ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

WORD SENSE DISAMBIGUATION FOR AMHARIC TEXT: A MACHINE LEARNING APPROACH

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

SOLOMON MEKONNEN

JUNE, 2010

A.A.U

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

WORD SENSE DISAMBIGUATION FOR AMHARIC TEXT: A MACHINE LEARNING APPROACH

BY

SOLOMON MEKONNEN

Approved by the Examining Board

Chairman, Examining Committee	Signature
Advisor	Signature
Examiner	Signature

Declaration

I, the under signed, declare that this thesis is my original work, has not been submitted as a partial requirement for a degree in any university and that all sources of materials used for the thesis have been duly acknowledged.

_____

Solomon Mekonnen

June, 2010

The thesis has been submitted for examination with my approval as university advisor.

Name: _____

Signature: _____

Date: _____

ACKNOWLEDGMENT

I would like to thank my advisor, Ato Wondwossen Mulugeta, for both his ideas and support in this research and his patience in helping me complete it within deadline. I am also grateful to our research group members Alemu Kumilachew, Abay Teshager and Zeleke Abebaw for sharing different resources and ideas.

I am especially thankful to my friend Ato Gebeyehu Kebede for initiating the research idea and giving me invaluable assistance in finishing this research. My thanks goes to my friends Daniel Yilma, Haftamu Atsbeha, Dawit Mulugeta, Nebyou Azanaw, Melkamu Beyene, Derib Erget, Fasica Tesfaye, Adamu Teshome and Alemayehu Tilahun for their support and encouragement.

Last and most importantly, to my mother Kulle Wordofa .She has been extremely supportive and encouraging in difficult times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ACRONYMS

AI	Artificial Intelligence
BNC	British National Corpus
CV	Cross Validation
IR	Information Retrieval
MRD	Machine Readable Dictionary
MT	Machine Translation
NLP	Natural Language Processing
SVM	Support Vector Machine

ABSTRACT

The theme of this thesis is Word Sense Disambiguation (WSD) for Amharic which addresses the problem of automatically deciding the correct sense of an ambiguous word based on its surrounding context. WSD is essential for many applications like Machine Translation and Information Retrieval. For the purposes of this research, we report experiments on five selected Amharic ambiguous words.

A corpus based approach to disambiguation is used, where machine learning techniques are applied to a corpus of Amharic sentences so as to acquire disambiguation information automatically. A total of 1045 English sense examples for the five ambiguous words are collected from British National Corpus (BNC) and the sense examples are translated to Amharic using dictionary. The sense examples are manually annotated and preprocessed to make it ready for experiment. Corpus based approach suffers from the so-called knowledge acquisition bottleneck. It needs large quantities of sense examples to learn disambiguation rules .This is very challenging for linguistic resource-deficient languages like Amharic.

Naive-Bayes classifier is employed from Weka 3.62 package in both the training and testing phases to perform the supervised learning on the preprocessed dataset using 10-fold cross-validation. We have evaluated the classifiers for the five ambiguous words and achieved accuracy within the range of 70% to 83% which is very encouraging but further experiments for other ambiguous words and using different approaches needs to be conducted.

CHAPTER ONE

INTRODUCTION

1.1 Background

Word Sense Disambiguation

Word Sense Disambiguation (WSD) has been of great interest and concern to the natural language and text processing community for the past fifty-years [32]. Fundamentally, WSD deals with choosing the correct sense (i.e., meaning) of a word in a given text from a list of possible senses based on the content [15]. Sense disambiguation is an "intermediate task" [61] which is not an end in itself, but rather is necessary at one level or another to accomplish most Natural Language Processing (NLP) tasks.

WSD is regarded as one of the most interesting and longest-standing problems in natural language processing. There are many uses for WSD. The most obvious application of WSD is Machine Translation. The machine translation process requires at least two stages: understanding the source language translation is done from and generating sentences in the target language. WSD is required in both stages since a word in the source language may have more than one possible translation in the target language. In order to be able to correctly translate a text, we need to know which sense is intended in the text.

Information Retrieval (IR) also benefits from WSD. Ambiguous words in the queries are problematic for information retrieval systems. Hence, retrieval engines need WSD for filtering out documents with senses irrelevant to the query.

Another potential area for WSD is Speech Processing. In speech synthesis, it is important to determine the correct pronunciations of words in order to generate speech that sounds natural. This process is difficult since there exists some words which are pronounced in more than one way depending on their content. WSD could help speech synthesis by identifying the correct sense of the word which will also provide the correct pronunciation. The reverse problem may occur in speech recognition for homophones, words that are spelled differently but pronounced in the same way. WSD can also be helpful in this situation by recognizing different senses of the same pronunciation based on its context.

Applications in text processing, grammatical analysis, content and thematic analysis also benefit from WSD techniques.

Over the years, there have been several robust, stand-alone WSD systems designed to operate with minimal assumptions about the type of information available from other processes [15]. Each of the systems has employed several common WSD approaches such as Artificial Intelligence (AI)-based [12, 33, 34, 41], knowledge-based [43, 52], and corpus-based [15, 28, 60] to perform the word sense disambiguation task.

AI methods began to flourish in the early 1960's and began to attack the problem of language understanding. As a result, WSD in AI work was typically accomplished in the

context of larger systems intended for full language understanding. In the spirit of the times, such systems were almost always grounded in some theory of human language understanding which they attempted to model and often involved the use of detailed knowledge about syntax and semantics to perform their task, which was exploited for WSD.

With knowledge-based approaches, the machine readable dictionaries (MRD) provide both the means for constructing a sense tagger along with the necessary target senses that will be employed in the system [14]. In 1986, Lesk [39] first implemented an approach in which all of the sense definitions of the word to be disambiguated were retrieved from the dictionary. Each of the senses was compared to the dictionary definitions of all the remaining words in the context. The sense with the highest overlap (i.e., common senses) with these context words was chosen as the correct sense.

In corpus-based approaches which are also called machine learning Approaches, the systems are actually trained to perform the task of WSD [14]. After the models have been trained on numerous examples, they are tested on unseen examples to determine the effectiveness of the trained classifier. There are three major corpus-based approaches, namely supervised learning, unsupervised learning along and bootstrapping. For the supervised learning method, the WSD system is constructed from a set of unbiased labeled instances drawn from the same distribution as the test set. The most common supervised learning approaches include Naive-Bayes classifiers, decision lists, decision trees, artificial neural networks (ANNs), logic learning systems, and nearest neighbor. In terms of the unsupervised learning approach, the WSD system is developed from a

clustering-based idea that attempts to discover representations of the word senses from unlabeled texts. The Bootstrapping method; it is a combination of supervised and unsupervised methods that deals with far few resources. In essence, the initial classifier is constructed with a small amount of labeled instances using any of the supervised methods and then is employed to extract a larger training set from the unlabeled instances.

Amharic Language

Amharic is spoken by about 30 million people as a first or second language, making it the second most spoken Semitic language in the world (after Arabic), probably the second largest language in Ethiopia (after Oromo), and possibly one of the five largest languages on the African continent[8]. Hudson [27] analyzed the Ethiopian census from 1994 and indicated that more than 40% of the population understood Amharic, while the current size of the Ethiopian population is about 80 million based on the preliminary reports from the census of May 2007.

As with the other languages, Amharic has many words that have multiple meanings, for example the Amharic word "**መሳል**" have three meanings in different contexts. It can be translated into English as "*to sharpen*", "*to cough*", or "*to vow*". When we look up a word in any dictionary, it can be seen that a word may have many meanings some of which are very different from the other. Given these complications, it is important for a computer to correctly determine the meaning in which a word is used.

1.2 Statement of the Problem

Amharic is the working language of the Federal Government of Ethiopia. Amharic is mother tongue for more than 30 million and second language for over 5 million people and it is one of the widely used languages in Ethiopia [8]. It is the most used language for text/document storage and media purposes in the country. Thus researches which are conducted in the language will benefit a significant number of the language speakers.

Ambiguities have been an issue in researches conducted in Amharic language. Yehenew[69] indicated that both lexical and structural ambiguities were challenges in his research on machine translation of English to Amharic. Yoseph[71] have also faced problems of synonym, polysemy and homonymy in his research on Amharic-English cross language information retrieval. The challenge has also been noticed as Atelach, et al[5] attempted to translate Amharic queries into English "Bags-of-words". They were required to perform manual disambiguation which misses domain specific senses and also often contain rare senses and also is time taking [40]. In addition like any other manual system the process of disambiguation may result in error.

There are also researches that were conducted to deal with ambiguities in Amharic language. Atelach[3] and Daniel[17] tried to resolve structural ambiguity using statistical approaches for parsing. Wube[63] also attempted to resolve structural ambiguities using a rule based approach.

As discussed earlier, There are many uses for word sense disambiguation. The most common are application of WSD in machine translation, Information retrieval, speech processing, text processing, grammatical analysis, content and thematic analysis. The absence of Automatic WSD would make it the development of such NLP and IR applications difficult. To the researcher's knowledge, Teshome[57] is the first research attempt in WSD for Amharic which tries to resolve lexical ambiguity .He demonstrated word sense disambiguation based on semantic vector analysis can improve the effectiveness of an Amharic Information Retrieval system. Machine learning approach has been used successfully for WSD in other language like English [15, 28, and 60], Chinese [64], Hebrew and German [16]. To the knowledge of the researcher the approach has not been experimented for WSD of the Amharic language.

1.3 Objective of the study

1.3.1 General Objective

The general objective of this research is to investigate the application of machine learning techniques to word sense disambiguation of Amharic texts.

1.3.2 Specific Objectives

To achieve the general objective, the study attempts to address the following specific objectives:

- Review the basic writing system, punctuation marks and syntactic structure of Amharic language;

- Study ambiguities in Amharic language so as to understand WSD issues in the language;
- Organize training and test corpus data;
- Review approaches of WSD adopted for other languages;
- Build and train WSD model using the selected classifier;
- Test the performance of the model;
- Forward conclusion and recommendations.

1. 4. Methodology

1.4.1. Literature Review

A study of available literature is done in the following areas:

- word sense disambiguation;
- Machine learning;
- classifier algorithms and their application;
- Amharic Writing system, punctuation marks and its syntactic structure;
- Ambiguities in Amharic language.

1.4.2. Data collection

In this study an approach that employs a supervised learning mechanism is selected to develop the WSD model. In this approach a significant number of sense examples are required to make training possible for classifier which was difficult to get for Amharic ambiguous words. For other languages like English, German and French a standard sense annotated data are available and used for WSD research. After reviewing available literatures, an approach that use monolingual of another language to acquire sense examples is used for this study (see section 2.2.1).An English corpora, British National Corpus(BNC) is used to acquire sense examples for Amharic ambiguous words and the examples are translated to Amharic(see section 4.1).

Thus a total of five ambiguous words are selected by a linguistic expert from the list of Homonyms collected by Girma[26]. As acquiring, translation and annotation of sense examples is very costly and time taking, in selecting the sample words attention was given to the usage of the words, the word class of sense of words and the representation of different word classes. Among the list of Ambiguous words, the most widely used and having the same word class of senses are selected. Because ambiguous words that have senses with different word classes can be resolved using part of speech tagger by their word class. The selected words are **አጠና** (eTena), **መሳል** (mesal) , **መሣሣት** (me`sa`sat) , **መጥራት** (metrat) , and **ቀረጸ** (qereSe) . In addition to the basic words their variations are also considered. Girma[26] compiled the senses of the selected words and it is included in Appendix B . The summary of the senses is presented in table 1.1.

Ambiguous Word	Senses		
	Sense1	**Sense2**	**Sense3**
eTena	strengthen	study	-
Mesal	cough	sharp	vow
me`sa`sat	taking care	thin	-
Metrat	call	clean	-
qereSe	record	shape	-

Table 1.1 Senses of selected ambiguous words

A total of 1045 sentences were acquired for the five ambiguous words and on average a total of 100 sentences is acquired for each senses of ambiguous words (see section 4.1).

1.4.3 Tools and techniques

In building the WSD model, Naive-Bayes classifier of Weka 3.62 package is used. Naive-Bayes classifiers are supervised learning approaches that have been employed to a wide-variety of problems [51]. Specifically, the Naive-Bayes classifier has been a very popular approach for WSD tasks with good performance [22, 49, 56, 59].That is why the technique is proposed for building the model. The Weka 3.62 machine learning tool is selected due to the familiarity of the researcher to the tool and because of its accessibility, processing capability and language independent features.

1.4.4. Experiments

1.4.4.1. Data preprocessing

The source data, which are English sense examples are translated to Amharic, stemmed, transliterated to English script, and manual annotation is done to enhance the training of classifier(see section 3.3). In the experiment 90 % of the dataset was used for training and the remaining 10% of the dataset for testing using 10-fold Cross Validation (CV) test split (see section 5.3).

1.4.4.2. Training and testing

Classifiers are trained using Naive Bayes classifier of Weka 3.62 package for the five ambiguous words using a set of annotated instances of the ambiguous words to create a statistical model. Five experiments are carried using 10-fold Cross CV with different features along the classifier and its parameters to train the model. And finally the performances of the classifiers are evaluated using the accuracy of their result.

1.5. Significance of the study

The results of the study are expected to produce experimental evidences that demonstrate the applicability of supervised machine learning methods to word sense disambiguation of Amharic texts. It will also contribute to future researches and development in the area of Natural Language Processing specifically in machine translation, speech processing, text processing, Information Retrieval, grammatical analysis, content and thematic analysis as those areas require word sense disambiguation as complement.

1.6. Scope and limitation of the study

Though there are supervised, unsupervised and bootstrapping machine learning techniques for WSD, due to time constraint to train, test and analyze the results, only Naive Bayes algorithm of Supervised WSD is used to build the WSD model. Owing to unavailability of sense annotated data and linguistic resources; the study was limited to the experimentation of five ambiguous words.

1.7. Organization of the Thesis

The thesis is organized into six chapters comprising Introduction, Word Sense Disambiguation, Amharic Language, Corpus preparation and System Design, Experimentation and Discussion, and Conclusion and Recommendations. This chapter gives the general introduction of the thesis. The second chapter reviews different literatures regarding Word Sense Disambiguation together with its different approaches. It also introduces machine learning with different algorithms. The third chapter reviews Amharic writing system and ambiguities in the language. The fourth chapter discusses the process of corpus preparation and the architecture of the system. The fifth chapter discusses the experimentation and discussion of the findings. Finally, chapter Six deals with the conclusion and the recommendations drawn from the findings of the study.

CHAPTER TWO

WORD SENSE DISAMBIGUATION

In this chapter literature in the field WSD is reviewed and discussed in brief. The chapter covers overview of WSD, application areas for WSD and discussion on major approaches that have been employed for WSD research with special focus on corpus based approach which will be used in this study. Moreover machine learning algorithms that are tested to perform well for WSD research including Naive Bayes Classifier which will be used in this research are discussed. The discussion on different approaches and algorithms would help the understanding of the central problem in WSD research and also facilitates the comparison of existing approaches to the specific solutions that are employed in this study..

2.1. Introduction

Many words in many natural languages including Amharic have multiple meanings or senses. For example, an Amharic word "*መሣሣት*" can mean *"to take care"* or *"to be thin"* in different contexts which leads to ambiguity. Humans resolve such ambiguity by understanding the context of surrounding each ambiguous word in a document and its sound. But for a machine, it will be difficult to determine the meaning of ambiguous words. For machines, there is a need for Word sense disambiguation (WSD) which is the task of automatically determining the meaning of an ambiguous word from its context. So in our example above, given the ambiguous word "*መሣሣት*", WSD involves interpreting

the surrounding context of the word and analyzing the properties exhibited by the context to determine the right sense of "**መሣሣት**".

According [32] WSD involves two steps. The first step is to determine all the different senses for every word relevant to the text or discourse under consideration, i.e., to choose a sense inventory from the lists of senses in everyday dictionaries, from the synonyms in a thesaurus, or from the translations in a translation dictionary. The second step involves a means to assign the appropriate sense to each occurrence of a word in context. All disambiguation work involves matching the context of an instance of the word to be disambiguated either with information from external knowledge sources or with contexts of previously disambiguated instances of the word. For both of these sources we need preprocessing or knowledge-extraction procedures representing the information as context features. However, it is useful to recognize that another step is also involved here: the computer needs to learn how to associate a word sense with a word in context using either machine learning or manual creation of rules or metrics. Unless the associations between word senses and context features are given explicitly in the form of rules by a human being, the computer will need to use machine learning techniques to infer the associations from some training material.

Word sense disambiguation (WSD) is an "intermediate task", which is not an end in itself, but rather is necessary at one level or another to accomplish many natural language processing tasks such as Information Retrieval (IR) [58], Machine translation (MT) [29] and question-answering [46]. For MT, WSD is important when it comes to selecting the appropriate target language word for an ambiguous source language word. For IR, sense

disambiguation would prevent the retrieval of irrelevant documents that contain query words of a different sense, while use of semantic tags could help in solving the prepositional phrase attachment problem. In question answering systems, WSD is used to retrieve the appropriate answer from document collection for a given query containing ambiguous words.

2.2. Approaches to Word Sense Disambiguation

All disambiguation work involves matching the context of the instance of the word to be disambiguated with either information from an external knowledge source (knowledge driven WSD), or information about the contexts of previously disambiguated instances of the word derived from corpora (data-driven or corpus-based WSD) [32]. Any of a variety of association methods is used to determine the best match between the current context and one of these sources of information, in order to assign a sense to each word which is called knowledge acquisition bottleneck in WSD literatures. Different WSD approaches have been used through the evolution of WSD research. The major ones are: Corpus-Based, Knowledge-based Approaches, and Hybrid Approaches. In this section the survey of these approaches will be presented.

2.2.1. Corpus-Based Approaches

A major challenge facing WSD research is the ability to acquire a large amount of words with their different contexts. Corpus-based approaches came up with alternate solution to the challenge by obtaining information necessary for WSD directly from textual data

which is called a corpus. A corpus provides a bank of samples which enable the development of numerical language models, and thus the use of corpora goes hand-in-hand with empirical methods [32].Corpus based approaches can be categorized into three sub classes based on the form of training : Supervised Word Sense Disambiguation, unsupervised Word Sense Disambiguation and Bootstrapping Approach to WSD.

Supervised Word Sense Disambiguation

Supervised Word Sense Disambiguation use machine-learning techniques to learn a classifier from labeled training sets, that is, sets of examples encoded in terms of a number of features together with their appropriate sense label (or class)[53]. The systems in the supervised learning approach category are trained to develop a classifier that can be used to assign a yet unseen example to one of a fixed number of senses. That means, there is trained corpus, where the system learns to classify and a test corpus which the system must annotate. So, supervised learning can be considered as a classification task.

Supervised learning requires a labeled training data that is every instance in the training data is associated with an output value or label that can be thought of as a special attribute or feature for each instance. For WSD, every instance in the training data should be assigned a label that corresponds to the correct sense of the ambiguous word that the instance contains or represents. Machine learning algorithms make use of the instance attributes or features in the training data and generate a model to predict the label of any given instance. This model can be applied to unseen instances to predict their labels. Algorithms that can learn to predict discrete valued labels are called classification

algorithms or classifiers, whereas the algorithms that can learn to predict continuous valued labels are called regression algorithms. As the task of WSD only involves discrete valued labels for word senses, we use only classification algorithms.

The main problem associated with supervised approach is the need for a large sense-tagged training set. Despite the availability of large corpora in some language, manually sense-tagging of a corpus is very difficult limiting the number of sense tagged words to be used and very few sense-tagged data are available now. To deal with this problem a variety of unsupervised WSD methods, which use a machine readable dictionary or thesaurus in addition to a corpus, have also been proposed [36], [66], [67]. Bilingual parallel corpora, in which the senses of words in the text of one language are indicated by their counterparts in the text of another language, have also been used in order to avoid manually sense-tagging training data [10]. In this method, bilingual corpora are used since different senses of some words often translate differently in another language. Parallel corpora, especially accurately aligned parallel corpora are rare, although attempts have been made to mine them from the Web [50].In [64] it is proposed to use Chinese monolingual corpora and Chinese-English bilingual dictionaries to automatically acquire sense examples for English ambiguous words and is reported that the result exceed previous state-of-the-art comparable systems. Their approach does not rely on scarce resources such as aligned parallel corpora or accurate parsers. In [16] the use of monolingual corpora of English for Hebrew and German language WSD is also tested and found the approach very useful for disambiguation.

For this study a supervised word sense disambiguation will be used. Sense example Amharic words will be acquired from English monolingual corpora as it worked for Hebrew [16] which is a Semitic language like Amharic and the senses will be translated back to Amharic as there are scarce resource like Wordnet, thesaurus, Sense-tagged data and large parallel corpus for Amharic language.

Unsupervised Word Sense Disambiguation

Unsupervised Word Sense Disambiguation methods are based on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word in context unlike supervised method [53]. Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck [23], that is, the lack of large-scale resources manually annotated with word senses. These approaches to WSD are based on the idea that the same sense of a word will have similar neighboring words. They are able to induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters. They do not rely on labeled training text and, in their purest version, do not make use of any machine–readable resources like dictionaries, thesauri, ontologies,etc. However, the main disadvantage of fully unsupervised systems is that, as they do not exploit any dictionary, they cannot rely on a shared reference inventory of senses [19].

Like the supervised learning, even the unsupervised WSD methods strive from the data sparseness problem, since enormous amounts of text are needed to ensure that all senses of a polysemous word are represented in the corpus.

Bootstrapping Approach to WSD

The bootstrapping approach is situated between the supervised and unsupervised approach of WSD. The aim of bootstrapping is to build a sense classifier with little training data, and thus overcome the main problems of supervision: the data scarcity problem specially lack of annotated data. The bootstrapping methods use a small number of contexts labeled with senses having a high degree of confidence. This could be accomplished by hand tagging with senses the contexts of an ambiguous word w for which the sense of w is clear because some seed collocations [32] occur in these contexts. These labeled contexts are used as seeds to train an initial classifier. This is then used to extract a larger training set from the remaining untagged contexts. Repeating this process the number of training contexts grows and the number of untagged contexts reduces. We will stop when the remaining unannotated corpus is empty or any new context can't be annotated.

2.2.2. Knowledge-based Approaches

Corpus based approach require considerable amount of work to create a classifier for each word in a language as discussed earlier. As a result researchers tend to work for a few words. Knowledge-based approaches use an explicit lexon like Machine Readable Dictionaries (MRD), thesauri, computational lexicons such as WordNet or (hand-crafted) knowledge bases as information source to resolve lexical ambiguities for many words[19].

Lesk [39] created knowledge bases which associate each sense in a dictionary with a signature composed of the list of words appearing in the definition of that sense. Disambiguation was accomplished by selecting the sense of the target word whose signature contained the greatest number of overlaps with the signatures of neighboring words in its context. Because of the fact that dictionaries are created for human use, not for computers, there are some inconsistencies [31]. Although they provide detailed information at the lexical level, they lack pragmatic information used for sense determination.

Thesauri provide information about relationships among words, most notably synonymy [19]. Thesaurus based disambiguation makes use of the semantic categorization provided by a thesaurus or a dictionary with subject categories. The basic inference in thesaurus-based disambiguation is that semantic categories of the words in a context determine the semantic category of that context as a whole [32]. And this category then determines the correct senses that are used. Similar to machine readable dictionaries, a thesaurus is a resource for humans, so there is not enough information about word relations.

Computational Lexicons are a large electronic database containing useful lexical relations in linguistic Psycholinguistic and computational research has led to a number of efforts to create large electronic databases of such relations [53]. Lexicon like WordNet is used for sense evaluation and for similarity measure in WSD. For example [55] created a knowledge base from WordNet's hierarchy and apply a semantic similarity function to accomplish disambiguation, also for the purposes of information retrieval.

2.2.3 Hybrid Approaches

Since they obtain disambiguation information from both corpora and explicit knowledge-bases, Hybrid Approaches do not fall into either knowledge or corpus-based. Hybrid systems aim to use the strengths of the both conquering specific limitations associated with a particular approach, to improve WSD accuracy. They base both on a 'knowledge-driven, corpus-supported' theme, utilizing as much information as possible from different sources. [67] is an example of Hybrid approaches. Yarowsky[67] used Bootstrapping approaches where initial data comes from an explicit knowledge source which is then improved with information derived from corpora. He defines a small number of seed definitions for each of the senses of a word (the seeds can also be derived from dictionary definitions or lexicons such WordNet). Then the seed definitions are used to classify the obvious cases in a corpus.

2.3. WSD for Amharic

Though there is clearly a need for WSD for Amharic, to the researcher's knowledge the only research attempt is Teshome[57].He has studied the use of WSD based on semantic vector for improving the precision and recall measurements of information retrieval for Amharic legal texts. The Ethiopian postal code which consisting of 865 articles was used as a corpus in the study. He developed his own algorithm based on distributional hypothesis stating that words with similar meanings tend to occur in similar contexts. For disambiguation of a given word, he computed the context vector of each occurrence of the words. The context vector was derived from the sum of the thesaurus vectors of the

context words. He constructed the thesaurus by associating each word with its nearest neighbors.

For evaluating WSD, he used pseudo words which are artificial words rather than real sense tagged words reasoning that it is costly to prepare sense annotation data. He compared his algorithm with Lucene algorithm and reported that the algorithm is superior over the Lucene's one.

The approach used in [57] and in this study is similar in the way that a corpus is used as a source of information for disambiguation. The difference is, in this study a Naive-Bayes algorithm which is a machine learning techniques is used where as in [57] an algorithm based on semantic vector is developed. In addition, in this study a real sense annotated data for ambiguous words is prepared unlike a pseudo words which is used in [57].Finally the domain for[57] is a specific legal text where as the corpus in this study is domain independent..

2.4. Machine learning

Learning can be defined as improving one's performance on a given task with the aid of prior experience [45]. One way of making computers learn involves training machine learning algorithms with the help of an initial set of training data. The experience that the machine learning algorithms gain from the training data can then be applied to make predictions about previously unseen data. One can train a machine learning algorithm such as the Naive Bayes classifier to disambiguate occurrences of the ambiguous a given

word. Such a trained classifier can then takes as input previously unseen sentences containing the word, and predict the correct sense of word in sentences.

Learning is further categorized as supervised or unsupervised. In supervised learning, a trainer provides the correct labels or outputs for the training data. In unsupervised learning, there is no trainer involved; the correct label for the training data instances is not available. The advantage of supervised learning is that high accuracy can be obtained on unseen instances given that a sufficient amount of manually labeled training data is provided to generate a good model. The drawback of the supervised learning approach is that manually labeled data is highly expensive to generate in terms of time as well as money. Unsupervised methods benefit from the fact that they do not require manually labeled data. However, they usually suffer from low accuracy values on unseen instances.

2.5. Machine Learning Algorithms

The following machine learning algorithms: Na¨ive Bayes Classifier, Decision Trees, Decision Lists, Support Vector Machines are widely used WSD research. The Na¨ive Bayes classifier, decision trees and decision lists have all been shown to perform well on the task of WSD [22,49,59] .Support Vector Machines have been recently shown to perform well on WSD and similar tasks [11,70].

2.5.1. Naive Bayes Classifier

Naive Bayes has proven effective in many practical applications, including Word Sense Disambiguation, text classification, medical diagnosis, and systems performance

management [29, 45, and 49].It is one of the simplest and most popular machine learning algorithms. It is based on the Bayes' rule for conditional probabilities, which states that:

$$P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{\sum_k P(X = x_j|Y = y_k)P(Y = y_k)} \quad \text{-1-}$$

This essentially states that the conditional probability or the posterior probability P(Y |X) can be found by taking the product of the conditional probability P(X|Y) and the unconditional probability or the prior probability P(Y), and dividing this product by the total probability that X has the given value over all possible values of Y . In the case of the Naive Bayes classifier, Y represents the output or the label for each instance of the dataset and X represents each instance in the dataset. Since X can have multiple attributes, we represent it as a vector hX1,X2,X3, ...,Xni with n features in general. Using this, the Bayes' rule equation for calculating the probability of any class value yi for a given instance X becomes:

$$P(Y = y_i|X_1, X_2...X_n) = \frac{P(X_1, X_2...X_n|Y = y_i)P(Y = y_i)}{\sum_k P(X_1, X_2...X_n|Y = y_k)P(Y = y_k)} \quad \text{-2-}$$

According to the equation above the class value yi assigned to an instance is the one which has the maximum probability. Since the denominator term remains the same for

probability calculation of all class values, we can assign the class value using the following equation:

$$Y_{out} = \arg\max_{y_i \epsilon Y} P(X_1, X_2 ... X_n | Y = y_i) P(Y = y_i) \quad \text{-3-}$$

that is, the class value is the one which maximizes the numerator of the Bayes' rule equation shown earlier.

The "Naive" part of the Naive Bayes classifier is that it makes the simplifying assumption that all the features of an instance are conditionally independent given its label Y . Therefore using the rule of conditional independence of probabilities, the above equation reduces for a Na¨ive Bayes classifier to:

$$Y_{nb} = \arg\max_{y_i \epsilon Y} P(Y = y_i) \prod_{j=1}^{n} P(X_j | Y = y_i) \quad \text{-4-}$$

Given a new unseen instance to classify, the Na¨ive Bayes classifier calculates the probability of each class value given the features of the new instance and then assigns it the class value that has the maximal probability.

In the context of Word Sense Disambiguation, Naive Bayes relies on the calculation of the conditional probability of each sense *Si* of a word *w* given the features *f j* in the context. The sense ˆ*S* which maximizes the following formula is chosen as the most appropriate sense in context:

$$\hat{S} = \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} P(S_i \mid f_1, \ldots, f_m) = \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} \frac{P(f_1, \ldots, f_m \mid S_i) P(S_i)}{P(f_1, \ldots, f_m)}$$

$$= \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} P(S_i) \prod_{j=1}^{m} P(f_j \mid S_i),$$

where *m* is the number of features, and the last formula is obtained based on the Naive assumption that the features are conditionally independent given the sense (the denominator is also discarded as it is constant and does not influence the maximization calculations). The probabilities *P(Si)* and *P(f j | Si)* are estimated, respectively, as the relative occurrence frequencies in the training set of sense *Si* and feature *fj* in the presence of sense *Si* . In the training set, there might not be occurrence of some senses for ambiguous words which is called zero count. The zero count affects the other non-zero conditional probabilities in the multiplication which will make the result zero. So it need to be smoothed: for instance, they can be replaced with P(Si)/N where N is the size of the training set [20,30]. However, this solution leads probabilities to sum to more than 1. Back off or interpolation strategies can be used instead to avoid this problem.

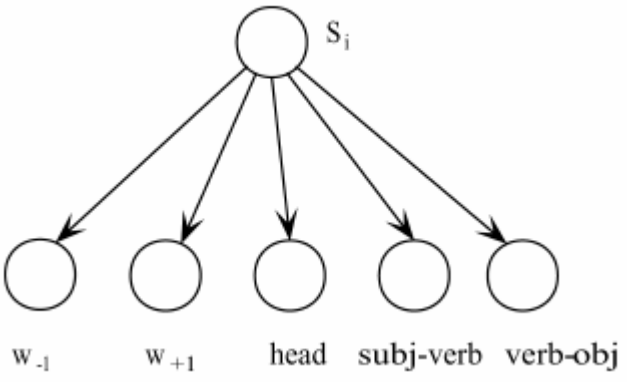


Figure 2.1 An example of Bayesian Network [53]

In Figure 2.1 we report a simple example of a Naive Bayesian network. For instance, suppose that we want to classify the occurrence of noun *bank* in the sentence "The bank cashed my check" given the features: {w−1 = the, w+1 = cashed, head = bank, subj-verb =cash, verb-obj= −}, where the latter two features encode the grammatical role of noun bank as a subject and direct object in the target sentence. Suppose we estimated from the training set that the probability of these five features given the financial sense of bank are P(w−1 = the | bank/FINANCE) = 0.66, P(w+1 = cashed |bank/FINANCE) = 0.35,P(head = bank | bank/FINANCE) = 0.76, P(subj-verb = cash | bank/FINANCE) = 0.44,P(verb-obj = − | bank/FINANCE) = 0.6. Also, we estimated the probability of occurrence of P(bank/FINANCE) = 0.36. The final score is

score(bank /FINANCE) = (0.36) ·(0.66) ·(0.35) ·(0.76) ·(0.44) ·(0.6) = 0.016.

This implies that the probability of the senses of the word *bank* to be *finance* in the specific context is 0.016.

2.5.2. Decision Trees

In the last decades, decision trees have been rarely applied to WSD (in spite of some studies), for example [9] and [37]. The decision tree classifier is one of the possible approaches to multistage decision making. It is one of the most intuitive and popular machine learning algorithms. They are based on the idea of information gain from information theory. A decision tree is a top-down hierarchy of test conditions on the

attributes of a dataset. Every node in a decision tree is a test of some attribute of the given instance, to categorize it into some subset depending on the value of the attribute for that instance. Every such non-leaf node in the decision tree (that tests an attribute) has as many branches or child nodes as the number of different values for the attribute being evaluated at that node. The tree is built starting from the root node, which tests the attribute that provides maximum information gain for the entire dataset, and the process continues recursively along each branch, until no further classification is required (usually within some tolerable level of error), so that the process can stop even if the dataset contains error).

Information gain is defined in terms of the entropy difference of a parent node in the decision tree and the weighted average of the entropies of its child nodes. Entropy of a node can be seen as a measure of "impurity" of a node in terms of the proportion of instances it contains of the different classes. The more balanced the proportion of different classes, the more the set is impure and hence the high entropy. So a set of instances with two class values will have maximum entropy if half of the instances are of one class and the other half are of the second class. The minimum entropy is achieved when all the instances are of the same class (either the first or the second). Mathematically, entropy is the weighted average of negative logarithms (to the base 2) of the probabilities of class values in the set of data instances at the given node. Therefore the entropy of a node *N* is:

$$E(N) = \sum -p_i log_2 p_i$$

Where p_i is the probability of class value *i*. For example, let us assume that for the node *N* being currently processed while building a decision tree there are *m* instances of the positive class (+) and *n* instances of the negative class (-). Then entropy at that node *N* is given by:

$$E(N) = -\left(\frac{m}{m+n}\right) log_2 \left(\frac{m}{m+n}\right) - \left(\frac{n}{m+n}\right) log_2 \left(\frac{n}{m+n}\right)$$

To define information gain, let us assume the following: node *N* has |*N*| instances and entropy *E(N)* is defined as above. Information gain is evaluated with respect to some attribute of the dataset.

Let us assume that the attribute being considered currently is *A* and that it has *v* distinct values in the dataset. Therefore as discussed earlier, if the current node evaluates attribute *A*, then we will have v branches and therefore *v* child nodes of *N*. Let us name these child nodes *Ai*, 1≤i ≤ v.

Depending upon their value for the attribute *A*, the |*N*| instances are divided among the child nodes *Ai*. Let us assume that the number of instances at child node *Ai* is |*Ai*| and the entropy at the child node *Ai* is *E(Ai)*. Now, we define information gain using attribute *A* at node *N* as:

$$Gain(N, A) = E(N) - \sum_i \frac{|A_i|}{|N|} E(A_i)$$

A decision tree is constructed recursively by evaluating the information gain of each attribute for the set of data instances at the current node. For the root node, the information gain of all attributes over the entire dataset must be determined. Then the attribute with the maximum information gain is selected as the attribute to be used as the test for the current node. For all non-root nodes, the information gain of only those attributes that have not already been used in the parent branch of current node is evaluated. The leaf nodes do not evaluate any attribute, and in the best case contain instances of just one class and are therefore "pure." In the event the dataset contains an error or is not separable using the decision tree algorithm, the leaf nodes may not be pure.

Given a new unseen instance to classify, a decision tree begins by evaluating the instance for the attribute at the root node and "passes" the instance down the appropriate branch in the decision tree, until it reaches a leaf node. If the leaf node is pure, then the instance is assigned the same class as that of all the nodes in the leaf node. If the node is not pure, then one approach can be to assign the new instance the class that is most frequent among the instances at the leaf node.

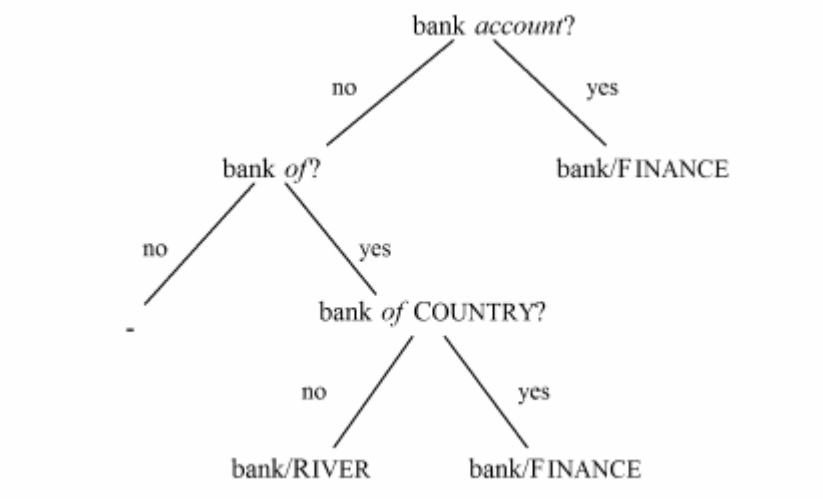


Figure 2.2 An example of Decision tree [53]

An example of a decision tree for WSD is reported in Figure 2.2 .For instance, if the noun bank must be classified in the sentence "we sat on a bank of sand," the tree is traversed and, after following the no-yes-no path, the choice of sense bank/RIVER is made. The leaf with empty value (-) indicates that no choice can be made based on specific feature values.

2.5.3. Decision Lists

Decision list learning is a rule based approach, and is similar in concept to decision trees. The aim of the decision list learner is to discover a set of "if....then" or "switch....case" conditions that test attributes of the data instances and assign them a class value based on the first rule that matches or covers the data instance. If none of the discovered rules matches a given instance, then the most frequently occurring class in the training dataset is assigned as its class value. The rules are learned in an iterative and incremental fashion using the features of the training data. One rule is learned at best accuracy which takes

precedence or information gain as in the case of decision trees. The final decision list is ordered, so that any new unseen instance to be classified is tested with each of the rules in the decision list in order from top to bottom, and the first rule that covers the instance decides the output class.

In WSD context, It can be seen as a list of weighted "if-then-else" rules. A training set is used for inducing a set of features. As a result, rules of the kind (feature-value, sense, score) are created. The ordering of these rules, based on their decreasing score, constitutes the decision list. Given a word occurrence w and its representation as a feature vector, the decision list is checked, and the feature with highest score that matches the input vector selects the word sense to be assigned:

$$\hat{S} = \text{argmax}_{S_i \in Senses_D(w)} score(S_i).$$

According to [67], the score of sense Si is calculated as the maximum among the feature scores, where the score of a feature f is computed as the logarithm of the probability of sense Si given feature f divided by the sum of the probabilities of the other senses given feature f:

$$score(S_i) = \max_f log\left(\frac{P(S_i \mid f)}{\sum_{j \neq i} P(S_j \mid f)}\right).$$

The above formula is an adaptation to an arbitrary number of senses due to [2] and [67] formula, originally based on two senses. The probabilities P(Sj | f) can be estimated

using the maximum-likelihood estimate. Smoothing can be applied to avoid the problem of zero counts. Pruning can also be employed to eliminate unreliable rules with very low weight.

Feature	Prediction	Score
account with bank	Bank/FINANCE	4.83
stand/V *on*/P ... bank	Bank/FINANCE	3.35
bank *of blood*	Bank/SUPPLY	2.48
work/V ... bank	Bank/FINANCE	2.33
the left/J bank	Bank/RIVER	1.12
of the bank	-	0.01

Table 2.1 An example of Decision List [53]

A simplified example of a decision list is reported in Table 2.1 the first rule in the example applies to the financial sense of bank and expects account with as a left context,the third applies to bank as a supply (e.g., a bank of blood, a bank of food), and so on (notice that more rules can predict a given sense of a word).

It must be noted that, while in the original formulation [49] each rule in the decision list is unweighted and may contain a conjunction of features, in Yarowsky[67] approach each rule is weighted and can only have a single feature. Decision lists have been the most successful technique in the first Senseval evaluation competitions (example: [68]). Decision lists applied in an attempt to relieve the knowledge acquisition bottleneck caused by the lack of manually tagged corpora [2].

2.5.4. Support Vector Machine

Support Vector Machines (SVMs) were developed by [13] for binary classification. SVMs are machine learning algorithms that have their roots in statistical learning theory and can be applied to classification as well as regression problems. The SVM formulation for classification is designed to handle only two-class problems, but there are extensions to this basic formulation that handle the multi-class classification problems such as WSD. The method is based on the idea of learning a linear hyperplane from the training set that separates positive examples from negative examples. The hyperplane is located in that point of the hyperspace which maximizes the distance to the closest positive and negative examples (called support vectors). In other words, support vector machines (SVMs) tend at the same time to minimize the empirical classification error and maximize the geometric margin between positive and negative examples.

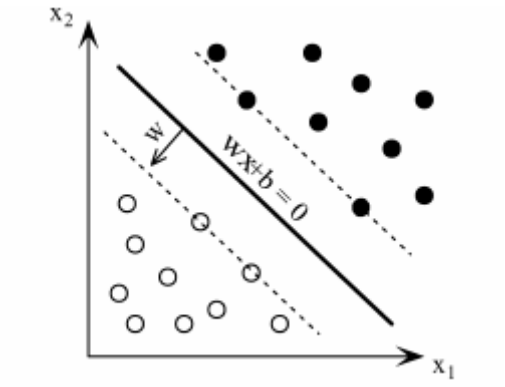


Figure 2.3 The geometric intuition of SVM [53].

Figure 2.3 illustrates the geometric intuition: the line in bold represents the plane which separates the two classes of examples, whereas the two dotted lines denote the plane

tangential to the closest positive and negative examples. The linear classifier is based on two elements: a weight vector w perpendicular to the hyperplane (which accounts for the training set and whose components represent features) and a bias b which determines the offset of the hyperplane from the origin. An unlabeled example x is classified as positive if f (x) = w·x+b ≥ 0 (negative otherwise). It can happen that the hyperplane cannot divide the space linearly. In these cases it is possible to use slack variables to "adjust" the training set, and allow for a linear separation of the space.

As SVM is a binary classifier, in order to be usable for WSD it must be adapted to multiclass classification (i.e., the senses of a target word). A simple possibility, for instance, is to reduce the multiclass classification problem to a number of binary classifications of the kind sense Si versus all other senses. As a result, the sense with the highest confidence is selected.

It can be shown that the classification formula of SVM can be reduced to a function of the support vectors, which—in its linear form—determines the dot product of pairs of vectors. In general, the similarity between two vectors x and y is calculated with a function called kernel which maps the original space (e.g. of the training and testing instances) into a feature space such that k(x, y) = _(x) ·_(y), where _ is a transformation (the simplest kernel is the dot product k(x, y) = x ?y). A nonlinear transformation might be chosen to change the original representation into one that is more suitable for the problem (the so-called kernel trick). The capability to map vector spaces to higher dimensions with kernel methods, together with its high degree of adaptability based on

parameter tuning, are among the key success factors of SVM.SVM has been applied to a number of problems in NLP, including WSD [20] .

2.6. Summary

A basic introduction to the field of WSD has been presented in this chapter. A survey of the major approaches to WSD has been presented, emphasizing the key WSD research problems that should be addressed by any type of solution. The field of Machine learning with associated algorithms that are used for WSD research has also been discussed. For this study a corpus based approach specifically supervised WSD is adopted. The classifier will be modeled using Naive Bayes classifier.

CHAPTER THREE

Amharic Language

Amharic is a Semitic language used as working language in Ethiopia. It is highly inflectional and quite dialectally diversified. With more than 20 million speakers, it is the second most spoken Semitic language in the World (after Arabic) and today probably one of the five largest on the African continent (albeit difficult to determine, given the dramatic population size changes in many African countries in recent years). [38]. Manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literature, journalism, education, and so on [4]. A wide variety of Amharic literatures including books, religious writings, fiction, poetry, plays, and magazines are available both in printed and machine readable format.

Like other Semitic languages, Amharic word variants are formed from roots that are bi-, tri- or quad radical. A phoneme represents a basic sound or unit of sound. Every glyph or consonant form is a phoneme or unit of word. A phoneme or collection of phonemes forms a morpheme, which is the smallest meaningful unit in a word[6].An Amharic root is a sequence of consonants and is the basis for the derivation of verbs; a stem, on the other hand is a consonant or consonant-vowel sequence[44].. The Amharic language makes use of prefixing, suffixing and infixing to create inflectional and derivational word form [44] .Words in Amharic can be divided into content-bearing and non content-bearing which are also called stop words.

3.1 The Amharic Writing System

Amharic belongs to the Semitic language family and is one of the most widely spoken languages in Ethiopia. Amharic has its own script that is borrowed from Ge'ez, another Ethiopian Semitic language [18]. The script consists of a core thirty-three characters (called 'fidel') each of which occurs in one basic form and in six other forms that may be described as orders. The seven orders represent different forms of a consonant. The non-basic forms are derived from the basic ones by somewhat regular modifications for the first four orders and for the last two words it is irregular [1]. These seven orders (the first basic order and the other six orders) represent the different sounds of a consonant-vowel combination (a characterization known as syllabic).. The 33 core characters then yield 231 distinct symbols. In addition to the 231 characters, there are others that contain special features usually representing labialization like ኳ(kwa) from ከ (ke) and ቈ(qwa) from ቀ (qe). etc. .For this study the writing system of [18] and [65] will be used and included in the Appendix A.

3.2. Amharic Punctuation Marks

Analysis of Amharic texts reveals that different Amharic Punctuations marks are used for different purposes. In [6] it is indicated that there are about 17 punctuation marks of which only a few of them are commonly used and have representations in Amharic software.

The Amharic writing system uses some indigenous and foreign punctuation marks (signs) in addition to the Amharic characters [18]. However, only few of them are practically used, especially in computer-written text. The word-separator (hulet neTb), two square dots arranged like colon (:), and sentence-separator (arat neTb), four square dots arranged in a square pattern (: :), are the basic punctuation marks in Amharic writing system that are used consistently. Today, the use of Hulet Neteb is not seen in modern typesetting. In typesetting its place is almost completely taken over by space. Lists in Amharic text are separated by an equivalent of comma, 'netela serez'(፣) followed by ASCII space and 'derib sereze' (፤), which is the equivalent of semi-colon. The use of '…' for question mark is not used rather a '?' which is borrowed from English is used. Table 3.1 lists the most commonly used Amharic punctuation with their equivalent in English which is adopted from [65].

Amharic	**English**
፡	**White Space**
፡፡	'
፤	;
፣	,
፧	**?**

Table 3.1 Most commonly used punctuation marks with their English corresponding marks

3.3. Syntactic Structure of Amharic

The syntactic structure is formed by combining different words. Since Amharic word formation follows its own structure, the syntax of the language also exhibit a unique structure. The syntactic structure of Amharic is generally SOV (Subject-Object-Verb).The modifiers in such structure generally precede the word or the phrases they modify. For example, the Amharic equivalent for the English sentence "she has understood mathematics." is "**እሷ ሂሳብ ገብቷታል ።** " In the sentence "**እሷ**" is the subject and the object is "**ሂሳብ** " and the verb is "**ገብቷታል**". But usually pronouns are omitted when used as a subject. For the above English sentence the common saying in Amharic is "**ሂሳብ ገብቷታል**" by implicitly understanding the pronoun.

3.4. Ambiguities in Amharic

Getahun[25] identified five types of ambiguity in Amharic: Phonological Ambiguity, Lexical Ambiguity, Structural Ambiguity, Referential Ambiguity, Semantic Ambiguity, and Orthographic ambiguity. We now summarize each type of ambiguity and the examples are adopted from [25].

3.4.1 Phonological Ambiguity

Phonological ambiguity is a result due to the sound used for the word from the placement of pause with in a structure which occurs in speech .It can be illustrated through the following example:

Example

ደግ ሰው ነበር

[deg + sew] neber

kind person was

In the above sentence '+' sign shows where the pause is. When the sentence is pronounced with pause it means "He was a kind man" but the meaning differs if it is pronounced without pause .It will mean "They had preparation for a banquet".

3.4.2. Lexical Ambiguity

Lexical ambiguity refers to a case in which either a lexical unit belongs to different part-of-speech categories with different senses, or to a lexical unit for which there is more than one sense, while these different senses fall into the same part-of-speech category [10].There are three different factors that can cause lexical ambiguity which are: Categorical Ambiguity, Homonymy and Homophonous Affixes.

Categorical Ambiguity

Categorical ambiguity is a result from lexical elements which have the same phonological form but belongs to different word class. This will be more described using the following ambiguous word:

Example

አክርማ ሰጠችኝ

ekrma seT-ec-N

? gave-she-me

In the above example the underlined word *"ekirma"* is ambiguous since it has both nominal and a verbal meaning. It has two interpretations:

i.She gave me Akirma(a kind of grass).[With nominal meaning]

ii.She gave me something after delaying it for sometime.[With verbal meaning]

Homonymy

Homonyms are those lexical items with the same phonological form but with different meanings which will cause ambiguity.It can be illustrated with the following example:

Example:

በወሬ አልፈታም

bewerE elfetam

In the example the underlined *"bewerE"* is an ambuigous word having the following two different structures and readings shown below:

i. be-wer-E *el-feta-m*

with-month-my *Neg-released-Neg*

In this sense it means that “I will not be released in a month”

ii. be-wer-E *el-fata-m*

Neg-release-Neg

It means that “I will not get frustrated by any rumor”

Homophonous Affixes

This ambiguity result when affixes serve as different word classes. The following example show how homophonous affixes cause ambiguity.

Example:

ቤቱ ፈረሰ

bEt-u ferese

The above sentence is ambiguous because the suffix /-u/ serves as a definite article or as a third person masculine marker. It has two different meanings:

i. The house is destroyed. And

ii. His house is destroyed

3.4.3. Structural Ambiguity

Structural ambiguity resulted when a constituent of a structure has more than one possible position. By a structure we mean the way syntactic constituents are organized. The following is an example of such ambiguity:

Example:

የሀበሻ ታሪክ አስተማሪ

ye-Hebexa tarik estemari

of-Abyssinia history teacher

The above sentence can have two different interpretations:

i.a person who teaches Abyssinian history

ii an Abyssinian who teaches history

It can be further illustrated using structural organization of the sub-constituent */tarik/ 'history'*. It is shown in the following labeled representation:

i. [[[ye-Hebexa tarik] [estemari]]]]

N N N

"Abyssinian history teacher"

ii.[ye-Hebexa [[tarik] [estemari]]]

N N N

3.4.4. Referential Ambiguity

This ambiguity arise, when a pronoun has more than one possible antecedents , thus having as many reading as there are antecedents .The following sentence is an example of such ambiguity.

Example

ካሳ ስለተመረቀ ተደሰተ

Kasa sletemereke tedesete

The above sentence has two different readings:

i.Kasa was pleased because he graduated and

Ii.Somebody was pleased because Kasa graduated

3.4.5. Semantic Ambiguity

Semantic ambiguity is caused by polysemic , idiomatic and idiomatic and metaphorical constituents.

The following sentence is an example Polysemic constituent which has multiple meanings.

Example:

መብራቱ ጠፋ

Mebrau tefa

The above sentences have two interpretations:

i. *The light went off.*

ii. *Mebratu(a person) disappeared*

Idioms refer to an expression that means something other than the literal meanings of its individual words. Idioms ambiguity can be illustrated using the following example:

Example:

በሬ ወለደ

berE welede

The literal meaning of the above example is *"An ox gave birth to a calf"* but the idiomatic expression refers to *"impossible* "to happen.

Metaphors have literal or non-literal (metaphoric) senses. The following is an example of metaphoric ambuigity:

Example:

አራስ ነብር

eras nebr

It has two different interpretations:

i.'inascible, hot tempered'

ii.'leopard with new-born cubs'

3.4.6. Orthographic Ambiguity

Orthographic Ambiguity is resulted from geminate and non-geminate sounds. The ambiguity can be resolved using context. Though in some cases it might not be possible like the following example:

Example

መኪናው ይሰራል

Mekinaw ysral

The word "*ysral* " is the cause of ambiguity. The sentence is ambiguous between the following meanings.

i.The car works*("ysral")*

ii.The car will be repaired*("yssaral")*

For this study, lexical ambiguity which is believed to be resolved by word sense disambiguation will be dealt among the type of ambiguities that are discussed.

CHAPTER FOUR

CORPUS PREPARATION AND SYSTEM ARCHITECTURE

4.1. Acquisition of Sense Examples

As discussed in the literature review part, one of the mechanisms to acquire sense examples is to use monolingual corpora of second language and translate the sense examples to the original language. For this study an English text corpus will be used for acquisition of sense examples. English is selected for the reason that it has been used successfully for Hebrew WSD research which is a Semitic language like Amharic [16] and the researcher is familiar with it. As the mechanism is used for this study, a total of 1045 sense example sentences for the five ambiguous words are acquired from the British National Corpus. BNC contains a total of 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. Its vast contents, having a variety of sources and its searchability are the reasons to use the corpus as a source of data.

The Acquiring process started from translating the senses of the Ambiguous words to their equivalent English words using Amharic-English Dictionary. Then using the translated English word sense example sentences containing the word is acquired from the English corpus .For example the Amharic ambiguous word "**መሳል**" has three senses that are *"cough", "vow", and "sharp"*. Using these three senses, sense example

sentences are acquired. The English sentences were examined thoroughly to check that it correctly represents the right sense of the Amharic word. For instance the Amharic word "ቀረጸ" has two senses "record" and "shape". But the English word "record" is ambiguous by itself. It has eight senses in English WordNet. But only sentences that have "record" senses that match to the word "ቀረጸ" are selected.

Agirre & Martinez [2] have reported that accuracy of classifiers degrade significantly when the training and testing samples have different distributions for the senses. In this study we tried to use a balanced distribution of senses for the ambiguous words to maximize performance when enough sense examples are available. On average, about 100 example sentences were acquired for each sense of ambiguous words with the exception of two senses on which enough example senses were not acquired from the corpus. The distributions of senses are summarized in table 4.1.

Ambiguous Word	Sense	Count	
eTena	strengthen	100	
	study	100	200
mesal	cough	100	
	sharp	72	245
	vow	73	
me`sa`sat	taking care	100	
	thin	100	200
metrat	call	100	
	clean	100	200
qereSe	record	100	
	shape	100	200

Table 4.1 Distribution senses of Ambiguous words

4.2. System Architecture

The architecture of the system is depicted in Figure 4.1.The system takes sentences that contain the ambiguous words as an input .The sentences will be preprocessed to make them suitable for further processing. Then the classification algorithm (Naive Bayes) builds classifier from the training set and applies the built classifier to test new instances and displays performance evaluation of the classifier. The detailed explanation of the processes is given in the next subsections.

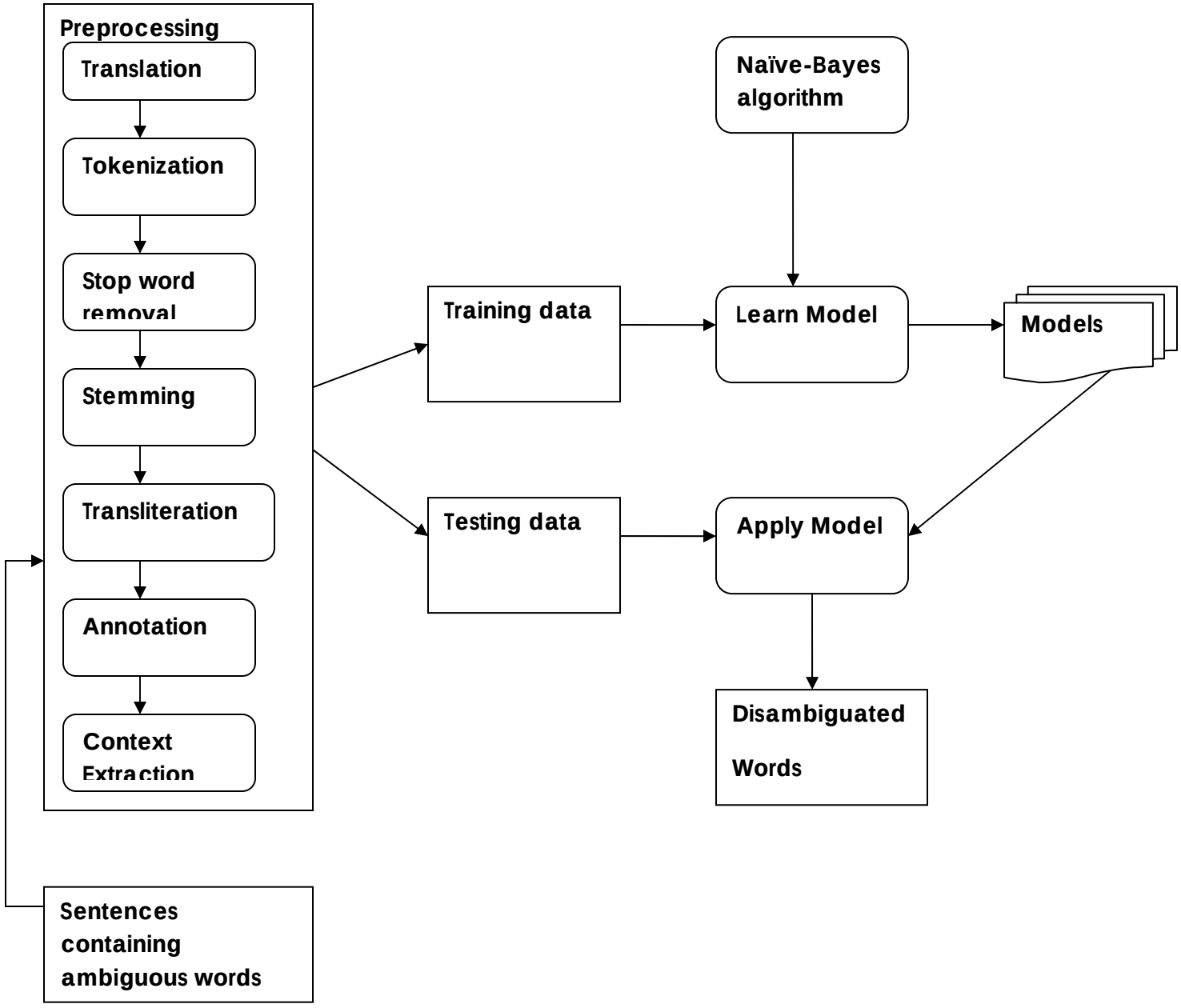


Figure 4.1 Architecture of WSD system for Amharic

4.3. Preprocessing

4.3.1. Translation

Once the English sense examples are acquired for the selected Amharic ambiguous words, the next step is to translate the sense examples to Amharic. Most researches that use second language corpora use either machine translation or machine readable bilingual dictionaries to translate the senses to the original language. However the researcher was not able to access to both resources for Amharic. Therefore a manual translation of the sense examples using English-Amharic dictionary is done by the researcher. The translation of the sense examples is carried out in such a way that the translated senses will have a clear meaning in Amharic. Except names of people, places, materials and words that don't have equivalent Amharic translation, all words are translated to Amharic.

The following is an example of translating the "cough" sense of "**መሳል**":

*The commonest symptom for the disease is **coughing** persistently, with frequent chest infection. (English sense)*

የበሽታው የተለመደ ምልክት ያለማቋረጥ መሳል ከሚደጋገም የደረት ህመም ጋር ነው።

(Translated Amharic Sense).

The word "**መሳል**" has the senses of *"vow", "sharp", and "cough"* but in the above sentence it refers to only the *"cough"* sense.

4.3.2. Tokenization

Tokenization refers to the process of splitting stream of characters in to raw terms or tokens. This process detects the boundaries of a written text. Tokenizing of a given text depends on the characteristics of language of the text which it is written. The Amharic language has its own punctuation marks which demarcate words in a stream of characters which includes 'hulet neTb' (:), 'arat neTb '(: :), 'derib sereze' (፤), 'netela serez'(፣) , exclamation mark '!' and question mark'?'. These punctuation marks don't have any relevance to identify the meaning of ambigous words using WSD. Therefore except 'arat neTb ' and ' question mark' which are used to detect the end of the sentence, all other punctuations are detached from words in tokenization process.

4.3.3. Stop Word Removal

Like other languages, Amharic has non content bearing words which are called stop words. Usually words such as article (e.g. '**ያ**ኛ**ው**', '**ይ**ሄ'), conjunctions ("'**እ**ና', '**ነ**ገርግን', '**ወ**ይም') and prepositions (e.g. **ውስጥ,ላይ**).Since stop words do not have a significant discriminating power in the meaning of ambiguous words , we filtered the sense examples with a stop-word list, to ensure only content words are included. In addition to stop words, names of people and places are also filtered from the sense examples as they are not related to the meaning of words.

4.3.4. Stemming

The Amharic language makes use of prefixing, suffixing and infixing to create inflectional and derivational word form. A stemmer is a system that tries to reduce various forms of a word to a single stem. In morphologically complex language like Amharic, a stemmer will lead to significant improvements in WSD systems [36].

For this study, automatic removal of suffix and prefixes is done using adopted algorithm[1]. And infixes are removed manually. The algorithm is applied in our corpus and worked well with about 12% error. After the semi-automatic removal, a manual inspection of the corpus is carried out in a way to correct a few errors in exceptional cases. For example, Amharic uses a suffix '**ኘ**' and in the algorithm this suffix is removed after checking whether the word is not in exception list that contains words that has '**ኘ**' at the end but which is not a suffix. And if it is in exception list the algorithm will not remove the suffix since it is part of the word. However there might be some words that contains '**ኘ**' at the end and not included in the exception list. In such cases the algorithm removes '**ኘ**' from the word which is not a suffix. Such errors are manually inspected and corrected. The suffix and affix removal algorithm is depicted in figure 4.2.

[1] The algorithm is developed by our research group members Alemu Kumilachew and Zeleke Abebaw.

1. open corpus, exception list and Normalization list files

2. While not end of corpus file is reached do

Read tokens

for each token in token list

If token starts with prefix

If token not in exceptional list then

Remove prefix

End if

else If token ends with suffix

If token not in exceptional list then

Remove suffix

End if

If token ends with sadis[2] alphabet

Normalize[3]

End if

End if

End for

3. end while
4. Write the list of tokens to corpus file
5. Close files

Figure 4.2 Prefix and Suffix removal algorithm

---

[2] Sadis refers to the sixth order in Amharic Alphabet.

[3] Normalization is used to correct a variant of a word to its stem after suffix is removed for some words (e.g. for a word "ሰዎች" "ች" will be removed as affix and "ሰዎ" will be normalized to "ሰው" which is the stem).

4.3.5. Annotation

Annotation is the process of tagging each training example with the label of its corresponding class. For this study different senses of ambiguous words are taken as a class and each training sentence is labeled manually with a class corresponding to its senses. Table 4.2 shows the list of classes for the selected ambiguous words:

Ambiguous Word	**Classes**		
	Class1	**Class2**	**Class3**
eTena	strengthen	study	-
mesal	cough	sharp	vow
me`sa`sat	taking care	thin	-
metrat	call	clean	-
qereSe	record	shape	-

Table 4.2 Classes of selected ambiguous words

The following example illustrates the annotation process:

Raw sentence

የበሽታው የተለመደ ምልክት ያለማቋረጥ <u>መሳል</u> ከሚያጋጥም የደረት ህመም ጋር ነው።።

In the example the ambiguous word "**መሳል**" can belong to classes cough, vow and sharp of mesal. But in the example the word has a sense of cough. So it will be annotated as follows:

የበሽታው የተለመደ ምልክት ያለማቋረጥ መሳል < cough> ከሚደጋገም የደረት ህመም ጋር ነው።።

4.3.6. Context Extraction

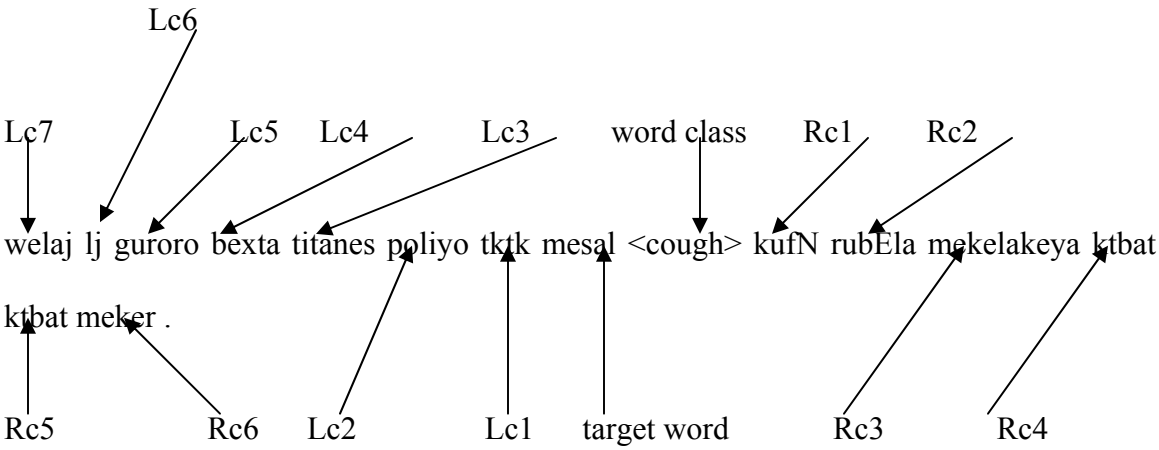
Context in WSD refers to the words surrounding the ambiguous words which are used to decide the meaning of the ambiguous word. For instance, for the sentence: "**የበሽታው የተለመደ ምልክት ያለማቋረጥ መሳል ከደረት ህመም ጋር ነው።።**" After stop word removal and stemming it will be "**ተለመደ ምልክት ማቋረጥ መሳል ደረት ህመም**".The contexts are the word surrounding the ambiguous word "**መሳል**" which are {**ተለመደ,ምልክት,ማቋረጥ,ደረት,ህመም**}. In this study the contexts of the ambiguous words are extracted using the algorithm in Fig 4.3 which is adopted from [24] and customized to fit to this study.

```
1. initialize array buffers of strings
2. j=1
3. open corpus file
4. while not end of file is reached do
        read sentence j

        i=1

        while end of sentence marker is not reached do

              read word i from sentence j

              if  word i is the target word

                assign word i in to an array buffer

                assign the meaning(label) of word i to an array buffer

                 if the n previous and following words from word i are within the sentence

                    read the n previous words  from the target word (word i) and assign
                    them to array buffer

                    read the n following words  from the target word (word i) and assign to
                    array  buffer

                 if not

                  assign empty value  to  array buffer

                 end inner while

            if not

            increment i by one

        end inner while

       increment j by one

5. end outer while
6. write the content of the array to file
7. close file
```

Figure 4.3 Context Extraction algorithm

4.4. Training and Testing Datasets

Once all the necessary preprocessing tasks are done on the corpus, 90% of the data is allocated for training and the remaining 10% is allocated for testing. In table 4.3 the description of attributes in the data set is presented. In the table Rcontext(i) and Lcontext(i) refers to the words that surrounds the ambiguous word to the right and left respectively where i ϵ {1, 2, …, 10}, the target word holds the ambiguous word and Word class takes the senses of the ambiguous word. If the ith left or right word from the target word doesn't exist, an empty value will be assigned to mean that there is no context. We have found that, the longest sentences in the corpus constitute a maximum of ten words to the left and the right of the ambiguous word. So we used 10 words to the left and the right of the ambiguous word as possible contexts. It can be further explained using the following example which is extracted from the corpus.



In the example the target word is "mesal" which is the ambiguous word and its word class is "cough" that is its sense in this context. *Lc* refers to the left context where as *Rc* refers to right context. There are seven left contexts and six right contexts surrounding

the target word which are labeled as is shown in the example. But there are no three right contexts (8,9,and 10) and there are no four left contexts (7,8,9,10) which will be assigned as empty.

NO.	Attribute	Description	value
1	Lcontext(i)	used to hold the ith left word from the ambiguous word	Any word in the corpus
2	Rcontexti(i)	used to hold the ith right word from the ambiguous word	Any word in the corpus
3	Target word	Holds the ambiguous word	Ambiguous words
4	Word class	Holds the label of the target word	Different senses of the ambiguous word

Table 4.3. Description of attributes used for this study

4.5 Summary

In this chapter, the design of the WSD system for Amharic is presented and discussed. Using the design, the process of corpus preparation of training and testing set for experimentation is illustrated. The next chapter deals with the experimentation and discussion on the results of the experiment.

CHAPTER FIVE

EXPERIMENTATION AND DISCUSSION

5.1. Introduction

As discussed in section 2.1.1 supervised word sense disambiguation is selected for this study. In the classification paradigm of supervised machine learning, a classification procedure is induced from a set of data for which the true classes are known, for a set of pre-defined classes. For WSD, learning such classification procedure requires the availability of sense-tagged data, where each training example is described by a feature vector and a corresponding class label. The feature vector comprises of attribute-value pairs, where the attributes are those contextual clues important for classification. The supervised learning task, as discussed in the preceding section, thus involves capturing important dependencies in the training data and representing these in a parametric model, from where the joint probability distribution can be defined. Once all the required model parameters have been estimated, the learned model can then be used as a classifier for WSD i.e. given a particular instantiation of the feature variables for a test sentence, the classifier predicts the value of the classification variable. It is the expectation that the learned classifier should perform well in classifying test examples, and its prediction accuracy is used to measure how well it has been able to generalize from the training data to unseen data.

For this study five classifiers namely Mesal, Qeretse, Atena,Metrat and Mesasat will be trained for each ambiguous words with their corresponding data sets that are defined in Chaper four of this thesis. In this chapter the experimental procedures with the analysis of the experiment results will be presented.

5.2. Experimentation Procedure

In this study a total of five experiments are conducted using Naive Bayes classifier of Weka 3.62 Package. The first experiment was conducted to check to what extent stemming of Amharic words in the corpus will affect the accuracy of the WSD classifiers. Then an experiment is conducted to find out the best training and testing split options that performs better for this study and 10-fold cross validation and 66% split test options, which are defaults in Weka package, are investigated. The third experiment was carried out to compare the accuracy of 10-fold cross validation with different percentile split option other than the default 66% split test option. The study also seeks to investigate the effect of different context sizes on disambiguation accuracy for Amharic, and to find out if the standard two-word window applicable for other languages and especially English [35] holds for Amharic. In this regard, different training data sets where the contextual information is obtained from 1-left and 1-right to 10-left and 10-right consequent surrounding words are prepared for each classifier. Finally experimentation is conducted to see the effect of sense distribution on the performance of the classifiers.

5.3. Discussion of Results

We now present and discuss the experimentation outputs for five of the experiments that are mentioned earlier.

Experiment I: The effect of stemming on the accuracy of the classifiers

As discussed earlier, stemming has been found a significant improvement on performance of WSD classifiers for morphologically complex languages. This experiment is performed to test whether this applies to WSD for Amharic. Both the default 10-fold cross validation(CV) and 66% split test options are used to test the experiment. In 10-fold CV the data is iteratively divided into training (90%) and testing (10%). Then the classifier is evaluated 10 times and the average result is reported. Whereas in 66% split test option as name indicates, it allocates 66% of the data for training and the rest 34% for validating the model.

The result of this experiment is presented as follows:

Classifiers	Accuracy	
	Before stemming	**After stemming**
eTena	65.3%	**66.8%**
mesal	70.2%	**78%**
Me`sa`sat	72.1%	**73.8%**
metrat	71.2 %	**73 %**
qereSe	69.8 %	**69.9%**

Table 5.1 The effect of stemming on accuracy using 10-fold CV test split option

Classifiers	Accuracy	
	Before stemming	**After stemming**
eTena	59.7%	**61.8%**
mesal	55.6%	**70.7%**
Me`sa`sat	71.1%	**72.7%**
metrat	69.7 %	**70.6 %**
qereSe	66.2 %	**69.7%**

Table 5.2 The effect of stemming on accuracy using 66% test split option

As is shown in table 5.1 and table 5.2, for all words, stemming improved the accuracy of all the five classifiers in both 10-fold CV and 66% test split options. The reason behind the enhanced accuracy might be that stemming brings variants of a word into their common stem which will minimize the consideration of the variants of a word as

different word by WSD model. As it is stated earlier, WSD models determine the meaning of a word by learning the pattern of surrounding words. So if stemming is done the variants of a word is taken as the same pattern which will improve the accuracy of the classifiers. For example, before stemming, a surrounding words "ሰዎቹ" and "ሰዎች" will be assumed as different words but basically they are the variants of the same word " ሰው".But after stemming these words will be taken as the same pattern that will increase the accuracy of the classifiers. Therefore in subsequent experiments the dataset that is stemmed will be used as it enhances the performance of the models.

Experiment II: Determining the appropriate test option

It is argued by [62] that 10-fold cross validation (CV) test split option is appropriate for small datasets. The dataset used in this study has on average 210 instances per classifier which is a small data set. In order to test whether 10-fold CV outperforms the default 66% test split option in Weka, an experiment is conducted ten times for each classifier and an average accuracy is taken.

Classifiers	**Accuracy**	
	10-fold CV	**66% split**
eTena	**66.8%**	61.8%
mesal	**78%**	70.7 %
Me`sa`sat	**73.8%**	72.7%
metrat	**73 %**	70.6%
qereSe	**69.9**	69.7%

Table 5.3 Comparison of 10-fold CV and 66% split test options

As presented in Table 5.3, 10-fold CV performs better than 66% split in all of the five classifiers. This is due to the appropriateness of 10- fold CV for small dataset like we used in this experiment. It is because it performs the training 10 times increasing the training dataset by ten fold.

Experiment III: Comparing 10-fold CV test option with other test split options

In experiment II it has been found that 10-fold CV works better relative to 66% test split option for dataset used in this study. A further experiment is conducted ten times and an average accuracy is calculated to compare 10-fold CV with other test split option. The test split options are randomly taken with five ranges which are 70%, 75%, 80%, 85%, and 90%.The percents refers to the portion of the data that is allocated for training and the rest is for testing.

Classifiers	**Accuracy**					
	70% split	**75% split**	**80% split**	**85% split**	**90% split**	**10-fold CV**
eTena	60 %	62 %	60 %	56.7 %	60 %	**66.8%**
mesal	68.1%	68.3%	70.8%	75 %	75 %	**78%**
Me`sa`sat	67.2%	69.4%	71.8%	72.9 %	68.4 %	**73.8%**
metrat	70 %	72 %	75 %	**76.7 %**	70 %	73 %
qereSe	70.7%	73.1%	**74.4%**	72.4 %	73.7 %	69.9%

Table 5.4 Comparison of 10-fold CV and different percentile split test options

As it is indicated in table 5.4 10-fold CV performs better in three of five classifiers (Mesal,atena,and mesasate).Where as for classifiers "queretse" and "metrat" 80% split and 85% split options achieved better result than 10-fold CV. But still 10-fold CV performs better for the majority of the classifiers (three out of five).So for the following experiments 10-fold CV will be used as it achieved a better result relative to other test split options.

Experiment IV: Determining optimal context window

In other languages an optimal context window size which refers to the number of surrounding words that is sufficient for extracting useful disambiguation is obtained through research. For example in English a standard two-word window on either side of the ambiguous word is found to be enough for disambiguation [35].But, this hasn't been established Amharic. For this study an experiment is carried out ten times for each

classifier to determine an average optimal window size from one-one window to ten-ten window on other side of the ambiguous word.

Window size	Accuracy				
	eTena	**mesal**	**me`sa`sat**	**metrat**	**qereSe**
1-1	**76.4%**	77.1 %	76.9 %	**81.5 %**	72.%
2-2	70.4 %	82.6 %	81.3	76.5 %	73.6%
3-3	70.1%	**83.2%**	**81.5%**	75 %	**76.2 %**
4-4	67.8%	82.4 %	77.4%	73 %	74.6%
5-5	68.3 %	80.9 %	75.4 %	73 %	75.7%
6-6	66.8%	81.3 %	74.4 %	74 %	74.6 %
7-7	66.3%	80.9%	74.6 %	72 %	73.1 %
8-8	66.8 %	80.5 %	72.8 %	72.5%	72. %
9-9	66.8%	78.4%	73.9 %	73 %	71.1 %
10-10	66.8%	78.%	73.8	73 %	69.9%

Table 5.5 Summary of experiment in different window sizes

As is shown in the table 5.5 for the three of the classifiers (mesal, qereSe, and me`sa`sat) the maximum accuracy is achieved on three-three word window. Where as for *"Metrat"* and ***"eTena*** " the highest accuracy is attained on one-one window .The result agrees with the findings in other language that the nearest words surrounding the ambiguous words give more disambiguation information than words far from the ambiguous word[35].For this study, since in all classifiers, the accuracy of windows after 3-3

window are less than that of a 3-3 window. Window size of 3 is considered to be effective.

Using a 3-3 window size the final accuracy of the classifiers can be summarized in table 5.6 :

Classifiers	**Accuracy (3-3 Window)**
eTena	70.1%
mesal	83.2%
Me`sa`sat	81.5%
metrat	75 %
qereSe	76.2 %

Table 5.6 Summary of Accuracy of classifiers using 3-3 window size

As indicated in table 5.6 the accuracy of the classifiers is achieved within the range of 70 to 83%.Mihalcea et al.[48] have shown that supervised WSD methods can yield up to 72.9% accuracy on words for which manually sense-tagged data are available. So the accuracy of this experiment can be considered very promising.

Experiment V: Effect of Distribution of training data on accuracy

WSD performance can be affected by the distribution of training data for each sense. In this study a balanced distribution of training data has been employed to maximize performance. But naturally the sense distribution might vary for each sense. To test the

effect of distribution of senses on accuracy an experiment using three-three window size has been conducted ten times and average accuracy is taken. We intentionally vary the distribution of senses and compare the result with balanced distribution of senses.

Classifiers	**Sense**	**Balanced Sense Distribution**		**Accuracy**	**unbalanced Sense Distribution**		**Accuracy**
eTena	strengthen	100		**70.1%**	100		69.5%
	study	100	200		40	140	
mesal	cough	100		**84.7%**	100		68.9 %
	sharp	72	245		40	160	
	vow	73			20		
Me`sa`sat	Taking care	100		**81.5%**	40		70.8%
	thin	100	200		100	140	
metrat	call	100		**75 %**	100	140	74.1%
	clean	100	200		40		
qereSe	record	100		**76.2 %**	100	140	74.7%
	shape	100	200		40		

Table 5.7 Effect of sense distribution on accuracy

As presented in table 5.7, the accuracy of unbalanced sense distribution resulted in less accuracy for all the five classifiers. The finding supports the findings in other studies that the accuracy of classifiers degrade significantly when the training sample have different

distributions for the senses as there will be bias to the high number of sense distribution [2].

5.4. Summary

In this chapter the experimentation procedure together with presentation and discussion of five experiments are covered. In the first experiment it has been shown that stemming significantly improved the accuracy of the classifiers. In successive experiment 10-fold cross validation test split option has been found to perform better relative to other percentile test split options. Using 10-fold CV an experiment was also conducted to determine optimal window size for the classifiers and three-three window size has been found as most favorable window size. Using three-three window, the final accuracy of the classifiers has attained within range of 70% to 83% which is a very encouraging in supervised WSD. Finally, experimentation has been carried out to test the effect of sense distribution on the accuracy of the models and it has been found that balanced sense distribution can give better result in increased accuracy than unbalanced sense distribution.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

The overall focus of this research is Word Sense Disambiguation (WSD) which addresses the problem of automatically deciding the correct sense of an ambiguous word based on its surrounding context's. WSD is essential tool for NLP and IR applications .WSD is considered to be one of the most challenging of all NLP research areas due to its reliance on a varied range of linguistic and statistical knowledge.

The problem of WSD is addressed for Amharic which is one of less studied language. Though Amharic has many ambiguous words due to knowledge acquisition bottleneck, five ambiguous words are selected and a classifier for each ambiguous word has been built. The words are eTena, mesal ,me`sa`sat, metrat , and qereSe.

The most popular approaches to WSD rely on supervised machine learning methods, where a machine learning classifier is required to be trained on manually labeled training instances, to generate a classifier model that can be used to classify future instances. In this study, supervised machine learning approach using Naive Bayes algorithm is used. These methods, however, face the problem of knowledge acquisition bottleneck, where the amount of labeled data provided to the classifiers is limited. For this study, a monolingual corpora of English language has been used to acquire sense examples and

the sense examples are translated back to Amharic which is one approach of tackling knowledge acquisition bottleneck.

Based on our Naive Bayes algorithm experiments on Weka 3.62 package, we conclude that Naive Bayes methods achieve higher accuracy on the task of WSD for selected ambiguous word, provided that the quality of the labeled data is good. We have achieved accuracy within the range of 70% to 83% for the five classifiers which is an impressive accuracy for supervised WSD.

We have also found that stemming of the Amharic words in the corpus can enhance the accuracy of the classifiers as Amharic is a morphologically complex language. The accuracy is increased after stemming is applied to words in the corpus.

For Amharic, there is no standard optimal context window size which refers to the number of surrounding words that is sufficient for extracting useful disambiguation. Based on our experiment we have found that three-word window on other side of the ambiguous word is enough for disambiguation for the five classifiers.

We also found that the sense distribution of an ambiguous word is a crucial factor in improving the accuracy using Naive Bayes methods. The best accuracy in our experiment is seen for words that have a balanced sense distribution.

In total, the chosen methodology which is supervised word sense disambiguation has been justified in terms of its theoretical foundations as well as the results obtained in our experiments for selected Amharic Ambiguous words.

6.2. Recommendations

Researches in Word Sense Disambiguation require a variety of linguistic resources like thesaurus, WordNet, machine readable dictionaries, and machine translation software in which we faced a significant challenge as Amharic lacks those resources. The other challenge was lack of sense annotated data for the language which makes the study to be limited for five ambiguous words. In this study we have only experimented with supervised machine learning (Naive Bayes) approach but there are other approaches which performed well for WSD in other language. Therefore; we have the following recommendations which include the development of resources and future research directions for WSD for Amharic text:

1. Researches in WSD for other language use linguistic resources like Thesaurus, Lexicon like WordNet, machine readable dictionaries and machine translation software. For Amharic those resources are not available. Taking into account their contribution to WSD and other researches concerned institutions should develop these resources.

2. For other language a standard sense annotated data are available for WSD research and also for testing a WSD systems. We don't have such data for Amharic. So there need to be an initiative to prepare the data for WSD research.

3. Future research directions for WSD in Amharic include:

i. Extending this experimentation using Supervised WSD for other ambiguous words in addition to those covered in this research.

ii. Due to time limitation in this study only Naive Bayes algorithm has been used. But other algorithms like decision tree, decision lists and support vector machines are used and found achieve impressive results for other languages .These algorithms should be experimented for WSD in Amharic.

iii. As supervised WSD requires manually labeled sense examples which is time taking. A research should be conducted using unsupervised WSD and bootstrapping approach for Amharic.

iv. In addition to corpus based approach there are also knowledge based and hybrid approach which are used for WSD for other language. These approaches need to be investigated for Amharic as well.

REFERENCES

[1] Abiyot, B. (2000). Developing Automatic Word Parser for Amharic Verbs and Their Derivation. Master Thesis, Addis Ababa University.

[2] Agirre, E. and Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web, in 'Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content'.

[3] Atelach, A.(2000).Automatic Sentence Parsing for Amharic Text, An Experiment Using Probabilistic Context Free Grammars, Masters Thesis, Addis Ababa University.

[4] Atelach, A., Askar, L., and Mesfin, G. (2003). Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward , in Proceedings of TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages, Batz-sur-Mer, France.

[5] Atelach A., Askar, L, Richard, C., and Jussi, K.(2004) ,Dictionary based Amharic-English Information Retrieval Proceedings of the third Workshop of the Cross-Language Evaluation (CLEF), Bath, England.

[6] Baye, Y. (1987 E.C). **የአማርኛ ሰዋሰው. አዲስ አበባ. ት.መ.ማ.ማ.ድ።።**

[7] Beletu, R (1982). A Graphemic Analysis of the Writing System of Amharic. Master Thesis, Addis Ababa University.

[8] Bj¨orn, G., Fredrik, O., Atelach A., and Asker, L.(2009), Methods for Amharic Part-of-Speech Tagging, Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages, Athens, Greece.

[9] Black, E.(1998). An experiment in computational discrimination of English word senses. IBM J. Research Development. Vol 32(2).

[10] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L.(1991). Word-sense disambiguation using statistical methods. In Proceedings of the29th Annual Meeting of the ACL.

[11] Clara, C., Philip, Resnik., and Jessica, S.(2001). Supervised Sense Tagging using Support Vector Machines. In Proceedings of SENSEVAL-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems. SIGLEX, Association for Computational Linguistics.

[12] Clara, Allan M., Lotofus., and Elizabeth, F.(1975). A spreading-activation theory of semantic processing, Psychological Review. Vol 82(6).

[13] Cortes, C. and Vapnik, V. (1995). Support-vector network. Machine Learning,vol 20(1).

[14] D. Jurasky and J. H. Martin (2000). Speech and Language Understanding. Upper Saddle River: Prentice-Hall, Inc..

[15] D.Yarowsky(1995).Unsupervised word sense disambiguation rivaling supervised methods, presented at Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge,MA.

[16] Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus, Computational Linguistics Vol 20(4).

[17] Daniel, A(2000). An integrated approach to automatic complex sentence parsing for Amharic text, Masters Thesis, Addis Ababa University.

[18] Dawkins, C.H. (1969). The Fundamentals of Amharic. A.A Sudan interior mission.

[19] Doina, T.(2004). Word sense disambiguation by machine learning approach: a short survey .Informatica, Vol XLIX(2).

[20] Escudero, G., M`arquez, l., and Rigau, G.(2000). On the portability and tuning of supervised word sense disambiguation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora Hong Kong, China.

[21] Ethiopian Central Statistical Authority (ECSA)(1998). The 1994 Population and Housing Census of Ethiopia: Results at Country Level. Vol. 1 Statistical Report48. Addis Ababa.

[22] Ezra, B. (1998). An Experiment in Computational Discrimination of English Word Senses. IBM Journal of Research and Development, Vol 32(2).

[23] Gale, W., Church, K. and Yarowsky, D. (1992).Estimating upper and lower bounds on the performance of word sense disambiguation programs, in 'Proceedings of the 30th Conference of the Association for Computational Linguistics', Newark, Delaware.

[24] Gebeyehu, K. (2009). the application of decision tree for part of speech (pos) tagging for amharic . Master Thesis, Addis Ababa University.

[25] Getahun A.(2001).Towards the Analysis of Ambiguity in Amharic,JES Vol XXXIV(2).

[26] Girma G.(2007). **በዐማርኛ ሥርዐተ-ጽሕፈት ውስጥ የድምፀ-ሞክሼ ሆሄያት አጠቃቀም ማስታወሻ**.Retrieved on 10 April, 2010 from: http://www.nlp.amharic.org/resources/lexical/word-lists/homonyms/homonym-collected-by-girma-getahun/

[27] Hudson, G.(1999). Linguistic analysis of the 1994 Ethiopian census. Northeast African Studies, Vol 6.

[28] Hearst and Marti A(1999).Noun homograph disambiguation using local context in large corpora, presented at Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research, Oxford, United Kingdom,.

[29] Hutchins, J. and Sommers, H(1992). Introduction to Machine Translation. Academic Press.

[30] Hwee, N.(1997) .Getting serious about word sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? .Washington D.C.

[31] Ide, N. and Véronis, J.(1990). Mapping dictionaries: A spreading activation approach, Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary, Waterloo.

[32] Ide, N. and Veronis, J.(1998), Word Sense Disambiguation: The State of the Art,Computational Linguistics.

[33] J. R. Anderson(1976). Language, Memory, and Thought. Hillsdale, NJ.

[34] J. R. Anderson(1983). A Spreading Activation Theory of Memory," Journal of Verbal Learning and Verbal Behavior, Vol 22.

[35] Kaplan, A. (1955). An experimental study of ambiguity and context, Mechanical Translation Vol 2(2).

[36] Karov, Y. and Shimon, E.(1998). Similarity-based word sense disambiguation. Computational Linguistics, Vol 24(1).

[37] Kelly, E. and Stone, P. (1975) Computer Recognition of English Word Senses. Vol. 3 of North Holland Linguistics Series. Elsevier, Amsterdam, The Netherlands.

[38] Lars A.,Atelach A., Bj¨orn ,G. ,Samuel, E. ,Asfeha, L., Nigussie, Habte.(2009). Classifying Amharic Webnews. Information Retrieval archive.Vol 12(3).

[39] Lesk, M. (1986), Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, in ‘Proceedings of the SIGDOC Conference’, Toronto, Ontario.

[40] Lin, D. and Pantel, P. (2002). Discovering word senses from text. In Proceedings of the 8th ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining (Edmonton, Alta., Canada).

[41] Masterman, M. and M. Masterman(1961). Semantic message detection for machine translation using interlingua," presented at International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London.

[42] Michael, L(1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, presented at Proceedings of the Fifth International Conference on Systems Documentation, Toronto, CA,.

[43] Michiels, A. (1982).Exploiting a large dictionary data base. Liege, Belgique: Universite de Liege, .

[44] Nega, A. and Peter, W. (2002). Stemming of Amharic Words for Information Retrieval Literacy and Linguistic computing, Vol 17(1).

[45] Nello C., John S., and Huma L. (2001). Latent semantic kernels, In Proceedings of 18th International Conference on Machine Learning .

[46] Pasca, M. and Harabagiu, S(2001). The informative role of WordNet in Open-Domain Question Answering. In Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources.Pittsburgh, PA.

[47] Patrick, P. and Dekang, L.(2002). Discovering Word Senses from Text, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada .

[48] Rada, M., Timothy C., and Adam K.(2004.) The Senseval-3 English lexical sample task. In Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3).

[49] Raymond, M.(1996). Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

[50] Resnik, Philip and David, Y. (2000). Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. Natural Language Engineering, Vol 5(2).

[51] Richard, O. Duda, Peter, E., Hart(1973). Pattern Classification and Scene Analysis. New York, NY: John Wiley and Sons.

[52] Robert, A.(1980). The structure of the Merriam-Webster Pocket Dictionary. Austin, TX: University of Texas at Austin.

[53] Roberto, N. (2009). Word Sense Disambiguation: A Survey .ACM Computing Surveys, Vol 41(2).

[54] Satanjeev, B. and Ted P.(2003). The Design, Implementation and Use of the Ngram Statistics Package. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics.

[55] Smeaton, A.F(1995). Linguistic Approaches to Text Management: An Appraisal of Progress. Journal Of Document & Text Management, Vol 2(2).

[56] T.Pedersen(1988).Learning Probabilistic Models of Word Sense Disambiguation, in School of Engineering and Applied Science: Southern Methodist University, .

[57] Teshome, K.(1999). Word Sense disambiguation for amharic text retrieval: a case study for legal documents. Master Thesis, Addis Ababa University.

[58] Voorhees, E. M.(1998). Using WordNet for text retrieval. In WordNet: An Electronic Lexical Database.MIT Press.

[59] William, G., Kenneth, Church, and David, Y.(1992). A Method for Disambiguating WordSenses in a Large Corpus. Computers and the Humanities, Vol 26(2)..

[60] W. A. Gale, K. W. Church, and D. Yarowsky(1993), A method for disambiguating word senses in a large corpus, Computers and the Humanities,Vol 26.

[61] Wilks, Yorick and Stevenson, Mark(1996). The grammar of sense: Is word sense tagging much more than part-of-speech tagging? Technical Report , University of Sheffield, Sheffield, United Kingdom,

[62] Witten, I. and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

[63] Wube, A.(2004). Rule Based Syntactic Disambiguation Parser for Amharic Sentence, Masters Thesis, Addis Ababa University.

[64] Xinglong, W. and John, C. (2005). Word Sense Disambiguation Using Automatically Translated Sense Examples .Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Association for Computational Linguistics

[65] Yacob, D. (1996). System for Ethiopic Representation in ASCII (SERA). Accessed on 12 March, 2010, from: http://www.abyssiniacybergateway.net/fidel/

[66] Yarowsky, D. (1992).Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, in 'Proceedings of the Fourteenth International Conference on Computational Linguistics', Nantes, France.

[67] Yarowsky, D. (1995).unsupervised word sense disambiguation rivaling supervised methods, in 'Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics', Cambridge, M.A.

[68] Yarowsky, D(2000). Hierarchical decision lists for word sense disambiguation. Comput. Human. Vol 34(1-2).

[69] Yehenew, S.(2004).Design and Development of Human-aided Rule-based English Amharic Machine translation prototype, Masters Thesis, Addis Ababa University.

[70] Yoong, K., Hwee , T., and Tee, K. (2004). Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. In Proceedings ofSENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.

[71] Yoseph S.(2004).Application of multilingual Thesauri for cross language information retrieval(CLIR)[Amharic –English CLIR for the legal Environment], Masters Thesis, Addis Ababa University.

APPENDIX A. The Amharic alphabet ('fidel') adopted from Dawkins [18] and Yacob [65].

	Ordinary characters							**Diphthong (‘diqala’) characters**				
1	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
	he	hu	hi	ha	hE	h	ho					
2	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ		
	le	lu	li	la	lE	l	lo			lWa		
3	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ			ሗ		
	He	Hu	Hi	Ha	HE	H	Ho			HWa		
4	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ		
	me	mu	mi	ma	mE	m	mo			mWa		
5	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ			ሧ		
	`se	`su	`si	`sa	`sE	`s	`so			`sWa		
6	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ		
	re	ru	ri	ra	rE	r	ro			rWa		
7	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ		
	se	su	si	sa	sE	s	so			sWa		
8	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ		
	xe	xu	xi	xa	xE	x	xo			xWa		
9	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቍ	ቋ	ቌ	ቍ
	qe	qu	qi	qa	qE	q	qo	qWe	qWu	qWa	qWE	qWi
10	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ		
	be	bu	bi	ba	bE	b	bo			bWa		
11	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ			ቯ		
	ve	vu	vi	va	vE	v	vo			vWa		
12	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ		

	te	tu	ti	ta	tE	t	to			tWa		
13	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ		
	ce	cu	ci	ca	cE	c	co			cWa		
14	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኊ	ኋ	ኌ	ኍ
	`he	`hu	`hi	`ha	`hE	`h	`ho	hWe	hWu	hWa	hWE	hWi
15	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ		
	ne	nu	ni	na	nE	n	no			nWa		
16	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ		
	Ne	Nu	Ni	Na	NE	N	No			NWa		
17	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ			ኧ		
	e	u	i	a	E	I	o			ea		
18	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኵ	ኳ	ኴ	ኸ
	ke	ku	ki	ka	kE	k	ko	kWe	kWi	kWa	kWE	kWu
19	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ					
	`ke	`ku	`ki	`ka	`kE	`k	`ko					
20	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
	we	wu	wi	wa	wE	w	wo					
21	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ					
	`e	`u	`i	`a	`E	`I	`o					
22	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ		
	ze	zu	zi	za	zE	z	zo			zWa		
23	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ			ዧ		
	Ze	Zu	Zi	Za	ZE	Z	Zo			ZWa		
24	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
	ye	yu	yi	ya	yE	y	yo					

25	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ		
	de	du	di	da	dE	d	do			dWa		
26	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ			ጇ		
	je	ju	ji	ja	jE	j	jo			jWa		
27	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ገሞ	ገሉ	ገሎ	ጓ	ጔ
	ge	gu	gi	ga	gE	g	go	gWe	gWu	gWi	gWa	gWE
28	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ				ጧ	
	Te	Tu	Ti	Ta	TE	T	To				TWa	
29	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ				ፏ	
	Ce	Cu	Ci	Ca	CE	C	Co				CWa	
30	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ					
	Pe	Pu	Pi	Pa	PE	P	Po					
31	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ				ጿ	
	Se	Su	Si	Sa	SE	S	So				SWa	
32	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ					
	`Se	`Su	`Si	`Sa	`SE	`S	`So					
33	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ				ፏ	
	fe	fu	fi	fa	fE	f	fo				fWa	
34	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ				ፗ	
	pe	pu	pi	pa	pE	p	po				pWa	

APPENDIX B. Selected ambiguous words and their Amharic meaning adopted from Girma (26).

ድምፀ-ሞክሼ ቃላት	ፍቺዎች
መሣሣት መሣሣት	መቅጠን፤ ዘርዛራ ወይም ሥሥ መኾን በሥሥት ወይም በጥንቃቄ መያዝ
መሳል መሳል መሳል	መለመን፤መጠየቅ ኡህ ኡህ ማለት መቀራረጫ ጠረዝን ማትባት
መጥራት መጥራት	ጥሪ ማረግ መጽዳት፤ ከጉድፍ፣ ወዘተ. መላቀቅ
ቀረጸ ቀረጸ	ድምጽን ቀዳ ምስል አወጣ
አጠና አጠና	ጥናት አደረገ፤ ተረዳ፤ መረመረ አጠነከረ፤ አበረታ

Appendix C. Sample list of English sense examples used with their Amharic equivalent translation.

1. I made a vow to St.Gabriel to fast for two days

   ለቀዱስ ገብርኤል ሁለት ቀን ለመፆም ተሳልኩ ።።

2. Over and above this, men might vow individuals or possessions to God as a thank-offering.

   ከዚህ በተጨማሪ ወንዶች ሰዎችን ወይም ያላቸውን ንብረት እግዚያብሔርን ለማመስገን ይሳሉ ነበር ።።

3. This harmony might be expressed as an offering which accompanies a vow of some kind or as a thank-offering or free-will offering.

   ይህ ስምምነት የሚገለፀው የሆነ ነገር በፍቃደኝነት ለመስጠት ወይም ለማመስገን በመሳል ነው ።።

4. Jacob vowed a vow, saying, “If God will be with me, and will keep me in this way that I go, and will give me bread to eat, and clothing to put on, so that I come again to my father’s house in peace, and Yahweh will be my God,

   ያእቆብ እግዚአብሔር በሄድኩበት ከጠበቀኝ፣ የምበላውን እና የምለብሰውን ከሰጠኝኝ፣ ወደ አባቴ ቤት እመለሳለው እግዚአብሔርም አምላኬ ይሁናል ብሎ ተሳለ ።።

5. I am the God of Bethel, where you vowed a vow to me. Now arise, get out from this land, and return to the land of your birth.

   እኔ ስለት የተሳልክልኝ የቤቴልሔም አምላክ ነኝ፤አሁን ከዚህ ምድር ተነሳና ወደ ተወለድክበት ምድር ተመለስ።።

6. Now, after years of hard work, we are in sight of immunizing all the world's children against polio, tuberculosis, diphtheria, whooping cough, tetanus and measles.

   ከብዙ ጠንካራ ሥራ በኋላ ሁሉንም ያዓለም ሕፃናት ከ ፖሊዮ ፣ የሳንባ ነቀርሳ፣ የጉሮሮ በሽታ፣ የትክትክ ሳል ፣ ቲታነስ እና ኩፍኝ ክትባት መስጠት ችለናል ።።

7. In the United Kingdom, parents are advised to have their children immunised against diphtheria, tetanus, polio, whooping cough, measles and rubella .

   በእንግሊዝ ወላጆች ልጆቻቸውን ከ ጉሮሮ በሽታ፣ ቲታነስ ፣ፖሊዮ፣ የትክትክ ሳል፣ ኩፍኝ፣ ሩቤላ የሚከላከል ክትባት እነዲያስከትቡ ተመከሩ ።።

8. The days of being forced to get out of it on cough medicine are well behind.

   የሳል መድኃኒት አልቆ የምንቸገርበት ጊዜ አልፏል ።።

9. Greenough and colleagues showed that babies who did not require respiratory support had a high prevalence of wheeze and cough in the first year of life.

   ግሪኖፍ እና ጓደኞቹ የመተንፈሻ አካላት እርዳታ የማያስፈልጋቸው ህፃናት ማቃተትና ሳል በመጀመሪያ ዓመታቸው እንደሚያጋጥማቸው አሳዩ።።

10. The commonest symptom for the disease is coughing persistently, with frequent chest infection.

    የበሽታው የተለመዱ ምልክቶች በተደጋጋሚ መሳል እና የሚደጋገም የደረት ህመም ናቸው ።።

11. Charlton applied a sharp knife, carving it into steaks in the kitchen.

    ቻርልተን ማብሰያ ቤት ውስጥ ጥብሱን ለመክተፍ የተሳለ ቢላ ተጠቀመ ።።

12. A terrific place to have breakfast in, not a knife sharp enough to cut a lemon.

    ቁርስ ለመብላት የማይመች ቦታ ነው፤ ሎሚ ለመቁረጥ የሚሆን እንኳን የተሳለ ቢላ የለም ።።

13. The film's sharp sword has many edges.

    ፊልሙ ላይ ያሉት የተሳሉ ጎራዴዎች ብዙ ጠርዞች አላቸው ።።

14. ‘Even now, the memories are sharp as broken glass.

    አሁንም ትዝታዎቹ ልክ እንደ ተሰበረ ብርጭቆ የተሳሉ ናቸው ።።

15. Again do not round over the sharp edges when sanding.

    አሁንም አሸዋ ስታፈስ በተሳሉ ጠርዞች ላይ አትዙር ።።

16. To call Graf and Kohde-Kilsch a team on this showing is a misnomer.

    ግራፍንና ኮህድ ከልሰችን በዚህ ትርይታቸው ቡድን ብሎ መጥራት ትክክል አይሆንም ::

17. Azeglio Vicini, can call on an almost full-strength squad for a game in which he hopes Italy will prove his assertion.

    አዚግልዮ ቪኒቺ የጣሊያንን ብቃት የሳያል ተብሎ የሚጠበቅ ጠንካራ ተብሎ ሊጠራ የሚችል ቡድን ይሟል ::

18. The house was fall of memories; but even to call them memories was to imply that Jack had put them behind him; and he had not.

    ቤቱ በትዝታ የተሞላ ነው ፤አንደውም ትዝታ ተብሎ የሚጠራው ጃክ ከጀርባው አርጓቸው ነው በሚል ነው፤ ግን አይደለም ።።

19. ‘All the words he uses are what you would call anti-feminist,’ said a police officer.

    ፖ'ሊሱ እንዳለው እሱ የሚጠቀማቸው ቀላት በሙሉ ፀረሴት ተብለው ሊጠሩ የሚችሉ ናቸው ::

20. Sorry, that's what we call the Monday morning meeting where we discuss what's going on.

    ይቅርታ ሁሉንም ነገር የምንወያበት የሰኞ ጠዋቱ ስብሰባ ብለን የምንጠራው ስብሰባ ይህ ነበር ::

21. A man is only ‘acceptable’ to females if he is ‘nice and clean’.

    ወንድ የሴቶች ተቀባይነት የሚያገኘው ጥሩ እና ጠራ ያለ ሲሆን ነው ::

22. It was poorly and sparsely furnished; a brave effort had been made to keep it tidy and clean.

    ጥሩ ባልሆነና በተራራቀ ሁኔታ ነው እቃዎቹ የተቀመጡት፤ የፀዱና የጠሩ ለማረግ ጥሩ ጥረት ተደርጓል ::

23. In our minds ‘eating everything that is placed in front of us’ is associated with ‘well done, that's a nice clean plate’.

    በዐዕምሯችን ፍለፊታችን የቀረበልንን ምግብ መብላት በጥሩ ከተሰራና ጠራ ያለ ብርድልብስ ጋር ይያያዛል ::

24. Her hair flew out behind her, and the clean air struck her face.

    ፀጉሯ ከኋላዋ ይውለበለባል፤ እነዲሁም የጠራ አየር ፊቷን ይመታዋል ::

25. It's quite a job, keeping the windows clean.

    ሥራው የመስኮቶችን ጥራት መጠበቅ ነው ::

26. Curled up on his armchair, thin as a wood shaving, he looks far too slight to carry this immense spectacle.

    በእጅ ወንበር ላይ ተጠቅልሎ፤ ልክ እንደ እንጨት መፈግፈጊያ ሣሥቶ፤ ግርማሞገሱን ለማሳየት ብዙ የሚቀረው ይመስላል ::

27. Above all the accounts, technical and economic, of the lift's operation, are quite incredibly thin.

    ከሁሉም በላይ ሁሉም አካውንቶቹ ቴክኒካዊ እና ኢኮኖሚያዊ የቀኝ ሥራዎች በሚገርም ሁኔታ የሳሱ ናቸው ::

28. armed with a thin red and white linen cloth.

    በሣሣ ቀይ እና ነጭ ላይነን ልብስ ታጥቀዋል ::

29. Higher than predicted gravity values occur over the oceans because they are underlain by thin and relatively dense crust.

ከተገመተው በላይ በባህሮች ላይ የመሬት ስበት የሚጨምረው በሣሣ እና በንፅፅር ጥቅጥቅ ባለ ቅርፊት ስለሚሸፈን ነው ።።

30. These colours are so strong that you have to thin quite a bit to gain softer tones.

እነዚህ ቀለሞች እጅግ ጠንካራ ናቸው፤ የለሰለሰ ቶን ለማግኘት ቲኒሽ መሣሣት አለበት ።።

31. 'You're very good at taking care of people,'

ለሰዎች በመሣሣት በጣም ጥሩ ነህ ።።

32. Wilson taking care of me and treating me like a lady — because there was a little something between us.

ዊልሰን ልክ እንደ ሴት እየሣሣልኝ እና እየተንከባከበኝ ነው ምክንያቱ በመሀካላችን ቲኒሽ ነገር ነበር ።።

33. It may be decorating a flat for a person, to taking care of the cat of an elderly hospitalised lady.

ሆስፒታል ያለችን የሴት ድመት መሣሣት፤ ለሰው ጓማን እንደሚያስጌጥ ሊሆን ይችላል ።።

34. Her husband Barry's taking care of the other two kids — he's a real capable boy.'

ባለቤቷ ባሪ ሌሎቹን ሁለት ልጆች እየሣሣላቸው ነው በእውነቱ አቅም ያለው ልጅ ነው ።።

35. She was given the special responsibility of taking care of me, and I owe her my life.

ለእኔ የመሣሣት ልዩ ሀላፊነት ትሰጥቷል፤ እና ሕይወቴን እሰጣታለው ።።

36. until we study the life cycles of animals in fine detail, we cannot know precisely which creatures depend upon what.

የእንስሳትን የሕይወት ዑደት በተብራራ ሁኔታ ሳናጠና፤ አንዱ ፍጥረታት በምን ጥገኛ እንደሁኑ በትክክል ማወቅ አንችልም ።።

37. their Social Class and Educational Opportunity gave a new impetus both to the study of these themes and to action upon them.

ማህበረሰቡ ያላቸው ቦታ እና ያገኙት የትምህርት ዕድል ይህን ዘርፍ እና የሚሰሩትን ድርጊት እንዲያጠኑ ኃይል ሰጥቷቸዋል ።።

38. If they can make us more aware of the Earth and our relationship with it then their study will have been worthwhile.

ስለምድር እና ከምድር ጋር ስላለን ግኑኝነት ብዙ እንድናውቅ ካረጉን ጥናታቸው ጠቃሚ ይሆናል ።።

39. The study of Scripture, he suggested, did nothing to hinder an inquisitive man's delight in the study of nature.

የመፅሀፍ ቅዱስ ጥናት የሰው ልጅ ስለ ተፈጥሮ መመራመር ፍላጎትን እንደማያደናቅፍ አሳሰበ ።

40. It is also particularly easy to study, because sound can be recorded and reproduced by a tape-recorder.

ላማጥናት በጣም ቀላል ነው፤ ምክንያቱም ድምፆች በቴፕ መቅጃ ተቀድተው ሊባዙ ይችላሉ ።

41. Setting out Labour's case for international co-operation to strengthen world security and combat environmental degradation

የሰራተኞችን ጉዳይ መወሰን ዓለምአቀፍ ትብብር የዓለምን ሰላምን እና ለማጥናት እና የአከባቢ ብክለትን ለመከላከል።

42. saving and investment, and to strengthen private and public institutions; and steps to protect the poor during the transition.

ቁጠባ እና ኢንቨስትምንት የግል እና የህዝብ ተቋማትን ለማጥናት እና በሽግግሩ ጊዜ ድሀን ለመከላከል።

43. Ministry of Higher Education, aims to strengthen teacher competence at local level by encouraging teachers.

የከፍተኛ ትምህርት ሚኒስቴር አስተማሪዎቹን በማበረታት የአስተማሪዎቹን አቅም በአከባበው ደረጃ ለማጥናት አቅዷል ።

44. all listed companies have an active audit committee will strengthen the auditor's position vis-a-vis client management.

የተዘረዘሩት ድርጅቶች የኦዲተሩን ቦታ ሊያጠና የሚችል ትጉ የኦዲት ኮሚቴ አላቸው ።

45. ‘He will strengthen our squad and give us a lot more options this season.

የኛን ስኳድ ያጠናልናል እንዲሁም በዚህ ዓመት ብዙ አማራጮች ይሰጠናል ።

46. Chances are that it won't make an ideal radio record any more than ‘Candle in the Wind.

ንፋስ ላይ ካለ ሻማ ላይ የተሻለ ጥሩ የሬዲዮ ቀረጻ መስራት አይቻልም ።

47. millions of journalists have begun a record Christmas break.

በሚሊዮን የሚቆጠሩ ጋዜጠኞች የገናን እረፍት መቅረጽ ጀመሩ ።

48. Since leaving the ranks of a solo career, John Cale has gone on to record a catalogue of solo work that is both voluminous and impressive.

የሶሎ ሙያውን ከተወ በኋላ ጆን ኬል በጠም ብዙ የሚመስጡ የሶሎ ሥራዎችን ሊቀረጽ ነው ።

49. During the hearing Mike Morley had tricked his way into the prison to record the interview with Nilsen.

ክሱ በሚሰማበት ጊዜ ሞርሊ የኒልሰንን ቃለመጠይቅ ለመቅረጽ መሄዱን ካደ ።

50. Ironically, David Guest's victory came on the day that television cameras were allowed for the first time to record proceedings in a Scottish court.

የሚደንቀው የዴቪድ ገስት ድል የመጣው ቴሌቪዥኖች የስኮትላነድን ፍርድ ቤት ሂደት እንዲቀርጹ በተፈቀደላቸው በመጀመሪያው ቀን ነበር ።

51. In spite of the many things it has achieved over the last hundred years — and we have all been shaped by that — it has got itself boxed in by one issue.

ባለፊት መቶ ዓመታት ብዙ ነገር ቢያሳካም፤ እንዲሁም በዛ ብንቀረጽም በአንድ ጉዳይ ብቻ ነው ራሱን የወሰነው ።

52. The rapid growth of film and media studies in colleges and schools has been dominated and shaped by the cultural theories of the left.

በኮሌጆችና በትምህርትቤቶች ፈጣኑ የፊልምና የሚዲያ ጥናት በባህላዊው የቀኝ ጽንሰሐሳብ የተቀረጸና ተፅእኖ ያረፈበት ነው ።

53. Drug laws are shaped by vested economic interest, and the real reason that some drugs remain illegal is to allow the law to intervene in the lives of those the state perceives as threatening.

የመድኃኒት ህጎች የተቀረጹት የኢኮኖሚውን ፍለጎት ባገናዘበ መልኩ ነው፤ እንዲሁም እውነተኛው አንዳነድ መድኃኒቶች ህገወጥ የሆኑበት እውነተኛ ምክንያት መንግስት አስጊ በሚላቸው ሰዎች ኑሮ ውስጥ ጣልቃ ለመግባት እንዲመቸው ነው ።

54. The stories were generally shaped by his values and his literary genius allowed every social detail to be authentic and appropriate.

ታሪኮቹ በአጠቃላይ የተቀረጹት በእሴቶቹና በያነዳንዱ የማህበረሰቡ ጥቃቅን ነገሮች ያለው የሥነፅሁፍ ችሎታ አስተማማኝና ትክክለኛ መሆኑ ነው ።

55. These pupils are shaped by many other factors than their schooling

ተማሪዎቹ ከሚማሩት ትምህርት በተጨማሪ በሌሎች ብዙ ተፅዕኖዎች ተቀርጸዋል ።