Addis Ababa University

Collage of Natural Science

Query Expansion for Tigrigna Information Retrieval

Tsadu Zeray

A Thesis Submitted to the Department of Computer Science in Partial Fulfillment for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

October, 2017

Addis Ababa University

Collage of Natural Science

Tsadu Zeray

Advisor: Yarega Asabie (PhD)

This is to certify that the thesis prepared by Tsadu Zeray titled: *Query Expansion for Tigrigna Information Retrieval* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	Dr. Yaregal Asabie	_____	_____
Examiner:	Dr. Ayalew Belay	_____	_____
Examiner:	Dr. Solomon Gizaw	_____	_____

Abstract

This research has been prepared to enhance the precision and recall of Tigrigna IR system by integrating query expansion mechanism. Query expansion is an effective mechanism to control the effect of polysemous and synonymous nature of query terms. The main reason for integrating query expansion is to increase retrieval of relevant documents as per user's query based on the correct sense of query terms. This study has a way to discriminate the various meanings of a polysemous term, based on word sense disambiguation (WSD) and find synonymous terms for reformulating user's query.

The proposed algorithm determines the senses of synonymous and polysemous words in user's query using Tigrigna WordNet. In this study, we experiment root form Tigrigna WordNet and Tigrigna morphological analysis in IR for the first time. Using the idea of N-gram model, word sense disambiguation is performed by comparing the existence of ambiguous query terms, associated with its synsets and related word using reference to Tigrigna WordNet. The notion of WSD is to identify the correct sense of ambiguous terms in user's query and select the synonyms of the word. Then the selected synonyms of the ambiguous query term added to reformulate the original users query and the modified query will be used for searching of final result.

The experimental result of this research gains in two different way, first prior IR system tested with morphological analysis instead of stemmer and second this IR system test by integrating query expansion model. The experiment shows encouraging result, the method of using morphological analysis before query expansion register a performance of 9%precision and 1.6 % recall, expanding query using synset expansion register an improvement of 12% precision and 4% recall on the overall performance. The number of words related to each polysemy terms is limited because of the lack of resource. Therefore, the uses of query expansion terms are limited to the information available on the WordNet.

Keywords: *Word Sense Disambiguation, WordNet, N-gram, Information Retrieval*

Acknowledgment

First, I would like to thank God, who makes everything possible. I would like to thank sincerely to my advisor Dr. Yaregal Asabie for his guidance, understanding and excellent supervision throughout this study. This thesis would not have been possible without the help and support of my friend Kaleab Girma, and my special thanks goes to Abrha for his critical support on my work. I would like to give my thanks to my Family for their support, endless love and encouragement. Especially my dad Zeray Gebru and my beloved sister Tiblets Zeray, you are the reason for the person I become today. I thank you all.

Table of Contents

List of tables

List of Figures

List of Algorithms

Acronyms/Abbreviations

BIM	Binary Independent Model
CONE	Conceptual Network Editor
CET	Candidate Expansion Term
IR	Information Retrieval
JDK	Java Development Kit
MIDF	Minimum Invers Document Frequency
MT	Machine Translation
NDC	Number of Document in a Class
NLP	Natural Language Processing
OCR	Optical Character Recognition
POS	Part Of Speech tagger
TC	Threshold Class
TCNDC	Threshold Class Number of Document in a Class
TIRS	Tigrigna Information Retrieval System
TREC	Text REtrieval Conference

Chapter 1: Introduction

1.1 Background

Since the 1940s the problem of information storage and retrieval has taken attention [15]. Researchers initiated because of vast amounts of information to which accurate and speedy access is becoming ever more difficult. The main problem is that relevant information may ignored since it is never uncovered, in turn this leads to much duplication of work and effort [15, 16]. The era of arrival of computers, give initiation to use them to provide rapid and intelligent retrieval systems [15].

Ideally, information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is return a set of documents satisfying the information need to the query [15]. User can obtain the document by reading all the documents in the store, retaining the relevant documents and discarding all the non-relevant. In fact, this create a perfect retrieval. But this solution is obviously impracticable by humans, because user either does not have the time spend reading the entire document collection, apart from the fact that it may be physically impossible for humans to do so [15]. An information retrieval (IR) system assists users to get the information need from the vast amount of document collection [1, 3, 15]. Even though the task of information retrieval systems is retrieving relevant information, unfortunately there is no any system, which is capable of retrieving only relevant documents as per user's query [3].

The purpose of information retrieval is to retrieve all relevant documents at the same time retrieving as few non-relevant as much as possible. When the document is relevant to a query, it should be retrieved in response to that query. Human indexers have traditionally characterized documents in this way when assigning index terms to documents. The indexer attempts to anticipate the kind of index terms a user would employ to retrieve each document whose content he/she is about to describe. Implicitly constructing queries for which the document is relevant. When the indexing is done automatically, it is assumed that by pushing the text of a document or query through the same automatic analysis, the output will be a representation of the content, and if the document is relevant to the query [15]. Even if its time and cost consuming, logically it is possible for a human to identify

the most relevance document to a query. But for a computer to do this we need to hypothesis a model within which relevance decisions can be conducted [9].

These days, the researches on IR domain focuses on the problem of making existing IR systems retrieve better results [3]. For this problem, different mechanisms have been introduced by different researches, to achieve a good solution. One of these solutions is query expansion [3, 4, 28]. Basically, the purpose of query expansion is to support users, to reformulate a more specific meaning of query terms, in order to get best results [3]. According to [3], findings show that, still there are many reasons for users' bad query formulation. The first one is ambiguity of words, this means the limit of user knowledge of single meaning of the query term. Second vocabulary mismatch, when the user doesn't have knowledge of the right vocabulary of their search domain area. Third lack of user's willingness on using alternative terms for their query reformulation, when they cannot get the wanted information on the first attempt. Information Retrieval focuses on finding relevant documents whose contents are exact match with a user query from a large document collection. Most of the time formulating good queries is difficult for users, for this reason it is necessary to use query expansion to retrieve relevant information [5]. Query expansion is an effective technique to improve the performance of information retrieval system. It is widely applied for improving the efficiency of the textual information retrieval systems [5]. These techniques help to control the polysemy and synonymy behavior of natural language, vocabulary mismatch issues by expanding the original query with additional relevant [3, 5, 28].

1.2 Motivation

Tigrigna is a member of the Semitic branch of the Afro-asiatic languages [2, 3]. Tigrigna is the working language of Tigray region in Ethiopia and one of the official languages in Eritrea. About 3.5 million in Eritrea and around 5 million in Ethiopia speaks Tigrigna language [2, 25].

Many recent researches have indicated that, electronic documents in Tigrigna language keep on growing every year [2, 6, 7]. The existence of synonym and polysemy terms in Tigrigna text create some difficulties on retrieving relevant documents. And due to morphological complexity, synonymy and polysemy behavior of the language, users

mostly formulate weak queries to retrieve documents. This may lead to the poor coverage of retrieving relevant documents in IR. This situation needs well-performing Tigrigna IR system that retrieves more relevant documents as per the information need. The research done by Ataly in 2014 recommended that, there is a need of query expansion model to handle the morphological complexity of Tigrigna language. The main reason for performing query expansion is to provide relevant documents as per user's query that can satisfy their information need. The above reasons motivate us to design query expansion model into Tigrigna IR system.

1.3 Statement of the Problem

Tigrigna language is sematic language spoken in Eritrea and northern Ethiopia [39]. About 3.5 million in Eritrea and around 5 million in Ethiopia speaks Tigrigna language [25]. Tigrigna document available in a web increase from time to time [7], currently text books, journal, magazines, newspapers, news, online education, books in Tigrigna language are available in electronic format on the web [28]. The increment of electronic documents create some difficulty on the task of IR systems. Finding relevant document of users need from entire collection is difficult.

Some researches has been done on the area of information retrieval systems for Tigrigna language [7, 27, 28], these IR systems attempts to retrieve relevant documents as per user's information need. Even if such IR systems exist, they cannot meet all users' expectations. The reason behind is different users expresses single query in different ways and different levels of writer's knowledge also lead to express single ideas using different terms. As a result, systems most of the time, lose the core idea of users query and retrieve poor results. This happens because of the ambiguity of words involved in the natural languages and expression mismatch among users and writers. However, the research done by atalay [28] recommended that, there is a need for improving the performance of information retrieval system by integrating mechanism to control the effect of synonymous and polysemous or ambiguous terms in Tigrigna language. The ambiguity nature of natural language, challenges IR system on satisfying user's information need.

Synonym means different words which has the same meaning. In other word, if two or more words refer to one meaning, these words are synonymous words [3]. For example, "ሃይለ እቲ ማዕፆ ዓፅይዎ" and "ሃይለ እቲ በሪ ዓፅይዎ" these two sentences are similar in meaning, which is "haile close the door" in English. The two words "ማዕፆ" and"በሪ" refers to the

same meaning which is “door”. Thus, these words are synonyms to each other. A word said to be polysemous if it has different contextual meaning in different sentences [3, 29]. For example, in the sentences “ሃይለ ብዙሕ ከይዱ” and “እቲ ጥንሲ ከይዱ” the word “ከይዱ” refers “gone” in the first sentence and refers “abortion” in the second sentence. Therefor the word “ከይዱ” is polysemous word, because it changes its meaning based on the context of the sentences. In addition, the above example there are many polysemy words in Tigrigna language. Some of them are “በልዐ” means “eat” or “diddle”, “ምስፍሕፋሕ” means “broadening” or “spreading out”, “በቕሊ” means “mule” or “germinal” they are different according the context users used. When users use these kinds of words on their query, the task of IR system will be difficult on retrieving relevant document. The existence of such synonymy and polysemous words leads to decrease in the performance of the system in both precision and recall [28] unless it handled in some mechanisms. One of the recommended solution to solve the problem is query expansion. Expanding a query by considering the synonymy and polysemy words, may make the system to retrieve relevant document. Therefore, according the above ambiguity, it is necessary to consider polysemous nature of terms in addition synonym terms in query expansion. So far of our knowledge there is no any research done in query expansion and sense identification based on root form of words for Tigrigna IR systems. Our research will be the first attempt of controlling the synonym and polysemy nature of Tigrigna words, to enhance the performance of Tigrigna IR systems. This research attempts to design query expansion using N-gram and designing a root based WordNet, in order to control the effect of synonymous and polysemous behavior of Tigrigna words.

In addition of query expansion, this research attempts to experiment the performance of IR system using morphological analysis instead of stemmer. According to [6, 28] the high inflection nature of Tigrigna language results the higher vocabulary mismatch problem, over-stemming and under-stemming problems which in turn, reduces the accuracy of information retrieval significantly. Our hypothesis is initiated to handle the above problem using morphological analysis. At the end the following research questions are explored and answered in this study.

- Does the morphological analysis technique help to enhance the existing IR system?
- What kinds of approaches or techniques are more suitable to handle the challenges of polysemous query terms?

- How synonymous and polysemous terms be controlled in order to enhance both precision and recall of the Tigrigna IR system?
- How N-gram technique can be effective on selecting the correct sense identification in the word sense disambiguation?
- To what extent our query expansion model improves the performance of Tigrigna IR system.

1.4 Objectives

General Objective

The general objective of this study is to design query expansion for Tigrigna IR system.

Specific Objectives

The following specific objectives are identified:

- To review related literatures on developing information retrieval system and query expansion mechanisms.
- To study the morphological property of Tigrigna language.
- To build lexical database (WordNet) for Tigrigna language.
- To design query expansion technique.
- To improve Tigrigna information retrieval system.
- To develop the prototype of the system.
- To evaluate the prototype of the system.

1.5 Methodology

In order to achieve the goals specified in specific and general objective of this study, the following research methods will be applied.

Literature Review

For clear understanding of the morphological behavior of Tigrigna language, concepts of information retrieval, models of query expansion, identifying appropriate tools and better techniques, some Tigrigna language books, articles, related research works done by local and international scholars, and other relevant publication will be reviewed and analyzed.

Data Collection

By consulting language experts in Tigrigna language, characteristics of the language will be studied and data collection is carried out from prior researches. The corpus will constitute polysemous and synonym words and word variants of the Tigrigna language.

Evaluation

A prototype of the system will be developed to see the difference on performance of the system. Testing will be performed using query terms with rich features in polysemous, synonyms etc. Finally experiment result on which documents are relevant as per all testing queries will be prepared.

1.6 Scope and Limitation

The goal of the study is to improve the performance of the research done by Ataly [28]. Based on the researcher recommendation, our research automatically expand a query coming from users in order to resolve ambiguity on retrieval of the relevant documents. However, this study will consider only synonym and polysemy nature of words with reference to WordNet, and identify the correct sense of polysemy words in particular. There is limitation in the present study, it doesn't consider other relations like hyponymy, meronym on the WordNet.

1.7 Application of Result

Many researchers indicate that information retrieval system is language specific, that's why different researches have been done on the area of IR for different languages. This implies IR systems need to consider the unique features of the language. Our query expansion model applied in different NLP applications such as, mainly on Tigrigna IR system to have better performance in retrieving most relevant document in ranked list. On question and answering system to analyze users question, on WSD in order to identify the correct sense of polysemous terms and in some forums frequently asked question in order

to analyze the given answers. Therefore, query expansion technique helps to improve the performance of some NLP systems on addressing better results.

1.8 Organization of the Thesis

The structure of this study is organized into six chapters and few sub chapters. Chapter one explains the research background, research problem, objective, scope and limitation, and application of the study are charted particularly.

Chapter two discusses the literature review of researches, thesis, books, published and unpublished journals related to IR system and query expansion techniques. This review helps to understand different theories, facts, techniques, methods of the research area. It includes overview of WordNet and sense identification methods, different expanding techniques, N-gram model expansion term selection technique and evaluation mechanisms. Additionally, background of Tigrigna language, Tigrigna writing system, and characteristics of synonymy and polysemy nature of Tigrigna terms and other related topics are discussed.

Chapter three discusses related works of the research area. It covers different studies of local and global scholars, it gives a clue to what have been done so far about IR systems and query expansion models.

Chapter four covers the design of proposed query expansion technique in detail. Our expansion model, WordNet, WSD, and the architectural view of the prior Tigrigna IR system is discussed. Techniques like morphological analysis and how the proposed techniques work are discussed. In addition, expressions and adopted supporting techniques and their algorithms are depicted in this chapter.

Chapter five discuses about the experimentation and results of the proposed technique. Experimental result measured in recall, precision and f-measure.

Finally, chapter six contains the conclusion made from the findings of the study and recommendations that should be conducted in future researches also indicated.

Chapter 2: Literature Review

2.1 Introduction

With the knowledge growth of human being, written materials representing different knowledge grows as well. From time to time electronic data also growth faster than expected [4]. According to [4] Information is recorded, stored and distributed in four physical Medias paper, film, magnetic and optical. For thousands of years people already understood the importance of archiving and finding information from these storages [16, 24]. In the era of computers technology, store large amounts of information and finding useful information from such collections became possible [16, 4].

This chapter is broadly divided into three sections. The first section discusses concepts related to information retrieval and query expansion operations with the series of activities involved in document and query expansion process. Concepts like user's relevance feedback, explicit and implicit feedback and other techniques based on different aspects of query expansion are briefly introduced. Performance evaluation of IR system to measures of effectiveness of IR system, recall and precision are discussed.

The second section discusses about WordNet, WSD, and N-gram model. It presents the basic tasks involved by WordNet in query expansion such as representing of term by synsets (synonym set) provides in root form, related words and some semantic relations between these synsets in query representation. In the last section, the features of Tigrigna writing system and challenges of the language related to information retrieval is discussed. The Tigrigna alphabets, numbers and punctuation marks are also presented in this chapter.

2.2 Brief History of Information Retrieval

At the time of mankind understands the importance of storing and retrieving written documents that was the period of invention of different mechanism to make it happen. Especially with inventions like paper and printing press. Then after the era of computers were introduced, people realized that they could be used for automatically storing and retrieving large amounts of information [15]. In 1945 Vannevar Bush published a ground-

breaking article titled “As We May Think” the stone corner for the idea of automatic access to large amounts of stored knowledge using web [15].

Despite of much success, the Web has introduced new problems of its own. Finding useful information from the Web is defiantly a tedious and difficult task. Specially, for native users the problem become tiredly frustrate all their effort lost on searching for information of their interest [1]. The main difficulty is the absence of well-defined underlying data model for the Web, this indicates that structure and information definition of knowledges designed in low quality. These difficulties have attracted researcher’s interest in IR and its technique as solution. Then-after the area of IR take attention as other technology [1, 3].

An information retrieval (IR) system assist users as a bridge between their information need and enormous amount of stored information [1, 3]. Even though the task of information retrieval systems is retrieving relevant information, unfortunately there is no such system which can retrieve relevant and only relevant documents as per user’s query [3].

2.3 Query Expansion

In the area of information retrieval, researchers realized that there is a difficulty for users to formulate good search query [1, 11, 17]. According to these researchers, most of users formulate weak queries without detailed knowledge of the document collection and the retrieval environment, this kind of queries lead information retrieval systems to poor coverage of desired result. Because of the above difficulty, most user’s query need to reformulate to obtain the results of their interest. In other word, original users query need to expand. Expanding a query means reformulating users query terms with similar meaning of words [1]. The main reason for developing query expansion is users query assumed to be ambiguous. To solve the ambiguity nature of query terms, users query need to be expanded or enhanced. This enhancement of user’s query, may help IR systems to retrieve relevant document and few non-relevant as much as possible. In this research, methods involving query expansion are investigated.

2.4 Users Relevance Feedback

Relevance feedback approach uses some ranked retrieved document to reformulate the original user's query [1]. Feedback phase's work in documents that are known to be relevant to the original query q are used to reformulate to q_m. The expectation is that the query q_m will return a good result of documents relevant to q. However, obtaining documents relevant to the original user's query is not easy and requires the direct interference of the user [1]. This approach pass through some steps with involvement of end users. First users formulate a query to the system and retrieve some ranked documents as a result second, from these ranked documents, users select some relevant documents to their query, the system uses the ranked documents to reformulate the original query of users, finally the system retrieve some relevant document based the reformulated query [13]. While the IR system expect users on deciding whether the top 10 results for a given query are relevant or not, unfortunately most users are unwilling to provide this information, mainly on the web[1]. Because of this problem, the idea of relevance feedback has been difficult to use for years. But if the information collected from users is related to the original query, it's expected that relevance feedback will produce good results [1]. Feedback approach is composed of two steps explicit feedback and implicit feedback.

2.4.1 Explicit Feedback

In explicit relevance feedback approach, the relevance judgment is provided directly by the users or by a group of human assistants. After users formulate a query as well as users review the top retrieved documents by the system and indicate some of themare relevant to the query [1]. To minimize misconceptions, feedback information is collected from various users and only information that is supported by majority of the users are considered [1]. Since users might be unwilling to corporate on providing feedback, an alternative is to use group of specialists to make the relevance assessments. In other word collecting feedback information from users is not easy and time consuming [1].

2.4.2 Implicit Feedback

In implicit relevance feedback, there is no involvement of user in the relevance judgment. Instead, the feedback information is derived implicitly by the system. Query expansion algorithms at first evaluate given query on collection of documents and then select from relevant documents appropriate terms. The original users query expanded with such selected terms from the first ranked document. The reformulated query is used to retrieve new set of relevant documents [1]. If the first set of relevant documents in the whole document collection is known, one can calculate the query vector that can fit the whole of the relevant documents in the corpus [30]. The best query vector technique used for differentiating the relevant documents from the non-relevant is stated by the Rocchio's algorithm which is called query expanding algorithm based on vector space model. This equation is given *Equation 1*[1].

$$\vec{q}_{opt} = \frac{1}{|C_r|}\sum_{\forall \vec{d}_j \in C_r}\vec{d}_j - \frac{1}{N-|C_r|}\sum_{\forall \vec{d}_j \in C_n}\vec{d}_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where C_r are set of relevant documents, C_n is the set of non-relevant document, N is the total number of documents in the total collection and d_j is any document in the total document collection, which sometimes can be relevant or non-relevant. The conceptq_{opt}holds an assumption that, term-weight vectors of relevant and non-relevant documents are dissimilar. Hence the core idea to reformulate the query is such that it gets closer to the term-weight vector space of the relevant documents and in turn away from the non-relevant once. As Baeza-Yates et al. [1] wrote that, the problem is the relevant documents are not known or it is an unrealistic situation. The way to avoid this problem is formulating an initial query q to q_m and to incrementally change the initial query vector. This incremental change is accomplished by restricting the computation to the documents known to be relevant according to the user's relevance judgment. There are three classic and similar ways to calculate the modified query q_m as follows [1].

Standard_rocchio:

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{N_r}\sum_{\forall \vec{d}_j \in D_r}\vec{d}_j - \frac{\gamma}{N_n}\sum_{\forall \vec{d}_j \in D_n}\vec{d}_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Ide_Regular:

$$\vec{q}_m = \alpha\vec{q} + \beta\sum_{\forall\vec{d}_j \in D_r}\vec{d}_j - \gamma\sum_{\forall\vec{d}_j \in D_n}\vec{d}_j \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Ide_Dec_Hi :

$$\vec{q}_m = \alpha\vec{q} + \beta\sum_{\forall\vec{d}_j \in D_r}\vec{d}_j - \gamma\max\_rank(D_n)\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

There is an understanding of these three techniques have similar results [1]. These techniques are carried out, based on multiple executions of their algorithm, until it is close enough to the ideal query proposed in *Equation 1* which is called query expanding method based on vector space model. The main advantage of the relevance feedback is introduced better results and simplicity but optimality (i.e. the number of iteration to produce vector q_m) remains a problem [3].

The relevance feedback model for probabilistic model which is dynamically ranks documents similar to a query q according the probabilistic ranking principle. Retrieved documents can also be made more relevant than initially retrieved once, using the similarity of document d_j to a query q in the probabilistic model. The formula for similarity in probabilistic model is given in *Equation 5*[1].

$$sim(d_j, q) = \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{P(k_i|R)}{1 - P(k_i|R)}\right) + \log\left(\frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})}\right) \ldots (5)$$

Where $P(k_i|R)$ and $P(k_i|\bar{R})$, are the probability of term k_i existing in the relevant document set R, and in the non-relevant document set R respectively. Unlike the vector space model based relevance feedback, the method based on probabilistic model doesn't expand queries rather it adjusts the weights of the query terms as they are. The problem encountered in this model is the values for $P(k_i|R)$ and $P(k_i|\bar{R})$ is not known, and thus initial assumptions are taken as 0.5 and $\frac{n_i}{N}$, respectively. Where, n_i is the number of documents that term k_i is found, and N is the total number of documents in the corpus. After retrieving documents based on the assumed values, a user judges the documents by marking them as relevant or not. Based on this judgment, the values of $(k_i|R)$ and $(k_i|\bar{R})$ are changed to $\frac{|D_{r,i}|}{|D_r|}$ and, $\frac{n_i - |D_{r,i}|}{N - |D_r|}$ respectively, where $|D_{r,i}|$ is the number of documents where term ki exists in,

and $|D_r|$ is the number of relevant documents as judged by the users. Although using this model enhances precision, it have three problems [1]. First, document term weights are not considered during the feedback loop. Second, weights of terms in the previous query formulations are ignored. And third, no query expansion is used (the same set of index terms in the original query is reweighted over and over again) [1].

2.5 Pseudo-Relevance Feedback

In relevance feedback the method of reformulating queries involves users. User's involvement on reformulating query is tedious and time wasting for them, which in turn reduces their interest of further searching [3]. Pseudo relevance feedback one the query reformulation approach which doesn't need the user's interference on the expanding process. The system automatically reformulates users query without their knowledge [14]. First retrieve some ranked documents based original users query, then by analyzing the first retrieved document, users query transform to the reformulated query to retrieve final results for users [14].

Implicit Feedback through Local Analysis

Implicit feedback through local analysis is a method that uses the documents which are considered as relevant by the system in the first retrieval attempt to reformulate the original query. This means, the documents retrieved for a given query q are examined at 20 query time to determine terms for query expansion [1]. There are two kinds of local automatic analysis techniques; query expansion based on local clustering and query expansion based on both local and global analysis [1].Local clustering is to build global term correlation structures that quantify term correlations and then use the correlated term for query expansion [1, 19]. The general idea is that, terms which co-occur in most documents are synonyms and they are talking about pretty much the same concept or meaning and the higher the number of documents in which the two term co-occur, the stronger this correlation. An alternative approach is to expand the query using information from the whole set of documents in the collection or corpus is called global analysis.

2.6 Local Context Analysis

Local context analysis approach is combination of global and local analysis. It is based on the use of noun groups, i.e., a single noun, two nouns, or three nouns in the document. The

local context analysis have three steps in order to expand a query.

First, retrieve the top n ranked documents using the original query Second, for each concept c in the passages compute the similarity sim(q, c) between the whole query q and the concept c. the query q is treated as a whole rather than as single concepts. The similarity sim(q, c) is calculated using *Equation 6*[1].

$$sim(q,c) = \prod_{k_i \in q} (\delta + \frac{\log(f(c,k_i) * idf_c)}{logn})^{idf_i} \dots\dots\dots\dots\dots\dots \quad (6)$$

Where n is the top ranked documents, and δ is a constant value having the value '0.1' which prevents the value of $sim(q, c)$ not to be 0 [1]. (c, k_i) is the association between concept c and term $k_i$found in query q, and is calculated using *Equation 7*[1].

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} * pf_{c,j} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (7)$$

Where$pf_{i,j}$, and$pf_{c,j}$are frequencies of term k and concept c in j^{th} passage respectively. The inverse document frequency for c and $k_i$are calculated using *Equation8 and 9*[1].

$$idf_c = \max(1, \frac{log_{10}N/npc}{5}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (8)$$

$$idf_i = \max(1, \frac{log_{10}N/npi}{5}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (9)$$

Where N is the total number of passages in the corpus, $np_c$and $np_i$are the number of passages containing concept c and term $k_i$respectively.

Finally, the top $n$ranked concepts according to $sim\ (q, c)$ are added to the original query. Then each added concept assigned a weight 1 - 0.9 * in, where i is the rank of the concept, while terms which were in the original query are assigned weight of 2.

Query expansion based on local context analysis has been castoff to for TREC corpus collections, so it may not perform well for other corpuses. Thus, it is recommended, that

it needs adjustments for different document collections in order to gain good performance [1].

2.7 Term Selection for Query Expansion

The main reason why query expansion is suggested as a solution for information retrieval is because users may not be satisfied at first attempt of retrieving, therefore expanding users query is good solution to retrieve more relevant documents [3, 19].

When a term is selected to expand a query, it is assigned weight ωt. In this case, before the inclusion of the new term t in the query, there were two document allocations, which are relevant and non-relevant according to the first retrieval. The assumption of the inclusion of term t doesn't affect the original relevant and non-relevant document distribution called the independence assumption [23]. Therefore, the new distribution for the relevant documents according to the independence assumption consists of a mixture of the original relevant documents and the original displaced upwards by ωt. This is done by making use of the probability p defined in *Equation 10*.

$$\omega t = \log \frac{p_t(1-q_t)}{q_t(1-p_t)} \quad \text{(10)}$$

Where $p_t$is a probability that a given relevant document is assigned to some expanding term t, and $q_t$is equivalent non-relevant probability.

$$(1-p_t)u + p_t(u_R + \omega_t) = u_R + p_t\omega_t \quad \text{(11)}$$

Where $u_R$is the mean value of the original relevant document and $u_N + q_t\omega_t$, the new mean for the non-relevant items. The effectiveness of the system can be calculated by the distance between the two means, given by *Equation 12*.

$$d' = u_R + p_t\omega_t - u_R - q_t\omega_t$$

$$= u_R - u_N + \omega_t(p_t - q_t)$$

$$= d + \omega_t(p_t - q_t) \quad \text{(12)}$$

Where d is, the distance between the original relevant and non-relevant documents. In other words, the inclusion of term t in the search formulation, with weight ω_t increases the effectiveness by $a_t = (p_t - q_t)$ [23]. Robertson [23] concluded that, as long as $a_t$and $\omega_t$are not the same, a decision rule for the inclusion of new terms should be based on $a_t$rather than ω_t.

2.8 Ontology Based Query Expansion

The concept of ontology indicates an association of some knowledge domain, which contains relevant entities and their relations on the knowledge domain. These entities are concepts or terms, which are linked to each other using relations like, subsumption and aggregation. Ontology can be considered as intentional semantic structure, which contains the implicit rules constraining the structure of a piece of reality [18].

Conceptual Network Editor (CONE), is a four step technique to construct ontology [18].

- Constructing dictionaries of common concepts from the existing repositories or documents at hand.
- Correlating each concept with terms enclosed by the ontology.
- Assigning a weight each term, articulating the level of accuracy of representing the concept.
- Finding and expressing relations between concepts.

The relationships found on ontology, have their own meaning. For example, in Taveter et.al [18], subsumption and aggregation means (is-a) and (includes-a) respectively. According to [18], subsumption and aggregation are assigned 0.7 and 0.3 respectively based on the model used to develop an ontology.

Query expansion based on ontology contains, matching query terms with concepts and select terms which have relations with query terms, according to the weights the relationship holds. In this case, those terms with higher weights are selected, because their level of similar to the query is higher than the rest of the terms, which have relations with the query terms. Therefore, expanding a query based on ontology is totally dependent on the quality of the developed ontology. This implies that the more the ontology is accurate, the better the results will gain are and vice versa.

2.9 Concept and Context

The other approach for information retrieval to retrieve relevant documents is retrieving based on meaning of terms. Some researchers [13, 20] shows that, a meaning can be expressed in terms of concepts and contexts. Contexts and concepts can be defined in a different ways. According [13, 44] all words are concepts except stop words and the context for a given word is all the words that co-occur in documents with that word. Concept is mutual associations between terms or words and documents [20]. For example suppose a_1,a_2…..a_n are set of documents in A and w_1, w_2,…w_n are set of words in W. If all members of set W describe all members of set A, then we can say the WxA is a concept. But if some of the terms in W describe some of the documents in A then, this can be context [20]. Based on the above relation, retrieval systems can use meaning based retrieval, this may lead to good coverage of retrieving relevant documents.

2.10 WordNet

Miller and Johnson-Laird (1976) proposed a research concerned with the lexical element of language called psycho-lexicology [21].Later, in 1985 group of researchers at Princeton University started to develop a lexical database (Miller, 1985). The starting point of these researcher's was offer searching dictionaries conceptually, instead of alphabetically, with an on-line dictionary. As a result, researcher's idea, they develop the first online conceptual dictionary called WordNet [21]. WordNet is lexical database used as knowledge base in natural language [21, 48, 49]. The unique quality of WordNet is network of a word based on their sense. Its structure is like dictionary stores words and meanings in relation to each other but it differs from traditional dictionary in some ways [49]. For example, words in WordNet are arranged semantically instead of alphabetically, the relationship between words in WordNet is the Synonym relation called Synset [47, 49]. Words in the same synset are synonymous at least in one sense. Word sense is the meaning of words depending on how contextually they used. For example, the word "mule" could mean an animal like horse in one sense and a germinal in another sense.

In WordNet, one word may occur in many synsets as it has different senses [46]. WordNet synsets also sometimes may contain compound words which are made up of two or more words but are treated like single words, but most of the time it contains single words.

There is a standard WordNet for English language (Miller, 1985).It contains 117,097 nouns, 22,141 adjectives, 11,488 verbs and 4,601 adverbs. The current version available for download is WordNet 3.0, which was released in December 2006 but there is a later release, 3.1, which is available for online usage only[26, 42]. Approximately 40% have one or more synonyms. WordNet stores information about words that belong to four parts–of–speech: nouns, verbs, adjectives and adverbs. Semantic and lexical relations between words and synsets are the basic relations of WordNet. Semantic relation is relationships between two synsets [48]. Lexical relations are relationship between two words within two synsets of WordNet [50, 51]. The semantic relations defined in WordNet are [29]:

- ***Synonymy*** **-**it is a semantic relation between synonym words.
- ***Antonym*** - is a symmetric semantic relation between antonym words.
- ***hypernymy (super-name)***- are transitive relations between synsets.
- ***Hyponymy (sub-name)***- semantic relation organizes the meanings of nouns into a hierarchical structure.
- ***holonymy (whole-name*****) -** are complex semantic relations.
- ***Meronymy (part-name)*** **-** semantic relation distinguishes component parts, substantive parts, and member parts.
- ***Troponymy (manner-name*****) -** is for verbs while hyponymy is for nouns, although the resulting hierarchies are much thinner.
- ***Pertainyms*** **-** is a relation which points the adjective to the nouns that they were derived from.
- ***Entailment*** **-**relations between verbs are also coded in WordNet

2.10.1 WordNet Structure

WordNet is a container of three separate databases, one for nouns, second for verbs and third one for adjectives and adverbs. The basic structure is synsets, synsets of WordNet contains a set of words, and these sets are labeled with the sense they represent [35, 42]. These senses can be said to be concepts, all of words can be said to express the same concept but different sense based context. Word forms which have different meanings appear in different synsets. Each of these synsets are also connected in some way to other synsets, expressing some kind of relation [42, 47]. The other concept doesn't consider in WordNet is Pronunciation. For instance, the noun **bass,** the pronunciation differs whether

talking about **bass** in the sense of the low tone or the instrument, or talking about the fish **bass** [42].

According to [42], Nouns have the richest set of relations of all parts of speech, which is 12 different relations represented in WordNet. As previously stated relation types, the hyponym/hypernym semantic relation is the most frequent used in noun. WordNet also separates the hyponyms between types and instances. An instance is a specific form of hyponyms, and these instances are usually proper nouns, describing a unique entity. These instances go both ways, just like the types [35, 42].

Verbs in WordNet are like nouns, have the hypernym relationship. The semantic relationship among verbs called troponyms. These drives from the event to a superordinate event. Troponyms can also be described as in which manner something is done, therefore explaining the difference of names. Antonym also other semantic relation exists for verbs in WordNet, and performs in the same way as ordinary, **go** is an antonym of **come** [26, 42].

Finally, the adjectives group are mostly organized in the terms of antonym. Like in the way of nouns and verbs, these are words which have meanings that are semantically opposed in adjective database. As all words in WordNet, they are also part of a synset.

2.10.2 Relations across Part of Speech

Most of the relations exist in WordNet are relations among words of the same part of speech. Additionally, there is some pointers across the subfields of part of speeches. The first pointer is pertainyms, which points from an adjective to the noun that was derived from. There are also pointers points to semantically similar words which share the same stem, called derivationally related form. For many of these pairs of nouns and verbs, the thematic role is also described. For instance, the verb **play** has a pointer to the noun **player**, and **player** would be the agent of **play** [42]. These are the relations exist between words in WordNet. Particularly there is a lack of part speech relations across to different part of speeches.

2.10.3 Query Expansion Using WordNet

Query expansion using WordNet is broadly used technique that attempt to improve retrieval system by adding semantically related term called expansion terms to a user query. The expanded query is expected to retrieve more relevant documents, thus improving overall performance of the IR system. In query expansion, the important thing is the source of the expansion terms. Researchers [16, 22, 46], experiment indicates that there is variety of sources for collecting expansion terms. These terms could be taken from the whole target collection or from a few documents ranked retrieved in response to the original query. Candidate expansion terms (CETs) are selected for including in to the original users query in order to expand [22].Sometimes lexical database like ontology or WordNet are used as source of expansion terms in different researches.

Researchers [21, 22, 42] stated that, the use of WordNet as a source for query expansion could be effective technique. By the sense of query words and sentiments. When we use WordNet as CETs, there are some point need to be considered:

- If a query word occurs in multiple synsets, which synsets would be selected?
- Once one synsets should be selected, which words should be added to the query?
- Should only synonyms contained on these synsets be added? Or others also considered?

As stated above in query expansion the very important decision is to choose the source of candidate expansion terms. A set of sample relevant documents for a given query would be a good source of CETs [47]. WordNet based query expansion typically starts with synonyms (possibly holonymes and meronyms etc.) of the query words as CETs. According [22] two terms tend to be strongly related if their WordNet definitions share many common terms. Thus, if the definition's of a term shares words with query word definition's, then the term may be semantically related to the query, even though it may not be a direct synonym of (or otherwise explicitly related via WordNet relations to) query words. This idea has been used to quantify the relationship between a CET t and a query word q_i as follows. The definition of t and q are considered as two sets of words, and the overlap between these two sets is taken as Rel (t, q_i), the semantic similarity between t and q_i is computed as follows

$$Rel_{t,q_i} = \frac{C_{t,q_i}}{c_t + c_{q_i} - c_{t,q_i}} \quad \text{.....(13)}$$

$$Rel_{t,q_t} = \frac{2 * c_{t,q_i}}{c_t + c_{q_i}} \text{.....(14)}$$

The overlap may be measured using either the Jaccard coefficient (26), or the Dice coefficient (27). Here, Cx represents the number of documents in which term x occurs, and Cx,y is the number of documents in which x and y co-occur [22].According to [22], experimental results show that Dice co-efficient performs slightly better than Jaccard co-efficient.

2.10.4 N-gram Model

Language models are useful for many NLP systems, such as OCR, WSD, IR and MT. One uses of statistical language model in IR is selecting correct sense of word among the given possible senses, based on the sequence of words in the local context of the ambiguous words. One of these statistical models we use in our study is N gram model. An N-gram is simply a sequence of preceding and successive n words along with their count. This considers the occurrences in collection of data [12]. In other hand Markov assumptions are applied which states that current word does not depends on the entire history of the word but at most on the last few words [12]. The number of words in the local context of ambiguous words makes a window. The size of window i.e. number of words to be considered at ± n positions is important because while constructing n size window following factors need to be considered [2].

When the value of n is large, the probability of getting correct word sense is high. i.e. in general, large collection of data will always improve the result. But on the other hand, most of the higher order n grams do not occur in large collection of data. This is the problem of sparseness of data.

As data size increase, the size of model also increase which can lead to models that are too large for practical use. The total number of potential n grams scales exponentially with n. Computer up to present could not calculate for a large n because it requires huge amount of memory space and time [2].

Based on [2], language model for word sense disambiguation have two evaluation metrics. These are entropy and sense disambiguation rate. In this study, we select the effect of size

n of window by associating them with perplexity and sense disambiguation rate. Word sense disambiguation rate defined as percentage of words which are correctly disambiguated in the query. In other hand entropy is a measure of information and it can be used as metric for how predictive a given N-gram model is about what the possible sense of word could be.

$$Entropy = -i/n \sum_{\frac{w}{n}} pr(w_1 \dots w_n) log_2\, pr(w_1 \dots w_n) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (15)$$

$$Preplexity = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{pr(w_i|w_1,w_2 \dots . w_{i-1})}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (16)$$

Where w1,w2,.....wn a word sequence, n=total number of words in the test set and pr is conditional probability. The goal the above equations are to obtain small values of these measures. Language model with lower perplexities and entropies tend to have higher word sense disambiguation rates [2, 53].

2.11 Performance Evaluation

Performance evaluations metrics are used to measure the degree of some objectives stated to achieve by the system [40]. Any system is developed to meet some targeted functionality. These functionalities are evaluated to make sure that, either the system performing as designed to meet the specified objectives. The evaluation metrics used for IR systems is known as retrieval performance evaluation [50]. There are various retrieval performance evaluation techniques [43]. In our research recall, precision and F-measure are used.

Recall and Precision

Recall and precision are statistical measurements used to evaluate the performance of information retrieval systems [43]. IR systems can be evaluated based how the system can retrieve relevant documents and few irrelevant documents as much as possible [1]. To evaluate recall and precision of a system first relevant documents need to be prepare according relevance judgment. And calculate recall and precision as follows:

Recall is some relevant documents retrieved by the system among the whole of the relevant documents identified by relevance judgment, and given by Equation 17 as stated in [1]:

$$Recall = \frac{Relevant\ document\ retrived}{Relevant\ document} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (17)$$

Precision means relevant documents retrieved by the system among the whole of the documents retrieved, and is given by Equation 18As stated in [1]:

$$Precision = \frac{Relevant\ document\ retrived}{Retrived\ document} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (18)$$

F-measure is average combines of both recall and precision and is given in Equation 19[1].

$$F - measure\ = \frac{2*\mathrm{P}*R}{\mathrm{P}+\mathrm{R}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (19)$$

These evaluation metric evaluate the system in two different ways [43]. Precision evaluates how much the system retrieve relevant documents among the whole document collection and recall evaluates how much relevant document retrieved among the whole relevant documents. And the average precision can be calculated on some recall level as follows:

$$\vec{P}(r) = \sum_{i=1}^{N_q} \frac{p_i(r)}{N_q} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (20)$$

Where $\overrightarrow{P}(r)$ is the average precision, N_q the number of queries used and (r) is the precision found for query i at recall level r. Since the recall levels for each query might be different from the 11 standard recall levels, use of an interpolation procedure is often necessary [1]. This is because average precision might disguise important anomalies in the retrieval algorithms and thus, it doesn't provide evaluation based on each single query [1]. For these situations, a single precision value for each query can be used, that is taking a precision at some recall level for each single query.

2.12 Tigrigna Writing System

Tigrigna is a Semitic language which is spoken by the Tigray people located in northern Ethiopia and Eritrea. For those regions, Tigrigna is a working language. There are around 8.5 million Tigrigna speakers worldwide [25]. The Ethiopic writing system is used to

represent the Semitic languages. The Ethiopic system for Tigrigna language consists of alphabets, numbers, and punctuation marks [23].

I. Alphabets

Alphabets are sets of letters arranged in fixed order used to write. They are also called phonemes which contain consonants and vowels. There are different alphabets representations in the world. The most alphabets representation is Latin or Roman alphabets which have been adapted by many languages. The Ethiopic writing systems have also their own writing systems. Similarly, Tigrigna alphabets used to write Tigrigna documents. There are seven representation of Tigrigna language alphabets [7, 24].

Tigrigna alphabets have equivalent translation in Latin representation by finding a Latin letter with similar sound Tigrigna letter. For example, the Tigrigna letter '𝙪' has a similar sound with Latin letter 'h'. Thus, the seven order of the letter '𝙪' can be represented by combining the Latin letter 'h' with vowels as shown in *Table 2.1*.

Table 2.1Sample of Tigrigna letter and their corresponding Latin letter

Order	1st	2nd	3rd	4th	5th	6th	7th
Tigrigna language	𝙪	𝙪ሁ	𝙪ሂ	𝙪ሃ	𝙪ሄ	𝙪ህ	𝙪ሆ
Equivalent Latin letter	He	hu	hi	Ha	hie	H	Ho

II. Punctuation Marks

Tigrigna writing system contains different kind of punctuation marks for word demarcation. Different punctuation marks have different usages in Tigrigna texts.

- The word separator mark (:) used in the old literature to separate one word from other words
- The end of sentence mark (: :) is used to shows when an idea is finished.
- The sentence connector mark (፣) is used to connect two sentences in to one sentence.
- The list separator mark (፡) is used to list things, separate parts of a sentence, and indicate a pause in a sentence or question.

- List mark (፡-) is used at the beginning of the lists.
- Question Mark (?) Is used at the end of a sentence which is a question,
- Exclamation mark (!) is used to indicate end of an emphatic declaration, or command and
- Double quotation mark (“) is used to quote some words or sentences taken from others [7].

III. Number Systems

Like other languages Tigrigna language have its own numbering system. According to [7] most of the time, Tigrigna number system are used for calendar purposes. Because of this European number system (1-10) are used for arithmetic computation in the many Tigrigna literatures.

2.13 Challenges in Tigrigna Language for IR

According to [7, 28] it was noted that information retrieval system face some challenges in Tigrigna language for text processing, some of these challenges are:

2.13.1 Spelling Variation of the Same Word

Tigrigna language have challenge of spelling variation in current literature. A word may be interpreted by different persons using different spelling variation. For instance, the word “television” can be written as “ተለቪዥን”, “ተለቭዥን” or “ተለቢጅን”. All these words are used to represent the word “television” in Tigrigna. This variation of spelling in Tigrigna words increases ambiguity and inconsistency in Tigrigna text documents.

2.13.2 Redundancy of Some Characters

In some of Tigrigna language literatures there are some characters which are redundantly used. These characters are the same in pronunciation and usage but different in character representation. Unlike humans, IR system face difficulties to determine these differences. For example, “ፀ and ጸ,” “ሠ and ሰ” these characters have the same usage and meaning in all Tigrigna text documents but have different symbolic representation. Thus, IR system challenges on determining the characters as the same meaning. Because IR systems work

by direct matching of query words with Tigrigna documents [3, 27]. Characters in Tigrigna language using interchangeably are given in *Annex C*.

2.13.3 Abbreviation

The abbreviations of Tigrigna words follow different formats. Either period '.' is or '/' symbol is used to abbreviate a word. The abbreviated words can be written without separators. For example, the word "society" in English "ሕብረተሰብ" (hbreteseb) can be written as "ሕ/ሰብ" (H/seb) or "ሕ.ሰብ" (H.seb). The variations of abbreviating Tigrigna words produce inconsistency in information retrieval system [7, 27]. Sample of abbreviated words given in *annex B*.

2.13.4 Compound Words

Tigrigna compound words are another challenge of Tigrigna IR systems. These words formation made in two ways, space and hyphens are used to separate them. When the hyphen is used the two words are treated as one word. However, when they are separated by space their meaning differ. For example, "ቤት-ትምህርቲ", "ስነ-ስርዓት", "መራሕቲ-ስድራ", and "ኣብያተ-ፅሕፈት" are compound words separated by hyphens [7]. There are also other words which prevails new meanings when combined with another word. For example, the word "ዐይኒ" which has an obvious meaning of "human eye", when combined with other terms or concepts it has different meanings, such as "ዐይኒ-መርፍእ" which indicates needle's hole" or "ዐይኒ-እንጀራ" which means, pores kind of things on the Ethiopian traditional food እንጀራ. Unless Tigrigna IR systems use some mechanisms to handle these problems, retrieving of relevant document could be challenging task to get better results.

Chapter 3: Related Work

3.1 Introduction

Developing good query expansion model for information retrieval can improve the performance of the system. Even if Tigrigna IR systems needs this model, but research area of query expansion not yet covered in Tigrigna language. The traditional way of retrieving documents is only based on matching of index terms with the query term. Because electronic documents are getting increased each day, the existing Tigrigna IR can't handle the burden independently. So query expansion model producing the meaning of the query term rather than finding relevant documents by direct matching only. So far to our knowledge, for reformulating users query, there is no research made yet on query expansion system for Tigrigna language. In this section the discussion proceeds with related IR Systems for Ethiopian and non- Ethiopian researchers made so far.

3.2 Tigrigna Information Retrieval System

Tigrigna is the Ethio-Semitic languages which belong to Afro-Asiatic super family [25].The sematic language Tigrigna is spoken in Eritrea and northern Ethiopia, there are also around 8.5 million speakers worldwide [25, 27].

Atalay Leul [28] proposed a probabilistic information retrieval system for Tigrigna language. This language specific retrieval system considers the language specific characteristics. In this research for Tigrigna IR system using the probabilistic model which has the mechanism to reweighting query terms using relevance feedback and in this IR system, the Binary Independent Model (BIM) is chosen and implemented. Additionally, the model does define uncertainty that exists in IR systems. It is a pioneer research on IR for Tigrigna text documents. This research is initiated to experiment the effectiveness of an IR system for Tigrigna using a rule-based Tigrigna stemmer developed by Yonas Fisseha [6]. This research, the potential of probabilistic model in Tigrigna text retrieval is investigated. 300 Tigrigna documents and 10 queries were used to test the approach. Then, the retrieval system is tested in two ways which is before pseudo relevance feedback and after pseudo relevance feedback. The system registered, after stemming and pseudo relevance feedback an average precision of 69.1%, recall 90% and F- measure 74.4%. This

result is face problem in controlling of synonyms and polysemous of terms that exist in Tigrigna text.

Hailay Beyene [27] proposed a general architecture for Tigrigna search engine. This research develops using Boolean search and tries to consider some characteristics of Tigrigna language. This study tries to handle the morphological variations of Tigrigna words using Tigrigna stemmer developed by Girma. To implement the search engine java programing language were used. The test environment of this search engine takes place on different Tigrigna sites like www.dmtsiweyane.com/ (website of Tigray regional state radio program), www.bahlitigrai.com/ (website of Tigray Culture Association) … etc. But this work as it's the first work of Tigrigna search engine it attempts to implement IR system not included query expansion mechanisms it only contains expanding short and long terms, in other words, short terms which have period or slashes to long terms or vice versa, IR system designers may face challenges on query reformulation issues, besides challenges that occur from the user's perspective. The result obtained was a precision value of 95% for AND operator and 47% for OR operator and a recall value of 73% for AND operator and 99% for OR operator [27].

3.3 Query Expansion for Amharic Language

Amharic is the official language of the federal government of Ethiopia. This sematic language is widely spoken by more than 17 million people as a first language and by more than 5 million people as a second language [10]. It is the official language of the state, the day-to-day language of the church, the language of primary education as well as a mother tongue of over fifteen million people [8]. There are some papers trying to investigate better method in query expansion on Amharic language.

Abey Bruck [3] propose semantic based query expansion for Amharic IR system this research goal was at improving the precision while preserving the original recall of an existing Amharic IR system developed by Alemayehu (2002). Statistical co-occurrence analysis, bi-gram analysis and bi-gram thesaurus methods have been introduced in this work. Experiments were done to identify the better technique which can improve precision of the system, without affecting the recall ability of the system. In this work, he uses Amharic bible old testament taken as a document corpus. It contains 21,000 steamed terms and 930 document and nine queries with polysemous terms are formulated for testing. The

results show that the statistical co-occurrence method out performed bi-gram and bi-gram thesaurus methods in terms of F-measure by 2%. It has also been observed that all the techniques improved the precision of the original system considerably. Statistical co-occurrence method augmented the precision by 14%, and the other two methods both augmented it by 18%. And test result of this work showed that statistical co-occurrence method outperformed other than the two approaches, and scored 10% improvement in total F-measure.

Abey Bruck and Tulu Tilahun [14] proposed bi-gram based query expansion technique for Amharic information retrieval system. The aim of this research is to increase precision of an Amharic information retrieval. This research studied the bi-gram technique for retrieving relevant document from large corpus. The bi-gram technique which uses the pseudo relevance feedback for expansion selects those terms which appear to the right or left of a query term. Therefore, there is a need for IR systems to adopt an approach to check whether expanding terms are polysemous or not.

The results show that the bi-gram technique improved in terms of F-measure by 8% in terms of precision in 16%. It has also been observed that the technique improved the precision of the original system considerably. The original F-measure of the information retrieval was 0.53 and precision was 0.39. The bi-gram technique improves to 0.61 and 0.57 respectively.

3.4 Query Expansion for English Language

Hui F. [26] proposed a re-examination of query expansion using lexical resources. The aim of this study was to show that query expansion using only manually created lexical resources can significantly improve the retrieval performance. In this research, axiomatic approaches and similarity functions were studied to develop retrieval functions and the comparison of similarity functions on two lexical resources (WordNet and dependency-thesaurus) were made, then incorporate these similarity functions into the axiomatic retrieval framework. In this research, they try to re-examine the problem of assign appropriate weights using recently proposed axiomatic approaches and find with appropriate term weighting strategy, they are trying to exploit the information from lexical resources to significantly improve the retrieval performance. Their empirical results tested are conducted over six TREC collections. Based on the experiment the similarity function

based on synset definitions is the most effective one [26]. QE_{def} significantly improves the retrieval performance for all the data sets. For example, in trec7, it improves the performance from 0.186 to 0.216, and in trec8, it improves from 0.225 to 0.266.

Xu and Croft [13] proposed Query Expansion Using Local and Global Document, in this study small database is built from pseudo-documents. A filtering step is used to remove words that are too frequent or too rare in order to control the size of the database [13]. This database is then used to retrieve related concepts, in this case, pseudo-relevance feedback documents and is combined with the original query and ranked. Phrases containing only terms in the original query are weighted heavily than those containing terms not in the original query. Local analysis as described in this research uses the originally retrieved documents (i.e. documents retrieved using the original query) for expanding purposes, and it doesn't literally expand the query but adjusts the weights of the query term based on the top ranked retrieved documents. But this method has biasness on the documents retrieved on the first attempt. Thus, queries that perform poorly and retrieve few relevant documents would seem likely to perform even worse after local feedback, since most words added to the query would come from non-relevant documents [13].

3.5 Query expansion for Chinese language

Zihihan li [11] proposed Improvement of Chinese Information Retrieval by Incorporating Word Segmentation and Query Expansion. In this research, different approaches were implemented in the first approach they propose a hybrid Chinese information retrieval model by in-cooperating word based techniques and traditional character based techniques. The aim of this approach is to investigate the influence of Chinese segmentation on the performance of Chinese information retrieval. Two ranking methods which is character-based ranking and word-based ranking were proposed to rank retrieved documents based on the relevancy to the query. But Chinese segmentation is not significant if it incorporates with character based approach, so the second query expansion approach was proposed, its novel query expansion method which applies text mining techniques to find patterns from the retrieved documents that highly correlate with the query term and then use the relevant words in the patterns to expand the original query. The author develops and implements Chinese information retrieval system for evaluating the proposed approaches. The NTCIR5 Chinese document collections were used in the experiments it contains 434,882 documents of news articles in traditional Chinese. The

experiment has been conducted to evaluate the performance proposed approaches Zhihan used with the benchmark Rocchio's query expansion method using 50 queries, results show that, by the hybrid model incorporating with character frequency and segmented word frequency query expansion methods was given slight improvement, when he compared with both it gives 30.1% in precision and 33.5% in recall. But the hybrid model incorporating with the text mining based query expansion given significant result, which was shown in precision from 32.5% to 36.4% and in recall from 36.8% to 41.7%.

3.6 Summery

Query expansion is an effective technique to improve the performance of information retrieval systems. Although hand-crafted lexical resources, such as WordNet could provide more reliable related terms, the usage of these lexical resource gets poor coverage. In different studies showed that query expansion technique leads to performance improvement. Research have been done for Tigrigna retrieval system without query expansion technique. Even if different researchers attempt to develop query expansion model for different languages, query expansion techniques depend on characteristic of the language. As far as our knowledge, there is no attempt of query expansion system for Tigrigna language. Therefore, Tigrigna language needs development of query expansion model which considers the characteristic of the language.

Chapter 4: Design of TIRS with QE

4.1 Introduction

IR systems retrieve documents based on some similarity measurement technique. In addition, an IR system measures their performance level in terms of recall and precision. Like any other systems, most of the time to score 100% on both precision and recall in the case of information retrieval systems performance is difficult. But good systems are designed to augment both precision and recall to the possible limit. Thus, the aim of this research is, to design a good system which satisfies the desirable retrieval constraints. That would be an IR system likely good system. This can be attained by integrating a query expansion model with the system, then words having similar meanings with user's query will be considered and retrieve relevant documents that satisfies the user's query.

In our research, we design a model of expansion for the system to consider synonymy and polysemy behavior of terms. The proposed method that use this assumption which states that, all query terms are short and ambiguous which has more than one contextual meaning based on the document corpus. This hypothesis helps the system to consider the nature of polysemous query term. Thus, the research question is how can a system selects a meaning for query terms? If a user inputs a query with more than one query term, then that query by itself is a context and those collections of terms are organized to part of a meaning. If these query terms are considered as polysemous terms, then they must share a meaning. Therefore, finding a similar meaning of the query terms can be the challenges for IR system. The proposed method selects expansion terms that can be given for the polysemous query terms as a synonym. In our research, we use a set of related words to select the synonym of polysemous query term. But the set of related words are expected to be related as preceding and succeeding words for the ambiguous word or words in document corpus. In information retrieval, adding appropriate synonyms to a query can improve retrieval effectiveness. However, most query terms have multiple meanings and adding a synonym of the query term which has a different meaning in the context of the query would cause deterioration in retrieval effectiveness. Therefore, determining the correct sense of each query term is important for effective retrieval. Once a query terms sense in a query context is determined, synonyms with the same meaning of query term

are added to the query, thus that documents having these synonyms words but not the actual term can be retrieved. In our research, we design WordNet to select correct sense of ambiguous words using n-gram model for query expansion to enhance Tigrigna information retrieval system.

The following few sub chapters illustrate the idea of using an expanded and good query to retrieve ranked relevant documents. The design of the proposed method and the discussion of expanding the given queries are presented.

4.2 Architecture of TIRS with QE

As shown in *Figure 4.1,* the IR system is expected to search and retrieve relevant documents based a given query from users. To improve the performance of IR system, there is a need to apply query expansion. In research, lexical resource like WordNet is developed as a reference for identifying the senses and meaning of the polysemy query terms using word sense disambiguation. The identified word senses are used during query reformulation. Finally, the query expansion module will be integrated to the Tigrigna IR system. Then the system retrieve relevant documents based on the formulated users query.

This chapter discusses the general architecture of query expansion model for Tigrigna information retrieval system. The proposed solution is designed to increase the number of relevant Tigrigna document to be retrieved by the system. Our query expansion model is represented in *Figure 4.1*.

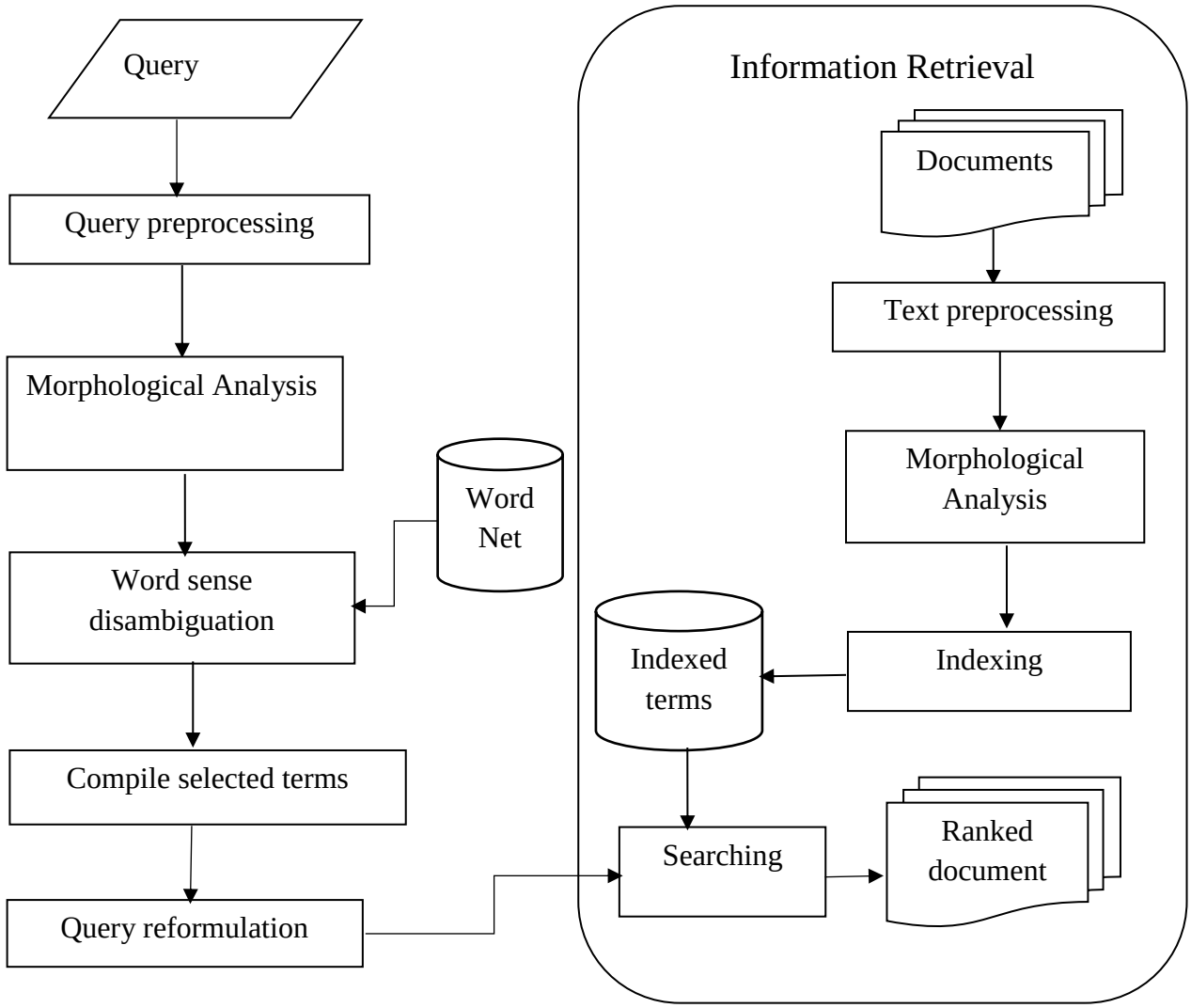

Figure 4.1Architecture of query expansion model

4.3 Query Preprocessing

For efficient search, our query expansion pattern contains preprocessing component that adopt from Atalay [28], to consider the unique features and characteristics of Tigrigna language. This component has pre-processing tasks before the query terms are forwarded to WordNet. Language specific features considered using tokenization, stop word removal and normalization. Users query preprocessing starts, by identifying the terms in the original query using tokenization process to prepare for removal of the stop words, then text is normalized from short forms and same words with different alphabet, finally these query terms return to their root form using morphological analysis process. The root form

of words forwarded to the WordNet for selecting synonym words to build the expanded query. This query preprocessing component has sub-components which have different responsibilities. The detailed architecture of this component represented in *Figure 4.2* and each sub-component is described.

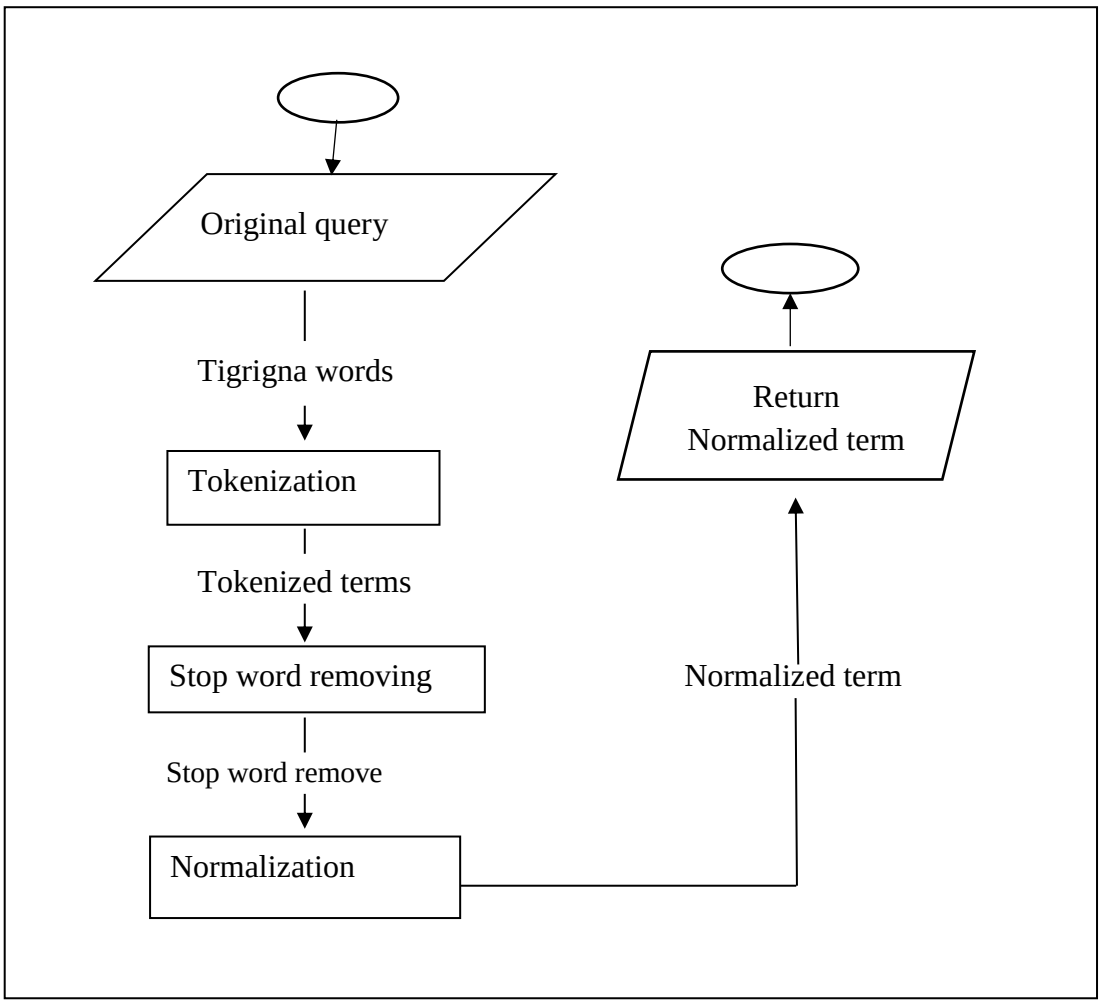

Figure 4.2Tigrigna preprocessing Component

Tokenization

Tokenization process is using to determining word boundaries in a text, it identifies the term suitable for further pre-processing [38]. Like other Semitic family Tigrigna language uses Ethiopic script. This Ethiopic script has its own punctuation marks such as ።, ፡, ፤, ፧ etc. The Tigrigna tokenizer identifies word boundaries using space and Ethiopic punctuation marks as split points. The numbers and punctuation marks in the text of each file were not considered. All punctuation marks, numbers and special characters are removed from the text before the query is processed further. All punctuation marks, special characters and numbers are converted to space and it is used as a word demarcation. Hence,

if a sequence of characters is followed by space, that sequence is identified as a word. A consecutive sequence of valid characters was recognized as a word in the tokenization process. *Algorithm 4.1*is written for tokenization process.

```
Input: original query

Output: tokenized query term

 Do

      Read the original query character by character

      If word contains punctuation marks, numbers, special
      characters (except "/" and "." in the middle)

      Then replace them with a space

      If query contains space

      Then get the string before the space as a word

      Else continue

End if
```

Algorithm 4.1Algorithm for the Tigrigna word tokenization sub component

Stop Word Removal

Stop words are characteristics of natural languages that doesn't have significant meaning when stand alone. In general Stop words are words which serve no purpose for NLP applications, but they are used very frequently in composing Tigrigna documents. Due to this feature, Tigrigna stop words must be removed to save disk space and speed up the searching process. Removing stop words helps to reduce the size of the query this also help from degradation of search result. In this study, stop words are compiled manually by consulting language experts, and previous Tigrigna researches. Stop word removal *Algorithm 4.2* adopted from [28].

```
Input: words

Output: stop word removed

   1. Receive query terms
   2. While (Not end of the query)
            Read the word

      If the word is found in the stop word list

      Remove it

      Else

      Continue

End if

End while
```

Algorithm 4.2Algorithm for the Tigrigna stop words remover sub component

Normalizer

Tigrigna language normalizer component have two responsibilities the first one is to check if the Tigrigna words written in short form or abbreviation. To handle this characteristic, check a word written in either by "."(Period) or "/" (slash) at the middle of word, if the normalizer get such kind of word, replace the word with appropriate long forms. The second responsibility of normalizer is to handle the existence of redundancy of some characters, to compile them in to one common direction [28]. The letters 'ሀ' (he) and 'ኀ' (he), letters 'ጸ' (Tse) and 'ፀ' (Tse) are some examples. These Tigrigna words have one of these forms, some writers use them interchangeably and may exist in different format to represent the same word. For example, ማዕጾ(ma'atso) and ማዕጸ(ma'atso) are two different representation of the same word meaning 'door'. This characteristic of natural language has negative effect on precision of information retrieval. Thus, a normalizer component is added to convert such difference in to a single form. Normalization component algorithm represented in *Algorithm 4.3*.

```
Input: Short form of words
Output: Normalized word
While (Not end of the word)
        Read the word character by character
        If the character is "ሓ","ሑ","ሒ","ሕ","ሔ","ሓ","ሖ"
            Then replace to "ሀ","ሁ","ሂ","ሃ","ሄ","ህ","ሆ"
            respectively
else if the character is "ሠ","ሡ","ሢ","ሣ","ሤ","ሥ","ሦ"

            Then replace to "ሰ","ሱ","ሲ","ሳ","ሴ","ስ","ሶ"
            respectively
        else if the character is "ጸ","ጹ","ጺ","ጻ","ጼ","ጽ","ጾ"
            Then replace to "ፀ","ፁ","ፂ","ፃ","ፄ","ፅ","ፆ"
            respectively
        If the word contains "/" or "." In the middle
           Then Read the characters before "/" or "." As a
           word
        If the word is found in the short words list
          Then Replace the short form with its expanded
          form
       End if
End while
```

Algorithm 4.3Algorithm for Tigrigna Normalizer sub component

Normalizing is very important task because it helps the system in minimizing memory consumption and optimize the time of retrieving of retrieval system by putting documents in a format suitable for a searching process.

4.4 Morphological Analysis

The main advantage of morphological analysis algorithms particularly for IR purposes to improve IR performance, by searching different morphological variants of search terms. It is also used in IR to reduce the size of index files. Since a single stem typically corresponds to several full terms, by storing stems instead of terms, compression factors of over 50 percent can be achieved. Reducing the dictionary size by rooting terms can be high in various NLP applications, especially for highly inflected languages. Therefore, it is important to reduce different variants of words to their corresponding single form before finding their synonyms in WordNet. Morphological analysis algorithm used in our study is depicted in *Algorithm 4.4.*

```
Input: Word from normalizer
Output: root words
Start
     Scan input word from left to right
     If the word is in morph list
     For each valid Word look for a possible root
     Get the root form of the word
     // list of stems to the WordNet
     Return Root
     Pass the possible root word in to WordNet
     //pass list of possible root to the WordNet //valid root
     Else
     Continue
End if
END
```

Algorithm 4.4Algorithm for the Tigrigna morphological analysis sub-component

The task of this component is to accept a list of words from normalizer component and then decompose each word into root words and morphemes of the terms to check against from the file we have and then pass them to the lexical resource or WordNet.

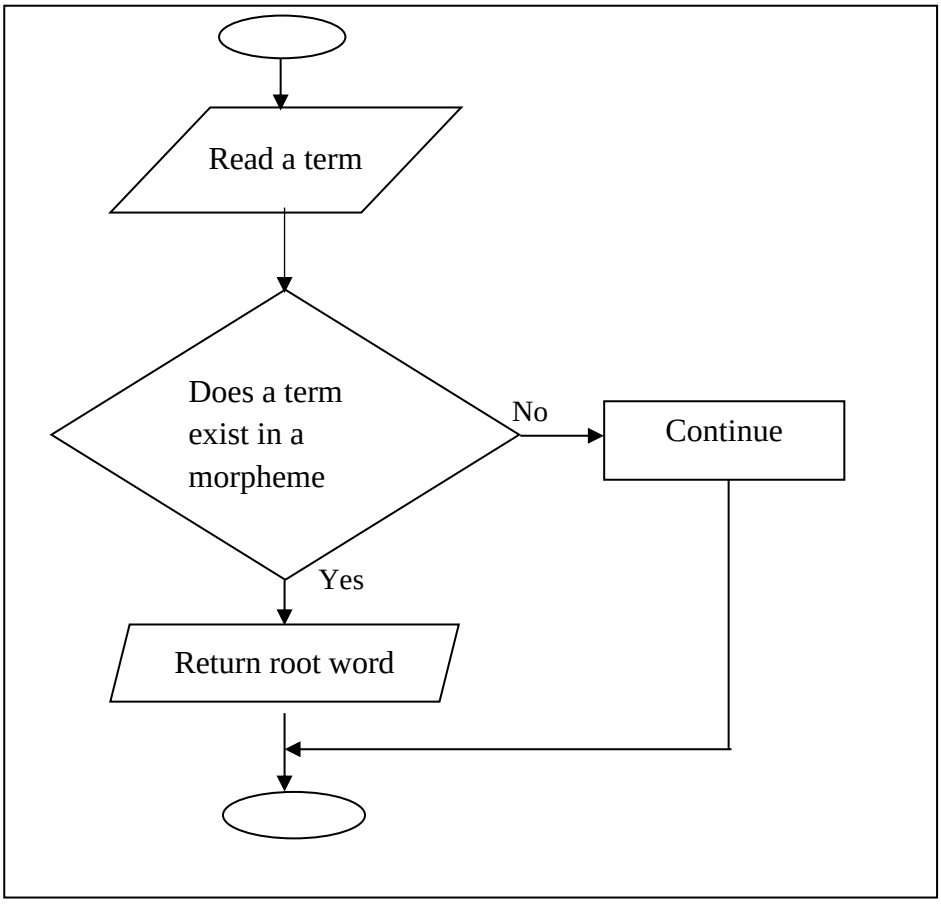

Figure 4.3Flowchart of the morphological analysis sub-component

4.5 Preparing Tigrigna WordNet

So far to our knowledge there is no standard WordNet developed for Tigrigna language, but in our study the WordNet is the core area of correct sense identification. WordNet, can be developed in two different ways, the merge approach and extended approach. Extended approach means of translating the synsets and relations in the English WordNet to your target language. Merge approach define your own synsets and relations in your own language and then line up your WordNet with the English WordNet using uniformity relations. According to [46], it was recommended the merge approach to develop WordNet of other language is good because of the following two reasons:

- Developing your own language WordNet from scratch gives opportunity to consider the language characteristics.
- It gives chance to fill some drawbacks of English WordNet like cross POS relation.

Tigrigna language WordNet can be applicable for different applications such as Tigrigna IR system, Text summarization and automatic text categorization to improve their performance. But so far to our knowledge Tigrigna language root form WordNet are not yet developed, that means our WordNet could be the first attempt in Tigrigna language. Since there is no standard WordNet for Tigrigna language, manually constructed root based lexical resource is used. Our lexical resource is constructed from corpus, by consulting language experts and with the use of Tigrigna dictionary [36] and Sewasiw Tigrigna [37]. This root based Tigrigna WordNet is limited to synonymic terms and related words which is information associated with each term only, which is words with list of synonyms that have different sense based on the context they appear. We choose synsets only for two reasons. First, to protect the WordNet from noise for the reason of a huge collection of words may lead to degradation of performance of the system. Second, we use morphological analysis which is a query terms coming to the WordNet in their root form only.

WordNet synset: -The basic relationship between words in WordNet is the Synonym relation called Synset. Words in the same synset are synonymous in a particular sense. Two words are synonymous if they are interchangeable one for the other in any sentence without changing the concept of the sentence. Word sense is the meaning a word can take

depending on how it is used. For example, the word "መጺጽ" could mean "bitter" in one sense and "player" in another sense. Each synset of a word contains one or more words including the word itself. Then, the synset of the word "መጺጽ" which is "bitter" in English፣ will be "መሪር፣ንምዛዛ" meaning "bitter in test" for the first sense and "ተናጋሪ፣ተጻዋቲ" meaning "player" for the second sense.

WordNet related word: - A related word for a word sense is the sequence of the word in that particular sense and typically succeeded and preceded word of an ambiguous query term. For example, the ambiguous word N and the words ±N is related words of the ambiguous word N, based on the word sequence position. For example, in the sentence ምስፍሕፋሕ ዘመናዊ ትምህርቲ if the ambiguous word N is "ዘመናዊ" then the related words of the ambiguous word which is ±N is "ምስፍሕፋሕ" and "ትምህርቲ". Based on the related words of the term, it may possible to identify the correct sense of terms.

4.6 Word Sense Disambiguation using N-gram

Retrieving document by determining the sense of each query term is better way for effective retrieval, this could be done by adding appropriate synonyms words to a query to improve retrieval effectiveness. However, most query terms have multiple meanings and adding a synonym of the query term which has a different meaning in the context of the query or adding polysemous word would cause deterioration in retrieval effectiveness. Once a query terms sense in a query context is determined, synonyms with the same meaning the query terms are added to the query, so that documents having these synonyms but not the exact term may be retrieved. In other word the logic of query expansion is better technique on improving IR. In this study, we design query expansion model to improve the performance of Tigrigna information retrieval.

The Word Sense Disambiguation (WSD) is performed on the pre-processed query terms of the user. Because it's assumed that user queries could be ambiguous. The task of disambiguating words in the query begins after initially entering the query to WordNet. The system takes this query and performs word sense disambiguation on each query word to disambiguate between query senses.

To understand in what sense the word is used and to identify the meaning of words in contextual manner a word sense disambiguation using n-gram is applied. Every node in

the WordNet is a synset, i.e, a set of synonyms. The word sequence of an ambiguous term, which is referred to as related word of the ambiguous word, is also provided. For a query term, all the synsets in which the term appears can be returned, along with the related word of the synsets.

We develop root based Tigrigna language WordNet that means synsets are a collection of root forms for every word or term. Now we discuss term relatedness based on the word sequence and information provided by root based WordNet. Since the sequence of word provides valuable information about the relatedness of a term, we can use the related word to identify the correct sense of the term. To understand in what sense the word is used and to identify the meaning of words in contextual manner a word sense disambiguation (WSD) using n-gram is applied.

The number of words in the local context of the ambiguous word makes a window. The size of window i.e. number of words to be considered at ± n positions. In this study the size of the window used is three. The preceded, succeeded and the ambiguous word at the middle. By comparing these pieces of information associated with the ambiguous term, it may possible to identify senses to these terms.

The expanded query that will be used for expansion is formed from the combination of these synonym words from disambiguation. The relatedness of each sense measured based on the sequence of words coming with ambiguous term. This helps to identify the correct sense of query terms. The new expanded query formed by comparing and take the one with the highest relatedness based on the occurrence of query terms.

Algorithm 4.5. Shows how our word sense disambiguation works using N-gram

```
Check if the query word is ambiguous against WordNet
If the word is ambiguous
     Then get the words in the left and right of the
     ambiguous word
If the words are related to the ambiguous word
     Then get the sense of the word
     Get the synonym term with the acquired sense
     Add the synonym to the original query
End if
Else if the word is not ambiguous
          Then return the original query word
```

Algorithm 4.5 Our word sense disambiguation works using N-gram

This technique highly increase the performance as much as the related word increases. When applying word sense disambiguation three kinds of answers are expected. The method might identify either the same sense for each ambiguous word, different senses or it might not identify at all. It depends on the quality of your ontology. The new query that will be used for expansion is formed from the combination of synset based method used for disambiguation.

4.7 Compilation of Selected Terms

Users may formulate short queries, even when the information need is complex. Irrelevant documents are retrieved as answers because of the ambiguity of the natural language. If we know that some of synonyms words in WordNet were relevant to the query, terms from those synsets can be added to the query in order to be able to retrieve more relevant documents. In this research, a technique using the Word Sense disambiguation to identify the correct sense or meaning of the term of the given query is developed. The next step is

to use the identified sense for expansion. For query expansion, we use synset expansion method only because synonym words to be able to determine the polysemous behavior of words. The user query is reformulated by adding the terms found on the synset of the selected sense term. Finally, the original users query is reformulated by adding the selected synonyms of the selected sense.

For example, if a user write a query

“ዘመናዊ ትምህርቲ ምስፍሕፋሕን ናይ ትምህርቲ ስርዓት ኣወቓቕራን ኣብ ኢትዮጵያ” then the query processed using preprocessing component, processed query is looks like, “ዘመናዊ ትምህርቲ ምስፍሕፋሕን ትምህርቲ ስርዓት ኣወቓቕራን ኢትዮጵያ

The processed query pass to morphological analysis to get root form of terms as follows፡ ዝምን፤ ምህር፤ ስፍህ፤ ስርእት፤ ውቅር፤ ኢትዮጵያ

After getting the root form of terms, query terms need to be disambiguated using WordNet and N-gram model, the words “ህድሸ፤እውን” is a synonym of the first word ዝምን፣ the words “ልምድ፤ምርእ”is a synonym of the second word “ምህር” the word “ልምእ፡ግልብት፡ህይሸ” is a synonymous of the third word “ስፍህ” based on these synonyms the users query will be reformulated.

Finally, the expanded query is ፡ “ህድሸ፤ እውን፤ ዝምን፤ ልምድ፤ ምርእ፤ ምህር፤ ልምእ፤ ግልብት፤ ህይሸ፤ ስፍህ፤ ስርእት፤ ውቅር፤ ኢትዮጵያ”. This expanded query will be used for final retrieval.

4.8 Indexing

Information Retrieval (IR) systems use inverted indexes for indexing and searching purposes. An inverted index contains a collection of terms and their corresponding occurrences in documents. However, a single word may occur in several morphological variants in a query and document. Thus, increasing the size of the index and decreasing retrieval performance. The need of indexing for decreasing this kind of problem. Before indexing terms with their documents, all terms preprocessed through the preprocessing component adopted from previous work. However, the preprocessing component sub-components are the tokenizer, normalizer and stop words remover. After preprocessing the filtered documents pass to morphological analysis to get the appropriate root form. Finally, query terms will be indexed using Tigrigna index engine.

4.9 Tigrigna Information Retrieval System

This research is developed to improve performance of former Tigrigna information retrieval system designed by [28]. As all retrieval systems, this Tigrigna IR system is fully functional and suitable to test the proposed system. We use this information retrieval system to test the current research.

The architectural view of the original Tigrigna IR system is shown on *Figure 4.4.*

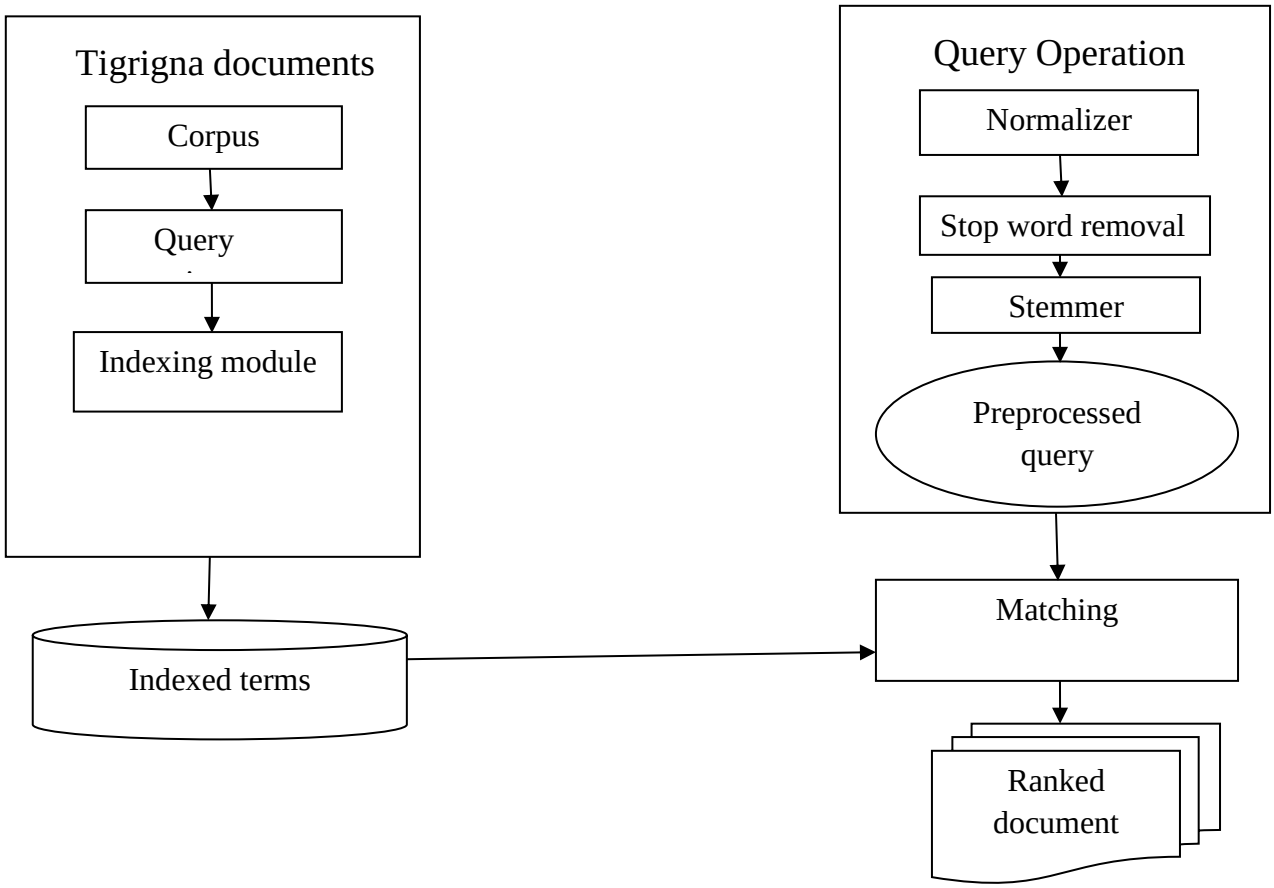

Figure 4.4Architecture of Tigrigna IR system

Chapter 5: Experimentation

5.1 Introduction

This chapter report the experiments, findings and the test environment conducted using the architecture designed in chapter four of this document. In this research, we design query expansion model for Tigrigna language information retrieval system. The experimentation phase holds implementation, testing and discussion of the challenges and findings that are recorded from the proposed technique. Data set selection and testing are also discussed throughout the chapter.

Our prototype developed with three basic roles. First, morphological analysis to get the correct root form of words. Second, WordNet which is used as reference to understand correct sense of ambiguous terms. Third query reformulation which helped to expand the query with the identified sense using word sense disambiguation from the WordNet. Finally, the query expansion module is integrated with information retrieval system to meet the designed target of enhancement of Tigrigna IR system performance. The searching and the prototype system has been developed using PHP programming language with MySQL database and JSON files based on the architectural view shown on *Figure 4.1*.

5.2 Document and Query Preparation

Experiments done on the current study based on sets of documents and queries set up by Atalay [28].To test the prototype system developed, 300 Tigrigna short documents were used as a corpus. The articles are taken from [28] for designing probabilistic information retrieval.

To test the performance of the system 10 (ten) queries were used which are previously used by [28], the adopted queries were rich with polysemous word and identification of these polysemous words were done by this research to select their contextual meanings (i.e. senses).

Table 5.1, the document corpus contains Seven (7) groups, which are health, education, Religion and philosophy, social, politics, sport and art related areas [28].

Table 5.1Types and sizes of corpus used for experiment

No.	**Types of news**	**No of documents**
1	Education	40
2	Sport	30
3	Health	50
4	Politics	80
5	Religion and philosophy	40
6	Social related	30
7	Art news	30
Total		**300**

In addition, *Table 5.2* shows the previous researcher selected 10 test queries to evaluate the performance of the system. These ten queries with polysemous words also used in this research for conducting the experiment [28].

Table 5.2test query terms

No.	Query	Query short-cuts
1	መርቹኒ ኣቦይ ብይርጋገብረመድህን	Q1
2	ምክልካል ተሓላላፍቲ ሕማማት ናይ ህፃናትን ኣዮታት ንጥዕና ኣገልግሎት መርሃግብሪ	Q2
3	ብዛእባ ምርመራ ኤችኣይ ቪ / ኤድስ	Q3
4	ዘመናዊ ትምህርቲ ምስፍሕፋሕን ናይ ትምህርቲ ስርዓት ኣወቓቕራን ኣብ ኢትዮጵያ	Q4
5	ዉሕሉል ኣገባብ ኣጸናንዓ መጽሓፍ ቅደስ	Q5
6	ስርዓት መርዓን ባህሊታትን ኣብ ገለገለ ክፍልታት ትግራይ	Q6
7	ታሪኻዊ ፣ ጅኦፖለቲካዊ ፣ ኢኮኖሚያዊን ሃይማኖታዊን ምትእስሳራት ኢትዮጰያን ግብፂን	Q7
8	ማእከላይ መሪሕነት ኤርትራዊ ዱሞክራስያዊ ኪዲንን ብረታዊ ቃልሲን ህዝቢን ኤርትራ	Q8
9	ታሪክ ካብ ባሕርና ህዝቢንትግራይ	Q9
10	ዓፈና ድምፂ ኣሜሪካን ሕገመንግስቲ ኢትዮጵያን	Q10

5.3 Implementation

5.3.1 Query Pre-processing

Implementing query preprocessing help to make the query expansion model as good as possible. Refine a user's query is an effective technique to help the user to address the

search terms to the specific information needed. In the preprocessing stage, this study addresses tokenization, normalization and stop word removal. Using morphological analysis provides a way to get root form of Tigrigna words in refining query and indexing Tigrigna documents. After preprocessing stage the indexing of document is done.

Tokenization

The first step of query expansion is tokenizing of user's query terms. To make further processing, boundary of query terms need to be determined using tokenizer. In our research to determine term boundaries separately, the tokenizer removes special characters', punctuation marks and numbers. Finally, tokenized Tigrigna words pass to the next step of processing.

```
function stringToknizer($string){
    $_array_string=explode(" ",$string);
        //remove single chracters and empty spaces listed as a word from the list
        $_array_string = array_values(
            array_filter(
                $_array_string,
                    function ($value) {
                    return mb_strlen($value) >= 2;
                    }
                )
            );
        return $_array_string;
}
```

Figure 5.1Screen-shot of Function for Tokenization

Normalization

In Tigrigna writing system, some characters can be used interchangeably without any difference in meaning. To convert such characters to their common form, normalization process is adopted from previous work. The other task of normalizer is to expand short form of words into appropriate format. For the purposes of this research abbreviated list are examined manually from corpus. The long form of abbreviated words is placed in JSON file in order to minimize the time complexity of the system.

```
function normalizer($string){
        global $db_con;
            $string=replaceCharacter($string);
            $stmt = $db_con->prepare("SELECT * FROM normalizer");
            $stmt->execute() or mysql_error();
            while($row=$stmt->fetch(PDO::FETCH_ASSOC))
                {
                    if (strpos($string, $row['word']) !== false) {
                        //replace un normalized forms of words to the normal form
                         $string=str_replace($row['word'], $row['normalform'], $string);
                    }
                }

            $string=str_replace('"', ' ', $string);
            $string=str_replace("'", "", $string);
    return $string;
}
```

Figure 5.2 Screen-shot of function for normalization

Stop Word Removal

After tokenization and normalization, the next step in the preprocessing stage is stop word removal. For the purposes of this research stop word list are selected manually (*Annex A).* The algorithm reads stop words list from JSON file and compared it with tokenized and normalized query terms. Then, if word is similar, it will be removed from the given query.

```
function removeStopWords($string){

    $json_data=file_get_contents("http://localhost/irtnew/stopwords_copy.json");
            $objs=json_decode($json_data,true);
            foreach($objs as $key=>$value){
                $string=mb_ereg_replace($value['stopword'],"",$string);
            }
            return $string;
}
```

Figure 5.3 Screen-shot of function for Stop word removal

5.3.2 Morphological Analysis

Morphological analysis used in information retrieval to resolve the vocabulary mismatch problem, in which user query words may not match document words. According to some researches [32, 33], it is clear that high inflection behavior of Tigrigna language results the higher vocabulary mismatch problem [45]. When researchers apply stemmer it also results some problem of over-stemming and under-stemming of words which in turn, reduces the precision of information retrieval significantly. The process of information retrieval will be improved greatly when morphological analysis is used to avoid the above problems. In our system manually compiled morphological analysis were used in indexing documents

and refining query terms. Query terms need to be returned in to their root forms to improve accuracy.

```
// morphological analizer
$root_words_morphs=$stmt->fetchAll();
    foreach($search_array as $search_terms){
        foreach($root_words_morphs as $root_words){
            if($search_terms==$root_words['morph']){
                $rootform[]=$root_words['lemma'];
                $chkwsdWordIds[]=$root_words['wordid'];
                if($counter==0){
                    $wheredata.="information like '%$search_terms%'";
                $searchTerms[$counter]=$root_words['lemma'];
                }else{
                    $wheredata.="or information like '%$search_terms%'";
                $searchTerms[$counter]=$root_words['lemma'];
                }
                $counter++;
            }

        }

    }
```

Figure 5.4 Screen-shot of code for morphological analizer

The algorithm takes a list of words and checks if the length of the characters is greater than two or not. If the length of the word characters is less than two, the word is returned without further processing. If the length of the word character is greater than two, the algorithm check characters if they matched with one of the root words compiled in JSON file. If the character is matched, the root word is returned.

5.3.3 WordNet Preparation

Word Sense Disambiguation

In our study user's query assumed to contain ambiguous or polysemous terms. The objective of this research is to determine the sense of each ambiguous word in the context of other query words. Thus, the identified senses used for expansion purpose. The senses of the polysemous query terms are prepared on the root form WordNet. Our WordNet is assembled from Tigrigna dictionary [36, 37], Tigrigna corpus and consulting Tigrigna language experts. In our WordNet words, synonyms and related words were stored in their root forms only.

The WordNet contains the ambiguous terms, synonym terms and the related words of each sense. For example, term "ምርቅ". First synset for the word "ምርቅ" is "ብርክ፡ፅልይ፡ትልብይ" then followed with related word which is "ኣቦይ፡ቀሺ", the second sense of the term "ምርቅ" is "ውስክ፡ብዝህ" with related word of "ግዝኣ፡ሽምት" and the third sense of the word is "ፍልጥ፡

ግብር፡ ግንይ፤” with related words “ትክል፡እቅሕ፤”. One sense of a single word may have group of two or three words but this means we can use them interchangeably because they are contextually synonyms. If the system get related word one in the local context of root word, then sense one is the correct sense. The structure of the WordNet is presented in *Figure 5.5*

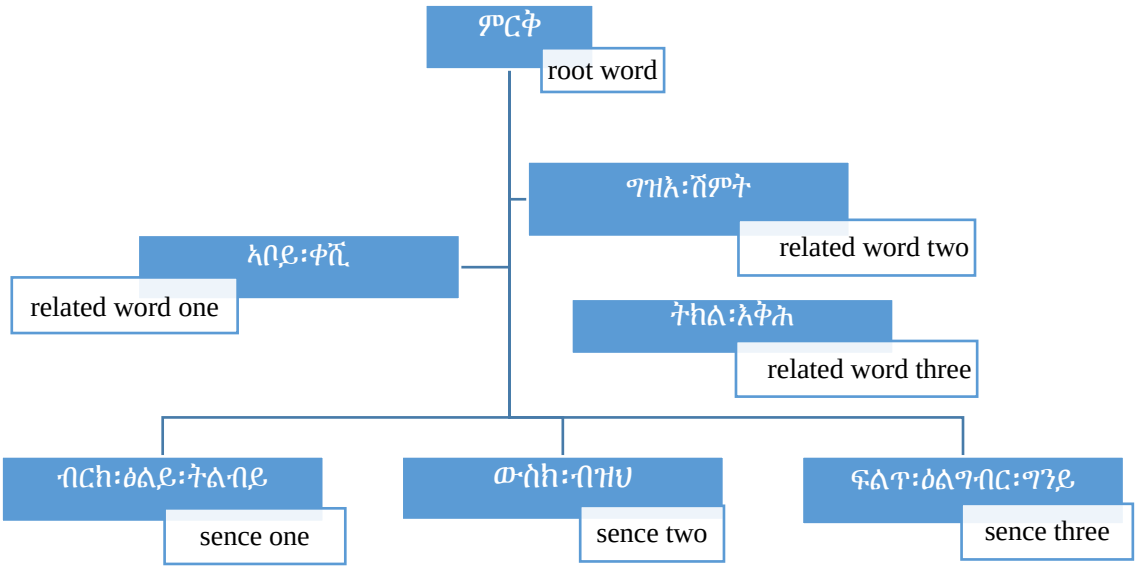

Figure 5.5the structure of term, synset, related word and sense on WordNet

In our study to select the correct sense of polysemous word, the related words play great role. Each related word relates with the query terms using contextual link. Sense one directly related with related word one. Sense two directly related with related word two. This indicates all sense have appropriate related words. Our WordNet contains only the synonymous and related word of the term in the local context or window. But design and developing of lexical resource like WordNet is difficult in terms of labor and time. It’s highly time and labor consuming.

Applying N-gram

Our query expansion model contains word sense disambiguation based on Tigrigna WordNet. The statistical model N-gram is the core model of word sense disambiguation. The WordNet include synonyms and related words of ambiguous words, the related words related with each ambiguous word are used for word sense disambiguation in our study.

The system checking the existence of these related words associated with the query terms in WordNet, finally the system can disambiguate the correct senses to these ambiguous terms using N-gram model. The algorithm of N-gram works to identify correct word sense by finding the relation among the ambiguous word and the preceded and succeeded words (related words) defined in the WordNet. This technique may help the system to identify the proper synonyms of the ambiguous word based on contextual meaning.

Example: suppose users query Q

Q = ምክልካል **ተሓላላፍቲ** ሕማማትን

For the word **ተሓላላፍቲ** the senses are:

Sense 1:ተላበድቲ: ተላባዕቲ: ተስፋሕፋሕቲ

Sense 2:ዘይራኸቡ: ዘይስማምዑ: ዘይቃደዉ

Sense 3:ተጓዓዝቲ: መንደኛታት: ኸየድቲ

Sense 4:ሸለኹቲ: ሰረፅቲ : ሓለፍቲ

Related words: ምክልካል፤ ሕማማትን፤ መርሃግብሪ፤ ሓሳብ፤ መንዲ፤ ፅርግያ፤ መርትዖ፤ ጨረራት፤

In the above example, there are three words from query Q but it may not always three, it could be more than three, this is depend based on the length of user's query. But each query term may have four and more senses. With the proposed way, the correct sense will be identified by the contextual relation of the words in the query terms. The relation starts from the ambiguous word found in the query to the preceded and succeeded words of the ambiguous word. These words put in WordNet like related words. In the above example, the word written in bold and underlined which is "ተሓላላፍቲ" is the ambiguous word and the related word of the word ተሓላላፍቲ is ምክልካል and ሕማማትን which is the preceding word ምክልካል and the succeeding word ሕማማትን. These words compared to the collection of related words in WordNet. When the related words exist in the related word list linked with one of the given senses, that sense will be the correct sense of the ambiguous word. መርሃግብሪ፤ ሓሳብ is the related word of sense two, መንዲ፤ ፅርግያ is the related word of sense three and መርትዖ፤ ጨረራት is the related word of the final sense. Then the algorithm founds

some similarity of words in the query term and collection of related words in WordNet. The mutual words found in the collection of related words are decides the weight of the sense. In the above example two words are found and the weight of the sense going to be 1, which indicates that sense could be the correct sense. Then the correct sense of the word ተሓላላፍቲ is goes to the first sense ተላበድቲ፤ ተላባዕቲ፤ተስፋሕፋሕቲ. In this case if there is no mutual word to be found for this sense the weight for the sense is goes to 0. The one with the weight of 1 is taken as the identified sense that will be used for the expansion. The same process is done for each term of the query and the one with the highest weight is assigned as the sense of the word based on the given query context. This is how our algorithm works.

In Tigrigna WordNet there are many polysemous Tigrigna words. We define some synonyms of these polysemous terms as synset. *Table 5.3* Represents14 ambiguous terms found in our queries and their correct sense. These words are ambiguous based on our corpus.

Table 5.3Ambiguous Words with their correct sense

No	Ambiguous Words	Query they found	No of sense	The correct sense from WordNet
1	መርቕኒ	Q1	3	ባርኹኒ : ተላቡይለይ :ፀለይለይ
2	ተሓላለፍቲ	Q2	5	ተላበድቲ :ተላባዕቲ: ተስፋሕፋሕቲ
3	ምርመራ	Q3	2	ፈተሻ: ምፅራይ :ምፍላይ
4	ምስፍሕፋሕን	Q4	3	ምልማዕ: ምጉልባት :ዝሓሸምግባር
5	ኣጸናንዓ	Q5	5	ኣረዳድኣ : ኣናብባ : ትዕዝብቲ
6	ስርዓት	Q6	2	ደንቢ : መምርሒ : ሕጊ
7	ምትእስሳራት	Q7	3	ምዝማድ : ምርኻብ :ምትእትታው
8	ኪዳን	Q8	2	እምነት :ቃል
9	ዓፈና	Q1O	2	መፅቀቲ : ጭቆና : መግዛእቲ
10	ትምህርቲ	Q4	2	ልምዲ : ተሞክሮ :ፍልጠት
11	መርሃግብሪ	Q2	2	ትልሚ : መደብ
12	ኣወቓቐራን	Q4	2	ኣሰራርዓ : ኣደራድራ : ኣቀማምጣ
13	ታሪኻዊ	Q10	2	ሕሉፍ : ጥንታዊ : ቀደም
14	ማእኸላይ	Q8	2	ኣተዓራቒ : ሞንጎኛ : ኣላዪ

In our study to test the developed lexical resource, we use sense disambiguation rate technique in order to answer the question, “how many ambiguous words disambiguated correctly?” The proposed method find the related words between the related words of the query terms and the related words list in WordNet. The system disambiguates 85.6% of the overall ambiguous terms. From all ambiguous words two words could not be identified because of the absence of related words with the other words in the query. Out of the 14 ambiguous terms 84.61% of them are identified correctly. Two terms “መርሃግብሪ” and “ጅኦፖለቲካዊ” disambiguated incorrectly using our technique.

The first word is “መርሃግብሪ”, the term “መርሃ-ግብሪ” Has two senses “ኣጀንዳ: ስብሰባ: ጉባኤ” and “ትልሚ:ቀፃሊንጥፈት:መደብ” While the first sense “ኣጀንዳ: ስብሰባ: ጉባኤ” which is not the right one based on the context of the given query, the second one is “ትልሚ:ቀፃሊንጥፈት:መደብ” which is correct sense. This shows that a term that is not disambiguated with our method because the word “መርሃ-ግብሪ” is compound word which is a combination of two words “መርሃ” and “ግብሪ”,since our WordNet support only word level disambiguation, this word pass without being disambiguated. The other ambiguous word which is “ጅኦፖለቲካዊ” is not disambiguated correctly, the reason is the word “ጅኦፖለቲካዊ” is a foreign language word which can’t be changed to root form in our WordNet. There is another challenge when we apply N-gram, every preceding and succeeding word of each term may not be defined in the related word list and some ambiguous term may doesn’t have synonym list. Because of this, the term with multiple senses may not be disambiguated using our method.

5.4 Performance Evaluation

Performance evaluations metrics are used to measure the degree of some objectives stated to achieve by the system. In our research ten queries are used to test the IR system with morphological analysis. These ten queries also expanded using our system automatically for evaluating the performance of the developed prototype. The performance of the system is evaluated before query expansion using IR system with morphological analysis and after applying query expansion. Systems evaluated to measure if the system meet the specified target. The objective of this research is to improve the performance of Tigrigna IR system using query expansion model. The performance of the prototype is evaluated before and after query expansion. These are:

- Evaluation of TIRS before query expansion with morphological analysis.

- Evaluation of TIRS after integrating query expansion.

The first evaluation is the evaluation of adopted IR system with modification of morphological analyzer instead of stemmer and the second evaluation goes to evaluation of the modified IR system with integration of query expansion module. The result expected from the system is after applying morphological analysis and query expansion on the query.

5.4.1 Evaluation Metrics

In our research recall, precision and F-measure are used. The aim of our study is to develop query expansion model that can help the IR system on improving the retrieval results by adding synonym terms to the user's query. But before query expansion applied, we try to experiment information retrieval with morphological analysis. In our test environment information retrieval without query expansion shows some improvement when compare to the previous work.

Figure 5.6 presents a screen-shot which shows the first list of retrieved document, using the query Q4 from *Table 5.2.*

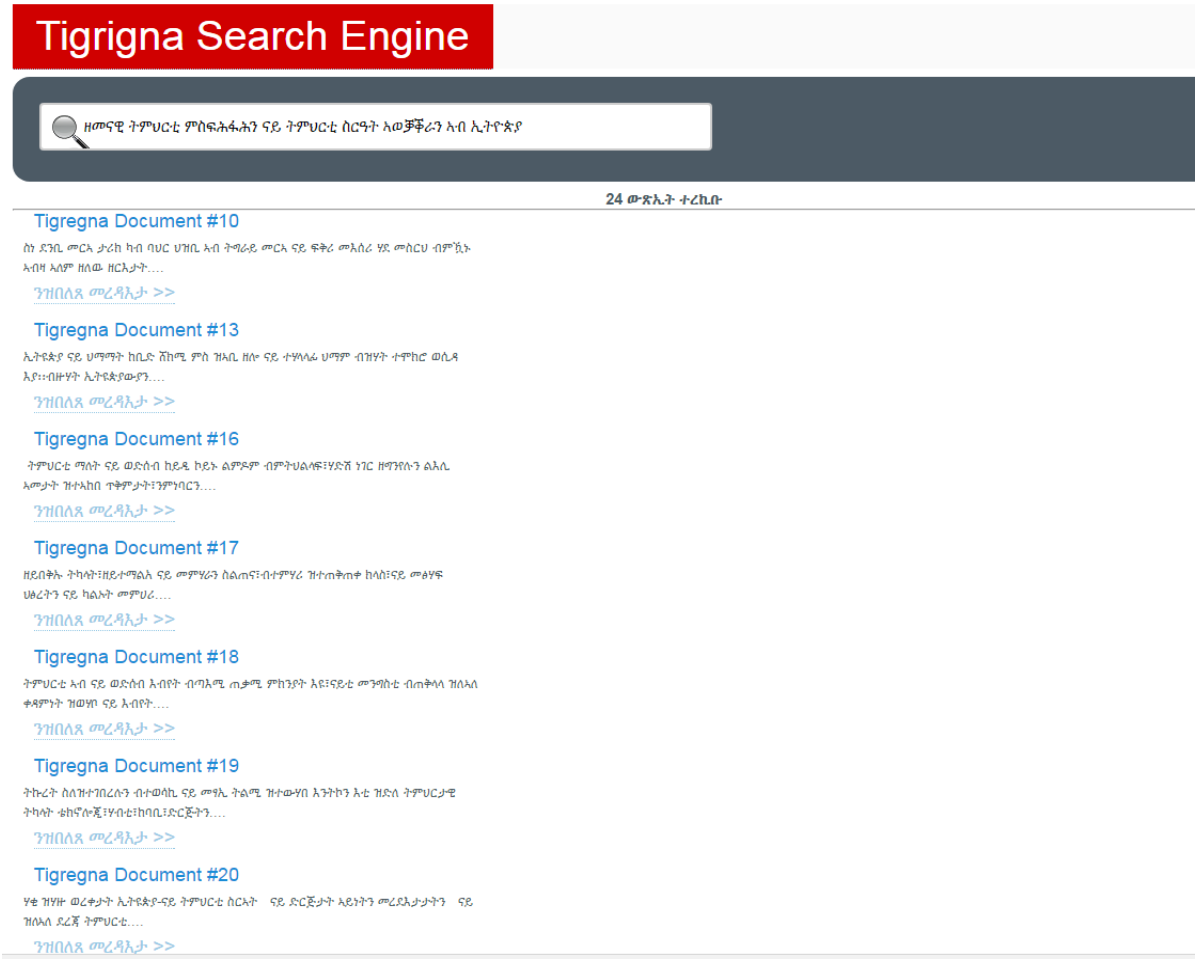

Figure 5.6Screen-shot of a prototype without query expansion

Figure 5.6 shows the result of the process without applying query expansion. For the query "**ዘመናዊ ትምህርቲ ምስፍሕፋሕን ናይ ትምህርቲ ስርዓት አወቓቕራን ኣብ ኢትዮጵያ**" this query retrieve 24 documents, from the whole retrieved14 of them are relevant. But in our corpus, there are 22 relevant documents.

5.4.2 Test Result of TIRS

I. Evaluation TIRS with Morphological Analyzer

In our study we use adopted IR system as test environment. But the stemmer of this IR system replaced with morphological analysis, to test which IR system is performed well in retrieving relevant document. Based on the modified system the initial result without expansion for those ten queries presented in *Table 5.4.*

Table 5.4 Experiment result before query expansion

Evaluation using morphological analysis						
Query	Relevant doc	Retrieved doc	Retrieved &relevant doc	P	R	F
Q1	11	10	10	1	0.909	1
Q2	15	14	12	0.857	0.8	0.733
Q3	10	10	8	0.8	0.8	0.778
Q4	22	23	22	0.88	1	0.808
Q5	20	22	19	0.86	0.95	0.90
Q6	5	7	5	0.714	1	0.667
Q7	12	15	11	0.733	0.916	0.857
Q8	14	20	13	0.65	0.928	0.6
Q9	7	8	6	0.75	0.857	0.775
Q10	7	10	7	0.7	1	0.411
Average				0.794	0.916	0.752

Table 5.4 shows result registered on improvement of IR system. The current study, uses **morphological analysis** to prevent query terms from high errors of over-stemming and under-stemming occur due to morphological complexity of the language. We try to solve the above problems by using morphological analysis to increases the relevancy of relevant documents and decrease non-relevant documents. Results register 78.8 % precision, 0.91 recall and 0.75 F-measure when using morphological analysis. Performance increase in percentage of precision9.7 %, recall1.6 %, and F-measure 8%. This experiment shows good improvements on the result.

II. Evaluation TIRS with Query Expansion

In this study reformulation of user's query Q done by adding terms with similar sense with each term in users query to get the expanded query $Q_{m.}$ The first step of the system is to get users query and processed using preprocessing component, the second step is getting the actual root form of user's query term and the final step is reformulating query using synonyms words of the ambiguous word found in user's query. Then the reformulated query will be submitted to the probabilistic IR and gives new retrieval result. For example,

query “ዘመናዊ ትምህርቲ ምስፍሕፋሕን ናይ ትምህርቲ ስርዓት ኣወቓቐራን ኣብ ኢትዮጵያ” is expanded to ህድሽ፤ እውን፤ ዝምን፤ ልምድ ፤ምርእ ፤ምህር ፤ልምእ ፤ግልብት ፤ህይሽ ፤ስፍህ፤ስርእት፡ውቅር፡ኢትዮጵያ” the words “ህድሽ፤እውን” isa synonymous of the first word ዝምን፣ the words “ልምድ፤ምርእ” is a synonymous of the second word “ምህር” the words “ልምእ፡ግልብት፡ህይሽ” is a synonymous of the third word “ስፍህ”. The rest query terms are not disambiguated using the word sense disambiguation method. Therefore, the system will not return any expansion terms for those words. *Figure 5.7* shows the modified query screen-shot using synonyms.

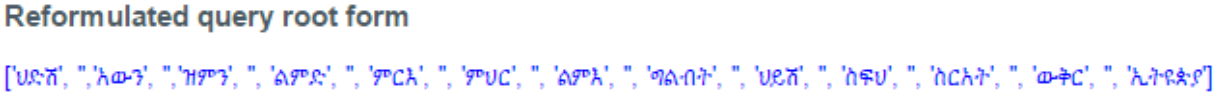

Figure 5.7 Screen-shot of modified query using synonyms

After query expansion of user’s query, the system gets reformulated query and gives different result as shown in *Figure 5.8* which is a screen-shot list of retrieved document using reformulated query.

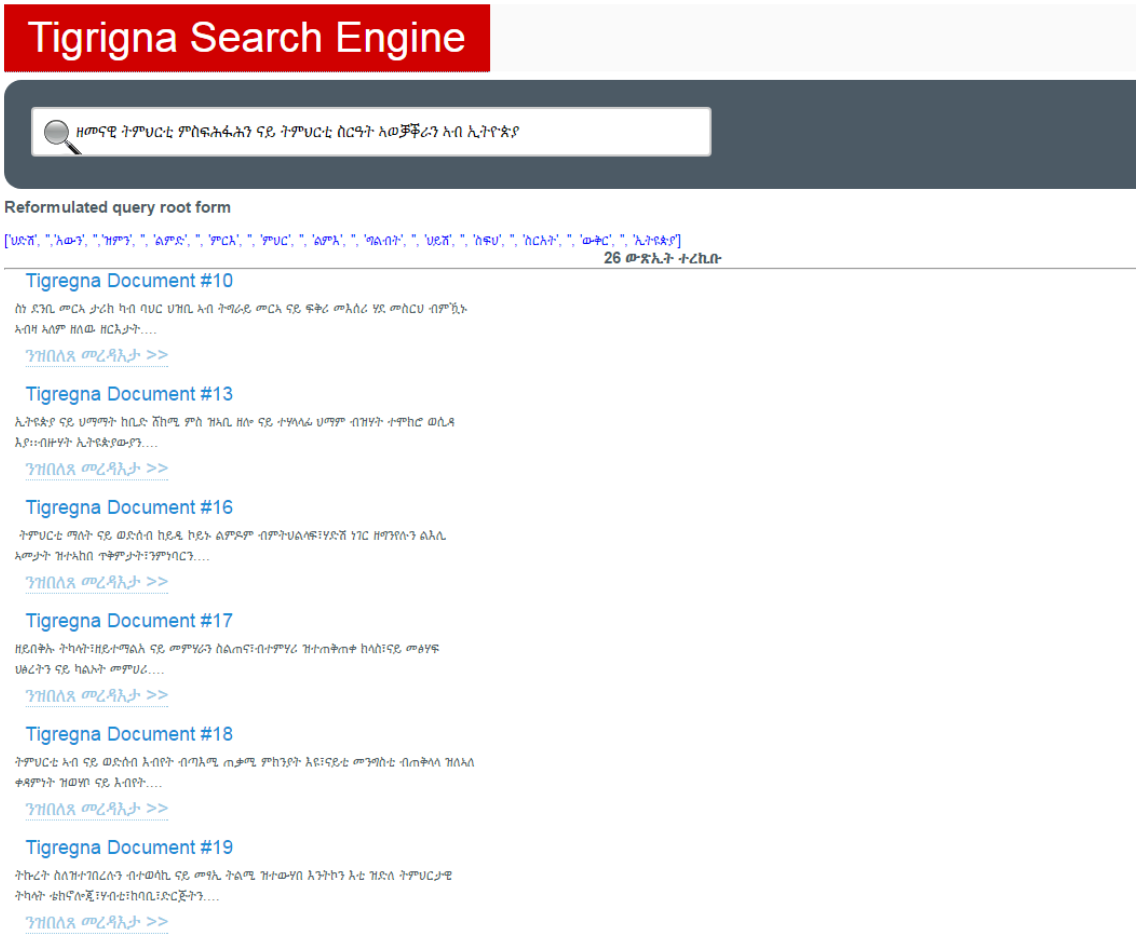

Figure 5.8Screen-shot of reformulated query result

Evaluation of the system also continue after expanding the original users query. This evaluation is done after query expansion method is applied on the ten queries given in *Table 5.2.*

Table 5.5 Experiment result after query expansion

Evaluation using query expansion with morphological analysis						
Query	Relevant doc	Retrieved doc	Relevant retrieved	Precision	Recall	F-measure
Q1	15	13	13	1	0.867	0.952
Q2	15	16	15	0.938	1	0.828
Q3	10	10	10	1	1	0.8
Q4	24	26	24	0.923	1	0.936
Q5	21	22	20	0.909	0.952	0.95
Q6	12	27	12	0.444	1	0.833
Q7	16	14	14	1	0.875	0.815
Q8	20	25	20	0.8	1	0.765
Q9	8	13	8	0.615	1	0.706
Q10	14	15	11	0.733	0.786	0.824
Average				0.812	0.948	0.847

Table 5.5 shows, the result of the system after query expansion and morphological analysis applied and registered result of81% precision, 94% recall and 84%F-measure. The main target of this study is to increase the performance of the probabilistic IR system. To achieve our objective, we use query expansion model and query terms expanded using synonymy, to increase the retrieval of relevant documents and decrease the retrieval of non- relevant document. This study shown improvement by 12%precision, 4% recall and 10% F-measure in overall performance.

5.5 Discussion

The above experimental result indicates, expanding a query using lexical resources registered good performance. Word sense disambiguation to determine the correct sense of ambiguous query terms gives good result. Applying this approach in large corpus, could give result close to 100%.The use of query expansion in information retrieval using our method shows improvement when it compares to the original system without expansion.

Table 5.5 presented the overall performance of the designed model. The goal of the system is to retrieve relevant documents as much as possible by increasing precision and recall of the system and decrease the retrieval of irrelevant documents to some extent. This study shows an effective use of morphological analysis for identifying root words and Word Sense Disambiguation on identifying the sense of query terms to form the reformulated query. Our root form WordNet and the algorithm used for n-gram performs well for the synset. But before query expansion is applied indexing documents using morphological analysis registered better performance in precision and satisfactory result in recall and F-measure. For query expansion, the technique using synset expansion registered a better performance when it comes to the overall performance, this study registers 81% precision, 94% recall and 84%F-measure. This increases by 12%in precision, 4% in recall and 8% in F-measure.

In general when our study compared with previous work of Atalay [28], from the approach Atalay used to determine the query terms using relevance feedback, the use of word sense disambiguation is a good approach. Additionally, the use of morphological analysis for information retrieval process gives good result.

On the other hand, there are some challenges encountered through the current study, and these challenges bounds the system from registering best performance expected from the proposed approach. The main problem is absence of standard resources.

So far to our knowledge there is no any developed root based WordNet for all Ethiopian languages, especially for Tigrigna language, there is no any standard WordNet. Designing lexical resource like WordNet is labor intensive, time consuming and difficult, especially with a single person and when there is lack of resources. Even if our research attempts to develop a WordNet, constructing standard WordNet accessible for researchers is necessary. In addition, it's known that relations of WordNet are limited to parts of speech tagger (POS) like, synonymy, hypernym …and there is limitation of cross POS relation. For example, there is no relation build between verbs and nouns.

We face a challenge on finding developed morphological analysis for Tigrigna language. Morphological variants of words have similar meaning and can be considered as equivalent in information retrieval. In natural languages, the need of narrow down words to their root form better instead of their stem. Even if the morphological analysis used in

our study is manually crafted still it’s difficult on finding morphemes and appropriate root words. Finding root word of compound word like መርሃ-ግብሪ and foreign borrowed words like ጅኦፖለቲካዊ also other difficulty in our study.

Another challenge of our study is finding large corpus and query for Tigrigna language. The developed approach was tested in small corpus compiled by Atalay, some weakness observed because there are no standard test queries, document content and size of corpus used for testing. Even if our study uses the same corpus and queries taken from the previous work for evaluating the performance, there is still a challenge of on using small size corpus information retrieval system, because there is no standard corpus for testing Tigrigna information retrieval system to be opened for every research.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

The main objective of this study is to improve the performance of Tigrigna information retrieval system. It's known that IR systems have different challenges such as, short queries of users, ambiguity of natural language and the vocabulary mismatch between query terms and relevant documents. But good IR system slightly pass the above challenges to retrieve best results.

To achieve the specified objective on increase the number of relevant documents retrieved, user's query need to be expanded by their context. The most recent query expansion approach involves the use of lexical resources like WordNet identify the sense of ambiguous queries. When IR systems works by determining the context of the user query, users may get more meaningful results. Our word sense disambiguation approach performs well on identifying the correct context of user's ambiguous query.

This study prepared morphological analysis that reduces the morphological variant of words in to a common form. The words using in the morphological analysis were gathered from our test corpus, Tigrigna dictionaries and lexicographers. It contains root forms and their morphemes. The indexing component replaced using morphological analysis to enable index terms for fast searching and decreasing memory space requirement. It is also used to speed up the access to desired information from indexed document collection as per user's query.

The prepared Tigrigna WordNet contains root form of words only, this means representation of words, synsets and relations are through root form of words only. The developed WordNet is bounded by our corpus and quires. The words used in the WordNet are limited to include the terms used for the prepared queries. The query reformulation is assembled by combining the senses identified using word sense disambiguation.

To summarize the findings of this research and recommend areas for future work, the experimental result show on using morphological analysis before query expansion registers in terms of precision 78%, recall 91% and F-measure by 75.2%. This method

registered an improvement of 9% in precision from original work. This shows the possibility of controlling vocabulary mismatch, over-stemming and below-stemming gives good improvement. After query expansion method applies, observed results shows improvement in terms of precision to 81%, recall to 94% and f-measure to 84% in overall performance of the system. Using word sense of disambiguation for query expansion registered an improvement of precision by 12%, recall by 4% and F-measure by 10% in overall performance of the system from the original IR system. Expanding users query using synonym words of query term registers encouraging result. In our study, the use of query expansion terms is limited based on the information available on the WordNet and test corpus.

Finally we conclude that there is a possibility of developing better IR system which retrieves relevant document based on context of user's query. This can be achieved by using good lexical resources or Tigrigna ontologies. We can also conclude the usage of morphological analyzer is more powerful when we use it instead of stemmer.

6.2 Contribution of the Thesis

This study shows the use of morphological analysis in the improvement of Tigrigna IR system. We prepared morphological analysis in order to reduce variant of words in to a common form. The indexing component of TIRS replaced using morphological analysis to enable index terms for fast searching and decreasing memory space requirement. It is also used to speed up the access to desired information from indexed document collection as per user's query. Performance increase in percentage of precision 9.7 %, recall1.6 %, and F-measure 8%. This experiment shows good improvements on the result.

The other contribution of our study is preparing Tigrigna WordNet contains root form of words only. This means representation of words synsets and relations are through root form of words. The developed WordNet helps to control the polysemy and synonym nature of words in Tigrigna language. In other word, TIRS can be improved by controlling the ambiguity nature of words.

This study is designed a well performed query expansion model to enhance the precision and recall of Tigrigna IR system. The main reason for integrating query expansion is to increase retrieval of relevant documents as per user's query based on the correct sense of

query terms. This study has a way to discriminate the various meanings of a polysemous term, based on word sense disambiguation (WSD) and find synonymous terms for reformulating user's query. *Table 5.6* shows the compiled result of our contribution. Experimental result increases significantly after using morphological analysis and query expansion. This implies some expanding terms are good for improving precision and recall.

Table6.1 Compiled result for proposed technique

	Average Precision	**Average Recall**	**Average F-measure**
Atalay's work	0.69	0.9	0.74
IR + morphological analysis	0.78	0.91	0.75
Reformulated query	0.81	0.94	0.84

6.3 Future Work

Finally, the researcher would like to recommend some future work of the research topics.

- In our study the WordNet contains only a set of synonyms and identifying the polysemous behavior of words. It also has a related word associated with each set of synonyms terms to identify the sense. On the other hand, it lacks hyponym synset. A hyponym is a set of words which have the same meaning but have specific meaning then the given word. This set of words have positive impact on retrieving of relevant document. Therefore, construct a standard Tigrigna WordNet is one research gup.
- Lack of effective Tigrigna morphological analysis contribute a major constraint on results of the system. To prevent words from over stemming and under stemming problem, morphological analyzer is a good solution. Therefore, it is recommended that future researches should consider the designing a fully functional morphological analysis for Tigrigna language.
- The test environment of our system is on the Probabilistic model. The nature of probabilistic also make initial guess of relevant documents despite of frequency of terms. This guessing based on Boolean expression doesn't consider the importance of terms with highest frequency, this also may lead ranking of relevant documents on the bottom. Therefore, there is a need to develop hybrid system that integrates

vector space model to guess relevant documents for user query using non-binary weighting technique.

- Developing a standard corpus, test queries and standard IR system with a good performance for testing the designed system is one recommended research area. Hence, future research need to consider the development of standard Tigrigna corpus, test queries and IR system that can be used by every researcher to evaluate effects of query expansion model on the IR system.

References

[1] Baeza-Yates R. and Ribeiro-Neto B., "Modern information retrieval", 2nded, Addison-Wesley-Longman Publishers, England, 1999.

[2] G S Josan and G S Lehal, "Size of N for Word Sense Disambiguation using N gram model for Punjabi Language", international Journal of Translation, Vol. 20, No. 1-2, 2008, pp. 47-56.

[3] AbeyBruck, "Semantic Based Query Expansion for Amharic IR'', Unpublished Master's Thesis, School of Information Science, Addis Ababa University, 2011.

[4] Lyman, P. and Varian, H.R, "How Much Information", Unpublished Paper, University of California, 2003.

[5] TewodrosHailemeskel, "An Experiment Using Latent Semantic Indexing (lsi) with Singular Value Decomposition (svd) On Amharic Text Retrieval", Unpublished Master's Thesis, School of Information Science, Addis Ababa University,2013.

[6] YonasFissha, "Development of Stemming Algorithm for Tigrigna Text", Unpublished Master's thesis, Department of Information Science, Addis Ababa University, 2011.

[7] GebrehiwotAssefa, "A Two Step Approach for Tigrigna Text Categorization", UnpublishedMaster's Thesis, School of Information Science, Addis Ababa University, 2011.

[8] S.Sarasawathi, Asmasiddhiqaa. M, kalaimagal. K, Kalaiyarasi M, "Bilingual Information Retrieval System for English and Tamil" journal of computing, Department of Information Technology, Pondicherry Engineering College, India, 2010.

[9] AtelachAlemu and Lars Asker, "Amharic-English Information Retrieval" White Paper, Department of Computer and Systems Sciences, Stockholm University, 2006.

[10] H. Moukdad, "Lost in Cyberspace How Do Search Engines Handle Arabic Queries" *The 12th International World Wide Web Conference*, Budapest, Hungary, 2003.

[11] Zhihan Li, "Improvement of Chinese Information Retrieval by incorporating word segmentation and Query Expansion", Master's Thesis, Faculty of Science and Technology, Queensland University of Technology, 2009.

[12] Gao J, Li M., Lee k., "N-Gram Distribution Based Language Model Adaptation", in *proceedings of ICSLP*, 2000.

[13] Jinxi Xu and W. Bruce Croft, "Query Expansion Using Local and Global Document Analysis", *Proceedings of the 19 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000.

[14] AbeyBruck, Tulu Tilahun, "Bi-gram based Query Expansion Technique for Amharic Information Retrieval System" Published Journal, Department of Computer Science and IT, Arba Minch University, 2015.

[15] Vannevar bush,"As We May Think" Published Paper, Atlantic Monthly, 176:101-108 July 1945.

[16] Van Rijsbergen, "Information Retrieval", Department of Computer Science, University of Glasgow, Second Edition, 1979.

[17] Mitra, M. &Singhal, A. &Buckley,C., "Improving Automatic Query Expansion", in *Proceedings of Annual ACM Conference on Research and Development in Information Retrieval*, 1998, pp 206-214.

[18] Taveter, K., Lehtola, A., Jaaranen, K., Sorva, J. and Bounsaythip, C., "Ontology Based Query Translation for Legislative Information Retrieval", Unpublished paper, available at VTT Information Technology, Finland, 1998.

[19] Kalmanovich, I,G and Kurland, O., "Cluster-Based Query Expansion", in *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval*, 2009, pp 646-647.

[20] Grootjen, F.A. &VanderWeide, Th. P. "Conceptual Query Expansion", Data & Knowledge Engineering, Vol. 56, 2004, pp. 174-193.

[21] George A. Miller, Richard B., Christine, Fellbaum, Derek G. & Katherine M., "Introduction to WordNet an On-line Lexical Database" Unpublished paper, 1993.

[22] Dipasree Pal, Mandar M. and Kalyankumar D., "Improving Query Expansion Using WordNet", Published Article, Indian Statistical Institute, Jadavpur University, 2013.

[23] Robertson, S.E., "On Term Selection for Query Expansion" Journal of Documentation, Vol. 46, No. 4, (1990), pp. 359-364.

[24] Frakes W., R. Baeza-Yates, "Information Retrieval Data Structures & Algorithms" Englewood Cliffs, NJ: Prentice-Ha, 1992.

[25] Antony A. and Louis H., "AFRICAN BOOK The encyclopedia of African and Africa American experience" 2nd ed., vol. 5, 2004.

[26] Hui Fang, “A Re-examination of Query Expansion Using Lexical Resources”, in *Proceedings of the ACL-08: HLT*, pp.139-147.

[27] HailayBeyene, “Design and Development of Tigrigna Search Engine”, Unpublished Mastsers Thesis, Department of Computer Science, Addis Ababa University, 2013.

[28] AtalayLuel, “A Probabilistic Information Retrieval System for Tigrigna”, Published Mastsers Thesis, Department of Information Science, Addis Ababa University, 2014.

[29] Germann, D.C., Villavicencio A., and Siqueira M. “An investigation on polysemy and lexical organization of verbs”, *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, Los Angeles California, 2010, pp. 52-60.

[30] Billhardt, H., Borrajo, D. and Maojo, V., “A Context Vector Model for Information Retrieval”, *Journal of American Society for Information Science and Technology*, Vol. 53, No. 3, 2002, pp.236-249.

[31] Omer Osman Ibrahim and YoshikiMikami, “Stemming Tigrigna Words for Information Retrieval*”,Proceedings of COLING 2012 Demonstration Paper*, Mumbai, 2012 pp. 345-352.

[32] Gasser M., “Horn Morpho a System for Morphological Processing of Amharic, Oromo, and Tigrigna” *Proceedings of the Conference on Human Language Technology for Development*, Alexandria, Egypt, 2011, pp. 94-99.

[33] Shuly W., “Morphological Processing of Semitic Languages”, *A Case Study in Natural Language Processing of Semitic Languages*, University of Haifa, Israel, 2014.

[34] T. Kassie, “Word Sense Disambiguation for Amharic Text Retrieval, A Case Study for Legal Documents,” Msc, Addis Ababa University, Addis Ababa, 2009

[35] D. Pal, M. Mitra, and K. Datta, "Improving Query Expansion Using WordNet", *Presented at CoRR*, 2013.

[36] G .Kassa, “A Tigrigna Language Dictionary”, EMAY Printers, Addis Ababa, Ethiopia 2003.

[37] S. Amanuel, “SewasiwTigrignaBisefihu,” Red Sea, Eretria, 1998.

[38] Omar O, and Yoshaki M., “Indexing Tigrigna language Documents” *Proceeding of the, International Conference on Information Processing Society*, Japan, 2013.

[39] Abreham Negash, "The Origin and Development of Tigrinya Language Publications", Published Desertion, Santa Clara: Santa Clara University, 2016.

[40] Ramesh N. and Chirag S., "Evaluating the Quality of Query Refinement Suggestions in Information Retrieval" Department of Computer Science,University of Massachusetts, 2006.

[41] C. D. Manning, et al., "Introduction to Information Retrieval," Cambridge University Press, vol. 1, Cambridge, England, 2008.

[42] D. Fensel, Ontologies: Springer, Springer Berlin Heidelberg, 2001

[43] LANCASTER, F.W., "Information Retrieval Systems Characteristics, Testing and Evaluation", Wiley, New York, 1968.

[44] M. Barathi and S. Valli, "Ontology Based Query Expansion Using Word Sense Disambiguation", *International Journal of Computer Science and Information Security*, IJCSIS, Vol. 7, No. 2, USA, February 2010.

[45] Savoy J. ,"Stemming of French Words Based On Grammatical Categories" *Journal of the American Society for Information Science*, Vol. 44, No 1,1993, pp. 1-10.

[46] SamrawitZewdneh,"Word Sense Disambiguation Using Semantic Similarity for Query Expansion in Amharic Information Retrieval", Unpublished Master's Thesis, School of Information Science, Addis Ababa University, 2014.

[47] S.Liu, C.Yu, W.Meng, "Word Sense Disambiguation in Queries" *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp 525-532.

[48] G.A.Miller "WordNet a Lexical Database for English", *Communications of the ACM*, Volume 38, 1995, PP. 39-41.

[49] S.Banerjee "Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet" Unpublished Master's Thesis, Department of Computer Science, University of Minnesota, Minnesota, 2002.

[50] B. He and I. Ounis, "Studying Query Expansion Effectiveness", in Advances in Information Retrieval, ed: Springer, 2009, pp. 611-619.

[51] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM computing surveys (CSUR)*, vol. 34, 2002, pp. 1-47.

Annex

Annex A: List of Stop Words

እታ	ዘሎ	የለዉን	ዘላ	ወይ
ከምዘለካ	በቶም	ነዚ	ይኹን	ኣሎ
እንታይ	ኣበይ	ዘሎ	የለዉን	ዘላ
ናብ	ዘለኒ	ግን	ጥራሕ	ካብቶም
ብቲ	እምበር	ብዘይ	ይኹን	እምበር
ናይዘን	እቲ	ንስኻትክን	ከምኡውን	በቲ
ናይቶም	ናይተን	ኣብዚኣም	እዙይ	ዘለዋ
ከምዘለኪ	ስለዝኾነ	ነይርወን	እዚኣም	ዘለወን
የብልካን	ክኾና	ከምዚኣተን	ኣይኮነትን	ዘለወን
ኣለ	ከምተን	ኢሉ	ኢላ	ኢለን
ከምቶም	ዘለና	ዘለካ	ዘሎ	የለዉን
ነይሮም	ከምቶም	እቶም	ምእንቲ	ስለ
ከምዘለኒ	ከምዘለኩም	እውን	ውን	ኣብዚ
ዘለኩም	ድሕሪ	ቅድሚ	ክሳዕ	በዚ
ዘለኪ	ምኽና	የለን	ኮይኑ	ምኽኖም
ኣይኮነን	ደኣ'ምበር	እታ	የብልናን	ከምዘለዎ
ንስኻትኩም	መን	እዞም	ከምዘለኺ	ዘላ
ስለዝኾነውን	የብለይን	ይኹን	እዛ	እሞ
ዘለዎም	እንተዝኾና	ዝኾነ	ንሕና	ነበሩ
ብኣኣም	ምስቲ	የብለንን	ነታ	ብዚ
ናይቶም	የብሎምን	ወዘተ	ናይ	ድማ
ክሳብ	ከምኡ	ዝባሃላ	ናይዞም	እዚ
ዝኾነ	ምኽን	ኣይኮኑን	ናይዘን	በኣኣም
ኸዓ	ነበረ	እኳ	ከለው	ከምዚኣቶም

Annex B: List of Abbreviated Words

ደ.ኣንስትዮ = ደቂኣንስትዮ

መ/ር = መምህር

ቤትት/ቲ = ቤትትምህርቲ

ቤትፍ/ዲ = ቤትፍርዲ

ክፍለት/ቲ = ክፍለትምህርቲ

ሃ/ስላሴ = ሃይለስላሴ

ቤ/ክርስትያን = ቤተክርስትያን

ወ/ር = ወታደር

ወ/ሮ = ወይዘሮ

ት/ቲ = ትምህርቲ

ቤትም/ሪ = ቤትምኽሪ

ኢ/ያ = ኢትዮድያ

ገ/ልምዓት = ገጠርልምዓት

ላ/ማይጨው = ላዕላይማይጨው

ታ.ማይጨው = ታሕታይማይጨው

ገ/ማርያም = ገብረማረያም

ገ/ዚሄር = ገረዚሄር

ወ/ሪት = ወይዘሪት

ወ/ስላሴ = ወልደስላሴ

ፍ/ስላሴ = ፍቅረስላሴ

ቤትፅ.ት = ቤትፅሕፈት

ፕ/ር = ፕሮፌሰር

ቀ.ሚንስትር = ቀዳማይሚኒስትር

ዶ/ር = ዶክተር

ገ/ጊዮርጊስ = ገብረጊዮርጊስ

ም/ኣቦወንበር = ምክትልኣቦወንበር

ሓ/ዓሰርተ = ሓለቓዓሰርተ

ሓ.ሚኢቲ = ሓለቓሚኢቲ

ሕ.ወኪል = ሕርሻወኪል

ሓ.ሽሕ = ሓለቓሽሕ

ሓ.ዘመን = ሓዲሽዘመን

ር/ምምሕዳር = ርእሰምምሕዳር

ማ/ሰብ = ማሕበረሰብ

ዓ/ግ = ዓድግራት

ዕ.ሓሙስ = ዕዳጋሓሙስ

ማ/ጨው = ማይጨው

ዓ.ዓ = ዓመተዓለም

ማ/ኮሚቴ = ማእኸላይኮሚቴ

ም/ኣቦወንበር = ምክትልኣቦወንበር

ቤትም/ሪ = ቤትምኽሪ

ተ/ሃይማኖት = ተክለሃይማኖት

ሚ/ር = ሚኒስቴር

ኮ/ል = ኮሌኔል

ሜ/ጄነራል = ሜጀርጄነራል

ብ/ጄነራል = ብርጋዴርጄነራል

ሌ/ኮለኔል = ሌቴናልኮለኔል

ኣ/ኣ = ኣዲስኣበባ

ሓ/ማሕበር = ሓረስቶትማሕበር

ዓ.ዓ = ዓመተዓለም

ማ/ኮሚቴ = ማእኸላይኮሚቴ

ር/መምህር = ርእሰመምህር

ፕ/ት = ፕሬዚዳንት

ሃ.ተፈጥሮ = ሃፍቲተፈጥሮ

ቤትፍ/ሒ = ቤትፍትሒ

ዓ.ም = ዓመተምህረት

ሚ/ሕርሻ = ሚኒስቴርሕርሻ

ቤትህ/ት = ቤትህንፀት

ር/ከተማ = ርእሰከተማ

Annex C: List of Characters with the Same Meaning but different structure

Order	1st	2nd	3rd	4th	5th	6th	7th
character	ጎ	ጐ	ጒ	ጓ	ጔ	ጎ	ጕ
replace to	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
character	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
replace to	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
character	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
replace to	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
character	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
replace to	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
character	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
replace to	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
character	ዐ	ዑ	ዒ	ዓ	ዐ	ዕ	ዖ
replace to	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
character	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
replace to	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ