QUERY EXPANSION FOR AFAAN OROMO INFORMATION RETRIEVAL BASED ON WORDNET

MSC THESIS

MELKAMU ABETU

OCTOBER 2017

HARAMAYA UNIVERSITY, HARAMAYA

Query Expansion for Afaan Oromo Information Retrieval Based on WordNet

A Thesis Submitted to the Department of Information Science
Postgraduate Program Directorate
HARAMAYA UNIVERSITY

In Partial Fulfillment of the Requirements for the Degree of
MASTER OF SCIENCE IN INFORMATION SCIENCE

Melkamu Abetu

October 2017

Haramaya University, Haramaya

HARAMAYA UNIVERSITY

POSTGRADUATE PROGRAM DIRECTORATE

I hereby certify that I have read and evaluated this Thesis entitled Query Expansion for Afaan Oromo Information Retrieval Based on WordNet prepared under my guidance by Melkamu Abetu. I recommend that it be submitted as fulfilling the thesis requirement.

Million Meshesha (PhD)	_____	_____
Major Advisor	Signature	Date

As a members of the Board of Examiners of the MSc Thesis Open Defense Examination, I certify that I have read and evaluated the Thesis prepared by Melkamu Abetu and examined the candidate. I recommend that the thesis be accepted as fulfilling the Thesis requirements for the degree of Master of Science in Information Science.

_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
_____	_____	_____
External Examiner	Signature	Date

Final approval and acceptance of the Thesis is contingent upon the submission of its final copy to the Council of Graduate Studies (CGS) through the candidate's department or Postgraduate committee (DGC or SGC).

DEDICATION

This work is dedicated to my lover Ayyuukoo for her nice counseling and encouragement next to almighty God who serves me to reach up on this stage.

STATEMENT OF THE AUTHOR

By my signature below, I declare and affirm that this Thesis is my own work. I have followed all ethical and technical principles of scholarship in the preparation, data collection, data analysis and compilation of this Thesis. Any scholarly matter that is included in the Thesis has been given recognition through citation.

This Thesis is submitted in partial fulfillment of the requirement for an MSc degree at Haramaya University. The Thesis is deposited in the Haramaya University Library and is made available to borrowers under the rules of the library. I solemnly declare that this Thesis has not been submitted to any other institution anywhere for the award of any academic degree, diploma or certificate.

Brief quotations from this Thesis may be used without special permission provided that accurate and complete acknowledgement of the source is made. Requests for permission for extended quotations from, or reproduction of this Thesis in whole or in part may be granted by the Head of the School or Department when in his or her judgment the proposed use of the material is in the interest of scholarship. In all other instances, however, permission must be obtained from the author of the Thesis.

Name: Melkamu Abetu Signature: _____

Date: 21/08/ 2017

School/ Department: Information Science

BIOGRAPHICAL SKETCH

The author was born in Jima Rare district, Horo Guduru Wollega Zone, Oromia region on January 10, 1991 from Ato Abetu Aga and W/ro Manalush Bitew. He attended his primary education at Bedaworke primary school and his secondary and preparatory school at Wayu preparatory school. After successful completion of his preparatory education, he joined Haramaya University in 2010 and graduated with Bachelor degree in Information Technology in July 2013. After graduation, he was employed at Haramaya University as Graduate Assistant and served there for about two years. Then in 2016, he has joined the Postgraduate Program at Haramaya University for regular program and continued his studies in Information Science.

ACRONYMS AND ABBREVIATIONS

IC	Information Concept
IDF	Inverse Document Frequency
IEEE	Institute of Electrical and Electronics Engineers
IR	Information Retrieval
MOE	Ministry of Education
NCAA	National Collegiate Athletic Association
ND	No Date
NSF	National Science Foundation
QE	Query Expansion
TF	Term Frequency
URL	Uniform Resource Locator
USA	United State of America
VOA	Voice of America
WSD	Word Sense Disambiguation

ACKNOWLEDGEMENTS

First of all, I would like to thanks almighty God who is the beginning and ending of my work. All things done by him I give glory to him.

Secondly, I would like to thank my advisor Dr. Million Meshesha for his critical comments on my work, for being my driving force throughout this thesis and his patience in helping me to complete my work and for freedom he gave me to pursue my own interests. I would also thank Tariku Mohammed for his comments and guidance on my work.

I wish to express my sincere thanks to Tilahun Shiferaw who is Head of department Information Science and Ashenafi Chalchisa who is Head of department Software Engineering for their advance helping and co-operation during my thesis work. I would also thank Dr. Imana and Mr. Hunduma for their nice guidance during preparing WordNet and Selamawit for her advance helping on Word Sense Disambiguation algorithms.

Last but not least, I would like to express heartfelt gratefulness to my lover Ayyuukoo for her nice counseling and encouragement on daily activity of my work.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

Figures **Pages**

LIST OF APPENDICES

ABSTRACT

Information retrieval enables to search for relevant documents from large corpus as per the information need of users. Query expansion is widely used technique for improving information retrieval effectiveness. Afaan Oromo is a Cushitic language spoken today by about 40 million people in Ethiopia. One of the major problems of Afaan Oromo text retrieval is its effectiveness in identifying relevant documents for users' query that satisfies their information need. The main objective of this study is to integrate query expansion for enhancing the effectiveness of Afaan Oromo text retrieval system. The designed query expansion for Afaan Oromo information retrieval system involves lexical resource like WordNet that is constructed as reference for identifying the senses and meaning of the user's query using word sense disambiguation by semantic similarity measure. Using the idea of original Lesk algorithm, word sense disambiguation is performed with gloss to gloss similarity measure by comparing information associated with its synonyms and gloss definition with reference to Afaan Oromo WordNet. The well-known word senses that are identified during word sense disambiguation from WordNet is used during query reformulation. Finally, the query expansion module is integrated with Afaan Oromo IR system to enhance the effective performance of the system after query expansion is applied. The experimental result shows that an integration of query expansion registers 56% F-measure which improves the performance by 5% from original query. The main challenges in this study are absence of standard well-crafted WordNet, effective stemmer algorithm and corpus for performance evaluation. It is therefore the researcher major recommendation for researchers to work in this line.

Keyword: Information Retrieval; Query Expansion; WordNet; Afaan Oromo Language

1. INTRODUCTION

1.1. Background of the Study

Human being has arranged and organized information for later retrieval and usage for just about 4000 years (Baeza-Yates and Ribeiro-Neto, 1999).

The era and long history of information retrieval does not start with the internet rather it is only in the last decade and a half of the IEEE's hundred years that web search engines have become persistent and search has become incorporated into the fabric of desktop and mobile operating systems. Earlier to the wide public day-to-day usage of search engines, IR systems were found in different commercial and intelligence applications as long ago as the 1960s.The first computer-based searching systems were built in the late 1940s and were stimulated by revolutionary innovation in the first half of the 20th century. As with many automated computer technologies, the potentialities of retrieval systems raised with increases in processor speed and storage capacity. The progress of such systems also reflects a fast progression away from manual library-based approaches of acquiring, indexing, and searching information to increasingly automated methods (Sandarson and Croft, ND).

The World War II in 1945 and cold war after the end of the actual war has the linkage for the development and increase of modern IR. Because of problems in Storing and retrieving bulky numbers of scientific papers publications for the duration of post war, the formulation of new mechanism become mandatory task. However, the storing and retrieving resources is difficult, most the papers are published for the attention of US military and it should be accessible to them easily. Accordingly, Bush (1945) defined the problem as Information explosion as mass production of information.

Information Retrieval (IR) is finding relevant resources on the net that satisfies user information needs. Computerized system which can perform information retrieval system is called information retrieval system (Manning *et al*., 2009).

Without detailed knowledge of document collection, most users find it difficult to formulate queries that are well designed for retrieval purpose. To illustrate, search engines often need users

to reformulate their queries to obtain the results that interests them. This difficulty suggest that the first query formulation should be treated as initial attempt to retrieve relevant information and that improved query reformulations could be written to retrieve additional useful documents.

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations (Ying, 2006). It is usable and necessary due to the ambiguity of natural language and also the problems in using a single term to represent an information concept (Smith *et al*., 2007). QE involves techniques such as finding synonyms of words and searching for synonyms and polysemy, finding all the various morphological forms of words by stemming each word in the search query, fixing spelling errors and automatically searching for the corrected form or suggesting it in the results and re-weighting the terms in the original query.

Query expansion is a methodology studied in the field of computer science, particularly within the area of natural language processing and information retrieval. Using relevance feedback, corpus dependent knowledge models and corpus independent knowledge models the query can be expanded (Smith *et al*., 2007).

1.2. Statement of the Problem

Afaan Oromo is a Cushitic language spoken today by about 40 million people in Ethiopia (about 40% of the country's population), in Kenya, Somalia and Djibouti and is the 3rd largest language in Africa after Arabic and Hausa. Afaan Oromo (meaning Oromo language) or Oromiffa, most probably rates second among the African indigenous languages. Even before two decades when Gada (1988) explored the history of this nation, Afaan Oromo was known as the mother tongue of about 30 million Oromo people living in the Ethiopian and neighboring countries such as Kenya, Somalia, and Djibouti.

Nowadays, journal, magazines, newspapers, online education, books and entertainment (Ager, 2012), Medias, videos, pictures, are available in electronic format both on the Internet and on offline sources. There is huge amount of information being released with this language, since it is the language of education and research, language of administration and political welfares, language of ritual activities and social interaction (Bush, 1945). The number of medias that uses

Afaan Oromo language and transfer information as primarily for speakers of language increases from time to time; for Example, Kallacha Oromiyaa, Bariisaa, Yeroo, Oromia Television and Radio (Web news),Voice of America (VOA) (web news), and different academic and recreational medias to mention a few (Bush, 1945). Technology has great role for the development of one language because it link the speaker and user of that language with easy system to access information in their daily activity. The fact that initiated this study is also enabling development of Afaan Oromo to grow with current information technology support. IR is not being optional technology, it is something that is very important to everybody and mandatory to use. In this Information Age, information is highly needed than anything else. However, searching this necessary information needs system support (Schatz, 1997).

There are few works done for Afaan Oromo language to facilitate text retrieval. Some of these works include; Afaan Oromo search engine (Tesfaye, 2010), Afaan Oromo text retrieval system (Gezehagn, 2012) and designing a rule based stemmer for Afaan Oromo text (Debela and Ermias, 2010), Word sense disambiguation for Afaan Oromo language (Tesfa, 2013), Hybrid word sense disambiguation approach for Afaan Oromo words (Yehuwalashet, 2016). As to the researcher knowledge there is no study that apply Query Expansion to enhance the performance of Afaan Oromo text retrieval system. However, there are works done on query expansion for Amharic information retrieval system. Iman, (2013) explored query expansion based on proper word sense disambiguation and Samrawit, (2014) applied word sense disambiguation using semantic similarity for query expansion in Amharic information retrieval.

According to Bendersky *et al*. (2012), query reformulation is a process during which the original query issued by the user is transformed into a structured query representation that is consumed by the search engine. The process modifies the original keyword query submitted by the user to the search engine in order to better represent the underlying intent of the query. The formulated query is then used as an input to the search engine's ranking algorithm. Thus, the primary goal of query reformulation is to improve the overall quality of the ranking presented to the user in response to their query.

The performance of an information retrieval system is mainly affected by polysemous and synonymous words. If a query contains a polysemous then for one meaning the precision is affected (Jain *et al*., 2013). For example, "walk" as in "The child started to walk" and "They live at 500 High Walk". Such senses may be more or less distant from one another: walk, "action", walk, "street" are relatively close, but crane, "bird" and crane, "machine" are much further apart.

According to Wei *et al*. (2010) explained, synonymous is challenging especially when the users search information from the web. Synonymous discovery is context sensitive. Thus, web search needs to comprehend different senses in different contexts. For example, “baby” and “infant” are considered as synonymous in many thesauri, however; “Santa Baby” has no relation with “infant”. Because “Santa Baby” is a song title, and the meaning of “baby” in this phrase is different than the usual meaning of “infant”. In general, polysemous and synonymous words affect the performance of information retrieval. For example, in Afaan Oromo, the word ‘sooressa’ and ‘duuressa’ are two different words. However; they are considered as synonymous words. The phrases ‘sirna gadaa’ and ‘bara gadaa’ have two different meanings: ‘sirna’ (system) and ‘bara’ (era). Although the phrases ‘sirna gadaa’ and ‘bara gadaa’ have the same word ‘gadaa’, their meaning is different.

One of major problems of Afaan Oromo text retrieval is its effectiveness in identifying relevant documents for users’ query that satisfies their information need. Gezehagn (2012) attempted to solve these problems by developing IR system that can enable to search for relevant Afaan Oromo text corpus and also used modern information retrieval techniques that enable to solve problems related with accessing information that satisfies information needs of Afaan Oromo users. However; the system cannot search effectively for relevant Afaan Oromo text corpus and also the techniques cannot satisfies the users’ information need because of query reformulation (query expansion) problem.

Therefore, the purpose of this research work is to integrate query expansion for enhancing the effectiveness of Afaan Oromo text retrieval system.

To this end, this study tries to answer the following research questions.

- Which query expansion technique is suitable for query reformulation?

- To what extent the proposed approach enhance the performance of Afaan Oromo IR system?

1.3. Objective of the Study

1.3.1. General objective

The general objective of this study is to apply query expansion for enhance the performance of Afaan Oromo text retrieval system.

1.3.2. Specific objectives

To achieve the above general objective, the following specific objectives of study are addressed.

- To review related works so as to have a conceptual understanding of the area and Study literatures, techniques and tools applicable to query expansion.
- To design an architecture for implementing query expansion for Afaan Oromo information retrieval.
- To integrate QE model as part of information retrieval system
- To evaluate the performance of the proposed prototype using retrieval effectiveness measures.

1.4. Scope and Limitation of the Study

The focus of this study is on designing text retrieval system and integrating query expansion for enhancing the effectiveness of Afaan Oromo text retrieval system and the focus of the study is on text corpus. Image, video and graphics are out of this research. In addition, the relevance feedback, artificial intelligence techniques to predict users' information need and profiling information behavior of users are out of the scope of this study. The corpus involves documents discussing issues such as, politics, sport, economic, social, accident, health, education, tourism and justice. Lexical resource is prepared manually as a source for expansion. After word sense disambiguation is processed from Afaan Oromo WordNet which is constructed manually, query

expansion using the gloss definition method is applied to enhance the Afaan Oromo information retrieval system.

Limited corpus and queries was used for evaluating the performance of the IR system developed in the study as a result of time factor. The reason is it takes a lot of time to prepare relevance judgment for queries and corpus with many documents.

1.5. Significance of the Study

For the development of one language, technology is very important means. When technology grows in one country, the users also choose to use it with their language. In general this study is used to enhance the performance of Afaan Oromo IR system and help to understand the possible use of information associated with each query expansion techniques and advantage of lexical resource in information retrieval area. This work also put stone for the future researcher to improve the better performance of Afaan Oromo IR system by applying the rest of query expansion techniques.

1.6. Methodology

In order to realize the objectives of this research, the following methodology are used.

1.6.1. Literature review

It is very important to review other related literatures to have a deep knowledge on the domain area and the problem solving methods. Therefore, different books, journal articles, conference proceedings and Internet resources are consulted for understanding the concepts and methods related to query expansion.

1.6.2. Method of data collection and data set preparation

To carefully study QE for Afaan Oromo information retrieval, an interview and consultation are made with the language experts from linguistic department. The appropriate corpus is also collected and normalized for Afaan Oromo words as representative dataset.

To develop QE for Afaan Oromo information retrieval, a standard and representative document corpus was selected. Therefore due attention are given on the preparation appropriate document corpus for query expansion. The test queries dataset are also appropriately be selected in order to perform exhaustive evaluation of the performance of query expansion for Afaan Oromo information retrieval system. At the end a person with comprehensive knowledge of the language is consulted.

1.6.3. Development tools and techniques

The development area used is windows environment and the programming language used is Python 2.7.2. Python is incredibly efficient language that the programs do more in fewer lines of code than many other languages can require. Python's syntax is also used to write "clean" code and the code is easy to read, debug, and extend and build upon compared to other languages. It is also used heavily in scientific fields for academic research and applied work (Matthes, 1972).

1.7. Evaluation

After developing the system for QE, the implementation phase evaluates the performance level of the system and the validity of dataset employed. The evaluation also includes precision, recall and F-measure to measure the performance of the prototype system (Sodanil and Ketmaneechairat, 2013). Precision is the fraction of the documents retrieved that are relevant to the users' information need; recall is the fraction of the documents that are relevant to the query that is successfully retrieved and F-measure is the mean of precision and recall.

2. LITERATURE REVIEW

2.1. Overview of Information Retrieval (IR)

According to Manning *et al*. (2009), Information retrieval (IR) is defined as the finding relevant documents of an unstructured nature (usually text) that satisfies an information need of users from large collections.

Information retrieval refers to the extraction of user-specified information from documents and files, ranging from books to online blogs, journals, and academic articles [Manning *et al*., 2008].

As Baeza-Yates and Ribeiro-Neto (1999) explained, "Information retrieval (IR) deals with the representation, storage, organization of, and access to information items". The representation and organization of the documents should provide the user with easy access to the information in which he/she is interested. Unfortunately, characterization of the information need of users is not a simple task.

Consider, for instance, the following hypothetical user information need: Find all documents containing information on college tennis teams which are maintained by a university and participate in the National Athletic tennis tournament. To be relevant, the page must include information on the national ranking of the team in the last three years and the email or phone number of the team coach. Clearly, this full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the user must first translate this information need into a query which can be processed by the IR system. In its most common form, this translation yields a set of keywords which summarizes the description of the user information need. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. .

2.2. The Retrieval Process

According to Baeza-Yates and Ribeiro-Neto (2011), given the documents of the collection, we first apply text operations to them such as eliminating stop words, stemming and selecting a subset of all terms for use as indexing terms. The index terms are then used to compose

document representations, which might be smaller than the documents themselves (depending on the subset of index terms selected).Based on the extracted document representations, it is necessary to build an index of the text. The steps required to generate the index compose the indexing processing and must be executed offline, before the system is ready to process any queries. The resources (time and storage space) spent on the indexing process are amortized by querying the retrieval system many times.

Once the document collection is indexed, the retrieval process can be initiated. Users first specify a query that reflects their information need. The query is then parsed and modified by operations that resemble those applied to the documents. Typically operations at this point consist of spelling corrections and elimination of terms such as stop words, whenever appropriated. Next, the transformed query is expanded and modified. For example, the query might be modified using query suggestions made by the system and confirmed by the user. The expanded and modified query is then processed to obtain the set of retrieved documents, which is composed of documents that contain the query terms. Fast query processing is made possible by the index structure previously built.

The steps required to produce the set of retrieved documents constitute the retrieval process. Next, the retrieved documents are ranked according to likelihood of relevance to the user. This is the most critical step because the quality of the results, as perceived by the users, is fundamentally dependent on the ranking. The top ranked documents are then formatted for presentation to the user. The formatting consists of retrieving the title of the documents and generating snippets for them, i.e., text excerpts that contain the query terms, which are then displayed to the user.

We can see the process from the following figure 1 (Baeza-Yates and Ribeiro-Neto, 2011).

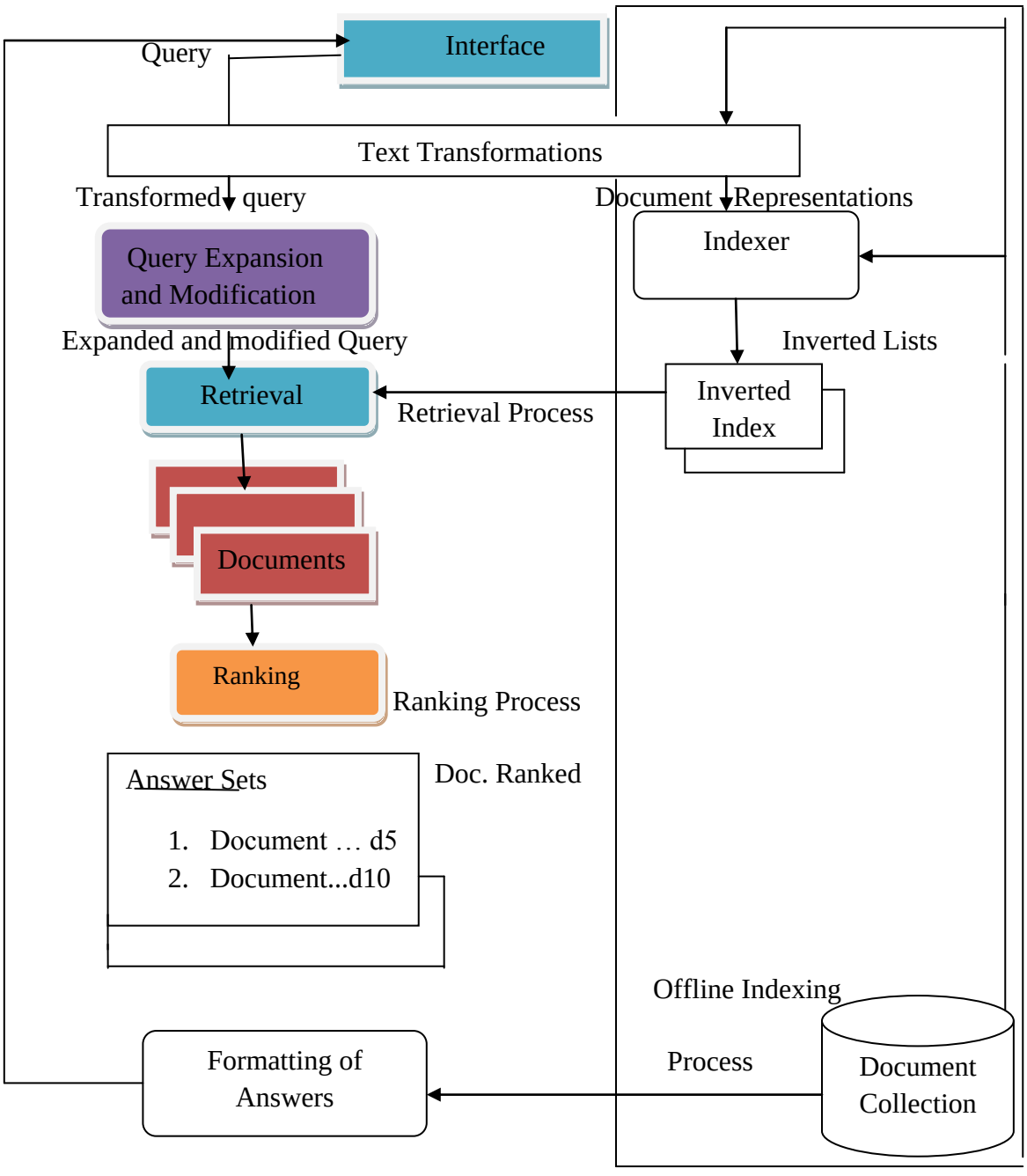

Figure 1. The process of indexing, retrieval, and ranking of documents

2.2.1. Text transformation

This step generates logical representation of documents and information need of users. To this end, the following tasks needs to be performed.

2.2.1.1. Tokenization

In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis. It is the process of splitting on white spaces and throwing away punctuation characters and tokenizes the text, turning each document into a list of tokens (Manning *et al*., 2008). For example, if the original document is: "Oromiyaan bineensotaa fi mukkeen hedduu qabdi." The tokens will be 'Oromiyaan', 'bineensotaa', 'fi' 'mukkeen,' 'hedduu,' 'qabdi'.

2.2.1.2. Normalization

Normalization operations include removal of punctuation and multiple spaces to just one space between each word, uppercase to lowercase letters, etc. Text normalization and the building of a thesaurus are strategies aimed at improving the precision of the documents retrieved (Baeza-Yates and Ribeiro-Neto, 1999).

2.2.1.3. Stop words

Stop word elimination used to be standard in older IR systems. Elimination of stop words might reduce recall (e.g. "To be or not to be" – all eliminated except "be" – no or irrelevant retrieval). According to Greengrass (2000), few terms occur frequently, a medium number of terms occur with medium frequency and many terms with very low frequency. This shows that writers use limited vocabulary throughout the whole document, in which even fewer terms used more frequently than others.

According to Debela and Ermias (2010), the compilation of stop words is also done statistically and frequently occurring content bearing words are also included. For example, barannoo,

barattoo, barnoota are varieties of the root barat (to learn), duree (rich), fayyadam (to use), dhiyeess (to approach), barsiisu (to teach), barsiisa (teacher), agarsiis (to show) and the like are included as stop word. This indicates that frequently occurring content bearing words of Afaan Oromo are not considered by the stemmer. More than 96 content bearing words that occur frequently are included as stop word.

2.2.1.4. Stemming

Stemming is language dependent that reduces tokens to their root form of words to recognize morphological variation. According to Nega and Willett (2002) explained, a stemmer that stem words without consideration of remaining stem, which removes words that are similar to prefix and suffix list but that are not actually affixes is called context-free stemmer. For instance, English word, regular‟ and „metal‟ will be “gular” and “met‟ respectively if context-free stemmer removes “re” and, “al‟.

A stemming algorithm is a procedure that reduces all words with the same stem to a common form, usually by stripping each word of its derivational and inflectional suffixes (Lovins, 1968).

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma (Manning *et al.*, 2009).

2.2.2. Indexing

Indexing is the subsystems of IR and it is an offline process of organizing documents using keywords extracted from the collection (Manning*et al.*, 2009). The indexing process involves preprocessing and storing of information into a repository (index). There are different indexing structure used for building the index file, such as inverted index, suffix tree and suffix array.

An inverted file (or inverted index) is a word-oriented mechanism for indexing a text collection in order to speed up the searching task. The inverted file structure is composed of: vocabulary,

occurrences and posting. The vocabulary is the set of all different words in the text(Baeza-Yates and Ribeiro-Neto, 1999).

According to Baeza-Yates and Ribeiro-Neto (1999), if the suffix tree indexes all text positions it can search for words, prefixes, suffixes and substrings with the same search algorithms and cost described for word search. However, indexing all positions makes the index 10 to 20 times the text size for suffix trees.

2.2.3. IR models

The IR model adopted determines the predictions of what is relevant and what is not (i.e., the notion of relevance implemented by the system) (Baeza-Yates and Ribeiro-Neto, 1999).

2.2.3.1. The Boolean model

Boolean model is a simple model based on set theory that terms are either present or absent (i.e. $w_{ij} \varepsilon \{0, 1\}$) (Baeza-Yates and Ribeiro-Neto, 1999).

The Boolean model is the first IR model that requires structural language for a query. It uses the logical operations “AND,” “OR,”and “NOT.” The Boolean model is categorized as an exact matching model. An exact matching model retrieves either all matching documents or no matching documents. In a large document collection the results might exceed one thousand documents, or there might be zero results if no matching documents 12 are found (Manning *et al*., 2008).

The Boolean model requires that the user have some knowledge of Boolean operations; therefore, not all users are able to attain satisfying results with this model. Additionally, the Boolean model has no way of favoring one document over another. In other words, it does not support document ranking. For that reason, it is less popular than other models. It is also time consuming, since it retrieves all matching documents, which in a large data collection could mean thousands of results. On the contrary, statistical models provide a score that indicates how well a document matches a query. In essence, the Boolean model builds a matrix; the rows of the matrix are the key terms, and the columns are the documents themselves. Each cell in the matrix

contains either a zero or a one. Zero indicates that the term does not occur in the specified document, and one indicates that the term is present (Singhal, 2001).

Boolean model is good model that it creates a sense of control to expert/user over the system. It is the user who is in charge for deciding what should or shouldn‘t be retrieved. Query reformulation is also simple because user is in charge of deciding what should be retrieved and should not. In contrast Boolean model may not retrieve anything if there is no matching document or, retrieves all documents if terms in query are matching with it. So there is no relevance judgment and ranking mechanism (Manning *et al*., 2009).

According to Baeza-Yates and Ribeiro-Neto (1999), since the concept of a set is quite intuitive, the Boolean model provides a framework which is easy to grasp by a common user of an IR system and the queries are specified as Boolean expressions which have precise semantics. However, retrieval strategy is based on a binary decision criterion (i.e., a document is predicted to be either relevant or non-relevant) without any notion of a grading scale, which prevents good retrieval performance (more of a data retrieval model) and while Boolean expressions have precise semantics, frequently it is not simple to translate an information need into a Boolean expression, in fact, most users find it difficult and awkward to express their query requests in terms of Boolean expressions.

2.2.3.2. Vector space model

The representation of a set of documents as vectors in a common vector space is known vector space as the vector space model and is fundamental to a host of information retrieval m o d e l (IR) operations including scoring documents on a query, document classification, and document clustering(Manning *et al*., 2008).

According to Baeza-Yates and Ribeiro-Neto (1999), Vector space model is the most commonly used strategy for measuring relevance of documents for a given query because of using binary weights is too limiting and non-binary weights provide consideration for partial matches. Its cosine ranking formula sorts the documents according to their degree of similarity to the query. However, it assumes that index terms are mutually independent and yields ranked answer sets

which are difficult to improve upon without query expansion or relevance feedback within the framework of the vector model.

$$sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right\|\vec{q}\,|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \quad \text{..... (2.1)}$$

Where $\vec{d}_j$ and $\vec{q}$ are the norms of the document and query vectors and w is the weight of the terms.

Vector space model is representation of index terms and query as vectors embedded in a high dimensional Euclidean space, where each terms is assigned as a separate dimension.

$d_j=(w_{1j},w_{2j},w_{tj})$ ……...(2.2)

$q= (w_{1q},w_{2q},w_{tq})$ ……(2.3)

Indexing of the document in the way that only content bearing terms represent the document, weighting the indexed terms to enhance retrieval of relevant document and ranking the documents to show best matching with respect to the provided query by user are the main procedures VSM (Manning *et al*., 2009).

2.2.3.3. Probabilistic model

According to Baeza-Yates and Ribeiro-Neto (1999), documents are ranked in decreasing order of their probability of being relevant. However, probabilistic model has disadvantages. For example: the need to guess the initial separation of documents in to relevant and non-relevant sets, The fact that the method does not take into account the frequency with which an index term occurs inside a document (i.e., all weights are binary) and adoption of the independence assumption for index terms. “Given a user query, there is a set of documents which contains exactly the relevant documents and no other. Suppose this set of documents as the ideal answer set. Given the description of this ideal answer set, we would have no problems in retrieving its documents. Thus, we can think of the querying process as a process of specifying the properties

of an ideal answer set (which is analogous to interpreting the IR problem as a problem of clustering). The problem is that we do not know exactly what these properties are. All we know is that there are index terms whose semantics should be used to characterize these properties. Since these properties are not known at query time, an effort has to be made at initially guessing what they could be. The initial guess allows us to generate a preliminary probabilistic description of the ideal answer set which is used to retrieve a first set of documents. An interaction with the user is then initiated with the purpose of improving the description of the ideal answer set: The user takes a look at the retrieved documents and decides which ones are relevant and which ones are not (in truth, only the first top documents need to be examined). The system then uses the information to refine the description of the ideal answer set. By the repeating this process many times, it is expected that such a description will evolve and become closer to the real description of the ideal answer set. Thus one should always have in mind the need to guess at the beginning the description of the ideal answer set. Further, a conscious effort is made to model this description in probabilistic terms" (Baeza-Yates and RibeiroNeto, 2011).

2.3. Query Expansion

Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations (Vectomova and Wang, 2006). QE majorly comprises of augmenting certain terms to the query such that the final query will reveal near to the desired results.

For example, users might want information about Information Retrieval and they give the query "Information Retrieval", then seeing the results they modify their query – to specify whether it is a course or research topic; accordingly, their new query would be something like "Information retrieval Course".

Query expansion can be categorized into two categories based on the source of expansion terms (Chowdhury *et al*., 2002). In the first category, query expansion is based on relevance feedback (Uzuner, 1998). The expansion is done using the processed result of relevance feedback. For example, query is expanded using most frequent terms occurring in a set of documents (implicit

WSD). In the second category, query expansion is based on knowledge structures like corpora (Grootjen and Van der Weide, 2006), thesauri (Voorhees, 1993), dictionaries, or the combination of them (Hozumi et al., 1998). For example, query is expanded using words from the same synsets in thesaurus (it needs explicit WSD). Most query expansion methods nowadays use relevance feedback (Manning *et al*., 2009).

2.4. Query Expansion Techniques and Approaches

As depicted below in figure 2 there are different techniques and methods available for designing query expansion.

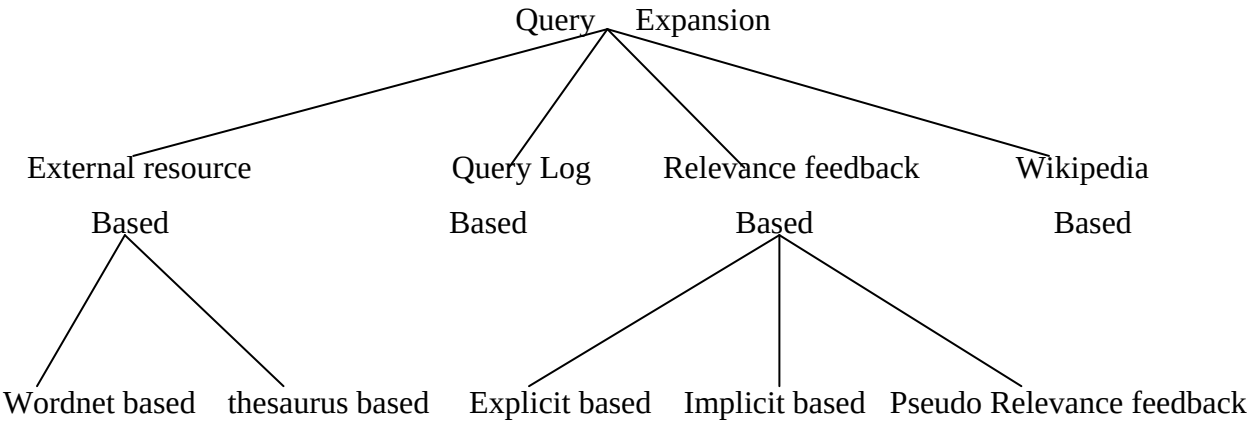

Figure 2. Query Expansion Techniques (Kankaria, 2005)

2.4.1. External resource based query expansion

External resource based query expansion helps users to find synonyms for query terms initially and later helps users fine good query terms. In these approaches, the query is expanded using some external resource like Word Net, lexical dictionaries or thesaurus. These dictionaries are built manually which have mappings of the terms to their relevant terms. There techniques involve look up in such resources and adding the related terms to query. Following are some of external resource based query expansion techniques (Carpineto and Romano, 2012).

Thesaurus based expansion

A thesaurus is a data structure that supports the automatic indexing and retrieval, and it is a structured dictionary, which is paying attention on the representation of a limited set of semantic relations between different concepts (Robert, 2009). It lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words and generally lists them in alphabetical order. In this approach thesaurus is used to expand the query terms and all the connected words of query terms are added to query. Thesaurus based system have been discovered and put to use by many organizations. A well known instance for such systems is Unified Medical Language System (Bodenreider, 2004) used with MedLine for querying the bio medical research literature.

WordNet based expansion

WordNet is a lexical database for English languages (Miller *et al*., 1990). It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser, its primary use is in automatic text analysis and artificial intelligence applications.

WordNet is also used by Smeaton *et al.* (1995) along with POS tagging for query expansion. Each query term was expanded independently and equally. The interesting thing is they completely ignored the original query terms after expansion. As results of this precision and recall dropped but they were able to retrieve documents which did not contain any of the query terms but are relevant to the query. WordNet is used for sense evaluation and for similarity measure in WSD (Smeaton, 1995).

In WordNet similarity (S_{def} (t_1, t_2)) is word overlap between glosses of all synsets divided by total of words in all synsets glosses and relation similarity gets value it terms are synonyms, hypernyms, holonoyms, or meronyms.

According to Sanderson (1996) explained, a synsets is a set of words that are synonyms of each other and together, these words define the synsets and its meaning. The synsets, to which a

particular word is assigned, constitute the individual senses of that word. These synsets are linked to form a semantic network, an example fragment of which is shown in Figure 3. As can be seen the links between synsets are formed by semantic relations, the most prevalent of which are the two complementary hierarchical relations, the hypernym or is-a relation (e.g. a cabin is a type of house), and the hyponym or instance-of relation (e.g. the class of houses has an instance of the type cabin). There are three other relations used to link synsets in the semantic network: the meronym or has-part/has-member relation (e.g. a house has a part attic); the holonym or member-of/part-of relation (e.g. an attic is a part of a house); and the antonym or is-opposite relation (e.g. black is the opposite of white).

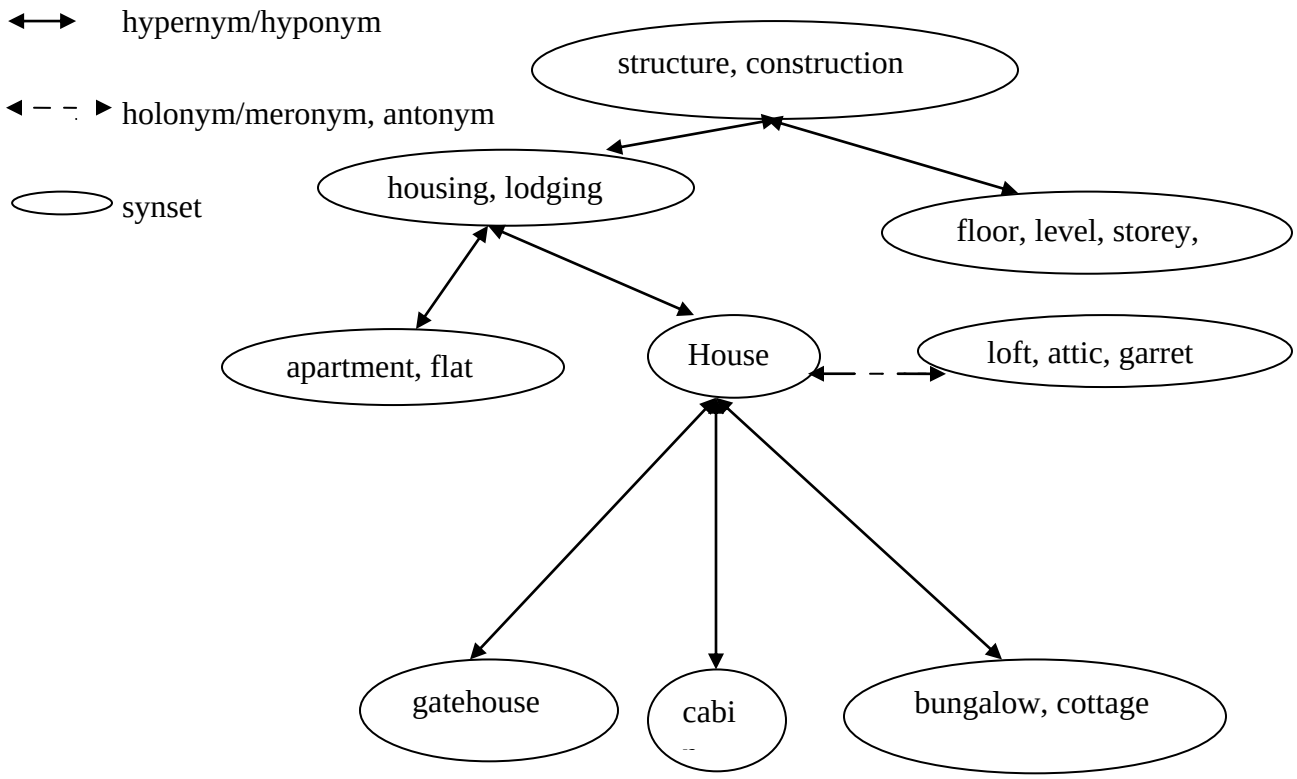

Figure 3. Fragment of the WordNet semantic network (Sanderson, 1996)

According to Banerjee (2002), WordNet is like a dictionary in that it stores words and meanings. However it differs from traditional ones in many ways. For instance, words in WordNet are arranged semantically instead of alphabetically. Synonymous words are grouped together to form synonym sets, or synsets. Each such synsets therefore represents a single distinct sense or concept. Thus, the synsets base, alkali represents the sense of any of various water-soluble

compounds capable of turning litmus blue and reacting with an acid to form a salt and water. Words with multiple senses can either be homonymous or polysemous. Two senses of a word are said to be homonyms when they mean entirely different things but have the same spelling. For example the two senses of the word bark tough protective covering of trees and the sound made by a dog are homonyms because they are not related to each other. A word is said to be polysemous when its senses are various shades of the same basic meaning. For example, the word accident is polysemous since its two senses – a mishap and anything that happens by chance are somewhat related to each other. Note that WordNet does not distinguish between homonymous and polysemous words, and therefore neither do we. Thus WordNet does not indicate that the two senses of the word accident is somewhat closer to each other in meaning than the two senses of the word bark. Words with only one sense are said to be monospermous. For example, the word wristwatch has only one sense and therefore appears in only one synsets. In WordNet, each word occurs in as many synsets as it has senses. For example the word base occurs in two noun synsets, {base, alkali} and {basis, base, foundation, fundament, groundwork, cornerstone}, and the verb synsets {establish, base, ground, found}.

WordNet stores information about words that belong to four parts–of–speech: nouns, verbs, adjectives and adverbs. Prepositions and conjunctions do not belong to any synsets (Banerjee, 2002).

WordNet defines a variety of semantic and lexical relations between words and synsets. Semantic relations define a relationship between two synsets (Banerjee, 2002). For example, the noun synsets {robin, redbreast, robin redbreast} is related to the noun synsets {bird} through the IS–A semantic relation since a robin is a kind of a bird. Lexical relations on the other hand define a relationship between two words within two synsets of WordNet. Thus whereas a semantic relation between two synsets relates all the words in one of the synsets to all the words in the other synsets, a lexical relationship exists only between particular words of two synsets. For example the antonym relation relates the words embarkation and disembarkation but not the rest of the words in their respective synsets which are {boarding, embarkation, embarkment}and {debarkation, disembarkation, disembarkment }.

2.4.2. Query logs based expansion

Query logs are maintained by each search engine in order to analyze the behavior of the user while interacting with search engine. These kinds of approaches use these query logs to analyze the user's preference and adds corresponding terms to query. This method can fail when user wants to search something which is not at all related to earlier searches.

Cui *et al*. (2002) developed a system which extracts the expansions terms based on user's behavior which is stored in form of query logs. They maintained a list of all the documents visited for a particular query. Probability of document being visited when a particular query word is present in a query is calculated to find the relevance of the document.

Yin *et al*. (2009) considered query log as bipartite graph that connects the query nodes to the Uniform Resource Locator (URL) nodes by click edges. Given a query node q and a URL node u, there will be an edge (q, u) if u is among the clicked answers for query q. They have recorded more than 10 percent improvement over the baseline.

2.4.3. Relevance feedback based expansion

The relevance feedback process, introduced in the mid- 1960s is a controlled, automatic process for query reformulation, that is easy to use and can prove unusually effective. The main idea consists in choosing important terms, or expressions, attached to certain previously retrieved documents that have been identified as relevant by the users, and of enhancing the importance of these terms in a new query formulation (Salton and Buckley, 1990).

Relevance feedback shields the user from the details of the query reformulation process because all the users has to provide is a relevance judgment on documents and breaks down the whole searching task into a sequence of small steps which are easier to grasp (Baeza-Yates and Ribeiro-Neto, 2011).

The relevance feedback loop requires the user to enter an initial query which results in a display of ranked documents (usually titles/abstracts). From this display, the user makes relevance judgments and selects the relevant documents. The relevant terms from these documents are

added to the initial query. To do such process the relevance feedback should consider the term selection, how to weight the new terms, whether to exclude the original query terms, whether to include all of the new terms or just some of them and if so how many new terms to include (Bhogal, 2011).

Explicit feedback from user

This interactive approach in which the initial retrieved documents are presented to user and the user is asked to select the relevant documents. These models are not much useful because users expect the system to be autonomous and retrieve the results for the user. User would ultimately get irritated by repeated interaction required from him for each search. These types of models can be used for testing search engines where developers are willing to interact with the system. In a classic relevance feedback cycle, the user is presented with a list of the retrieved documents and, after examining them, marks those that are relevant (Baeza-Yates and Ribeiro-Neto, 2011).

Implicit feedback

This is a type of model in which the user's feedback is inferred by the system. The feedback can be inferred from user's behavior like: The pages which user opens for reading, or pages on which user clicks once the results are displayed back to the user. In an implicit relevance feedback cycle, there is no participation of the user in the feedback process (Baeza-Yates and Ribeiro-Neto, 2011).

Pseudo relevance feedback

In Pseudo Relevance Feedback based models initial query is fired and top k results are obtained. Then important terms, mostly based on co-occurrence, from these documents are extracted and added to query. Then this expanded query is re fired to retrieve final set of documents which are made available to the user. The relevancy of expansion terms depends upon the initial retrieved documents. In PRF the system simply assumes that its top-ranked documents are relevant and uses these documents to augment the query with a relevance feedback ranking (Baeza-Yates and Ribeiro-Neto, 1999).

2.5. Measures of Word Semantic Similarity

The algorithm bases its disambiguation decisions on the semantic similarity of words. This similarity is measured by utilizing the IS–A relationships defined between synsets in WordNet. Two words are said to be most similar when they are synonymous, that is when some sense of one words in the same synsets as some sense of the other word. Word is called an extended synonym of word if some sense of belongs to a synsets that is the hypernym of the synsets that some sense of word belongs to. Word is considered hyponymous with word when some sense of is in a synsets that is the hyponym of the synsets that some sense of belongs to. Finally is said to have a coordinate relationship with if some sense of and some sense of belong to synsets that have a common hypernym (Banerjee, 2002).

Meng and Gu (2012) used different approaches in semantic similarity. For example, making use of a large corpus and gathering statistical data from corpus to estimate a score of semantic similarity is the one approach and the use of the relations and the hierarchy of a thesaurus, which is generally a hand-crafted lexical database such as WordNet, is another approach they used.

According to Torres and Gelbukh (2009), determining the degree of semantic similarity, or relatedness between two words is an important problem in NLP. Similarity measures are used in such applications as word sense disambiguation, determining discourse structure, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and automatic correction of word errors in text.

Semantically related terms that are found in each category are listed out from the list of similar synsets. Their frequencies are then obtained from the index file. Therefore, in the example shown, there are two features selected to represent the category. They are ‘corn’ and ‘maize’. The rationale behind choosing synonymous terms in representing a category is that these terms have a stronger discriminative power in representing a category; as they are conceptually close (Chua and Kulathuramaiyer, 2004).

Chua and Kulathuramaiyer (2004) used the following principle in semantic similarity method. Appropriate senses of the words (which are in context) can be determined by finding the semantic distance between different senses of words (in context) and choosing those which are at

least distance. It is important to point out that the context, talked above, can be local or global. In local context, the window is restricted to few words around the target word. And in global context it consists of lexical chains which are sequence of semantically related words and corresponds to coherence of discourse. The meaning of the words is identified based on their membership of lexical chain.

Liu *et al*. (2011) classified WordNet based semantic similarity measures as the following categories. The first category is node-based methods. Node-based methods estimate semantic similarity by computing the amount of information contained by related words in WordNet and in which the main data sources are the nodes and their properties. This kind of method is also called as information-based methods. According to the definition in the information theory, the Information Content (IC) of a concept c can be quantified by the following formula.

IC(c)=-log(P(c)), where P(c)is the probability of *c* appearing in a corpus.

The drawbacks of node-based methods include, first it is a time-consuming work to analysis the corpora for estimating the IC values and second unbalanced contents of the employed corpora may significantly decrease the accuracy of the IC values. The second category is edge-based methods. Edge-based methods assess the semantic similarity by calculating the length of edges on the shortest path between the words in WordNet and use the edges and their types as the data source. Edge-based methods utilize the shortest path between concepts (i.e., *c1* and *c2*) in WordNet to estimate the semantic relatedness between *c1* and *c2*. Lengths of all edges on the shortest path are accumulated to quantify the semantic similarity. It is the way of calculating the length of edges that differentiates methods in this category. The accuracy of the edge-based methods is significantly affected by the lack of considering the varieties of semantic distances between adjacent words, which is caused by the uneven word densities in WordNet. The third category is hybrid methods. Hybrid methods combine the information from different resources to estimate the semantic similarity between concepts. For example, combining the *IC* of concepts with the structure information retrieved from WordNet to conduct the estimation.

2.5.1. Resnik's information content

There are two main Resnik's measures (Resnik, 1995). These are information content based relatedness measure and higher information content specific to particular topics, lower ones specific to more general topics.The related value is equal to the information content (IC) of the Least Common Subsumer (LCS) (most informative subsumer). This means that the value will always be greater-than or equal-to zero. The upper bound on the value is generally quite large and varies depending upon the size of the corpus used to determine information content values. To be precise, the upper bound should be ln (N) where N is the number of words in the corpus.

Resnik (1995) brings together ontology and corpus and guided by the intuition that the similarity between a pair of concepts may be judged by "the extent to which they share information". He defined the similarity between two concepts lexicalized in WordNet to be the information content of their lowest super-ordinate. He argues that the links in the hierarchy of WordNet representing a uniform distance in the edge-counting measurement cannot account for the semantic variability of a single link. He assumed that for a concept c, let p(c) be the probability of encountering and instance of concept c. The IC value is obtained by considering the negative log likelihood.

2.5.2. The Jiang and Conrath Measure

Jiang and Conrath (1997) approach captures the Information content (IC) of the two concepts along with the Information content of most specific common Subsumer. It basically calculates the distance between two concepts:

$$\text{Distance}_{jcn} = IC(C1) + IC(C2) - 2*IC(LCS)(C1, C2))$$

Where IC is information concept, LCS is Least Common Subsumer, C1 and C2 are concepts, LCS (C1, C2) is lowest node in hierarchy that is a hypernym of C1, C2 and jcn is Jiang and Conrath.

Distance $_{jcn}$ measure give the measure of un-relatedness between the two concepts, high score indicate low similarity and low score indicate high similarities (Gupta and Yadav, 2014).

The edge length between concept c (a node in the shortest path) and concept p (the parent node of c in the shortest path) is calculated by:

$length(c, p) = \log(P(p)) - .\log(P(c))$.

They also considered the link type, depth, conceptual density, and information content of concepts.

2.5.3. The Lin Measure

According to Lin (1998), the similarity value returned by the Lin measure is a number equal to 2 * IC (lcs) / (IC (concept1) + IC (concept2)). Where, IC (x) is the information content of x. One can observe, then, that the similarity value will be greater-than or equal-to zero and less-than or equal-to one. If the information content of any of either concept1 or concept2 is zero, then zero is returned as the similarity score, due to lack of data. Ideally, the information content of a concept would be zero only if that concept were the root node, but when the frequency of a concept is zero; we use the value of zero as the information content because of a lack of better alternatives. His measure of semantic relatedness of concepts is based on his Similarity Theorem. It states that the similarity of two concepts is measured by the ratio of the amount of information needed to state the commonality of the two concepts to the amount of information needed to describe them. The commonality of two concepts is captured by the information content of their lowest common subsumer and the information content of the two concepts themselves.

2.5.4. The Lesk measure

The Lesk algorithm is a classical algorithm for word sense disambiguation (Lesk, 1986). Lesk proposes to measure the relatedness between two concepts by the overlap between the corresponding definitions of them, as provided by a dictionary. The application of the Lesk similarity measure is not limited to semantic networks, and it can be used in conjunction with any dictionary that provides word definitions (Torres and Gelbukh, 2009).

The Lesk Measure works by finding overlaps in the extended definitions of the two concepts. The relatedness score is the sum of the square of the overlap lengths. For example, a single word overlap results in a score of 1. Two single word overlaps result in a score of 2. A two word

overlap (i.e., two consecutive words) results in a score of 4. A three word overlap results in a score of 9.

2.6. Word Sense Disambiguation

Word Sense Disambiguation(WSD) is the task of separating the sense of word in textual context and the process of association of given word with meaning (sense) in context that is distinguishable from other meaning attributable to that word (Jain *et al*., 2013).

Jain *et al*. (2013) put WSD task in two steps: The first step is Sense Repository. The Sense Repository identifies all the different meanings of all the words relevant to the text under consideration. It may be from list of senses in dictionaries, from synonyms in thesaurus, from translations in a translation dictionary. The second step is Sense Assignment. Sense Assignment involves the assignment of appropriate sense to each occurrence of word in textual context.

According to Chua and Kulathuramaiyer (2004), Word sense disambiguation (WSD) is defined as "the process of disambiguating words by telling which sense an ambiguous word belongs to".

According to Paskalis and Khodra (2011), ambiguity in information retrieval can bring the problem of the retrieval of irrelevant documents, while different words which represent the same concept can bring the problem of the retrieval system cannot find all of the relevant documents. These problems can decrease the information retrieval performance system. If the query expanded is using wrong sense information, the chance in which the result would be wrong is big. Because ambiguity is one of the factors that decrease retrieval system performance, retrieval performance can be improved if ambiguity can be solved (Krovetz, 1997). Also, if ambiguities in query can be improved, the problem occurring from different words with the same concept can also be solved using query expansion. Thus, query should be disambiguated before query expansion is done. This process is called word sense disambiguation (WSD). After the senses are determined, senses information is used for query expansion (Liu *et al*., 2005).

2.6.1. Word sense disambiguation approaches

There are two basic approaches of WSD (Jain and Nathawat, 2012), such as dictionary based and machine learning approaches.

2.6.1.1. Machine learning approaches

In these approaches, what is learned is a classifier that can be used to assign as yet unseen examples to one of a fixed number of senses and systems are trained to perform the task of word sense disambiguation.. These approaches vary as the nature of the training material, how much material is need, the degree of human intervention, the kind of linguistic knowledge used, and the output produced. But the system accuracy can definitely be improved by machine learning methods. These approaches can be mainly classified into two (Kumar *et al*., 2012), supervised and unsupervised methods.

Supervised approaches use the sense-annoted corpora to train from and generate the classifier system. Classifier is used for the classification of word to assign appropriate sense to each instance of that word. The training set is used to learn the classifier. Target word is manually tagged with sense from sense inventory. Supervised methods give better results than unsupervised approaches. Bayesian classification and Information Theory are the supervised algorithms applied to WSD in statistical language (Agirre and Edmonds, 2006).

Unsupervised methods work on raw unannoted corpora. Word senses are induced from input text by clustering word occurrences and then classifying new occurrences into the induced clusters. These methods do not use any dictionaries, thesauri, ontologies etc. In these sense is labeled externally to target word. It is the greatest challenge for WSD researchers. The underlying assumption is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context, a task referred to as word sense induction or discrimination. Then, new occurrences of the word can be classified into the closest induced clusters/senses. Performance has been lower than for the other methods described above, but comparisons are difficult since senses induced must be mapped to a known dictionary of word senses (Schutze, 1998).

Semi-Supervised methods allow both labeled and unlabeled data. These use small annoted corpus as seed. Seed is used to train initial classifier using any supervised method. Initial classifier extract large training set from the remaining untagged corpus. Thus on repeating this process we will get a series of classifier until the whole corpus is consumed or given maximum number of iteration is reached.

The Yarowsky (1995) algorithm was an early example of such an algorithm. It uses the 'One sense per collocation' and the 'One sense per discourse' properties of human languages for word sense disambiguation. From observation, words tend to exhibit only one sense in most given discourse and in a given collocation.

2.6.1.2. Dictionary and knowledge based methods

Dictionary and Knowledge based methods depend primarily on the dictionaries, thesauri, ontology and lexical knowledge bases to retrieve different senses of word in context. Main knowledge based techniques are: the overlap of sense definition, selection restrictions and structural approaches. Most approaches use the WordNet as sense inventory.

According to Kumar *et al*. (2012), knowledge based approaches uses Lesk algorithm, walker's algorithm, random Walk algorithm for similarity measures.

Lesk algorithm

"The Lesk algorithm (Lesk, 1986) is one of the first algorithms developed for the semantic disambiguation of all words in unrestricted text" (Agirre and Edmonds, 2006). According to him, a word is disambiguated by comparing the gloss of each of its senses to the glosses of every other word in the phrase. The sense whose gloss shares the largest number of words in common with the glosses of other words is selected as the correct sense. The only resource required by the algorithm is a set of dictionary entries, one for each possible word sense, and knowledge about the immediate context where the sense disambiguation is performed.

(1) for each sense *i* of W_1
(2) for each sense *j* of W_2
(3) compute Overlap (i, j), the number of words in common between the definitions
of sense *i* and sense *j*
(4) find i and j for which Overlap (i, j) is maximized
(5) assign sense *i* to W_1 and sense *j* to W_2

Figure 4. Dictionary-based Lesk algorithm

Although traditionally considered a dictionary-based method, the idea behind the Lesk algorithm represents the starting seed for today's corpus based algorithms. The main idea behind the original definition of the algorithm is to disambiguate words by finding the overlap among their sense definitions. Namely, given two words, W1 and W2, each with NW1 and NW2 senses defined in a dictionary, for each possible sense pair W1 *i* and W2j, i = 1..NW1, j = 1..NW2, first determine the overlap of the corresponding definitions by counting the number of words they have in common. Next, the sense pair with the highest overlap is selected, and therefore a sense is assigned to each word in the initial word pair. Fig 2.4 illustrates the main steps of the algorithm (Agirre and Edmonds, 2006).

Walker's algorithms

The algorithm of Walker (Walker II, 1990) is widely used for drawing trees of unbounded degree, and it is widely assumed to run in linear time, as the author claims in his article. But the presented algorithm clearly needs quadratic run time.

According to Kumar *et al*. (2012), the Walker's algorithm is put as following. The first stepis seeking the thesaurus category to which that sense belongs. Then calculate the score for each sense by using the context words. Contexts will be added to the score of the sense if the thesaurus category of the word matches that of the sense. By using WordNet, it is possible to find the conceptual distance by analyzing the hyponyms. Once we find out the conceptual distance, conceptual density can be measured. If the conceptual distance is smaller, conceptual

density will be higher. Let w be the word to be disambiguated. w1, w2, w3wn etc are the words in context. Each symbol represents the different senses of the word in context. Highest density will be obtained for the sub hierarchy containing more senses.

Random walk algorithm

In random walk approach, a vertex is formed for each possible sense of each word in a text. By using definition base similarity, we can add weighted edges. A graph based ranking algorithm is then applied to find score of each vertex. Then the highest score vertex is selected as the correct sense (for each word) (Kumar *et al*., 2012).

2.7. Afaan Oromo Language

Afaan Oromo is a Cushitic language spoken by about 40 million people in Ethiopia (about 40% of the country's population), in Kenya, Somalia and Djibouti and is the 3rd largest language in Africa after Arabic and Hausa. Afaan Oromo (meaning Oromo language) or Oromiffa, most probably rates second among the African indigenous languages. Even before two decades when Gada (1988) explored the history of this nation, Afaan Oromo was known as the mother tongue of about 30 million Oromo people living in the Ethiopian and neighboring countries such as Kenya, Somalia, and Djibouti.

2.7.1. Alphabets and sounds -qubeelee fi sagaleewwan

Afaan Oromo is phonetic language in which its characters sound is the same in every word in contrast to English language. Afaan Oromo was written with either the Ge'ez script or the Latin alphabet until the 1970s. Starting from 1974-1991 writing any piece of writing using any script by Afaan Oromo language was protected to be official. Since 1991 Latin alphabet is used as official alphabet of Oromo Language (Ager, 2012). Afaan Oromo uses Latin character (Roman alphabet) but with some modifications on sound of consonant and vowels. It has 28 letters called 'qubee'. However, later on a new letter 'Z' was included in the alphabet as there are words which require the letter. Alphabets sound is also included with how it pronounced in English words ((Ager, 2012)).

Table 1. Upper Case, lower case and their sounds (Gezehagn, 2012)

Alphabets	**Sounds**	**Alphabets**	**Sounds**	**Alphabets**	**Sounds**	**Alphabets**	**Sounds**	**Alphabets**	**Sounds**	**Alphabets**	**Sounds**
A a	*[aa]* like **ask**	**B b**	*[baa]* like **bi**rd	**C c**	*[Caa]* like **ca**t	**D d**	*[daa]* like **da**m	**E e**	*[ee]* like **a**te	**F f**	*[ef]* like **fu**ngi
G g	*[gaa]* like **gu**n	**H h**	*[haa]* like **ha**t	**I i**	*[ie]* like **I**ndi**a**	**J j**	*[jaa]* like **Ju**st	**K k**	*[kaa]* like **Ca**st	**L l**	*[la]* like **li**fe
M m	*[ma]* like **ma**n	**N n**	*[naa]* like **na**sty	**O o**	*[oo]* like **o**ld	**P p**	*[pee]* like **pa**st	**Q q**	*[quu]* like **quit**	**R r**	*[ra]* like **ra**t
S s	*[saa]* like **sa**lad	**T t**	*[taa]* like to**ta**l	**U u**	*[uu]* like **u**rge	**V v**	*[vau]* like **va**ry	**W w**	*[wee]* like **wa**nt	**X x**	*[taa]* like ___
Y y	*[y]* like **you**th	**Z z**	*[Zay]* like **Tha**t	**CH ch**	*[chaa]* like **cha**t	**DH dh**	*[dhaa]* like ___	**SH sh**	*[shaa]* like **sh**y	**NY ny**	*[nyaa]* like _____
PH ph	*[phaa]* like										

2.7.2. Vowels -dubbachiiftuu

Afaan Oromo vowels are similar with English vowels and represented by the five letters, **a, e, o, u and i.** All vowels are pronounced basically the same way throughout Oromia. These vowels when stressed may be opened: deemuu (go), nyaadhu(eat) or closed: b*a*da, r*a*fi.

The Afaan Oromo vowels always are pronounced in sharp and clear fashion which means each and every word is pronounced strongly, for example:

A: Amala, Gamna, Jabaa

E: Eegee, Ijoollee, Roobalee

I: Amajji, Ilaali, Deemi, Dhiisi

O: Oromoo, Haaloo, Haroo

U: Utubuu, Guddadhu, Beekuu

Table 2. Dubbiistoota (vowels) (Gezehagn, 2012)

Vowels			
	Front	**Central**	**Back**
Close	i /ɪ/, ii /iː/		u /ʊ/, uu /uː/
Mid	e /ɛ/, ee /eː/		o /ɔ/, oo /oː/
Open		a /ʌ/	aa /ɑː/

2.7.3. Consonants -Sagaleewwan dubbifamtootaa

Most Afaan Oromo constants do not differ greatly from Italian, but there are some exceptions and few special combinations.

i. The consonant “g” has a hard sound. Gaarii, gad-bayi, gargaari.

ii. The combinations NY and DH have a hard sound. e.g Nyaadhu, Dhugi.

Table 3. Dubbifamtoota (consonants) (Gezehagn, 2012)

Consonants						
		Bilabial/ Labiodental	**Alveolar/ Retroflex**	**Palato-alveolar/ Palatal**	**Velar**	**Glottal**
Stops and Affricates	Voiceless	(p)	T	ch /tʃ/	K	' /ʔ/
	Voiced	B	D	j /dʒ/	G	
	Ejective	ph /p’/	x /t’/	c /tʃ’/	q /k’/	
	Implosive		dh /ɗ/			
Fricatives	Voiceless	F	S	sh /ʃ/		H
	Voiced	(v)	(z)			
Nasals		M	N	ny /ɲ/		
Approximants		W	L	y /j/		
Rhotic			R			

2.7.4. Double consonants - qubee dubbifamaa dachaa

All Afaan Oromo consonants except the combination consonants ny, dh, ph, and sh have double consonant combinations if the syllable is stressed. Failure to make this distinction results in miscommunication. Examples: Arba, Joortuu, Malaammmaltummaa, walqixummaa, walabummaa.

2.7.5. Stress - qubee jabaataa

Afaan Oromo words do have stress, or emphasis, which is placed within its syllables. Some Afaan Oromo words are pronounced with the stress on the last syllable: e.g fannoo, harree, gaarree.

On the other hand, few words are stressed on the first syllable. These words always have a combination consonant: e.g nyaadhu, dhayi, nyaara and nyaapha (foreigner). There are a few words where this is not the case, but for the most part the second to last syllable should always carry the emphasis when speaking.

2.7.6. Grammar- Seer-luga

Afaan Oromo language has its own grammar as other languages which are called ‘seer-luga’.

2.7.6.1. Numbers

Numbers come after the noun they modify, so that “two mangoes” is “maangoo lama”, just as “five birr” is “qarshii shan” and 200 is dhibba lama. Ordinal numbers are formed by adding the suffix -ffaa or -affaa to the number. Fractions can be expressed by saying the numerator as a cardinal number and then the denominator as an ordinal number (MYLANGUAGES, 2011)

2.7.6.2. Definiteness

Where English uses “the” to indicate definiteness (a specific something of shared knowledge), Oromo drops the final vowel and uses the suffix -*(*t)icha for masculine nouns and -*(t)ittii* for feminine nouns. Making a noun definite is less common in Oromo than in English, and is used

only for objects known to both the speaker and the listener. A noun can be either definite or pluralized, but not both. A definite noun is therefore ambiguous in number, and context determines if it is singular or plural. Definite nouns are not modified by demonstrative pronouns or possessive pronouns. If modified by an adjective, the definite marker is attached to the adjective (discussed in the next chapter). Example:

Base noun (dictionary form)	Definite noun
- Nama-man	- namich-the man (men)
- Muuzii-banana	- muuzicha-the banana (s)
- Durba-girl	- durbartittii-the girl (s)

2.7.6.3. Personal pronoun

Oromo pronouns include personal pronouns (refer to the persons speaking, the persons spoken to, or the persons or things spoken about), indefinite pronouns, relative pronouns (connect parts of sentences) and reciprocal or reflexive pronouns (in which the object of a verb is being acted on by verb's subject) (Gezehagn, 2012).

Table 4. Afaan Oromo personal pronouns (Gezehagn, 2012)

English	**Base**	**Subject**	**Dative**	**Instru Mental**	**Locative**	**Ablative**	**Possessive Adjectives**
I	ana, na	ani, an	naa, naaf, natty	Naan	Natty	Narraa	koo, kiyya [too, tiyya (f.)]
You (sg.)	Si	Ati	sii, siif, sitti	Siin	Sitti	Sirraa	Kee [tee (f.)]
He	Isa	Inni	isaa, isaa(tii)f, isatti	isaatiin	Isatti	Isarraa	(i)saa
She	isii, ishii, isee, ishee	isiin, etc.	ishii, ishiif, ishiitti, etc.	ishiin, etc.	ishiitti, etc.	ishiirraa, etc.	(i)sii, (i)shii
We	Nu	nuti, nu'i, nuy, nu	nuu, nuuf, nutty	Nuun	Nutty	Nurraa	keenna, keenya [teenna, teenya (f.)]
You (pl.)	Isin	Isini	isinii, isiniif, isinitti	Isiniin	Isinitti	isinirraa	keessan(i) [teessan(i) (f.)]
They	Isaan	Isaani	isaanii, isaaniif, isaanitti	isaaniitiin	Isaanitti	isaanirraa	(i)saani

2.7.6.4. Adjectives

Oromo Adjectives are words that describe or modify another person or thing in the sentence and are very important in Afaan Oromo because its structure is used in every day conversation (MYLANGUAGES, 2011). Examples:

Table 5. Afaan Oromo adjective

Colors-bifa	size-hamma Qulqullina	shape-boca	taste-dhamdhama	quality-
English-Afaan Oromo	English-Afaan Oromo	English-Afaan Oromo	English-Afaan Oromo	English-Afaan Oromo
Black-gurracha	Big-gudda	Circular-geengoo	Bitter-hadhaawaa	Bad-hamaa
Blue-baluu	Deep-gadi fagoo	Straight-sirrii	Fresh-asheeta	Clean-qulqulluu
Brown-bifa bunaa	Long-lafarra dheeraa	Square-golarfee	Salty-sogiddawaa	Dark-dukkanaawaa
Gray-daalacha	Narrow-dhiphaa	Triangular-golsadee	Sour-ogonnawaa	Difficult-ulfaataa
Green-magariisa	Short-gabaabaa	E.g.	Spicy-kaurgoo baayyatu	Dirty-xuraawaa
Red-diimaa	Small-xinnaa	The circular house-manicha geengoo	Sweet-mi’awaa	Dry-qooraa
White-adii	Tall-dheeraa		E.g.	Easy-salphaa
E.g. green trees – muka magariisa	Thick-yabbuu		A salty expensive food-nyaata hadhaawaa	Empty-duwwaa

2.7.6.5. Adverbs- Ibsa xumuraa

Oromo Adjectives are words that describe or modify another person or thing in the sentence and they are the words that modify any part of language other than a noun (MYLANGUAGES, 2011).

Table 6. Adverbs in Afaan Oromo (Gezehagn, 2012)

Adverbs of time			
English	**Afaan Oromo**	**English**	**Afaan Oromo**
Yesterday	kaleessa	**Adverbs of manner**	
Today	harr'a	very	baayyee
tomorrow	bor	quite	baayyee
Now	amma	really	dhugumaan
Then	gaafas	fast	dafee
Later	eger	well	gaarii
tonight	edans	hard	cimaa
right now	amma isa ammaa	quickly	dafee
last night	eda	slowly	suuta
this morning	ganam kana	carefully	qalbiidhan
next week	torban dhufu	absolutely	matuma
recently	dhiyeenya kana	together	walii wajjin
Soon	dhiyootti	alone	qophaa
immediately	hatattamaan	**Adverbs of frequency**	
Adverbs of place		always	yeroo hunda
Here	as	sometimes	gaaffii gaaf
There	achi	occasionally	gaaffii gaaf
over there	gara sana	seldom	darbee darbee
everywhere	iddoo hunda	rarely	darbee darbee
nowhere	eessayyu	never	yoomiyyuu
Home	mana		
Away	fagoo		
out	ala		

2.7.6.6. Prepositions

Oromo prepositions link nouns, pronouns and phrases to other words in a sentence. The word or phrase that the preposition introduces is called the object of the preposition (MYLANGUAGES, 2011).

Table 7. Afaan Oromo Prepositions (Gezehagn, 2012)

English Prepositions	Oromo Prepositions
about	*waa'ee*
above	*gubbaa / gararraa*
across	*Gama*
after	*booddee / booda*
against	*Faallaa*
among	*jara giddu*
as	*Akka*
at	*Itti*
before	*Dura*
behind	*dudduuba / dugda duuba*
below	*jala / gajjallaa*
beneath	*gajjallaa*
beside	*bira*
between	*Gidduu*
beyond	*Garas*
but	*Garuu*
by	*..dhaan*
despite	*ta'uyyuu*
down	*Lafa*
during	*Utuu*
except	*Malee*
for	*F*
from	*Irraa*
in	*Keessa*
inside	*Keessa*
into	*keessatti*
near	*Bira*
next	*ittaanee*
of	*Kan*
on	*Irra*
opposite	*fuullee / faallaa*
out	*Ala*
outside	*Alla*
over	*Irraan*
per	*..tti*

English Prepositions	Oromo Prepositions
since	*ergii*
than	*mannaa*
through	*gidduu*
till	*hamma*
to	*tti*
toward	*garas*
under	*jala / gajjallaa*
unlike	*faallaa*
until	*hamma*
up	*gubbaa*
via	*karaa*
with	*wajjin*
within	*keessatti*
without	*malee*
two words	*jechoota lama*
according to	*akka kanaatti*
because of	*kanaaf*
close to	*bira*
due to	*kanaaf*
except for	*kana malee*
far from	*irraa siqee / iraa fagaatee*
near to	*itti aanee*
next to	*itti aanee*
outside of	*kanaa alatti*
prior to	*kanaan dura*
three words	*jechoota sadii*
as far as	*hamma*
as well as	*fi*
in addition to	*dabalatees*
in front of	*fullee isaa*
in spite of	*ha ta'uyyu malee*
on behalf of	*maqaa …*
on top of	*kana irraan*
demonstrative prepositions	*Agarsiisoo*

2.7.6.7. Negation

Oromo negation is the process that turns an affirmative statement (I am happy) into its opposite denial (I am not happy). Negation and negative expressions have a very important role in Oromo (MYLANGUAGES, 2001).Examples:

I don't eat-hin nyaadhu	I don't wait-hin eegu
I don't go-hin deemu	I don't write-hin barreessu
I don't sleep- hin rafu	I don't speak-hin dubbadhu

2.8. Challenges of the Language in IR System Design

There is challenge in building similarity models, making assumptions when programming and taking into consideration each word comparison situation. Unfortunately, at this point, there have not been many algorithms that accurately deal with similarity matching and comparing words in every scenario. For example, the complexity of similarity matching in each synonyms and polysemy:

Synonyms: Dealing with synonyms as part of unstructured data analytics at first may seem to be a simple task. The definition of "synonym" itself is widely known and understood by most people. But the level that two words are synonymous adds several layers of complexity when trying to find two documents with comparable meaning. A "true synonym" is defined as two words that have the exact same meaning. If "true synonyms" do not exist, then the definition of synonyms should be clarified as two words that have similar meaning. The challenges with synonyms then becomes trying to determine on indexing, how similar two synonyms are to each other (CSLD, ND). Additionally, there is not a reliable way to establish an accurate similarity index because the level of similarity between two synonyms may differ depending on context. For example: the synonyms "walqabata" (link) and "riqaa" (bridge). In some cases, sentences using the two words can be very comparable and more similar as demonstrated in this example:

"Yaadni Atiyeezimii fi Aginootizimii walqabata" (There is a link between the concepts of atheism and agnosticism).

“Garaagarummaa hubannoo Atiyeezimii fi Aginootizimii gidduu jiru riqaa taasisuu nan danda’a” (“I can bridge the gap between what an atheist believes and what an agnostic believes”).

However, based on context sentences using the same two words in different sentences may have completely different meaning and treating the words as synonyms would result in less accurate results as conveyed in these two sentences:

“Dameen kuni cancala walqabata 62 qaba” (This chain has 62 links).
“Ani riqaalee baay’ee qaxxaamureen dhufe yeroon karaa lageewwaniirraa deemutti” (I crossed several bridges while walking besides the river).

Polysemy: has the same spelling but has different meanings or defined as any two words that has different meaning but either sounds the same when spoken, has the same spelling, or both (CSLD, ND). For example, the word “qoraan” (wood):

“qoraan muka sana irra jiru ni gubate” (The wood on that tree is burnt). In this sentence the word “qoraan” (wood) indicates the dry tree that is cut for fire.
“galgala qofaa kee qoraan keessa akka hin deemne of eeggadhu ” (Becarefull not to go walking at night alone in the wood). In this sentence the word “qoraan” (wood) indicates the forest that tree is dense at one place.

Additionally, it is difficult to distinguish if the word is homonym or polysemic (CSLD, ND).For example, its difference is also considered, if the word is two separate words with different meanings (a homonym) or one word with the same meaning but used in different ways (polysemic). Here are some examples:

Polysemy that is also homonymous: “hirkatanii”. Example:
“Isaan namarratti hirkatanii jiraatani” (Depend on).
Polysemy Example: “Haati manaa fi abbaan manaa wal irratti hirkatanii jiraatu” (Helping each other).
Homonymy Example: “Isaan gola ciisichaa keessa hirkatanii jiru” (They are sleeping).

2.9. Related Works

2.9.1. Global researches

Sodanil and Ketmaneechairat (2013) improved the quality of search using QE. The experiment is on the query expansion technique using keyword-based query. The results shown are higher than using only original query regarding to effectiveness of information retrieval system and also compared to the base-line system, the method provides higher performance in terms of recall and precision.

The Original Lesk algorithm (Lesk, 1986) uses dictionary definitions (gloss) to disambiguate a synonymous word in a sentence context. The major objective of its idea is to count the number of words that are shared between two glosses.

Jothilakshmi *et al*. (2013) analyzed the various QE approaches such as user feedback, local analysis (from initially retrieved documents) and Global analysis (global information such as thesaurus and Ontology) and improved that the Word Sense Disambiguation techniques and Ontology (Upper or Domain) plays a vital role to expand the user queries to increase the precision and recall of the system.

Alhroob *et al*. (2013) determined the query effectiveness two query expansion techniques (global and local query). As the study shows although, local context analysis has some advantages over the similarity thesaurus, Association thesaurus which is global is generally the most effective one.

Smith *et al*. (2007) analyzed various query expansion approaches include relevance feedback, corpus dependent knowledge models and corpus independent knowledge models. Case studies detailing query expansion using domain-specific and domain-independent ontology are also included in order to examine the reasons for the success or failure of ontology based query expansion.

Rivas *et al*. (2014) developed and evaluated preprocessing and query expansion techniques for retrieving documents in several fields of biomedical articles belonging to the corpus Cystic Fibrosis, a corpus of MEDLINE documents. The Studies were carried out to compare the

weighting algorithms Okapi BM25 and TF-IDF available in the Lemur tool, concluding that TF-IDF with TF formula given by BM25 approximation is superior in its results.

Paskalis and Khodra (2011) implemented Word Sense Disambiguate (WSD) using WordNet and searching using some query expansion methods on Apache Lucene and evaluated the performance of query expansion using some methods. The researchers described as relevance feedback plays a big role in the query expansion using corpus' information gave bigger increase in performance than using thesaurus' terms. When the researchers conclude their study, the effort of increasing retrieval performance does not need to be focused on query disambiguation using complicated ways rather it should be better to focus on the relevance feedback and how to make use of it.

Hong-Zhao *et al*. (2002) realized Query expansion with experimental results on TREC-9 collections in which query expansion method result is significant improvements over the IR without query expansion and shown that the decaying co-occurrence model is effective on query expansion for Chinese IR.

Reshma *et al*. (2013) developed the system that improves effectiveness by considering synonyms and negations of the terms specified in the query. As the researchers evaluated and put the results, the relevant documents are returned to the user based on the similarity measure. It can be observed that, the retrieved documents are the most relevant documents with respect to the given query. It can also be noted that the retrieval process is not subjective and show promising results.

Willet *et al*. (1992) evaluated three methods for the expansion of natural language queries in ranked-out put retrieval systems. The methods are based on term co-occurrence data, on Soundex codes and on a string similarity measure. The results shown that there is no statistically significant difference between any of four searches, in terms of the numbers of relevant documents retrieved in ranked output searches. Generally the expansion methods evaluated are not guaranteed to result in a significant increase in search performance.

2.9.2. Local researches

Some researches attempts to develop information retrieval system for Afaan Oromo language. A number of IR systems developed so far for retrieving Afaan Oromo texts. Debela and Ermias (2010) designed a rule based stemmer for Afaan Oromo texts using n-gram and rule based approach methods. This research mainly conducted to solve the problem of inflectional and derivational words that have similar meanings with performance of 95.73% results and concluded that stemming is important for highly inflected languages such as Afaan Oromo for many applications that require the stem of a word. Gezehagn (2012) developed Afaan Oromo text retrieval system using inverted index method. This research mainly conducted to solve the problem related with accessing information that satisfies needs of Afaan Oromo users with performance of 57.5% precision and 62.64% recall results and concluded that indexing and searching modules with vector space model are very important in text retrieval system to retrieve text documents.

Researches for query expansion with local languages are very few and there is no research done for query expansion for Afaan Oromo language. However, the following researches have been done on query expansion for Amharic language.

Nega (2003) identified significant causes for variation that can help the researcher to focus on opportunities for improvement that underlay the averages by showing the potential of statistical repeated measures analysis of variance for testing the significance of factors in retrieval performance variation. The result shown on retrieval method, topic and their interaction are all significant and the observed retrieval performances of query expansion runs are truly significant improvements for most of the topics and also the analysis of the effect of query expansion on document ranking confirm that affects ranking positively.

Samrawit (2014) attempted to extend the application of query expansion using semantic similarity measure towards designing an effective word sense disambiguation. This research mainly conducted to solve the problem of word sense disambiguation that is involved with context based query expansion with performance of 59% F-measure and concluded that to

increase the number of relevant documents retrieved, queries need to be disambiguated by looking at their context.

Iman (2013) attempted to expand query for Amharic language in information retrieval system based on proper sense disambiguation. This research mainly conducted to solve the problem of word sense by identifying the correct senses of terms and expand the query using the term's definition or synonym sets of terms with performance of 59% F-measure for synsets expansion and 30% F-measure for gloss expansion and concluded that WordNet is important to find the word sense and used accordingly for sense disambiguation to identify the correct senses of terms and expand the query using the term's definition or synonym sets of terms.

Although there is no research done on query expansion for Afaan Oromo language, Gezehagn (2012) mentioned the gap of query reformulation (query expansion) on Afaan Oromo IR system and also recommended that the synonym and polysemy nature of Afaan Oromo language greatly affect the retrieval performance that initiates the integration of query operations. Starting from the Gezehagn's research gap and recommendation, the researcher developed QE for Afaan Oromo IR based on WordNet that enhance the retrieval performance.

3. METHODS

The proposed system pre-process the document collection and user information query for effective Afaan Oromo information retrieval system. The pre-processing involves the following IR processes to extract index terms. These are tokenization, normalization, stop words removal and stemming. This is followed by Indexing the selected index terms from document corpus using inverted index file. Searching for relevant documents for users query follows vector space IR model. The users query also expanded to enhance relevant document retrieval to satisfy information need of users.

3.1. Description of the Study Area

Afaan Oromo is a Cushitic language spoken today by about 40 million people in Ethiopia (about 40% of the country's population), in Kenya, Somalia and Djibouti and is the 3rd largest language in Africa after Arabic and Hausa. Afaan Oromo (meaning Oromo language) or Oromiffa, most probably rates second among the African indigenous languages. Even before two decades when Gada (1988) explored the history of this nation, Afaan Oromo was known as the mother tongue of about 30 million Oromo people living in the Ethiopian and neighboring countries such as Kenya, Somalia, and Djibouti.

Information retrieval is not being an optional technology, it is very important to everybody and mandatory to use, especially in different languages and also very important to retrieve information from internet effectively from the huge collection available. In this Information Age, information is highly needed than anything else. However, searching this necessary information for Oromo language needs system support called information retrieval system (Schatz, 1997).

3.2. The Proposed Architecture for Afaan Oromo Information Retrieval

Figure 5 below depicts the architecture of query expansion based Afaan Oromo information retrieval system.

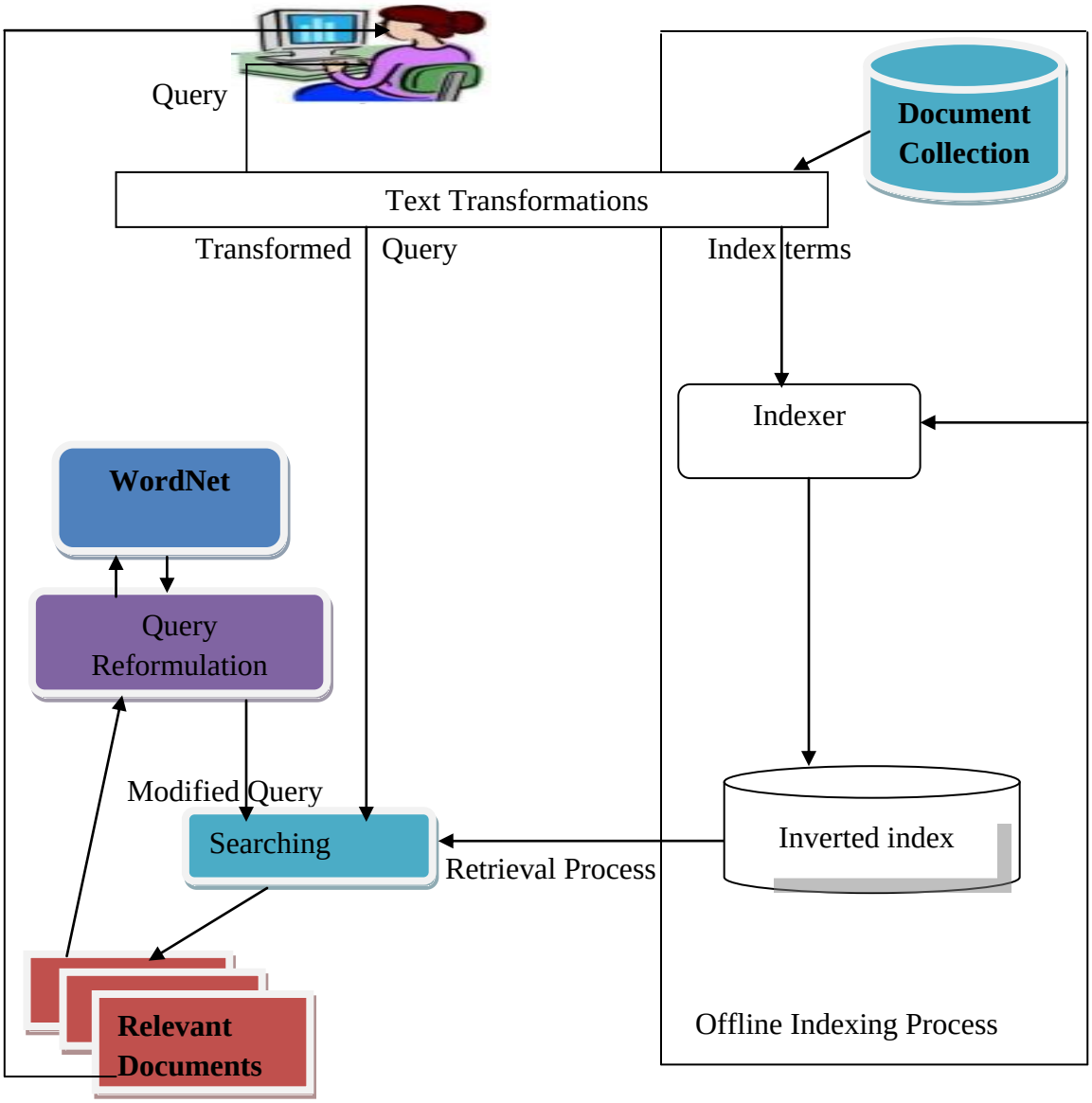

Figure 5. Architecture of query expansion for Afaan Oromo information retrieval system

The above architecture is the combination of retrieval processing and word sense disambiguation based query expansion. Given the document collection, the first step is applying text operations such as tokenization, normalization, stop words removal and stemming for selecting a subset of all terms for use as indexing terms. The index terms are then used to compose document representations, which might be smaller than the actual themselves (depending on the subset of index terms selected). The identified index terms are then organized using inverted index file. Once the document collection is indexed, the retrieval process can be initiated. User first specifies a query that reflects his/her information need. This query is then parsed and modified by operations that resemble those applied to the documents. Next, the transformed query is expanded and modified after word sense disambiguation algorithm is applied and the synonyms of queries words are identified in WordNet. The expanded and modified query is then processed to obtain the set of retrieved documents, which is composed of documents that contain the query terms. Fast query processing is made possible by the index structure previously built. The steps required to produce the set of retrieved documents constitute the retrieval process. Next, the retrieved documents are displayed as search result in ranked order.

3.3. Word Sense Disambiguation

The main task of word sense disambiguation in information retrieval is disambiguating words by telling which sense an ambiguous word belongs to (Chua and Kulathuramaiyer, 2004).Thus, word sense disambiguation is performed on the original query terms of the users because of the user queries are assumed to ambiguous while searching relevant documents. This means, the synonymous and polysemous of words bring the ambiguity problems while searching relevant documents from collection of documents. In original Lesk algorithm (Lesk, 1986), word sense disambiguation is processed using gloss to gloss by comparing information interrelated with its synonyms and gloss definition. This algorithm is also used for Afaan Oromo word sense disambiguation in retrieval processing. The algorithm is applied on Afaan Oromo WordNet which is prepared manually. The major objective of using this algorithm is to count the number of words that are shared between two glosses. The more overlapping the words, the more related the senses are. To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words.

3.3.1. Semantic similarity and word sense disambiguation

Word can have more than one sense that can lead to ambiguity. For example: In Afaan Oromo, the word "guddina (growing)" has different meanings in the following two contexts:

- "Guddina" – dheerina (length).
- Guddina" – dinagdeen guddachuu (growing in economy).

The algorithm bases its disambiguation decisions on the semantic similarity of words (Banerjee, 2002).In word sense disambiguation, to use semantic relations for retrieval of textual documents, the correct WordNet sense has to be assigned to words in the text. For example, from manually constructed Afaan Oromo lexical resources; the WordNet incorporates synonyms sets and gloss definitions. Therefore, it is possible to assign the senses meaning semantically using Lesk algorithm.

3.3.2. Word sense disambiguation with original Lesk algorithm

Disambiguation is the process of finding out the most appropriate sense of a word that is used in a given sentence. The Original Lesk algorithm (Lesk, 1986) uses dictionary definitions (gloss) to disambiguate a synonymous word in a sentence context.

The Lesk algorithm exploits the similarity or relatedness between the sense definitions of the ambiguous word (MA) and the definitions of the words of its context {M1, M2, …, Mi, …, Mn}. Figure 6 below provides the architecture of the Lesk algorithm, where S Ai is the gloss definition corresponding to the i^{th} sense of the ambiguous word. The original algorithm uses a dictionary as a resource. For every possible meaning of the word to disambiguate Sj, a definition D (Sj) is attributed. The word M (belonging to the context of the ambiguous word) is represented by the dictionary definitions E (M).

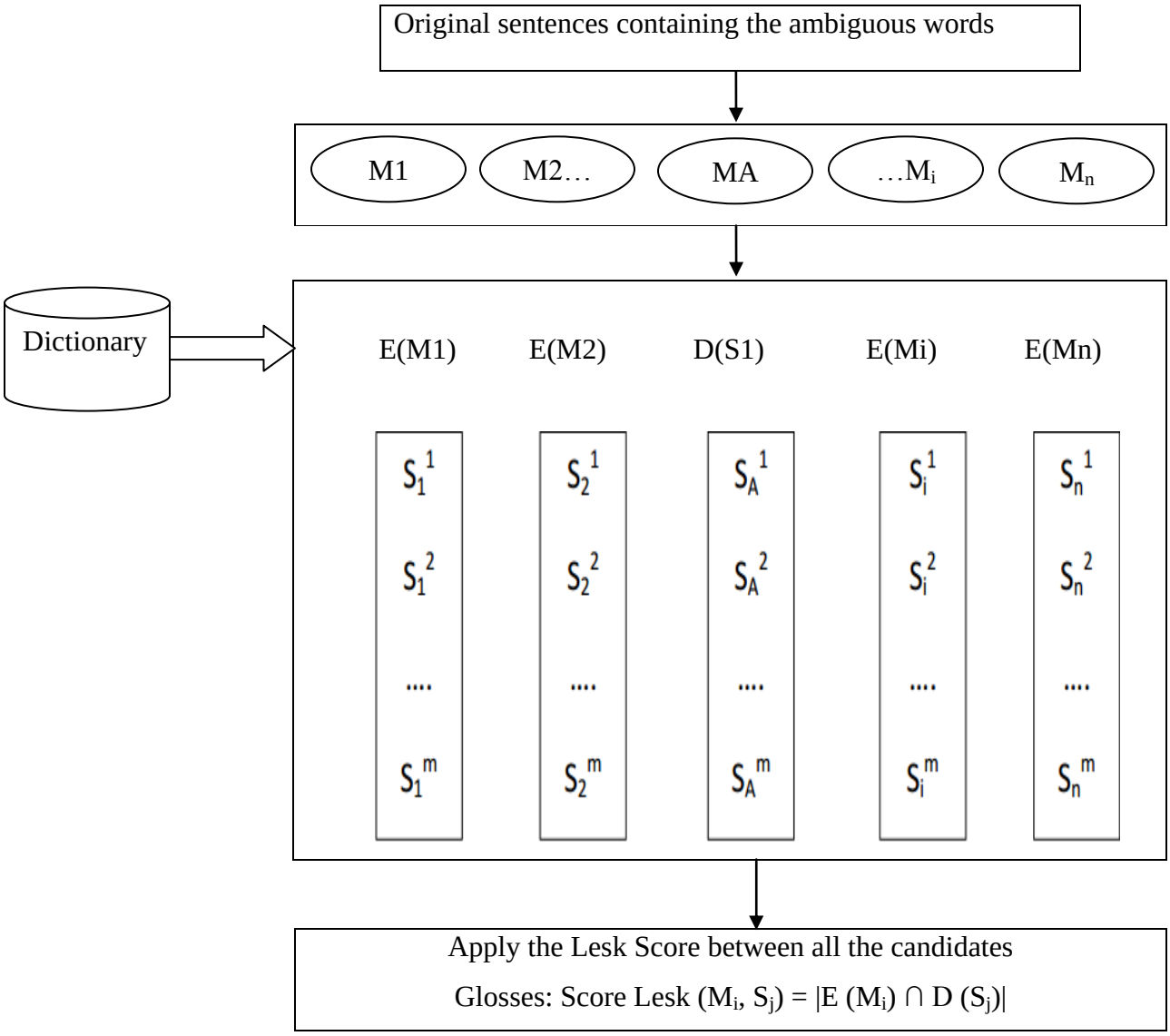

Figure 6. Original Lesk algorithm architecture

The original Lesk algorithm calculates the overlap between each possible definition of the ambiguous word and the definition of words contained in the same sentence as the ambiguous word. For example, the below equation 3.1 calculates the overlap of two words.

$$\text{Score Lesk } (M_i, S_j) = |E(M_i) \cap D(S_j)| \quad (3.1)$$

The major objective of its idea is to count the number of words that are shared between two glosses. The more overlapping the words, the more related the senses are.

- Sense 1: kind of evergreen tree with needle-shaped leaves.
- Sense 2: waste away through sorrow or illness.

The word "cone" has three senses:

- Sense 1: solid body which narrows to a point.
- Sense 2: something of this shape, whether solid or hollow.
- Sense 3: fruit of a certain evergreen tree.

By comparing each of the two gloss senses of the word "pine" with each of the three senses of the word "cone", it is found that the words "evergreen tree" occurs in one sense in each of the two words. So, these two senses are then declared to be the most appropriate senses when the words "pine" and "cone" are used together.

3.3.3. Afaan Oromo WordNet

There is no standard lexical resource prepared for Oromo language. Therefore, lexical WordNet is prepared manually for this research study. The WordNet is prepared with consultation ofAfaan Oromo linguistic experts and with the use of Afaan Oromo dictionary ‘Galmee Jechoota Afaan Oromo’ (Tesfaye, 2003). The WordNet is limited to contain only two information associated with each term that are words with list of synonyms that have similar senses and the gloss definition those groups of words with similar sense. The WordNet synsets group Afaan Oromo words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members and WordNet glosses are a sense disambiguated corpus. This means, words from the definitions (glosses) in synsets are manually linked to the context-appropriate sense in WordNet.

3.3.3.1. Overlaps of senses definitions

The sense overlaps includes the gloss overlaps or synsets overlaps. The original Lesk algorithm performs WSD by calculating the relative word overlap between the context usage of a target

word, and the dictionary definition of each of its senses in a given machine readable dictionary. The sense with the highest overlap is then assumed to be the correct one. Lesk algorithm identifies the sense of a word w whose textual definition has the highest overlap with the words in the context of w. Formally, given a target word w, the following score is computed for each sense S of w. Score Lesk Var (S) =|context (w) ∩ gloss (S)| where context (w) is the bag of all content words in a context window around the target word w and gloss (S) is the bag of words in the textual definition of sense S of w.

For example, the two senses or meanings for the word **guddina** are listed below and words which overlap with the following input sentences are marked in bold.

Naatol **guddina** hatattamaarra jira.

1. Guddina – guddaa yookiin baay'ee ta'iinsa:qaamaan dheerachuu.
2. Guddina – qabeenya dinagdeen sokkaa jira: misooma **hatattamaarra** jira

Sense 2 of **guddina** has one overlap, whereas the other senses have zero, so the second sense is selected.

This study used the original Lesk's algorithm that is word sense disambiguation using gloss definition.

3.3.3.2. Gloss definition

There are content words in common between the definition of one sense of w_i ={D (w_i)} with one sense of w_j= {D (w_j)}.

It is not uncommon that multiple pairs of the definitions of w_i and w_j have words in common since the definition of a term contains quite a few words. The supposition in this gloss definition is if the two terms are semantically related and have similar context, the gloss used to define those terms‟ senses can contain at least one same word as opposed to other words used by different senses. For example:

sooressa@nama qabeenyaan of gahe:duuressa;badhaadhaa:soorumaan kan ciccite

beekamaa@beekamtii kan qabu:inni ogummaa harkaan beekamaadha;ulfaataa:duuressakabajama

From the above fragment of Afaan Oromo WordNet the synonyms of word sooressa is duuressa and also it is found on the definition of the word beekamaa.

The word sense disambiguation steps used query expansion for Afaan Oromo information retrieval:

- Get the query word
- Disambiguate using gloss to gloss definition method
- Calculate frequency of sense using gloss to gloss
- Select the sense with the highest frequency
- Combine to formulate the new query

Let Q= (w_1, w_2,……w_i….w_j….w_k,), the original query contains a number of words. Those are the words needed to be disambiguated if they are ambiguous words. Therefore, the words should be found from the WordNet having their own synsets and gloss definition. If the word is ambiguous it has two or more senses.

The first word found on the given query with its multiple sense sets W_i={(S_1,G_1), (S_2, G_2)….} the second word found on the given query with its multiple sense sets W_j={(S_1, G_1), (S_2, G_2)…}.The synsets of each word can have one or more term sets. S= {t1, t2…} the synsets of the first sense with the terms (t1, t2= set of synonyms terms). The gloss of each word is also a combination of words. It is always more than one word as long as it is the definition of the word. G= {t1, t2, t3……..tn} the gloss of the first sense with the (t1, t2...= terms that defines the synonyms (s) and (w)).

3.3.4. Query expansion

For query expansion the researcher uses gloss expansion because of many Afaan Oromo words have different meanings from zone to zone. Their synonyms are mainly explained by example in gloss definition.

Gloss is the short definitions providing proper meaning of words. The researcher expands the original query depending on the gloss definition.

For example, for original query: sooressa beekamaa

For sooressa the senses are:

1. nama qabeenyaan of gahe:sooressa duuressa;
2. badhaadhaa:soorumaan kan ciccite

For beekamaa the senses are:

1. beekamtii kan qabu:inni ogummaa harkaan beekamaadha
2. ulfaataa:duuressa kabajama

The original query ‘sooressa beekamaa’ has two words and each word has two senses with a total four comparisons required to identify similar sense and gloss to gloss is applied using Lesk algorithm. In the above example, the words that are found after colon are gloss definitions. In this proposed way (gloss to gloss), the first sense definition of the word ‘sooressa’ that is **“**sooressa duuressa**”** is compared with the two gloss definition of the term “beekamaa”. Then it gets an overlap with gloss definition of the second sense of the word ‘beekamaa’ that is “duuressa kabajamaa”. Finally, the common words found in each sense definition are taken from the intersection of each definition. In this example, one word is found (i.e., ‘duuressa’). Then the second sense of the word **‘**sooressa**’** gloss definition which is **“**soorumaan kan ciccite**’** continued the same process. In this case there is no common word to be found for this sense. Then, the one with common intersection is identified for the expansion. The same process is done for each term of the query and the one with common intersection is assigned as the sense of the word based on the given query context.

3.4. System Evaluation

While retrieving relevant documents from the collection to the users, the IR performance and effectiveness should be evaluated to measure to what extent the users are satisfied. Thus, the Afaan Oromo information retrieval system is evaluated before query expansion and after query expansion by using effectiveness measures, such as precision, recall and F-measures (Sodanil and Ketmaneechairat, 2013).

Precision: it is the fraction of the documents retrieved that are relevant to the user's information need as defined in the following formula.

Precision = # of relevant docs retrieved …… …….. (3.2)
# of docs retrieved

Recall: It is the fraction of the documents that are relevant to the query that is successfully retrieved

Recall = # of relevant docs retrieved …… (3.3)
# of relevant docs

F-Measure: is the harmonic mean of precision and recall which is defined as follows.

F-M = 2 *Precision * Recall …… (3.4)
Precision + Recall

4. EXPERIMENTATION AND DISCUSSION

This research attempts to come up with a retrieval system that searches relevant documents from Afaan Oromo text by using query expansion so as to satisfy information need of users. To design the system four major steps are taken. The first one is construction of Afaan Oromo WordNet which is used as knowledge based to understand the meaning of concepts. The second one is word sense disambiguation for the purpose of identifying the word sense of the given query. The third one is Query reformulation which helped to expand and modify the original query with the identified sense from the lexical resource. Finally, query expansion module is integrated with Information Retrieval system to show its contribution to the enhancement of the performance of Afaan Oromo IR system.

4.1. Data Preparation

Finding a large size and standard corpus for Afaan Oromo language is one of the challenges faced in this research. The state of the art in the area of text processing indicates that, there is no any developed standard corpus for Afaan Oromo language and also as a result of time factor it takes a lot of time to prepare large size of corpus with many documents. Thus, in this study the researcher uses small size corpus (100 different textual documents) which were collected from different news and media, such as VOA, Bariisaa and online educational resources that are written in Afaan Oromo and available on the web. The news articles used involve different subjects like health, politics, religion, sport, economy, culture and social as shown in table 8 below.

Table 8 . Types and size of news articles used for experiment

No.	Types of news	No. of documents
1	Health	15
2	Politics	15
3	Religion	15
4	Sport	15
5	Economy	15
6	Culture	15
7	Social	10
Total		**100**

Additionally, in order to test the experiment eight (8) queries are identified. These queries are selected depending on content of files in the corpus. For each query terms there is at least one (1) relevant document and a maximum of thirteen (13) relevant documents. These queries are marked across each document as either relevant or irrelevant to construct relevance judgment as a reference during IR effectiveness measure. The main importance of having identified queries is to evaluate the performance of the system. These queries are selected subjectively by the researcher after reviewing the content of each article manually.

According to that the relevance judgment each document is identified to be as a relevant or non-relevant toward each query terms.

Table 9. List of queries with their relevance judgment

QNo.	Queries	Relevant	Non Relevant	List of relevant documents for query
Q1	Gaaffii uummataa	11	89	1,2,3,4,5,6,7,96,97,98,99,100
Q2	Guddina dinagdee	13	87	8,9,10,11,12,13,14,15,16,17,64, 65,66
Q3	Waldaa Ispoortii Itoophiyaa	12	88	18,19,20,21,22,23,24,25,26,61, 62,63
Q4	Karaa amantii	12	88	27,28,29,30,31,32,33,34,35,71, 72,73
Q5	Qulqullina barnootaa	12	88	36,37,38,39,40,41,42,43,44,87, 88,89
Q6	Rakkoo nageenyaa	13	87	45,46,47,48,49,50,51,52,53,54, 84,85,86
Q7	Sirna gadaa	10	90	55,56,57,58,59,60,67,68,69,70
Q8	Sooressa beekamaa	10	90	73,74,75,76,77,78,79,80,81,82

4.2. Word Sense Disambiguation

The main objective of this study is to apply query expansion to enhance the performance of Afaan Oromo text retrieval system. Thus, to expand the query the senses of the queries must be prepared from Afaan Oromo WordNet.

4.2.1. Preparation of Afaan Oromo WordNet

Sample WordNet is constructed in consultation with Linguists using the Afaan Oromo to Afaan Oromo dictionary (Tesfaye, 2003). For example, the following figure 7 shows the sample of Afaan Oromo WordNet.

beekamaa@beekamtii kan qabu:inni ogummaa harkaan beekamaadha;ulfaataa:duuressa kabajamaa

gaaffii@gaafachuu:daa'imni xiqqoon gaaffii gaafachuu guddisti;dhaqanii fira ilaaluu:warra kee gaafachuu yoom deemta

gadaa@sirna ittiin bulmaata uummata oromoo:sirna dimokiraatawaa uummanni oromoo sadarkaa umurii irratti hundaa'ee ofii isaa ittiin of bulchu;seera:ittiin bulmaata sirna gadaa keessatti seerri tumamee jira;amantii:sirni gadaa amantii waaqeffannaa ti

guddina@guddaa yookiin baay'ee ta'iisa:guddina nafaa caalaa guddina sammuu waaqni sii yaa kennu;dheerina argachaa:guddina qaamaan ol dabaluu ispoortiin guddinaafi;misoomaan guddachuu:dinagdeen dabaluu

ispoortii@tapha adda addaa kan qaama ofii ittiin ho'ifatan:atileetiksii biyoolessaa

ispoortii@tapha adda addaa kan qaama ofii ittiin leenjisan:tapha kubbaa harkaa tapha kubbaa miilaa;shaakala qaamaa:qaama ofii ispoortii barsiisuu

sooressa@nama qabeenyaan of gahe:sooressa duuressa;badhaadhaa:soorumaan kan ciccite

Figure7. Sample WordNet with basic words and their sense of meaning

The WordNet contains the term, synonyms term and the definition of the synonyms. The term before '@' is a reference term about which the different senses are given, for example, term "gaaffii". The synsets are defined between '@' and ':' in this case the first synsets for the word "gaaffi" is gaafachuu" then followed with gloss definition which is daa'imni xiqqoon gaaffii gaafachuu guddisti*"*. If the given term has multiple senses, it is separated with ';' on the WordNet. The second sense for the term "gaaffii" is "dhaqanii fira ilaaluu:warra kee gaafachuu yoom deemta" is the second sense and the rest is the gloss definition.

The format for the term, synsets, gloss and sense on WordNet are: root word, synsets of first sense, gloss of first sense, synsets of second sense and gloss of second sense. For example, the format for sooressa@nama qabeenyaan of gahe:sooressa duuressa;badhaadhaa:soorumaan kan ciccite:

Root word: sooressa

Synsets of first sense: nama qabeenyaan of gahe
Gloss of first sense: sooressa duuressa
Synsets of second sense: badhaadhaa
Gloss of second sense: soorumaan kan ciccite

Generally, constructing Afaan Oromo WordNet is very challenging and time consuming. The constructed WordNet contains only the synonymous and gloss definition information of the term.

4.3. Afaan Oromo Information Retrieval before Query Expansion

Afaan Oromo IR is developed before query reformulation; then after, query expansion module is integrated. Tokenization, normalization, stemming and stop word removal are the main text operations take part in indexing and searching to develop IR system.

4.3.1. Tokenization and normalization module

In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for indexing and searching. It is the process of splitting on white spaces and throwing away punctuations and tokenizes the text, turning each document into a list of tokens (Manning *et al*., 2008). Figure 8 shows tokenization algorithm used in this study.

1. assign the characters to variable characters
2. define the tokenize function by giving the parameter document

    3. split document in to lower case using split and lower built functions and assign it to variable

    4. return [variable. strip (characters)for variable in variable

Figure 8. Algorithm for tokenization

4.3.2. Stop word remover module

According to Greengrass (2000), few terms occur frequently, a medium number of terms occur with medium frequency and many terms with very low frequency. This shows that writers use limited vocabulary throughout the whole document, in which even fewer terms used more frequently than others. Figure 9 depicts stop word detector and removal algorithm used in this study.

```
1. Open stop word list
2. for each word w in Token List
     if a word w is in stop word list then
        Remove a word
      else
            add to non-stop word list
  end for
```

Figure 9. Algorithm for stop word remover

4.3.3. Stemming module

Stemming is language dependent that reduces tokens to their root form of words to recognize morphological variation. According to Nega and Willett (2002), a stemmer that stem words without consideration of remaining stem, which removes words that are similar to prefix and suffix list but that are not actually affixes is called context-free stemmer.

A stemming algorithm reduces all words with the same stem to a common form, usually by stripping each word of its derivational and inflectional suffixes (Lovins, 1968).

Figure 10 shows stemming algorithm (Debela and Ermias, 2010) used in this study for conflating suffix.

```
1. read a word to be stemmed
2. if word matches with one of the rules
remove the suffix and do the necessary adjustments
go to 3
else
Go to 4
      end if
3. Return the stemmed word to the calling portion
4. if there is no applicable condition and action exist
            return the word as it is
5. End if
```

Figure 10. Algorithm for stemmer

4.4. Afaan Oromo IR System

The prototype for Afaan Oromo IR system is developed to measure the recall, precision and F-Measure and also used to compare and contrast the effectiveness of Afaan Oromo IR system after query expansion for Afaan Oromo IR system.

The first step of this system is to get query from the user. Figure 11 presents a screen shot which shows list of retrieved document using the query, ‘sooressa beekamaa’.

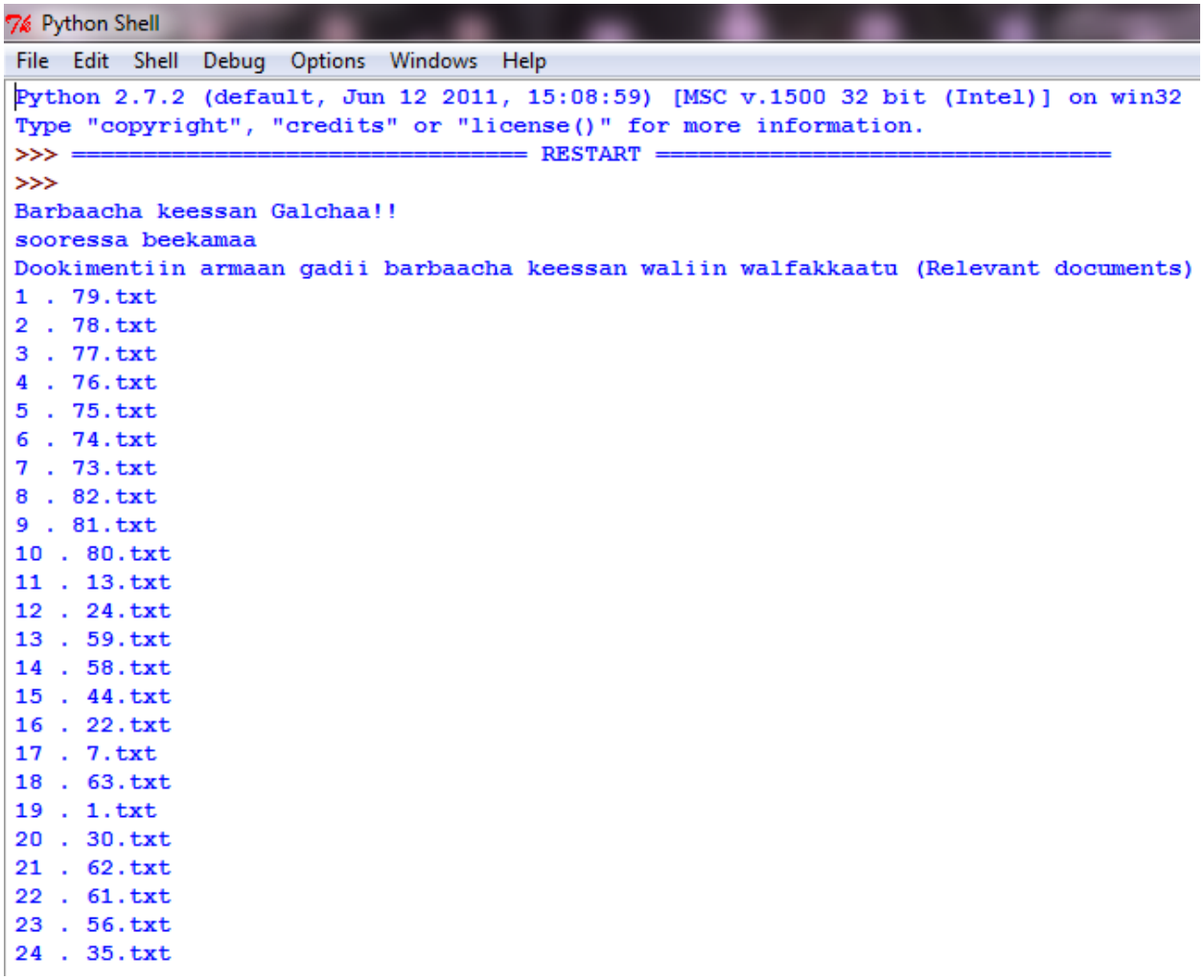

Figure11. Retrieved documents for a given query ‘sooressa beekamaa’

The information retrieval system used in this study is developed based on the vector space model retrieval model. The procedure of the system is that first documents are pre-processed for removing stop words, stemming and normalization in order to extract content bearing index terms. The final step is ranking the documents to show best matching with respect to the provided query by user. The above figure 11 shows the result of this process without applying query expansion. For the query ‘sooressa beekamaa’, it retrieves 24 documents, out of them 7 of them are relevant (document numbers 73, 74, 75, 76, 77, 78, 79), however, in the corpus there are 10 relevant documents for the query. The modified query for expansion is formed from the terms found in gloss to gloss method that is used to disambiguate the word. There are terms of original query. These terms have one or more senses. Each of the sense of the term is compared

with each of the other senses of the original query term. When it compares gloss to gloss the target is to find the common words to identify the correct sense of that term based on the context given from the original query. The number of common words found to query terms is declared to be the score of the sense of the target word and assign as the weight of the term.

4.4.1. Performance evaluation before query expansion

While retrieving relevant documents from the collection to the users, the IR performance and effectiveness should be evaluated to measure to what extent the users are satisfied. Thus, the Afaan Oromo information retrieval system is evaluated before query expansion and after query expansion by using effectiveness measures, such as precision, recall and F-measures (Sodanil and Ketmaneechairat, 2013).

The researcher used VSM for this work. Vector space model is the most commonly used strategy for measuring relevance of documents for a given query because of using binary weights is too limiting and non-binary weights provide consideration for partial matches. Accordingly TFIDF weighting is used for knowing term importance. Cosine similarity measure identifies relevant documents that are fully or partially matching with the query. Finally, ranked relevant documents based on their degree of similarity to query (Baeza-Yates and Ribeiro-Neto, 1999).

The below table 10 presents the initial result without query expansion for the eight queries used to test the system.

Table 10. Initial retrieved result before query expansion

QNo.	Queries	Corpus	Retrieved	Relevant	R	P	F-M
Q1	Gaaffii uummataa	11	26	7	0.64	0.26	0.38
Q2	Guddina dinagdee	13	29	9	0.69	0.31	0.43
Q3	Waldaa Ispoortii Itoophiyaa	12	14	9	0.75	0.64	0.69
Q4	Karaa amantii	12	21	6	0.50	0.29	0.37
Q5	Qulqullina barnootaa	12	15	9	0.75	0.60	0.67
Q6	Rakkoo nageenyaa	13	14	10	0.77	0.71	0.74
Q7	Sirna gadaa	10	23	7	0.70	0.30	0.42
Q8	Sooressa beekamaa	10	24	7	0.70	0.29	0.41
				Total	**0.69**	**0.43**	**0.51**

From above table 10,the average result of recall, precision and F-measure of the system using the initial retrieve made by the model for the relevance of documents are 69%, 43% and 51% respectively. This shows that all relevant documents in the collection are not retrieved. For this reason the percentage of recall is low. And also many documents retrieved are not relevant to query. For this reason the percentage of precision is very low. Finally, the harmonic mean of recall and precision is also low which indicates that the performance of the system is not satisfactory because of Afaan Oromo synonymous and polysemous words. For example for query ‘guddina dinagdee’(economically growing) three relevant documents are not retrieved because of synonymous word of guddina which is written as ‘misooma’ in document 64,65,66 respectively. And also 16 non relevant documents are retrieved from the collection because of polysemous word ‘guddina’ (growth) which is written in these documents with different meaning “’dheerina (length)’, ‘dabalu (increasing)’, ‘baay’ee (many)’”.

4.4.2. Performance evaluation after query expansion

The new query is reformulated by adding terms with similar sense with each term in original query. After all senses of the words are disambiguated, the highest overlap is chosen (Lesk, 1986).

For example, for query ‘sooressa beekamaa’ the words sooressa and beekamaa have two senses and two senses respectively.

The senses of the word sooressa are:

1. nama qabeenyaan of gahe:sooressa duuressa
2. badhaadhaa:soorumaan kan ciccite

The senses of the word beekamaa are:

1. beekamtii kan qabu:inni ogummaa harkaan beekamaadha
2. ulfaataa:duuressa kabajamaa

From the above query ‘sooressa beekamaa’; each word has two senses with a total four comparisons required to identify similar sense. Using Lesk Algorithm each sense is compared with each other. The method is gloss to gloss by using original Lesk algorithm. From the above example the phrases written before colon are the synsets and the rest is gloss definition of a term.

In the above example, the gloss definition of the first sense of the word ‘sooressa’ is compared with gloss definition of the first and second sense of the word ‘beekamaa’. Then it gets an overlap result (i.e., 1). The common word found in ‘sooressa#1’ and ‘beekamaa#2’ is ‘duuressa’. Then the gloss definition of the second sense of the word ‘sooressa’ is continued with the same process with gloss definition of the first and second sense of the word ‘beekamaa’. In this case there is no common word to be found for this sense. In this, the overlap result is 0. Thus, the one with the highest overlap result is taken as the identified sense that will be used for the expansion. The same process is done for each term of all queries and the one with the highest overlap result is assigned as the sense of the word based on the given query context.

The below figure 12 shows the retrieved documents after applying query expansion for the query: ‘‘sooressa beekamaa’’.

```
Barbaacha Haaraa (New Query)


['sooressa', 'beekamaa', 'duuressa', 'kabajamaa']
barbaachi keessan erga jabaatee booda dookimentii funaanaman (After expansion)
1 . 80.txt
2 . 25.txt
3 . 79.txt
4 . 78.txt
5 . 77.txt
6 . 76.txt
7 . 75.txt
8 . 74.txt
9 . 73.txt
10 . 98.txt
11 . 82.txt
12 . 81.txt
13 . 24.txt
14 . 70.txt
15 . 69.txt
16 . 68.txt
17 . 67.txt
18 . 4.txt
19 . 83.txt
20 . 72.txt
21 . 97.txt
22 . 1.txt
23 . 61.txt
24 . 34.txt
25 . 99.txt
26 . 13.txt
27 . 58.txt
28 . 3.txt
29 . 44.txt
30 . 22.txt
31 . 7.txt
32 . 63.txt
33 . 59.txt
34 . 30.txt
35 . 62.txt
36 . 56.txt
37 . 35.txt
38 . 60.txt
>>>
```

Figure12. List of retrieved relevant documents after query expansion for the query ‘‘sooressa beekamaa’’

Thus, using gloss to gloss method ‘sooressa beekamaa’ is expanded in to ‘duuressa kabajamaa’. As it is shown from the above figure 12, the relevant documents that are not retrieved in figure 11 are retrieved after query expansion.

The below table 11 shows search result after query expansion for the eight queries used to test the effectiveness of the IR system designed for Afaan Oromo text.

Table 11. Retrieved search result after query expansion

QNo.	Queries	Corpus	Retrieved	Relevant	R	P	F-M
Q1	Gaaffii uummataa	11	43	11	1	0.26	0.41
Q2	Guddina dinagdee	13	33	13	1	0.39	0.56
Q3	Waldaa Ispoortii Itoophiyaa	12	37	11	0.92	0.30	0.55
Q4	Karaa amantii	12	28	9	0.75	0.32	0.45
Q5	Qulqullina barnootaa	12	20	12	1	0.60	0.75
Q6	Rakkoo nageenyaa	13	17	13	1	0.76	0.86
Q7	Sirna gadaa	10	26	9	0.90	0.35	0.50
Q8	Sooressa beekamaa	10	38	10	1	0.26	0.41
				Total	**0.95**	**0.41**	**0.56**

From the above table 11, the average result of recall, precision and F-measure of the system from the retrieved result after expansion made by the model for the relevance of documents are 95%, 41% and 56% respectively. This shows that the percentage of recall after query expansion is increased by 26% when it is compared with initial retrieved documents before query expansion which is 69%. This indicates that relevant documents in the collection are almost retrieved except some document. These relevant documents that are not retrieved after expansion is because of some new queries are found on synsets rather than gloss. For example, the new query added on the original query ‘ispoortii’ is ‘atileetiksii’ but it is not expanded because of the word ‘atileetiksii’ that is found on synsets. For such reason some relevant documents are not retrieved. And also the average result of precision of the system from the retrieved result after expansion made by model for the relevance of documents are decreased by 2% when it is compared with initial retrieved documents before query expansion which is 43%. This indicates that retrieved documents are not relevant to the query because of polysemous words are also increased after query expansion. Finally, the harmonic mean of recall and precision is increased. This indicates

that the performance of the system is increased after query expansion because of the recall is increased. Thus, the performance of the Afaan Oromo IR system after query expansion is satisfactory when it is compared with the performance of Afaan Oromo IR system before query expansion.

In general, this study shows that the query expansion for Afaan Oromo IR system is effective when it is compared to Afaan Oromo IR system. The algorithm used for semantic similarity and query expansion is gloss to gloss method. This techniques registered better performance for Afaan Oromo IR system after query expansion when it is compared to Afaan Oromo retrieval system without query expansion. This gloss to gloss expansion made that the performance of the system is increased in 5% F-score.

4.5. Discussion of Results

The obtained result shows that the application of query expansion for Afaan Oromo IR system based on WordNet enhances system performance. This further shows that the Afaan Oromo WordNet is much useful for determining the correct sense during searching. The below table 12 shows that query expansion in information retrieval enhance search result as compared to without query expansion.

Table 12. Summarized result of the overall performance of Afaan Oromo IR

Measures	Original query	Modified query
Recall	0.69	0.95
Precision	0.43	0.41
F-Measure	0.51	0.56

The above table 12 shows query expansion based on WordNet achieved high recall, low precision and high F-measure when it is compared to the system without expansion. The reason for high recall and low precision is synonym and polysemy words respectively. For example, in the first initial retrieved results before expansion and the retrieved results after expansion, the recall of ‘sooressa beekamaa’ is 0.70 and 1 respectively. This shows, after query expansion the

recall of ‘sooressa beekamaa’ is increased because of synonym words. This means, the query ‘sooressa beekamaa’ is expanded to ‘duuressa kabajamaa’. This means, the number of relevant documents retrieved for the new item (‘duuressa kabajamaa’) is increased after query expansion. Therefore, when the number of relevant documents retrieved per the number of relevant documents in the collection is increased, the recall is also increased. However, the precision is decreased because of polysemous words. For example, in the new item (‘duuressa kabajamaa’) added to original query (‘sooressa beekamaa’) the word ‘kabajamaa’ has different meaning such as ‘sodaatamaa’ (fearful), and ‘aangoo’ (authority). Therefore, when the total number of irrelevant documents retrieved is increased, the precision is decreased.

Even if this research attempts to show the possible use of WordNet and its information associated with each term for query expansion in IR system, there is a need of constructing WordNet by including different information associated with the given term. Some relevant documents are not retrieved after applying query expansion because of some new queries are found on synsets rather than gloss. For example, the new query added on the original query ‘ispoortii’ is ‘atileetiksii’ but it is not expanded because of the word ‘atileetiksii’ that is found on synsets. And also because of polysemous words many irrelevant words are retrieved. For example, for the query ‘guddina dinagdee’ (growing in economy) many irrelevant documents are retrieved. The reason is that the polysemous word ‘guddina’ is found in many non-relevant documents with the meaning of “’dheerina (length)’, ‘dabalu (increasing)’, ‘baay’ee (many)’”. After query expansion also irrelevant documents retrieved are increased because of polysemous of new words. For example, for original query ‘karaa amantii’ (the way to God) the new terms added are ‘daandii waaqayyoo’ (the road to God). Because of the word ‘daandii’ the irrelevant document ‘daandiin konkolaataa Finfinnee irraa gara Adaamaatti geessu hojjetamee xumurame’ (the car road which links Addis Ababa to Adama is completed) is retrieved.

For the reason that finding standard corpus and query for Afaan Oromo language is a challenge in this study, the comparison made in this research is not with the previous researchers work whom did on Afaan Oromo IR system rather the system is compared with itself. Thus, it is difficult to compare and construct the performance result of the system beside of other researchers. Another challenge is there is no standard stemmer for Afaan Oromo language to be used for every research.

5. CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The key goal of an IR system is to retrieve information which is relevant to the user and query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval system (Vectomova and Wang, 2006).

The developed prototype involves constructing WordNet, word sense disambiguation and query reformulation using query expansion.WSD using semantic similarity identifies the correct sense from lexical resource like WordNet and to use it in query expansion techniques for Afaan Oromo language. The Afaan Oromo lexical resource built from two dictionaries available for Afaan Oromo language that has a format of WordNet. The words used in the WordNet are limited to include the terms used for the prepared queries. The information associated with each terms in the WordNet contains the synsets, synonymous terms of phrases of the word and the gloss definition of the word. The query expansion module is integrated with Afaan Oromo IR system to reformulate the original query. The query reformulation is applied by merging the sense's identified using gloss to gloss method and Lesk sense similarity measure for word sense disambiguation. The experiment is carried out to test the performance of IR system by expanding the queries using terms found from gloss.

According to the experiment made, the IR system registered, on the average 95% recall, 41% precision and 56% F-score. The effectiveness of IR system shows that in this study a promising result is obtained. However, still precision of the system needs due consideration, which is greatly affected because of polysemous nature of Afaan Oromo language. Also unavailability of well-constructed lexical resource like WordNet is another challenge which affects the result obtained.

5.2. Recommendation

Based on experimental result and analysis the researcher recommends the following observed points to be taken in to consideration for future work in order to enhance the effectiveness of Afaan Oromo text retrieval system.

- Lexical resources such as WordNet are a must to implement semantic based information retrieval. Hence the researcher recommends Afaan Oromo Standard WordNet should be constructed for the future.
- In this study the precision of the system is very low and also decreased after query expansion because of polysemous words. Thus, to increase the precision of the system and decrease retrieval of irrelevant documents there is a need to undertake further research on the way to control the effect of polysemous words in the retrieval process.
- In this study the researcher made an experiment gloss-to-gloss method for semantic similarity measure. To optimize effective of query expansion for Afaan Oromo IR system the researcher recommend further study that attempts to combine synsets to gloss and gloss to gloss for sense disambiguation.
- Finding a standard corpus, test queries with relevance judgment, standard IR system with a better stemmer algorithm for testing the designed system is one of the challenges faced in this research. Therefore, future research need to consider the development of standard Afaan Oromo language corpus, test queries and IR system that can be used by every researcher to evaluate progress made in designing techniques for enhancing effects of Afaan Oromo IR system.
- The current study tries the integration of query expansion along with the well-known vector space model. These days probabilistic retrieval model is also one of the promising directions for designing an IR system. The model allows to integrate query expansion with the IR system and it also has inbuilt query reformulation mechanisms using term reweighting. So the researcher recommends scholars to compare and contrast its performance with the vector space model.

6. REFERENCE

Agirre, E. and Edmonds, P. 2006. *Word sense disambiguation*: *Algorithms and Applications*. University of the Basque Country, New York.

Aysh Alhroob, Hayel Khafajeh and Nisreen Innab. 2013. Evaluation of different query expansion techniques for Arabic text retrieval system. *American Journal of Applied Sciences*, 10 (9): 1018-1024.

Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*, ACM Press, Addison-Wesley New York.

Baeza-Yates, R. and Ribeiro-Neto, B. 2011. *Modern Information Retrieval*: *The Concepts and Technology behind Search, 2nd Edition,* ACM Press, Addison-Wesley Longman Limited, New York.

Banerjee, S. 2002. Word sense disambiguation to WordNet. MSc Thesis, University of Minnesota Duluth, U.S.A.

Bhogal, J. 2011. Investigating ontology based query expansion using a probabilistic retrieval model. City University, London.

Bhogal, J., Macfarlane, A. and Smith P. 2007. A review of ontology based query expansion. *Information Processing & Management*, 43 (4): 866-886.

Bodenreider, O. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32 (1): 267–270.

Bush, V. 1945. As we may think. *The Atlantic Monthly*, 176 (1): 101 -108.

Carpineto, C. and. Romano, G. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys* (*CSUR*), 44 (1):1-50.

Christopher D.Manning, Prabhakar Raghavan and Hinrich Schütze . 2009. *An introduction to information retrieval, Online Edition*, Cambridge University Press, Cambridge England.

Chua, S and Kulathuramaiyer, N. 2004. Semantic Feature Selection Using WordNet. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,* 166-172 September 2004. Washington, DC, USA.

Cui, H., Wen, J.R. ,Nie, J.Y., and Ma, W.Y. 2002. Probabilistic query expansion using query logs. *Proceedings of the 11th International Conference on World Wide Web*, 325–332 May 2002. Honolulu, Hawaii, USA.

Debela Tesfaye and Ermias Abebe.2010. Designing a rule based stemmer for Afaan Oromo text. *International Journal of Computational Linguistics (IJCL*), 1 (2): 1-11.

Dian Paskalis , F.B. and Khodra, M.L. 2011.Word sense disambiguation in information retrieval using query expansion. *International Conference on Electrical Engineering and Informatics,* 17-19 July 2011, Bandung, Indonesia.

F. Cuna Ekmekcioglu, Alexander M. Robert son and Peter Willet. 1992. Effectivenes of query expansion in ranked –output document retrieval systems. *Journal of Information Science*, 18 (2): 139-147.

Gada Melba. 1988. *An Introduction to the History of the Oromo People*. Khartoum, Sudan.

Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science,* 41 (4):288-297.

Gezehagn Gutema. 2012. Afaan Oromo text retrieval system. MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Greengrass, E. 2000. Information retrieval: *A survey Information Retrieval,* 1-224.

Grootjen, F.A. and Van der Weide, Th.P.. 2006. Conceptual query expansion. *Data and Knowledge Engineering*, 56 (2): 174- 193.

Gupta, A. Dharamveer kr. Yadav . 2014. Semantic similarity measure using information content approach with depth for similarity calculation. *International Journal of Scientific & Technology Research*, 3 (2): 165-169.

HE Hong-Zhao, HE Pi-lian, GAO Jian-feng and HUANG Chang-ning. 2002. Query Expansion for Chinese information retrieval by using a decaying co-occurrence model. *Transaction of Tianjin University*, 8 (3): 83-86.

Iman M. Yusuf. 2013. Query expansion based on proper sense disambiguation. MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Jain, R. and Nathawat, S. 2012. Sense Disambiguation Techniques: A Survey. *International Journal of Advances in Computer Science and Technology*, 1 (1):1-6.

Jiang, J. and. Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the International Conference on Research in Computational Linguistics*, 1-15 September 1997, Taiwan.

Jothilakshmi, R., Shanthi, N. and Babisaraswathi, R. 2013. A survey on semantic query expansion. Journal *of Theoretical and Applied Information Technology (JATIT*), 57 (1): 128-138.

Kankaria, A. 2005. Query expansion techniques. Indian Institute of Technology Bombay, Mumbai.

Krovetz, R. 1997. Homonym and Polysemy in information retrieval. Proceeding of the 35th Annual Meeting of Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistic, July 1997, Madrid, Spain.

Kumar, S., Sharma, N., and Niranjan, S. 2012. Word sense disambiguation using association rules: A Survey. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, 2 (2): 93-98.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation*, 24-26, Toronto, Ontario, Canada.

Lin, D.1998. An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning,* 296-304 July 1998.

Liu, G. ,Wang, R. , Bucklay, J.and Helen M.Zhou. 2011. A WordNet based semantic similarity measure enhanced by internet-based knowledge. *Proceedings of the 23rd International Conference on Software Engineering and Knowledge Engineering,* 1-9 July 2011.*Eden Roc Renaissance, Miami Beach*, USA.

Lovins, J.B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11 (1): 22-31.

Mandala, R., Tokunaga, T. and Hozumi, T. 1998.The use of WordNet in information retrieval, in *use of* WordNet in natural language processing systems*: Proceedings of the Conference*. 31-37.

Manning, C. D., Raghavan P. and Schutze H.2008.*Introduction to Information Retrieval, Online*. Cambridge University Press, New York.

Manning, C. D., Raghavan P. and Schutze H.2009.*An Introduction to Information Retrieval, Online Edition*. Cambridge University Press ,Cambridge, England.

Matthes, E. 1972. *Python Crash Course: A Hands-on, Project-Based Introduction to Programming*. No Starch Press, San Francisco.

Meng, L. and Gu, J. 2012. A new method for calculating word sense similarity in WordNet. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 5 (3): 197-206.

Michael Bendersky, Donald Metzler and W. Bruce Croft. 2012. Effective query Formulation with Multiple Information Sources. *Proceedings of the 5th ACM International Conference on Web Search and Data Mining,* 443-452 February 2012. Washington, USA.

Miller, G. A., Beckwith, R. Fellbaum, C. D. Gross, D. Miller, K. 1990. WordNet: An online lexical database. *Int. J. Lexicograph.* 235–244.

Mylanguages.org. 2011: Oromo numbers, available:www.mylanguages.org/oromo_numbers.php. Accessed: March 8, 2017.

Nega Alemayehu. 2003. Analysis of performance variation using query expansion. *Journal of the American Society for Information Science and Technology*, 54 (5):379 –391.

Nega,A. and Willett, P. 2002.Stemming of Amharic words for information retrieval, in Literary and Linguistic Computing,17 (1): 1- 18.

Rekha Jain, Sulochana Nathawat and Purohit, G.N. 2013. Modified page rank algorithm to solve ambiguity of polysemous words. *International Journal on Cybernetics & Informatics (IJCI*), 2 (2): 13-19.

Reshma, O.K., Sreejith, C and Reghu Raj, P.C. 2013. An effective Malayalam information retrieval system using query expansion. *International Conference on Control Communication and Computing* (*ICCC*). 265-270.

Resnik, P. 1995. Using information content to evaluate semantic similarity. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

Rivas, A.R., Iglesias, E.L. and Borrajo, L. 2014.Study of query expansion techniques and their application in the biomedical information retrieval. *Hindawi Publishing Corporation the Scientific World Journal*.

S. Liu, C. Yu, and W. Meng, 2005. Word sense disambiguation in queries, in *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 525-532.

Samrawit Zewdnesh. 2014. Word sense disambiguation using semantic similarity for query expansion in Amharic information retrieval. MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Sandarson, M. and Croft, B**.**, n.d. The History of Information Retrieval Research. Unpublished article.

Sanderson M. 1996. Word sense disambiguation and information retrieval. PHD dissertation, University of Glasgow, Glasgow.

Schatz, B.R. 1997. Information retrieval in digital libraries: Bringing search to the Net. *Journal of Science*, 275 (5298): 327-334.

Schütze, H. 1998. Automatic word sense discrimination. Computational Linguistics, 24 (1): 97–123.

Shiri, A.A. Revie, C and Chowdhury, G. 2002. Thesaurus-assisted search term selection and query expansion: A review of user-centered studies. *Knowledge Organization*, 29 (1): 1-19.

Simon Ager. 2012. Oromo language, available:www.sas.upenn.edu/Africanstudies/Hornet/Afaan Oromo 19777. Html. Accessed: October 12, 2016.

Singhal, A.2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull*, 24 (4): 35-43.

Smeaton, A. F. , Kelledy, F. and O'Donnell, R.. 1995. Trec-4 experiments at Dublin City University: Thresholding posting lists, query expansion with wordnet and pos tagging of Spanish: 373–389.

Smeaton, A.F.1995. Linguistic approaches to text Management: an appraisal of progress. *Journal of Document & Text Management,* 2 (2):67-80.

Sodanil, M. and Ketmaneechairat, H. 2013 .Information retrieval experiment on subjective words query expansion. *International Conference of Information and Communication Technology* (ICoICT), 161-165.

Sussna, M. .1993. Word sense disambiguation for free-text indexing using a massive semantic network, *Proceedings of the International Conference on Information & Knowledge Management* (CIKM), 2(1): 67-74.

Tesfa Kebede. 2013. Word sense disambiguation for Afaan Oromo language. MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Tesfaye Firisa. 1996. *Galmee Jechoota Afaan Oromoo*. Addis Ababa, Ethiopia.

Tesfaye Guta. 2010. Afaan Oromo search engine. MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Torres, S. and Gelbukh, A. 2009. Comparing similarity measures for original WSD Lesk algorithm. Advances in Computer Science and Applications. Research in Computing Science,155-166.

Uzuner, O. 1998. Disambiguation applied to information retrieval. MSc Thesis, Department of EECS, MIT, M. Eng.

Vectomova Olga, Wang Ying . 2006. A study of the effect of term proximity on query expansion. *Journal of Information Science*. 32 (4): 324–333.

Voorhees, E. M. 1993. Using wordnet to disambiguate word Senses for text retrieval. *Proceedings of the 16th ACM-SIGIR Conference*. 171-180.

Walker II, J. 1990. A node-positioning algorithm for general trees. Software – Practice and Experience, 20 (7):685–705.

www.csldsolutions.com/synonymy-vs-homonymy-vs-polysemy/. Accessed: August 11, 2017.

Xing Wei, Fuchun Peng, Huishin Tseng, Yumao Lu, Xuerui Wang, Benoit Dumoulin. 2010. Search with synonyms: Problems and Solutions. Beijing. 1318–1326.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, 189 196.

Yehuwalashet Bekele. 2016. Hybrid word sense disambiguation approach for Afaan Oromo words. MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Yin, Z., Shokouhi, . M. and Craswell, N. 2009. Query expansion using external evidence. In Advances in Information Retrieval. 362–374.

7. APPENDICES

Appendix I- Qubee fi Dubbiftuu (Qubees and their phone)

Sagalee gaggabaaboo (Short sounds)				
`a	`e	`i	`o	`u
ba	be	bi	bo	bu
ca	ce	ci	Co	cu
cha	che	chi	co	cu
da	de	di	Do	du
dha	dhe	dhi	dho	dhu
fa	fe	fi	Fo	fu
ga	ge	gi	go	gu
ha	he	hi	Ho	hu
ja	je	je	jo	ju
ka	ke	ki	Ko	ku
la	le	li	lo	lu
ma	me	mi	Mo	mu
na	ne	ni	no	nu
nya	nye	nyi	Nyo	nyu
ra	re	ri	Ro	Ru
sa	se	si	so	su
sha	she	shi	Sho	shu
ta	te	ti	to	tu
va	ve	vi	Vo	vu
wa	we	wi	wo	wu
xa	xe	xi	xo	xu
ya	ye	yi	yo	yu
za	ze	zi	zo	zu

Sagalee dhedheeroo (Long sounds)				
`aa	`ee	`ii	`oo	`uu
baa	bee	bii	boo	buu
caa	cee	cii	coo	cuu
chaa	chee	chii	coo	cuu
daa	dee	dii	doo	Duu
dha	dhe	dhi	dho	dhu
faa	fee	fii	foo	Fuu
gaa	gee	gii	goo	guu
haa	hee	hii	hoo	Huu
jaa	jee	jee	joo	juu
kaa	kee	kii	koo	Kuu
laa	lee	lii	loo	luu
maa	mee	mii	moo	Muu
naa	nee	nii	noo	nuu
nyaa	Nyee	nyii	nyoo	Nyuu
raa	ree	rii	roo	Ruu
saa	see	sii	soo	suu
shaa	shee	shii	shoo	Shuu
taa	tee	tii	too	tuu
vaa	vee	vii	voo	Vuu
waa	wee	wii	woo	wuu
xaa	xee	xii	xoo	Xuu
yaa	yee	yii	yoo	Yuu
zaa	zee	zii	zoo	zuu

Appendix II: Afaan Oromo Stop Words

	Aanee	Gararraa	Ishiirraa	Koo	Siin
	agarsiisoo	garas	ishiitti	kun	silaa
	akka	garuu	ishiitti	lafa	silaa
	akkam	giddu	isii	lama	simmoo
	akkasumas	gidduu	isiin	malee	sinitti
	akkum	gubbaa	isin	manna	siqee
	akkuma	ha	isini	maqaa	sirraa
	ala	hamma	isinii	moo	sitti
	alatti	hanga	isiniif	na	sun
	alla	henna	isiniin	naa	tahullee
	amma	hoggaa	isinirraa	naaf	tana
	ammo	hogguu	isinitti	naan	tanaaf
	ammoo	hoo	ittaanee	naannoo	tanaafi
	an	hoo	itti	narraa	tanaafuu
	ana	illee	itumallee	natti	ta'ullee
	ani	immoo	ituu	nu	ta'uyyu
	ati	ini	ituullee	nu'i	ta'uyyuu
	bira	innaa	jala	nurraa	tawullee
	booda	inni	jara	nuti	teenya
	booddee	irra	jechaan	nutti	teessan
	dabalatees	irraa	jechoota	nuu	tiyya
	dhaan	irraan	jechuu	nuuf	too
	dudduuba	isa	jechuun	nuun	tti
	dugda	isaa	kan	nuy	utuu
	dura	isaaf	kana	odoo	waa'ee
	duuba	isaan	kanaa	ofii	waan
	eega	isaani	kanaaf	oggaa	waggaa
	eegana	isaanii	kanaafi	oo	wajjin
	eegasii	isaaniitiin	kanaafi	osoo	warra
	ennaa	isaanirraa	kanaafuu	otoo	woo
	erga	isaanitti	kanaan	otumallee	yammuu
	ergii	isaatiin	kanaatti	otuu	yemmuu
	f	isarraa	karaa	otuullee	yeroo
	faallaa	isatti	kee	saaniif	yommii
	fagaatee	isee	keenna	sadii	yommuu
	fi	iseen	keenya	sana	yoo
	fullee	ishee	keessa	saniif	yookaan
	fuullee	ishii	keessan	si	yookiin
	gajjallaa	ishiif	keessatti	sii	yoolinimoo
	gama	ishiin	kiyya	siif	yoom

Appendix III: Afaan Oromo WordNet

aaaga@oduu gaarii kan nama gammachiisuu:oduu dhugaa ta'e

aadaa@sagalee rakkinaa:aarii;akkkaataa namoonni biyya tokkoo keessaa jiraataan:ittiin bulmaata uummata oromoo;duudhaa:aadaa oromoo keeesssatti marqaan cuukkoon caccabsaan baay'ee jaallatamu

abdii@waan gara fuulduraatti argadha jedhanii eeggatani:ollaa abdatteeti dhirsa heexoo obafte

adabuu@nama balleessetti tarkaanfii fudhachuu:murtii dabarsuu;

amantii@dhugaa ta'iinsa waa tokkoo beekumsa:waanta dhugaadha jedhamee fudhatame;karaa waaqayyoon itttin qunnaman:daandii dhugaa waaqayyoo

ambaa@mandara namni irra qubate:safara

barnootaa@beekumsa barumsa irraa argatan:manni barnootaa bakka beekumsaati;barmoota:beekumsa daa'imni oggaa shan guunnaan mana barnootaa galchanii barumsa qulqullina qabu akka argatu gochuudha

beekamaa@beekamtii kan qabu:inni ogummaa harkaan beekamaadha;ulfaataa:duuressa kabajamaa

beela@qoonqoo nama qabiisa:barbaachii nyaataa namatti dhaga'amina;waan nyaatan dhabanii rakkoon namarra gayiisa;oongee:gadadoo;bara beelaatti dhabamee dhibeen hin dhibamu:abjuun baraa beela biddeen biddeen

beera@dubartii dulloomte:jaartii;dubartii heerumte:durba durbummaa hin qabne

bu'uu@gadi deemuu yookiin wayi irraa uutaaluu:muka koruuun nama hin dhibu irraa bu'uutu nama dhiba malee;keessa gadi seenuu:lixuu

caaluu@bira darbuu:taruu;irra dheerachuu:baay'achuu

cafaqa@maxinoo dhadhaadhaan laafee sukkuumame:aannanii fi cafaqni nyaata dargaggeessati

dhaabbata@dhaabbata atileetiksii biyyoolessaa:gurmaa'ina tapha adda addaa

dinagdee@qabeenya:horii lafa qabaachuu guddina dinagdee;duuniyaa:qabeenya horii

dira'uu@waan karaa irra hin jirre:waan fokkuu qabu hojjechuu;namni nama hiriyaa ofii hin ta'iinitti yookiin hotiitti taphachuu:ijoolleetti dira'uun cubbuu dha

duwwaa@kan keessi isaa homaa hin qabne:ona

eegee@qaama horii kan duuba irraan gad rarra'u:eegee jabaatteef re'een hormaata hin dhokfattu;duuba waan ta'ee:adabadhuu taa'i akka eegee duukaa wajjinii hin kaatiinii

faallaa@karaa hin taane:dalga;karaa malee deemuu:karaa horiin deemu dhiiseeti faallaa deeme

facaasaa@yeroo midhaan facaafatan:waqtii midhaaan itti facaafatan;kibxata:lammaffoo;kenna ateetee:har'i ateetee Caaltuuti

falaxaa@qoraan babbaqaqfame:muka babbaqaqfamee dallaaf qophaa'ee

gaaffii@gaafachuu:gaaffii namaa dhiyeessuu qorannoo daa'imni xiqqoon gaaffii gaafachuu guddisti;dhaqanii fira ilaaluu:warra kee gaafachuu yoom deemta

gadaa@sirna ittiin bulmaata uummata oromoo:sirna dimokiraatawaa uummanni oromoo sadarkaa umurii irratti hundaa'ee ofii isaa ittiin of bulchu;seera:ittiin bulmaata sirna gadaa keessatti seerri tumamee jira;amantii:sirni gadaa amantii waaqeffannaa ti

guddina@guddaa yookiin baay'ee ta'iisa:guddina nafaa caalaa guddina sammuu waaqni sii yaa kennu;dheerina argachaa:guddina qaamaan ol dabaluu ispoortiin guddinaafi;misoomaan guddachuu:dinagdeen dabaluu

ispoortii@tapha adda addaa kan qaama ofii ittiin ho'ifatan:tapha kubbaa miilaa;atileetiksii biyoolessaa

itoophiyaa@biyya sabaa fi sablammiin beekamtu:abisiiniyaa

karaa@lafa namni yookiin konkolaataan irra deemu:kan bakka tokkoo qabee bakka biraatti nama geesssu;akkaataa haala:karaa hojiin kun ittiin hojjetamu irratti duddubachuu barbaachisa;roga dhimma namaa ilaallatu:ani karaa kootiin waanan dubbachuu fedhu raawwadheera;daandii:karaa amantii daandii waaqayyoon itti qunnaman

nageenyaa@nagaa:dhukkubaan kan hin qabamne;fayyaa:akkami ati fayyaadhaa nagaa galata waaqayyoo;rakkoo waraanaa dhabamsa:waraanni biyya keenya keessa hin jiru biyyi keenya nageenya qaba

qulqullina@qulqulluu:xurii yookiin huba kan hin qabne bishaan quruu hundi qulqulluu miti;cubbuu kan hin hojjenne:namni qulqulluun badii namaa hin jaallatu;haala sirrii ta'een adeemsisuu:qulqullinaan barsiisu fakkeeyaaf qulqullina barnootaa

rakkoo@rakkina:qabeenya dhabaa hiyyummaa eega horiin jalaa dhumtee rakkina guddaa irra jira;dhiphina:mucaa dhalattuu dhiistee rakkina natti uumte;rakkoo nageenyaa:sababa dhukkubaan fayyaa dhabuu jeequmsa dhala namaarratti raawwatamu

sirna@hojii yookiin dhimma:haala hojii;naqata qopheeffatan:haala adeemsaa mijeeffatan;akkaataa ayyaana yookiin waan biraa itti hojjetan yookiin waaqeffatan:sirna waaqeffannaa yookiin amantii fakkeenyaaf sirna gadaa

sooressa@nama qabeenyaan of gahe:sooressa duuressa;badhaadhaa:soorumaan kan ciccite

uummataa@uummata bal'aa:nama baay'ee qorannoo hawaasaa;ilmaan namaa:oromoon uummata addunyaa keessaa isa tokko;saba heddumina qabu:uummanni bal'aan oromoo yeroo dheeraadhaafi heeddumina isaatiin hin boonne

waldaa@iddoo lagni walitti makamu:waldaa lagaa irratti bishaan namicha nyaate;dhaabbata:dhaabbata atileetiksii

Appendix IV: Python code for QE for Afaan Oromo IR based on WordNet

```
import re
import os
import math
import sys
from search_corpus2 import qv_List# Relevant documents are printed at 'search_corpus2' before query expansion
from search_corpus2 import weight
from operator import itemgetter
g=[]
c=0
wn=open('wordnet.txt','r')
while wn.readline()!='':
    c=c+1
wn.close()
wnn=open('wordnet.txt','r')
wnn.readline()
for i in range(c):
    ll=wnn.readline()
    g.append(ll)
ww=[]
www=[]
for i in range(len(g)):
    ww.append(g[i])
    www.append(ww)
```

```
    ww=[]
 #print (www)
wordnet=[]
word=""
 #print("\n\n\n\n")
for i in range (len(www)):
    word=www[i][0]
    ww=word.split('@')
    wordnet.append(ww)
for i in range(len(wordnet)-1):
    word=wordnet[i][1]
    ww=word.split(';')
    wordnet[i][1]=ww
for i in range(len(wordnet)-1):
    for j in range(len(wordnet[i][1])):
        word=wordnet[i][1][j]
        ww=word.split(':')
        wordnet[i][1][j]=ww
for i in range(len(wordnet)-1):
    for j in range(len(wordnet[i][1])):
        word=wordnet[i][1][j][0]
        ww=word.split(' ')
        wordnet[i][1][j][0]=ww
        word=wordnet[i][1][j][1]
        ww=word.split(' ')
```

```
        wordnet[i][1][j][1]=ww
print (wordnet)
sense=[]
for i in range(len(qv_List)):
    for j in range(len(wordnet)):
        if qv_List[i]==wordnet[j][0]:
            sense.append(wordnet[j])
print(sense)
l3=[]
l4=[]
x=""
y=""
for i in range(len(sense)):
    print ("\n")
    for j in range(len(sense[i][1])):
        for k in range(len(sense[i][1][j][1])):
            w=sense[i][1][j][1][k]

            for l in range(len(sense)):
                if i==l:
                    continue
                else:
                    i
                    for ll in range(len(sense[l][1])):
                        for lll in range(len(sense[l][1][ll][1])):
```

```
                    ww=sense[l][1][ll][1][lll]
                    if w==ww:
                        x=str(i)+str(1)+str(j)+str(1)
                        y=str(l)+str(1)+str(ll)+str(1)
                        l4.append(x)
                        l4.append(y)
                        l3.append(l4)
                        l4=[]
expandedQ=qv_List
d=l3[0][0]
c=1
for i in range(len(l3)):
    if l3[i][0]==d:
        continue
    else:
        c=c+1
        d=l3[i][0]
w=[]
for iii in range(c):
    for i in range(len(l3)):
        w.append(0)
        for j in range(i+1,len(l3)):
            if l3[i]==l3[j]:
                w[i]=w[i]+1
    x=w
```

```
    xx=0
    for i in range(len(w)):
        if x<w[i]:
            x=w[i]
            xx=i
        else:
            continue
    print ("Barbaadaa haaraa (new query)")
    aa=int(l3[xx][0][0])
    ab=int(l3[xx][0][1])
    ac=int(l3[xx][0][2])
    aaa=int(l3[xx][1][0])
    aab=int(l3[xx][1][1])
    aac=int(l3[xx][1][2])
    expandedQ=expandedQ+sense[aa][ab][ac][1]
    print expandedQ
    temp=l3[0][0]
    l=len(l3)
    for j in range(len(l3)):
        for i in range(len(l3)-1):
            if l3[i][0]==temp:
                del l3[i]
                break
l=0
f=0
```

```
key=0
for j in range(len(expandedQ)):
    for i in range(key+1,len(expandedQ)-1):
        if expandedQ[key]==expandedQ[i]:
            del expandedQ[i]
            break
    key=key+1
print '\n'
print 'Barbaacha Haaraa (New Query)'
print '\n'
print expandedQ
Equery=d
sim={}
for key in Equery:
    for docword in weight:
        if key in docword:
            sim[docword]=Equery[key]*weight[docword]
x= sim
dicc={}
for s,v in x.iteritems():
    qo=s.split()
    a=qo[0]
    dicc[a]=v
d=dicc
items = [(v, k) for k, v in d.items()]
```

```
items.sort()
items.reverse() # so largest is first
items = [(k, v) for v, k in items]
i=1
print 'barbaachi keessan erga jabaatee booda dookimentii funaanaman (After query expansion)'
for x in items:
    print i,'.', x[0]
    i=i+1
```