can be added to a root word to make a new or existing meaning. In this architecture we have collected sample affixes, or morphemes to stem and put them in the affixes component in the architecture. We have also added a component for stop words, which holds identified most occurring Somali words which don't give a meaningful idea.

4.3 Document Preprocessing

Document preprocessing module in the proposed system is initiated by the Somali text corpus component. This component consists of collection of Somali language documents from the different domains used for testing. Document preprocessing is a significant module of the design of the system. It consists of sub-components such as stemming, tokenizing, stop word removal and normalization. The module reads documents from Somali text corpus which holds and contains collections of documents. After it reads documents from the collection, then it tokenizes the lines of text read from the corpus. The next task of the module is normalization which removes special characters from the documents. In the third phase the module removes the stop words from the documents read, after removing the stop words the module stems Somali words into their root form. The stems are then submitted to the indexing module, to make the searching process faster.

4.3.1 Tokenization

During preprocess of the texts, this component makes Somali language texts as tokens. It takes sequence of sentences and breaks them into tokens. During the tasks of this component we have also removed all special characters. Because of the reason that compilers understand texts or any other strings in the form of tokens. We have used this component to break the texts into tokens, so that compilers can understand them. Tokenization depends on white spaces as it tokenizes the texts based on the white spaces which are between them. Algorithm 4.1 is used to tokenize the texts read from Somali text corpus:

```
Input: Text corpus
Output: Tokenized Text
Set corp to corpus directory
For
Files in corp
    Read the content of each file
    split the lines of the file after white space
    Store the result in a container
    if container
        return container
end for
end if
```

Algorithm 4.1: Tokenization Algorithm

4.3.2 Normalization

As stated in Section 2.7.1 of this study Somali language has some consonants that are doubled. The doubled consonants bring a mismatch of some words, as some people might write them not doubled while others write them doubled. This component normalizes these letters by fixing the differences between them. It changes doubled letters into the mostly used format of the letters which is single format. The doubled Somali consonants are summarized in the statement "ma nala garaadbaa" which means do they have the same thinking with us in English. Examples of the doubled consonants are such as m, n, g, l, d, r, b. the doubled form of these consonants are written as: "ammaah",

"bedell", "dawladda", "daggan". We also normalize the upper case and lower-case characters and also some words which can have different forms of writing for different people. Upper case and lower-case letters might cause an issue into a Somali language writing system. Although it might not bring an effect into some Ethiopian languages such as Amharic. Algorithm 4.2is used to implement the normalization process.

Input: tokenized Text
Output: Normalized set of documents

```
accept tokens from tokenizer
for word in tokens
   if word contains "ll", "gg", "mm", "dd"
      set ll="l",gg="g",mm="m",rr="r",bb="b",nn="n"
      and dd="d"
      return word
   else
      return no normalization needed
end for
end if
end else
```

Algorithm 4.2: Normalization Algorithm

Table 4.1 shows sample normalized words encountered during the experimentation.

Table 4.1: List of Normalized Words

List of words	After Normalization
dawladda	dawlada
daggan	dagan
dallalka	dalalka
Eebbe	Eebe
Ammaah	Amaah
dharraar	dharaar
bannaan	banaan
Hadda	Hada
yay	yey
lay	ley

4.3.3 Stop Word Removal

This component removes the stop words from the documents as well as the query automatically. It reads stop words from a file and checks them against the query and in addition the documents. If the query or the document contains stop word, then the stop word

is removed. This section allows the disk to be free and fast as it frees some space by not counting certain words. Based on the mechanisms used to identify stop words, which will be discussed later in Chapter 5. The identified stop words in Somali are mostly categorized under prepositions, articles and particles. Some of the identified stop words in Somali language are: waa, baa, ayuu, ayay, kii, kaa, kale, uu, u, sidii, iyada, isaga, la, etc. the following simple Algorithm is developed to remove stop words in Somali language.

```
Input: set of normalized Text
Output: stop word free documents

Set s to stopword.txt
Set Q to query terms
Set D to documents terms
if Q, D contain in S
    Remove S
else
    return S
end if
end else
```

Algorithm 4.3: Stop-word Removal

4.3.4 Stemming

This component stems the words, it converts inflected and derived words into their original stems. It allows reduce the word mismatch, which is created because of the derivation of words. Somali language is an agglutinative language which means that the grammatical formation of its words is determined by the prefixing and suffixing of its affixes [33]. However, the language is morphologically rich yet there are not natural language processing tools such as stemmer available. For that reason, we have implemented simple stemmer to remove and transform inflected words into their stem words. This minimizes mismatch of Somali language texts during Retrieval process. The stemming component finds the stems of words by removing additional affixes appended to the words. In addition, the stemming component is used to identify the affixes by checking if the words end certain predefined words or not. After identification this component processes the affixes by checking them against the user query and documents read from the corpus. Some of the possible affixes in Somali are: “san”, “ka”, “ku”, “kii”, “ay”, “da”, “niin”, “taa”, “ah”, “ga”, “ma”, “yaa”, “aa” etc. The sentence “Itoobiyada Cusubaa Fiican” can be stemmed as “itoobiya cusub fiican”.

Algorithm 4.4 is a simple algorithm used to remove the encountered affixes during processing:

```
Input: set of documents
Output: stemmed documents
_____________________________________

Set affix to defined affixes
DO
  for affix in affixes
      if word ends or starts with affix
          remove the affix
          return word
      else
        search new terms
end for
end if
end else
```

Algorithm 4.4: Stemmer

stemming Somali language is different than, other Ethiopian languages and as well English language. Because the language is agglutinative language. This means that grammatical formation of words is determined by attaching morphemes into another root words. In this case combinations of words make a meaning. We have selected stemmer for its simplicity than morphology. In the stemming process we have collected set of morphemes that can make a meaning when added to other root words.

4.4 Indexing

This section creates inverted indices or inverted index and then stores it in the index. During indexing there are, three types of file structures which can be used, such as sequential, signature and inverted file structures. From these we have selected, inverted file structure. This file structure divides the documents into two parts. These are record-level and word-level inverted indices. As stated in section 2.2, the record-level holds list of references to a document. While the word-level deals with posting of the words in addition to the references. Understanding that every one of the file structures has its advantages, yet we have selected this type of indexing, based on its fast speed and less memory consumption.

Our aim of using indexing is to make our search fast and less time-consuming during search. In this module the whole document read from the corpus is converted into inverted tokens. After it converts documents into tokens, it stores the tokens with some additional information such as: document name, path of the document, term frequency, and postings of the terms. Number of times in which every term occurs in the document is counted. Its positions in the document is calculated. Doing all these tasks in the indexing process, the component does them with help of Lucene library which is very powerful and fast library. The process of indexing is used as an offline after it is created.

4.5 Query Preprocessing

The query preprocessing module accepts query from the user and then processes it. This module removes special characters from the query such as full stops, single quotes, double quotes etc. After removing special characters this component also tokenizes the strings accepted from the user. Same as the document preprocessing, this module comprises components such as: stop word removal, tokenization, normalization and stems. This means that every task of document preprocessing is carried out here. Somali language uses full stops called **joogsi,** it also uses commas and colon called **hakad** and **wardhawr** respectively [29]. Example of Somali sentences in which the special characters used are sentences such as **la.aan**. Which means nothing and the sentence **ra'isul wasaare** which means prime minister in English.

4.6 Query Expansion and Thesaurus

In this section, we have focused on adding additional terms to the original user's query from manually constructed thesauri containing Somali language synonym words. Because users can formulate short queries which doesn't fit to their information need using the system. This section implements Automatic query expansion which adds extra meaningful words to the query of the user without user interaction. The expansion terms in the system are based on synonyms or near-synonyms which are closely related to the query terms. This doesn't make the system complicated as it displays only the top most relevant documents. Words which have similar meaning to the query terms are added as an expansion terms to further information need of the user. The expanded query is fired to the system after the initially

accepted query of the user is refined and appended more terms. The constructed thesaurus holds alphabetically ordered group of semantically related words. The words in thesaurus are arranged based on their similarity in meaning. Query terms are matched against set of predefined terms in the thesaurus. The words in the query are stemmed, for this reason, the thesaurus holds also predefined set of stemmed words. This is done in order to avoid word mismatch, between the query terms and the thesaurus words. Algorithm 4.5 is a simple algorithm to expand the query and extract synonyms of query terms from the thesauri.

Input: initial query terms
Output: set of Expanded query terms

```
load the thesaurus
if a word in the thesaurus is similar to query terms
     Add the word to the query terms
else if
      match not found
Break
else
   continue
      searching new terms
end if
end else if
end else
```

Algorithm 4.5: Query Expansion

4.7 Searching

In order to match and retrieve, information needed from unstructured data in the index file searching module is important. This section does different tasks; these can be categorized as searching a matching document from the index file. It searches and then matches documents using the term frequency (tf), and inverse document frequency (idf). This module ranks documents based on the cosine similarity of the documents. It brings the document with the highest cosine similarity value at the top. The module accepts tokens from the query expansion component. The accepted tokens are then compared with the tokens in the index file. Documents are then selected based on the matching tokens in the index file. This selection of documents is based on the similarity of documents to the tokens received from the query expansion component. In addition to the matching the module has a component

which ranks documents based on cosine similarity. The ranked documents are then finally displayed to the end user of the system. Components in the result process module are presented as:

4.7.1 Matching

From the architecture shown in Figure 4.1 the matching component is where the searching function takes place. It is the component in the searching module that accepts the expanded query of the user and sends it to the index module of the system. It retrieves set of relevant documents from the corpus through the index module. Those retrieved documents from the corpus are transferred to the next component, which is normally the document ranking component. The matching component is a key component for that it is the component that matches the user's query and inverted indices in the index file. This simple algorithm is used to match tokens from the user and those in the index file:

Input: Expanded query terms
Output: Ranked set of documents

```
Qet initial query from the user
DO
       perform document preprocessing
       removing any special character
       accepting only alphanumeric and numeric.
       tokenize the sentence by breaking them into terms.
       send the tokens into the index module to check whether
                        there is a   match or not
  if there is a match
      return the match
  else
     Return there is no match.
end if
end else
```

Algorithm 4.6: Matching Algorithm

4.7.2 Document Ranking

Documents are ranked based on their relevance to the query terms. The rank is calculated using the cosine formula, as the library used is based on cosine similarity which uses this approach. The document with the highest relevance to the query or the one which the query terms repeatedly happen are first brought to the user. In ranking documents, we have used cosine similarity of the documents. Which is one of the most used ranking function for

electronically published papers reviewed. It fetches the top k ranked documents from the matched in the searching process of documents from the corpus. The document ranking component is an intermediate component which is between the matching component and final result to the user. The final set of results is displayed to the user based on the information need of the user. Documents are listed in decreasing order with respect to their relevance of the refined query. This result is displayed to the user with respect to the expanded query. This component helps information searchers, in that it puts wrongly retrieved non-relevant documents at the bottom. Our aim of adding and analyzing this component is to fix problems related to document relevance.

Chapter five: Experiment and Implementation

5.1 Introduction

In this chapter we present tools and environments used to conduct the experiment. We also report the findings of the experiment based on the architecture drawn in Chapter four. We have created simple prototype to measure the system. With the prototype we have implemented the preprocessing tasks in NLP. The Chapter will contain two parts, the first part of the chapter will present the data set preparation, query formulation, tools used. The second part of the Chapter will present screen shots and results of the system.

5.2 Test Collection

Languages like English language have standard test collection corpuses such as ACM and the TREC. However, In Somali there is no a standard test collection corpus for Information Retrieval. For this reason, we have constructed our own test collection corpus to test the system. We have used three specific domains for document collection such as: politics, sport and education. The sites used as document collection include: BBC news Somali, VOA news Somali, caasimada.net, hiiraanonline, and educational textbooks of grades 7,8,9,10,11 and 12. The reason why we have used these specific sites and textbooks is because, as stated in section 2.7.1, some Somali words might be written in different forms. Because of the variation in accent between Somalis, while using the language. Thus, we use these sites and textbooks as they use the standard Somali writings. The collected document is 2007 in total in which 108 docs are relevant to the queries. Table 5.1 shows the collected documents:

Table 5.1 Test Collection

Domains	**Number of Documents**
Politics	1580
Sport	120
Educational	307
Total Documents	2007

Stop word Identification

To identify stop words, we have used a threshold formula, which the words are recognized as stop words if they are above certain criterion. According to Tomas Mikolov *et al.* [67] stop words give less information than rare words. For this reason, we have used a subsampling approach to collect frequent words from the corpus. In [67] the probability for which a word can be counted as a stop word is calculated using the formula:

$$p(w_i) = 1 - \sqrt{\frac{t}{f(W_i)}} \qquad (9)$$

Where $p(w_i)$ is the probability of a word, $f(W_i)$ is the frequency of the word and t is the threshold where t is set in our case to t=1, according to section 2.2 stop words are commonly appearing words in the document which don't have distinguishing power. Figure 5.1 shows sample of the identified stop words:

had	iyo
waa	waxaa
ah	oo
mar	wa
wax	aad
ugu	in
si	u
sidoo	kale
la	dib
sidaas	soo
ay	ayaa
mid	ahaa
sheeg	lagu
loo	hore
ama	ee
hor	lahaa
kala	duwan
waxaad	wuxuu
uu	ahayd
ma	tah
wixii	aan
iya	loogu
inuu	yah
ha	reer
hadii	nin

Figure 5.1: sample common stop words

Stemming

Somali language didn't have any stemmer so far, thus, we have created simple stemming pseudo-code and implemented the pseudo-code using small java program. In any case stemming words might bring some wrongs, by stemming words into the wrong way. To reduce stemming words wrongly, we have collected list of stems from the domains used. For that reason, when a word is to be stemmed. It should first be checked from the list of stems collected. If the word is in the list, then that word may not be stemmed, if it is not then that word will be stemmed. Figure 5.2 shows stems collected:

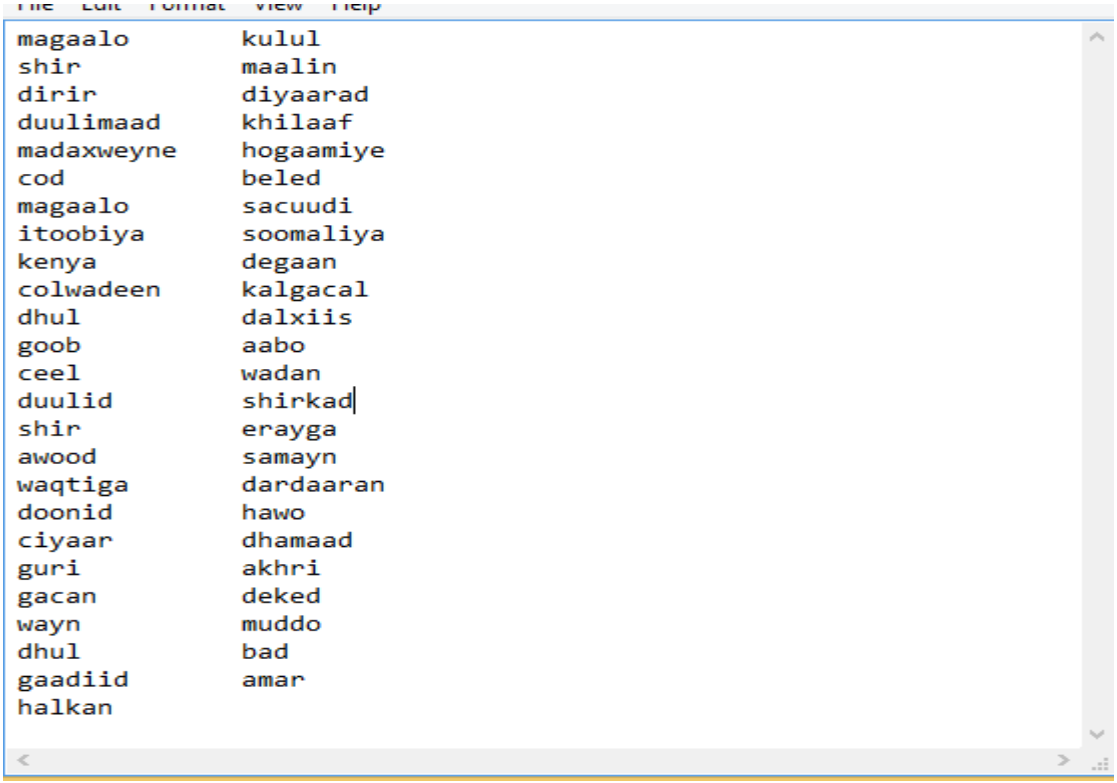

Figure 5.2: Stem Words

5.3 Implementation Tools

The prototype of the system has been implemented using java programming language to design the graphic user interface of the system. We have also used open source java library called Lucene, version 6.4 for indexing and searching. Lucene is a fast and powerful

searching library. As a development environment the prototype is carried out under windows pc with a processing speed of 2.4 Hz.

5.4 Thesaurus Construction

Based on the three domains used, and the 10 queries formulated, we have constructed manual thesaurus for Somali language. We have selected the thesaurus, because of its simplicity then other query expansion Techniques. The thesaurus contains the synonyms and near synonyms of the words. It is constructed with the help of Somali dictionary (Qaamuuska Af-Soomaaliga). The terms are constructed as term to term relationship. The constructed thesaurus is shown in Figure 5.3:

```
map.put("buuq", "qaylo");
map.put("badelmid", "badelan doorsoon");
map.put("badelan", "doorsoon badelmid");
map.put("buubid", "duulimaad");
map.put("buug", "waraaqo maqaal");
map.put("baarlamaan", "xeerbeegti");
map.put("baadigoob", "raadin dayday");
map.put("baacin", "caydhin");
map.put("biyo xireen", "dam niil wabi");
map.put("baraare", "barwaaqada");
map.put("beled", "degaan");
map.put("booqasho", "booqad, siyaaro taariikhi");
map.put("booqad", "booqasho siyaaro");
map.put("barwaaqada ", "reynreyn");
map.put("baansiin", "batrool");
map.put("bardoodan", "daray");
map.put("baashaal", "sheeko kaftan");
map.put("turaanturin", "riixid");
map.put("tuurid", "xoorid");
map.put("tamashlayn", "warwareeg");
map.put("jid", "wado");
map.put("jaar", "daris");
map.put("jaamacada", "goob waxbarasho xarun");
map.put("jacayl", "caashaq");
map.put("xisbi", "axsaab urur");
map.put("xaashi", "warqad");
map.put("xidhid", "xanibid");
map.put("xanibid", "xidhid");
map.put("xeerbeegti", "baarlamaan");
map.put("xushmad", "qadarin");
map.put("xarayn", "gelin");
```

Figure 5.3: Thesaurus

5.5 System Performance Evaluation

Information Retrieval systems are evaluated using different evaluation metrics [46]. However, the most usual techniques are: recall, precision and f-measure. In this system we have used precision, recall and f-measure for system evaluation. Recall is defined as the ratio of relevant documents retrieved, by the total number of relevant documents in the corpus. In the same way precision can be defined as the ratio of the number of relevant documents retrieved, by the total number of documents retrieved. Precision weights the ability of the system to retrieve top-ranked documents that are most relevant to user's query, and it is defined to be the percentage of the retrieved documents that are truly relevant to the user's query. F-measure also called harmonic mean, balances both the recall and precision of the system. It is a performance measure.

The formula used to obtain the f-measure in an Information Retrieval system is as:

$$F - measure = \frac{2PR}{P+R} \tag{10}$$

Where p is the precision and R is the recall and F is the F-measure.

$$Precision = \frac{Relevant \cap Retrieved}{Retrieved} \tag{11}$$

Where Relevant is the number of documents which is Relevant to the query, retrieved is the number of documents retrieved is response to the query of the user and P is the precision.

$$Recall = \frac{Relevant \cap Retrieved}{Relevant} \tag{12}$$

Where R is the recall, Relevant is the number of documents which is relevant to the query and Retrieved is the total document retrieved.

5.6 Query Selection

According to the documents collected, the authors have prepared 10 queries to test the system. The queries have been selected with respect to their relevance to the document and the help of language experts. Thus the relevance of the queries to the corpus collected has been tested manually across all documents. Table 5.2 shows the Constructed queries for the corpus.

Table 5.2 Selected Queries

No.	
1	Buugga Ardayga Fasalka afraad
2	macnaha fikirka Suufinimada
3	jacaylkii qays iyo leyla
4	daraasadaha doorsoonka cimilada
5	xisbiga barwaaqada ee itoobiya
6	ceelasha shidaalka ogaden
7	goobaha dalxiiska itoobiya
8	shirkadda diyaaradaha itoobiya
9	Biyo xireenka abay
10	jaamacadda addis ababa

Searching

To search and use the system, the user is expected to submit a meaningful query, by using the GUI. The GUI has one text field intended for the end users to use as query submission field.

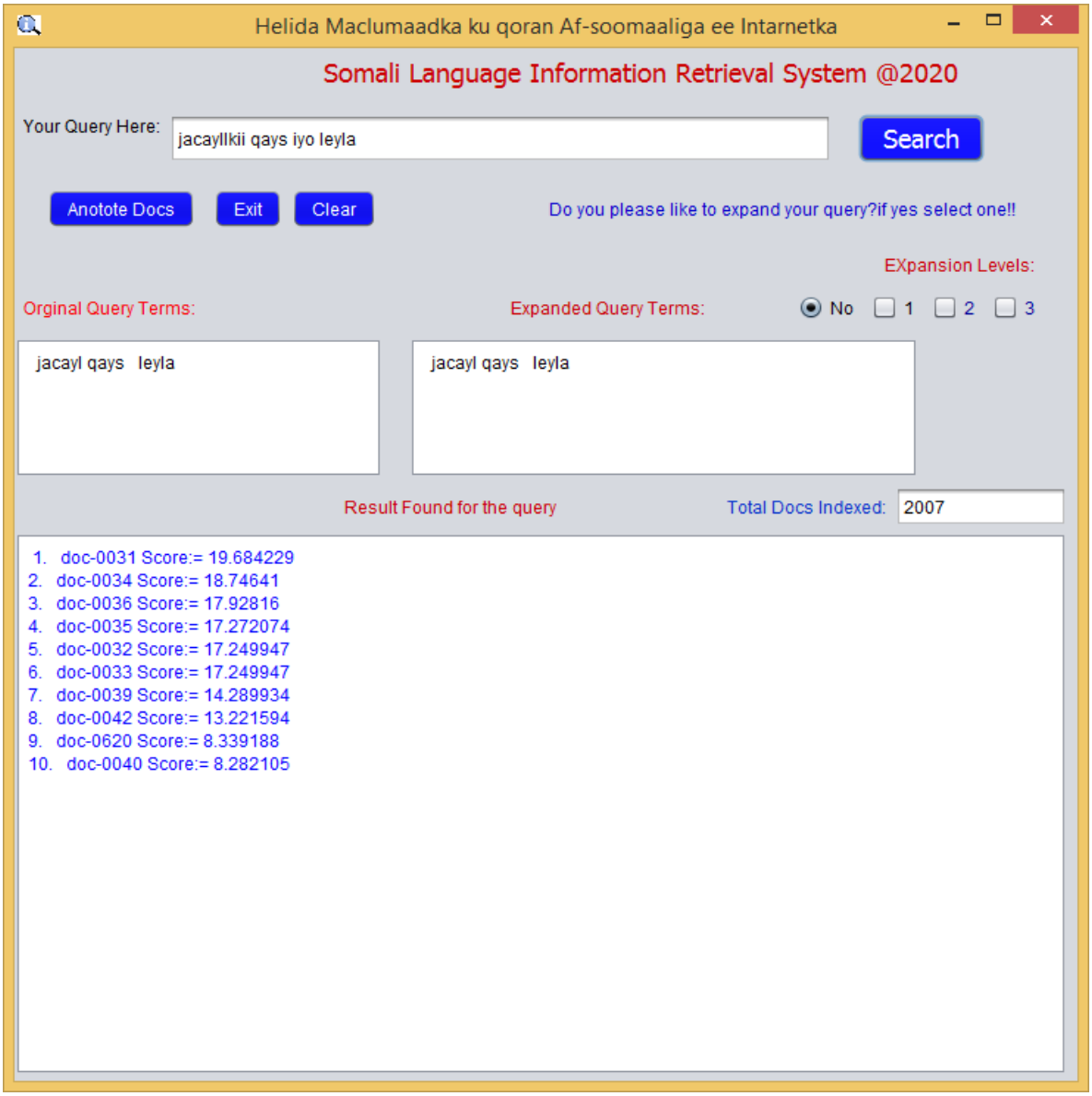

Figure 5.4: Searching Without Expansion

Figure 5.5: Searching with Expansion

Table 5.3 Evaluating Recall and Precision without Query Expansion

Query No.	**Relevant**	**Retrieved**	**Retrieved Relevant**	**P**	**R**	**F**
Query-1	13	14	9	0.64	0.69	0.82
Query-2	16	20	12	0.6	0.75	0.73
Query-3	12	14	9	0.64	0.75	0.69
Query-4	10	14	8	0.57	0.8	0.67
Query-5	10	14	7	0.5	0.7	0.58
Query-6	9	14	6	0.43	0.67	0.52
Query-7	10	14	8	0.57	0.8	0.67
Query-8	12	16	7	0.44	0.58	0.5
Query-9	9	18	6	0.33	0.67	0.44
Query-10	7	10	6	0.6	0.86	0.71
Average				**0.53**	**0.72**	**0.63**

Table 5.3 shows the precision, recall and f-measure of 10 queries using keyword based retrieval. The system recorded average of 53% precision, 72% recall and 63% f-measure. It has achieved this result with simple stemming and stop word removal. As shown in Table 5.3, the precision is lower than the recall in unranked set of retrieval. It is always true that recall is high whenever the retrieved relevant is high. Precision depends on the N randomly selected retrieval documents. In this case Table 5.3 shows that in query-2 the recall is 0.75 while precision is 0.6, that means this is unranked retrieval in which the selected number of documents to be displayed is greater than the total relevant documents.

The original queries in Table 5.3 are expanded using thesaurus. This technique of query expansion is selected for its simplicity. The terms in the query are added more other terms which are synonyms or near-synonyms to the query terms. The word **jacayl** may be expanded to: **caashaq ishqi rabitaan xubi kalgacal kalgaceyl lexejeclo.** Table 5.4 shows the list of expanded 10 queries using thesaurus for Somali language retrieval experiment. The added terms are the bold ones.

Table 5.4 List of Expanded Somali Queries Using Thesaurus

Query No.	**Original Queries**	**Expanded Queries**
Query-1	Jacaylkii qays iyo leyla	jacayl qays leyla **caashaq cishqi kalgacal kalgaceyl lexejeclo rabitaan kalgaceyl rabitaan**
Query-2	fikirka Suufinimada	fikir suufinima **dhabanahays fakar ujeedo aamin**
Query-3	Buugga Ardayga Fasalka afraad.	buuga arday fasal 4aad **galaas dugsixarun wax barasho**
Query-4	daraasadaha doorsoonka cimilada	daraasadaha doorsoon cimilada **badelmid badelan cilmi baadhis hawada cilmilo-gooreed**
Query-5	xisbiga barwaaqada	xisbi barwaaqada **axsaab urur urur-siyaasadeedreynraynbaraarebarwaaqo**
Query-6	Shidaalka ogaden	shidaal ogaden **baansiin batrool**
Query-7	goobaha dalxiiska itoobiya	goobaha dalxiis itoobiya **xarumaha tamashlayn booqosho magaalo warwareeg meelaha meel beled**
Query-8	shirkadda diyaaradaha itoobiya	shirkada diyaaradaha itoobiya **sharikad sharikad urur caymis axsaab xisbi**
Query-9	Biyo xireenka abay	Biyo xireenka abay **dablaawe barafo dareere dam niil wabi haro biyo fadhiisin wabiga itoobiya masar mareen Hoor midablaawe ah oo ku barafooba barafooba dareere**
Query-10	jaamacadda addis ababa	jaamacada ababa **waxbarasho xarun waxbarasho xarun tacliin goob waxbarasho Aqoon habaysan oo la barto**

Table 5.5: Evaluating Recall and Precision with Query Expansion

Query No.	**Relevant**	**Retrieved**	**Retrieved Relevant**	**P**	**R**	**F**
Query-1	13	14	10	0.71	0.77	0.74
Query-2	16	20	13	0.65	0.81	0.72
Query-3	12	14	11	0.78	0.92	0.84
Query-4	10	10	9	0.9	0.9	0.9
Query-5	10	14	8	0.57	0.8	0.67
Query-6	9	14	8	0.57	0.89	0.69
Query-7	10	14	9	0.64	0.9	0.7
Query-8	12	16	8	0.5	0.67	0.57
Query-9	9	18	7	0.39	0.78	0.52
Query-10	7	10	7	0.7	1	0.82
Average				**0.64**	**0.84**	**0.71**

In Table 5.5 the average recall, precision and f-measure with thesaurus based query expansion has been illustrated. The recall and precision of unranked set of retrieval with randomly selected number of items has been shown in the table. The system registered better precision and recall after applying this query expansion technique. After query expansion is conducted the performance of the system is improved 12% of recall and 11% of precision, by enhancing precision from 53% to 64% and recall from 72% to 84%. It also shows 8% improvement of f-measure. Overall thesaurus-based query expansion shows an important improvement of overall system performance. In Table 5.5, when all relevant docs are retrieved, then recall becomes high. Unlike the recall the precision depends on the randomly selected number of retrieval docs. In this case the precision is always better when ranked retrieval is used.

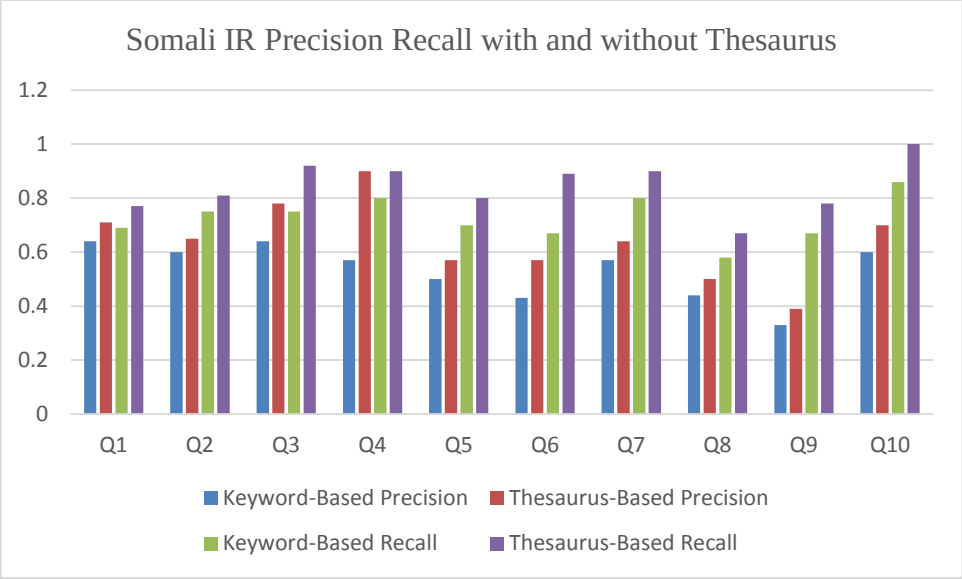

Figure 5.6: Precision Recall with and without query expansion

In Figure 5.6 the precision and recall with and without query expansion is shown. In both measures of precision and recall there is enhancement after applying thesaurus as query expansion. The 10 queries used to test the system show changes with the usage of thesaurus. It has been tested, that generated graph of the data shows that the precision and recall of the thesaurus based expansion perform better than the precision and recall in the keyword based expansion.

In Information Retrieval a confusion matrix is used to show the differences between the true and predicted values [68]. It shows whether the result is as expected or not. The confusion matrix of four queries is shown in Table 5.6.

Table 5.6 Confusion Matrix

#Query	N	Actual	Predicted no	Predicted yes	Total Predicted
Query-1	14	No	TN=1	FP= 3	14
		Yes	FN=0	TP= 10	
Query-2	20	No	TN=4	FP=3	20
		Yes	FN=0	TP=13	
Query-3	14	No	TN=2	FP=1	14
		Yes	FN=0	TP=11	
Query-4	10	No	TN=0	FP=1	10
		Yes	FN=0	TP=9	
Total					

where TP is true positive, the actual label is yes and model also predicted yes, TN is true negative. Actual label is no and model predict no, FP is false positive actual label is no and model predict yes, FN is false negative, actual label is yes and model predict no. As illustrated in Table 5.6, N is a randomly selected number of documents. The actual is the true values, predicted yes is total predicted relevant documents and predicted no is total predicted irrelevant documents. Example in Query-1, the total retrieved documents are 14, the relevant docs retrieved are 10, the expected irrelevant docs is 1 and unexpected irrelevant docs is 3.

Table: 5.7: precision Recall for 3 Levels of Expansion with 4 queries

Query no.	Query Expansion Levels					
	Level-1		Level-2		Level-3	
	P	R	P	R	P	R
Query-7	0.57	0.8	0.64	0.9	0.64	0.9
Query-8	0.56	0.75	0.5	0.67	0.19	0.25
Query-9	0.39	0.78	0.39	0.78	0.39	0.78
Query-10	0.7	1	0.2	0.29	0.1	0.14

Thesaurus is constructed from synonyms or words which can have similar meaning. Thus two synonym words in the thesaurus can have two different synonyms. For this reason, when the level of expansion increases, the dissimilarity of the words might increase as well. As shown in Table 5.7 the precision and recall of some queries show decreasing values as the level in expansion increases. For example, the precision in level-1 query 10 is 70%, where the precision in level-2 with the same query is 20%. When level increases variations between synonym words becomes large.

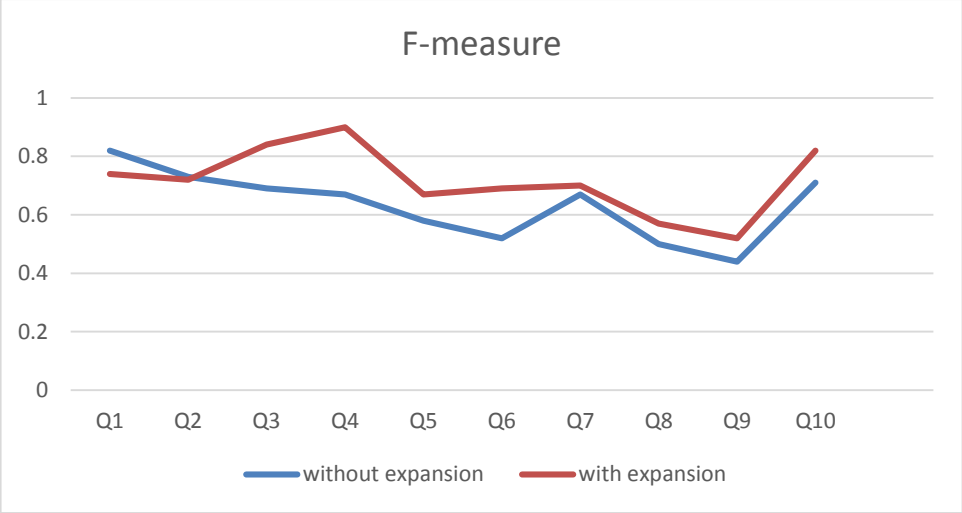

Figure 5.8: f-measure with and without expansion

Figure 5.8 illustrates graph of the f-measure values before and after query expansion is applied. In this graph except queries one and two all other queries show better performance when thesaurus based query expansion is used. Thus f-measure registers importance of recall and precision with the usage of thesaurus based technique of query expansion, compared to keyword based.

Discussion

Our evaluation has been conducted using Lucene open source library with a simple graphic user interface. To test the performance of the system, 10 queries have been submitted to the system. The system has registered promising performance of the evaluation metrics with the use of thesaurus. This approach of query expansion, expands query terms and adds additional words to the original query terms as shown in Table 5.4. It has shown enhanced performance compared to the keyword based retrieval. As indicated in Table 5.3 the system

recorded 53% and 72% precision and recall respectively. Also as illustrated in Figure 5.6 improvements are shown in recall and precision with the use of reformulated query. After query refinement the system improved precision from 53% to 64% while it has improved recall from 72% to 84%. This means that the probability of retrieving documents increases when query is reformulated. In addition to the recall precision the f-measure registers enhancement of 8% by improving it from 63% to 71%. Based on the results found, it is clear that Somali Information Retrieval brings better improvements by applying thesaurus query expansion. Because thesaurus lists the synonyms it scores better results by find exact words. In other hand it sometimes becomes difficult for thesaurus to be constructed. As thesaurus construction is manual and time taking task. Although we can see that good results have been achieved yet to improve performance further researches need to be conducted. Researches which can further improve recorded results could include: parts of speech tagging, stemming and morphological analyzing as Somali language is morphological complex.

Chapter Six: Conclusion, future work and contribution

6.1 Conclusion

Textual data as well as other types of data has been digitized and electronically made available. To access such textual data Information Retrieval systems became important systems. Information Retrieval systems depend on language characteristics; thus it is necessary to make search engines suitable for every language. To minimize research gaps between the various types of languages, studies has been conducted on different languages. This research attempts to investigate the first Somali language Information Retrieval using thesaurus. The study is marked to be the first IR using thesaurus for this language so far. We have conducted a literature review and figured out the characteristics of Somali language, its writings and origins.

We have developed a prototype to evaluate the characteristics and performance of the system. The developed prototype has been conducted using java programming language and it has a simple Graphic user interface. In the experiment to further the performance of the system a query expansion technique has been used. Our query expansion technique used is thesaurus based. Thesaurus is a manually constructed set of alphabetically ordered set of words. Somali language didn't have any standard testing collections, thus this research arranged a manually collected set of 2007 text documents in Somali. To index the document, we have used lucene library, which is fast searching open source library. Before indexing the test collection is cleaned by using text preprocessing, the preprocessing tasks done include: normalization, tokenizing, stop word removal and stemming. In general, the preprocessing tasks in Somali language need their own algorithms, for this reason we have tried to develop simple stemmer and normalizing algorithms. After the document is preprocessed we have performed measuring technique to retrieve matching documents to the user query.

In this work to retrieve matching documents we have used VSM as a retrieving model. The VSM uses document matrix in which terms in the documents are represented as matrix, based on their importance. To rank the documents according to their relevance, we have used the cosine similarity measurement. This ranking function became an important because lucene library is by default based on cosine similarity function. After the model and ranking

function proposed are used the system registered an encouraging result. It has recorded 53% and 72%, precision and recall respectively without query expansion. It has achieved 64% and 84%, precision and recall respectively after query expansion is applied. This result shows that thesaurus based query expansion improves text retrieval in Somali language.

6.2 Contribution

As a contribution, this research has contributed the following bulleted points:

- We have designed an architecture for Somali Language Information Retrieval using thesaurus.
- We have experimented the first Somali language Information Retrieval using thesaurus.
- Test collection has been collected for the first time in Somali language from different sites
- Manual thesaurus has been constructed with the help of Somali qaamuus.

6.3 Future work

The following future research directions has been forwarded, based on research findings.

- In this research simple stemmer has been developed, fully functional stemmer for Somali language is needed.
- An establishment of standard corpus for Somali language is necessary to do further research.
- Parts of speech tagging can improve, the achieved results to further the result, POS is needed.
- In this research only textual data retrieval is considered, other types of data retrieval such as video, audio, and image need to be performed as future work.

References

[1] Michael Lesk Bellcore, "THE SEVEN AGES OF INFORMATION RETRIEVAL," *International Federation of Library Associations and Institutions*, 1996.

[2] Arpit Deo, Jayesh Gangrade and Shweta Gangrade, "A SURVEY PAPER ON INFORMATION RETRIEVAL SYSTEM," *International Journal of Advanced Research in Computer Science* , vol. 9, no. 1, pp. 778-781 , 2018.

[3] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, An Introduction to Information Retrieval, Cambridge,England: Cambridge University Press , 2009.

[4] Tapan P Gondaliya and Hiren D Joshi, "Journey of Information Retrieval to Information Retrieval Tools - IR&IRT A Review," *in 11th International CALIBER-2017 Anna University, Chennai*, 2017.

[5] William B. Frakes and Ricardo Baeza-Yates, Information Retrieval: Data Structures & Algorithms, 2004.

[6] Gezehagn Gutema Eggi, "Afaan Oromo Text Retrieval System," SCHOOL OF INFORMATION SCIENCE , ADDIS ABABA UNIVERSITY , 2012.

[7] E.Iyswarya and M. Balamurugan, "A Trend Analysis of Information Retrieval Models," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, pp. 531-534 , 2017.

[8] Namita Mittal, Richi Nayak, MC Govil and KC Jain, "Dynamic Query Expansion for Efficient Information Retrieval," in *International Conference on Web Information Systems and Mining*, MNIT Jaipur, 2010.

[9] omar osman ibrahim and yoshiki mikami, "indixing tigrinya language documents," information processing society of japan, nagaoka university of technology, 2013.

[10] "Somalia: Language and Culture," NYS Statewide Language Regional Bilingual Education Resource Network (RBE-RN) at New York University, new york, 2012.

[11] "Language situation and dialects," LANDINFO , Report Somalia, 2011.

[12] Diana Briton Putman and mahamood cabdi Noor, "The Somalis Their History and Culture," The Refugee Service Center center for applied linguistics , washington DC, 1993.

[13] Ethiopia, "nationalstatistics," 19 6 2012. [Online]. Available: http://www.csa.gov.et/images/documents/pdf_files/nationalstatisticsabstract/2012/population. [Accessed 24 9 2019].

[14] Lisa Peters and Chris Mayer, "Somali and English: Some Differences and the Implications for Writing Tutors and Instructors," MinneTESOL Journal, Fall 201, 2016.

[15] Eyob Nigussie Alem, "Afaan Oromo – Amharic Cross Lingual Information Retrieval: Acorpus Based Approach," Unpublished Masters Thesis, School of Information Science Addis Ababa University, 2013.

[16] Xiaolian Li, Kunying Li, Dexin Qiao, Yu Ding and Daiming Wei, "Application Research of Machine Learning Method Based on Distributed Cluster in Information Retrieval," in *International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2019.

[17] tewodros abebaw chekol, "applying thesaurus based semantic compression for improving the performance of amharic text retrieval," University of Gondar, Unpublished Masters Thesis Department of Information Technology, 2014.

[18] jaltu fita, "afaan oromo Question Answering System," Unpublished Masters Thesis, Department of Computer science Addis Ababa University, 2017.

[19] Rong Yan and Guanglai Gao, "Pseudo-Based Relevance Analysis for Information Retrieval," in International Conference on Tools with Artificial Intelligence, Hohhot, 2017.

[20] Venkat N. Gudivada, Dhana L. Rao and Amogh R. Gudivada, "Information Retrieval: Concepts, Models, and Systems," Department of Computer Science , East Carolina University, Carolina , 2018.

[21] D. Hiemstra, "Information Retrieval Models," J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., Goker, A. and Davies, 2009.

[22] Gesare Asnath Tinega, Waweru Mwangi and Richard Rimiru, "Text Mining in Digital Libraries using OKAPI BM25 Model," International Journal of Computer Applications Technology and Research , vol. 7, no. 10, pp. 398-406, 2018.

[23] Stephen Robertson and Hugo Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, p.

333–389, 2009.

[24] E. E. Ogheneovo and R. B. Japhet, "Application of Vector Space Model to Query Ranking and Information Retrieval," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 5, 2016.

[25] Deepa Yogish, Manjunath T N and Ravindra S Hegadi, "Variants of Term Frequency and Inverse Document Frequency of Vector Space Model for Effective Document Ranking in Information Retrieval," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, pp. 2278-3075, 2019.

[26] Vaibhav Kant Singh and Vinay Kumar Singh, "VECTOR SPACE MODEL: AN INFORMATION RETRIEVAL SYSTEM," in International Journal of Advanced Engineering Research and Studies, Chhattisgarh, 2015.

[27] Taras Shevchenko, "Efficient Search in Short Documents," National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Institute for Applied System Analysis.

[28] Youngjoong Ko, "A New Term Weighting Scheme for Text Classification using the Odds of," Department of Computer Engineering, Dong-A University, Busan, 2015.

[29] Anastasios Tombros and C.J. van Rijsbergen, "Query-Sensitive Similarity Measures for Information Retrieval," in Knowledge and Information Systems , London, 2004.

[30] Manoj Chahal, "Information Retrieval using Jaccard Similarity Coefficient," International Journal of Computer Trends and Technology (IJCTT), vol. 36, no. 3, pp. 2231-2803 , 2016 .

[31] Welde janfa, "ontology based query expansion for enhancing the performance of amharic information retrival:," Unpublished Masters Thesis, school of information science, addis ababa university , 2014.

[32] AhmedAbdoAzizAhmedAbdulla, HongfeiLin ,BoXu and Santosh Kumar Banbhrani, "Improving biomedical information retrieval by linear combinations of different query expansion techniques," in *From 12th Annual Biotechnology and Bioinformatics Symposium(BIOT-2015),* Provo,UT, 2015.

[33] Fang w ul., Guosbi wu and Xiangling Fu., "Design and Implementation of Ontology-Based Query Expansion for Information Retrieval," *FlP International Federatioll fOT Information Processing.,* vol. 254, p. 293·298, 2007.

[34] A. Kankaria, "Query Expansion techniques," Indian Institute of Technology Bombay, Mumbai, 2012.

[35] Jentrisi Priyatno and Moch Arif Bijaksana2, "Clustering Synonym Sets in English WordNet," *in 7th International Conference on Information and Communication Technology (ICoICT)*, 2019.

[36] Tsadu Zeray, "Query Expansion for Tigrigna Information Retrieval," Unpublished Masters Thesis, Department of computer Science, Addis Ababa University ,2017.

[37] Yejun Wu, "Enriching a thesaurus as a better question-answering tool and information retrieval aid," *Journal of Information Science*, vol. 44, p. 512–525, 2018.

[38] Tewodros abebaw chekol, "applying thesaurus based semantic compression for improving the performance of amharic text retrieval," Unpublished Masters Thesis, department of information technology , University of Gondar, 2014 .

[39] "somalia:language and culture," NYS Statewide LanguageRegional Bilingual Education Resource Network (RBE-RN) at New York University, New York, 2012.

[40] Sunita Shah, "SOMALI A PROFILE," Clinical Lead Speech & Language Therapist Bilingual Specialist Tanvi Shah Speech & Language Therapist Bilingual Specialist London SIG Bilingualism, London , 2007.

[41] A. m. Handulle, "Tema mosrmal," Tema mosrmal, 5 11 2010. [Online]. Available: http://morsmal.no/so/morsmal-somali-1-4/1aad-4aad/632-shibbaneyaasha-labalaabma-ee-af-soomaaligadobbelt-konsonanter.. [Accessed 27 2 2020].

[42] mauro Tosco, "short notes on Somali previous scripts," intergovernmental academy of Somali language, 2010.

[43] WWJDFI, Af-Soomaali Tilmaame Bare Fasalka 4aad, Addis Ababa: Wasaaradda Waxbarashada ee Jamhuuriyadda dimuqraadiga Federaalka Itoobiya, 2013.

[44] Lisa Peters and Chris Mayer, "Somali and English: Some Differences and the Implications for Writing Tutors and Instructors," in MinneTESOL Journal, Fall, Northern Arizona, 2016.

[45] Deqa M. Hassan, "Somali Dialects in the United States: How Intelligible is Af-Maay to Speakers of AfMaxaa?," Unpublished Masters Thesis, department of English, Minnesota State University, Mankato, Minnesota, 2011.

[46] Diana Briton Putman & Mohamood Cabdi Noor, The Somalis Their History and Culture, Washington DC: The Refugee Service Center Center for Applied Lingtdstics 1118 22nd Street NW Washington DC 20037 (202) 429.929, 1993.

[47] Ahmed Mohamed, "AN OVERVIEW OF THE INTERFACE BETWEEN ASPECTS OF SOMALI,"Unpublished Thesis,College of Education,St. Cloud State

University,Minnesota ,2013.

[48] J. W. C. KIRK, B.A., A GRAMMAR OF THE SOMALI LANGUAGE WITH EXAMPLES IN PROSE AND VERSE AND AN ACCOUNT OF THE YIBIR AND MIDGAN DIALECTS, Westmead, Farnborough, Hants., England: CAMBRIDGE AT THE UNIVERSITY PRESS, 1969.

[49] ETTIEN KOFFI, SHUKRIA OMAR, HASSAN YUSSUF and MOHAMMED DAHIR, "Somali Verb Conjugation Paradigms: Present, Past, and Future," Linguistic Portfolios, vol. 6, pp. 2472-5102, 2017 .

[50] Sanjeev K Sunny and Mallikarjun Angadi, "Potential Roles and Applications of Thesauri in Digital Information Retrieval Systems," *in 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services*, 2018.

[51] Rong Yan and Guanglai Gao, "Pseudo-Based Relevance Analysis for Information Retrieval," in *International Conference on Tools with Artificial Intelligence*, Hohhot, 2017.

[52] M. Uma Devi and G. Meera Gandhi, "Wordnet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain," *Journal of Computer Science*, vol. 361.371 , 2014.

[53] Bhavadharani M, Ramkumar M P and Emil Selvan G S R, "PERFORMANCE ANALYSIS OF RANKING MODELS IN INFORMATION RETRIEVAL," *in Proceedings of the Third International Conference on Trends in Electronics and Informatics, Madurai*, 2019.

[54] R. C. S. G. A. W. N. a. D. Radev, "Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations," Yale University, 2019.

[55] Martin Braschler and Brbel Ripplinger, "Stemming and Decompounding for German Text Retrieval," *Springer*, no. 2633, pp. 177-192, 2003.

[56] Kareem Darwish and Walid Magdy, "Arabic Information Retrieval," *Foundations and Trends in Information Retrieval*, vol. 7, no. 4, p. 239–342, 2014.

[57] Ibrahim Moawad, Waseem Alromima and Rania Elgohary, "Bi-Gram Term Collocations-based Query Expansion Approach for Improving Arabic Information Retrieval," *Arabian Journal for Science and Engineering*, 2018.

[58] Kutlu Emre Yılmaz, Ahmet Arslan and Ozgur Yilmazel, "TURKISH TEXT RETRIEVAL EXPERIMENTS USING LEMUR TOOLKIT," Computer Engineering Department, Anadolu Universit, Eskisehir, 2014.

[59] Abey Bruck and Tulu Tilahun, "Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique," Journal of Computer and Communications, vol. 3, pp. 67-76, 2015.

[60] Tewodros hailemeskel gebermariam, "amharic text retrieval: an experiment using latent semantic indexing (lsi) with singular value decomposition (svd)," department of information science,addis ababa university , 2003.

[61] Atelach Alemu Argaw and Lars Asker, "Amharic-English Information Retrieval," Department of Computer and Systems Sciences, Stockholm University, 2006.

[62] Berhanu Anbase, "Applications of Information Retrieval for Afaan Oromo text based on Semantic based Indexing," Unpublished Thesis, Department of information Technology, Jimma University, 2019.

[63] Teklay Birhane and Birhanu Hailu, "Design and Implementation of IR System for Tigrigna Textual Documents," I.J. Modern Education and Computer Science, vol. 11, pp. 31-38 , 2019.

[64] Elizabeth Boschee, Joel Barry, Jayadev Billa, Marjorie Freedman, Thamme Gowda, Constantine Lignos, Chester Palen-Michel, Michael Pust, BanriskhemK, Khonglah, Srikanth Madikeri, JonathanMay and ScottMiller,"SARAL:ALow-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, 2019.

[65] Hamed Bonab, James Allan and Ramesh Sitaraman, "Simulating CLIR Translation Resource Scarcity using High-resource Languages," College of Information and Computer Sciences, University of Massachusett, Amherst, 2019.

[66] Abdillahi Nimaan, Pascal Nocera, Frédéric Bechet and Jean-François Bonastre, "Information Retrieval Strategies for Accessing African Audio Corpora," Laboratoire Informatique d'Avignon - UAPV, Avignon, France, Institut des Sciences et des Nouvelles Technologies - CERD, Djibouti, Antwerp, Belgium, 2007.

[67] Tomas Mikolov, Kai Chen, Jeffrey Dean, Greg Corrado and Ilya sutskever, "Distributed Representations of Words and Phrases and their Compositionality," 2013.

[68] Keneilwe Zuva and Tranos Zuva, "EVALUATION OF INFORMATION RETRIEVAL SYSTEMS," International Journal of Computer Science & Information Technology, vol. 4, no. 3, 2012.

Annexes

Annex A: Sample java Code to Normalize Somali text

```
for (int i = 0; i < qu.size(); i++) {

        if (qu.get(i).contains("dd")) {

            stmp = qu.get(i).replace("dd", "d");

            sentence = sentence.replace(qu.get(i), stmp);

        } else if (qu.get(i).contains("ll")) {

            stmp = qu.get(i).replace("ll", "l");

            sentence = sentence.replace(qu.get(i), stmp);

        } else if (qu.get(i).contains("gg")) {

            stmp = qu.get(i).replace("gg", "g");

            sentence = sentence.replace(qu.get(i), stmp);

        } else if (qu.get(i).contains("bb")) {

            stmp = qu.get(i).replace("bb", "b");

            sentence = sentence.replace(qu.get(i), stmp);

        } else if (qu.get(i).contains("mm")) {

            stmp = qu.get(i).replace("mm", "m");

            sentence = sentence.replace(qu.get(i), stmp);

        } else if (qu.get(i).contains("nn")) {

            stmp = qu.get(i).replace("nn", "n");

            sentence = sentence.replace(qu.get(i), stmp);

        } else if (qu.get(i).contains("rr")) {

            stmp = qu.get(i).replace("rr", "r");

            sentence = sentence.replace(qu.get(i), stmp);

        }


    }
```

Annex B: Somali stop words

uun	een	kali	kara	kaas	kee	ah
kuwee	kuwaas	ku	ayaa	waxa	iyo	ee
kii	kaa	kan	eegayo	dibada	daray	loo
kale	sidoo	hore	badan	inta	dhamaan	kali
horjeeda	lasocda	kahor	markale	mar	kabacdi	saamaynaya
saameeyey	saamayn					
inkastoy	ayuu	inkastuu	iyada	iyaga	isaga	aad
mid	in	dhan	fadhiya	badhtanka	dhexjira	anaga
waxay	dhex	waxaan	inaga	ee	idinka	aniga
meel	laga	uun	xag	xagii	reebo	walba
halka	la	si	saas	sidii	sida	meelaha
aan	meelahaa	ahayn	goobo	goob	yihiin	haboon
noqday	ahaa	sababtu	sa	noqdeen	noqonaysa	noqonaya
sab	noqotay	sababtii	sababta	sababtoo	sabab	noqonayaan
bilaabid	horaysay	hore	bilaabeen	bilow	bilaabata	ahaan
aamin	kafiican	ugufiican	fiican	gu	u	fiican
hoose	aaminaysa	hoos	aaminaya	aamintay	aaminay	aaminid
ahaansho	waynaa	hooseeya	waynaan	waynaatay	waynaanaysa	daba
iib	laakiin	wac	wayn	iibsadeen	iibsatay	wacaysa
kara	iibsanaya	kartaa	kartaan	karaysa	karaya	karin

Annex C: Sample stemmed Somali words

cimilada	beled	degaan	baraare	batrool
koob	shirkada	abay	booqad	daris
kalaas	bare	addis	taariikhi	jaar
fiican	xeerbeegti	jaamacada	siyaaro	wado
baarlamaan	macalin	dawlada	booqad	jid
tartan	galaas	waraaqo	booqasho	warwareeg
duulimaad	Suufinimada	wabi	dayday	tamashlayn
buubid	barwaaqada	niil	raadin	xoorid
buug	dowlada	dam	baadigoob	tuurid
badelan	ababa	xireen	baarlamaan	riixid
hawada	duulimaadyada	caydhin	maqaal	turaanturin
fasalada	dalka	baacin	barwaaqada	kaftan
sheeko	baansiin	bardoodan	daray	baashaal
waxbarasho	goob	jacayl	kalfadhi	ishciq
cimilo	siyaasad	caashaq	iskudhexyaac	cimilada
hadaba	doorsoon	doorasho	duulimaadyadeeda	xarun
dagaal	erayga	buuga	rabitaan	kicid
kulul	shir	shirkad	duulid	wadan
ceel	magaalo	isururis	xeelad	raadin
waqtiga	dhamaad	dardaaran	doonid	ciyaar
hawo	akhri	guri	gacan	muddo
deked	dhul	wayn	bad	gaadiid

Signed Declaration Sheet

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other universities and that all sources of materials used for the thesis has been duly acknowledged.

Declared by:

Name: Abdisalam Mahamed Badel

Signature: _____

Date: June ,2020

Confirmed by Advisor:

Name**:** Yaregal Assabie (PhD)

Signature: _____

Date: June, 2020