St. Mary's University

Faculty of Informatics

Department of Computer Science

WORD SENSE DISAMBIGUATION FOR AFAAN OROMO: USING KNOWLEDGE BASE

Shibiru Olika Gonfa

August 2018

St. Mary's University

Faculty of Informatics

Department Of Computer Science

WORD SENSE DISAMBIGUATION FOR AFAAN OROMO USING KNOWLEDGE BASE

By

Shibiru Olika Gonfa

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

Addis Ababa, Ethiopia

August 2018

ACCEPTANCE

WORD SENSE DISAMBIGUATION FOR AFAAN OROMO USING KNOWLEDGE BASE

By

Shibiru Olika Gonfa

Accepted by the Faculty of Informatics, St. Mary's University. In partial fulfillment of the requirements for the Degree of Master of Science in Computer Science

Thesis Examination Committee

____

Mr. Hafte Abera

Internal Examiner

____

Dr. Solomon (Ph.D)

External Examiner

____

Mr. Asrat Mulatu
Dean, Faculty of Informatics

August 2018

Declaration

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

Shibiru Olika Gonfa

____

Addis Ababa, Ethiopia

This thesis has been submitted for examination with my approval as an advisor

Mr. Michael Melese

____

Addis Ababa, Ethiopia

Acknowledgement

First of all I would like to thank the almighty God who was on my side from the start to the end in smoothening everything that was complex during the course of this work. My next gratitude goes to my advisor, **Michael Melese**, for his valuable and constructive comments and in showing me clear research directions.

I would also like to thanks Oromia cultural Bureau Director, who had provided and facilitate for me on ways of getting "**Galmee Jechoota Afaan Oromo**" which is prepared by Addis Ababa University "**Wiirtu Qo'anno fi Qoranno Afaanota Itoophiyaatiin**."

My Friends**,** ***Moti***, ***Henock***, ***Amanuel*** and others, I always recall you for your jokes and debates we make on various social, economic and political issues.

Acronyms

WSD -Word sense Disambiguation

NLP-Natural Language Proceessing

AOWSD-Afaan Oromo Word sense Disambiguation

OROWORDNET-Oromo WordNet

MRD-Machine Readable Dictionary

ML-Machine Learning

List of Figures

List of Tables

List of Appendix

Abstract

Word sense disambiguation (WSD) is important and difficult problem that requires to be solved in Nature Language Processing. Afaan Oromo words have many meanings based on the context with which the word is used.

In Afaan Oromo there is much ambiguous word in which there meaning is changing with the context. This creates the user of the language to be confused about the meaning of those words.

In this paper we apply Knowledge based WSD method which is based on the database developed from scratch that uses Afaan Oromo Dictionary to disambiguate polysemous words in the sentence. The disambiguation process becomes accomplished based on words and sense relations developed in the database. The word sense disambiguation consists of preprocessing, morphological analysis, Afaan Oromo WordNet and disambiguation components to disambiguate ambiguous words of the language. Preprocessing component becomes the first stage to preprocess the input sentences to be used by morphological analysis to reduce the words to its root form or stem. The wordnet database stores words and it's Synsets with their relation and concepts to disambiguate the polysemous words. Finally the disambiguation component disambiguates the ambiguous word using information from other components of word sense disambiguation that we use in this paper.

Lastly, we conduct two experiments. The first experiment is with and without morphological analyzer that uses Afaan Oromo WordNet databases. The result of the experiment shows that an accuracy of 50.75% and 63.95% obtained. The second experiment becomes experiments that we conduct using various windows sizes to determine appropriate window sizes. According to the experiment window size of three- three becomes appropriate for Afaan Oromo.

Keywords: Natural Language Processing, Afaan Oromo WordNet, Word Sense Disambiguation,

Knowledge Based Approach.

1. Introduction

Currently, in 21st century, many individuals use web technology for searching and reading documents and texts to get what we want. During search process the user notice that the result of search is not appropriate as expected. The reason behind this is because of ambiguity in the search (query) words. Almost every word in natural languages is polysemous, that is, they have numerous meanings or senses. For instance the word "**Mirga**" has two meaning in the following sentence in Afaan Oromo context.

a. **Mirga** namoota eeguu.

b. **Mirga** qabadhuu deemi.

In the first sentence, the word "**Mirga"** means "**waan tokko gochuuf dandeettii seeraan qaban yookiin mirga namaa kabajuu**" when translated to English 'respect human right'. In the second sentence, it means "**kara harka mirga ofii qabatanii deemu**" when translated to English "go having right direction". These words sharing same spelling and pronunciation but have different senses. Human can easily understand which sense of "**Mirga**" is intended. In today's technology it is good if software could also detect which sense of "**Mirga**" was intended. Thus, various researchers conduct a research on natural language processing on word sense disambiguation which picks the intended sense of a word for a pre-defined set of words, using resources like a machine-readable dictionary, such as WordNet. Because, analysis of lexical and semantic words is necessary for computers to make sense of the words to return expected result [2].

Computer system process data based on fixed rules. Even though, computers are best at following fixed rules it becomes difficult to accurately disambiguate any words in any context. Thus, word sense ambiguity is a hard problem for the developers of Natural Language Processing (NLP) systems. A word has various contexts in various sentences [4].

Human language is ambiguous; so that many words can be interpreted in multiple ways depending on the context in which they occur, the identification of the specific meaning that a word assumes in context is only apparently simple.

The problem of WSD is a difficulty of associating the ambiguous word with its sense. As a result of this, there should be methods to associate ambiguous word to its sense. In order to do this matching, first, an inventory of the senses associated with each word to be disambiguated must be available; second, a mechanism to associate word senses in context to individual senses must be developed, and thirdly, an evaluation procedure to measure how well this disambiguation mechanism performs must be adopted [1]. According to [5] automatic disambiguation of lexically ambiguous words is generally a two-tiered problem. First, a dictionary containing information necessary for the disambiguation is needed like wordnet. In this dictionary all meanings for each word are listed. Second, this dictionary is used to determine which word sense is the appropriate one in a given piece of text. To disambiguate the ambiguous word it is must to have a rule for the construction of WSD and their subsequent application to a real disambiguation problem, achieving WSD. Generally, ambiguous word is unconscious to people because human are very good at resolving them using context and their knowledge of the world [1]. However, computer systems does not have cognitive knowledge like human being, and do not do a good job of making use of the context to disambiguate ambiguous word [4].

Lexical disambiguation in its broadest definition is nothing less than determining the meaning of every word in context, which appears to be a largely unconscious process in people. As a computational problem it is often described as “AI-complete”, that is, a problem whose solution presupposes a solution to complete natural-language understanding or common-sense reasoning
In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the problem of computationally determining which “sense” of a word is activated by the use of the word in a particular context. Words are assumed to have a finite and discrete set of senses from a dictionary, a lexical knowledge base, or ontology [6].

WSD has obvious relationships to other fields such as lexical semantics, whose main endeavour is to define, analyze, and ultimately understand the relationships between “word”, “meaning”, and “context”. Although word meaning is at the heart of the problem, WSD has never really found a home in lexical semantics. It could be that lexical semantics that has been more concerned with representational issues and models of word meaning and polysemy so far too complex for WSD [7].

Word Sense Disambiguation (WSD) is a process of automatically identifying the correct meaning of a word that has multiple meanings. In WSD, these meanings are referred to as senses, or concepts, which are obtained from a sense-inventory. The ambiguous word is referred to as the target word and the context in which the target word is used is called an instance. WSD is an enabler for other tasks and applications of computational linguistics and natural language processing (NLP) such as parsing, semantic interpretation, machine translation, information retrieval, question answering, text mining, Computational Advertising and the like. The computational identification of meaning for words in context is called word sense disambiguation [6].

Word Sense Disambiguation was considered as an important sub-problem of Machine Translation during the late 1940s.During that researchers recognized the essentials of WSD such as the local context in which a target word to be disambiguated occurs, the statistical distribution of words and senses, and the role of knowledge bases. As a result of lack of available computational resources, a bottleneck was reached and not much progress was made. But with the accessibility of lexical resources in the 1980s, WSD saw a revival, with people turning to AI based approaches to tackle the problem. The advancement of statistical modeling and Machine Learning, in the 1990s saw three major developments: WordNet became available, the statistical revolution in NLP swept through, and Senseval began. The purpose of Senseval is to evaluate the strengths and weaknesses of computer programs designed to automatically determine the sense of a word in context with respect to different words, different varieties of language, and different languages. SENSEVAL (currently renamed SEMEVAL) is an international competition on WSD organized by the Association for Computational Linguistics (ACL) **S**pecial **I**nterest **G**roup on the **LEX**icon (SIGLEX) [1, 9].

There are many research's conducted on word sense disambiguation on different languages using various methodologies. Among this WordNet is now widely used in the Natural Language Processing (NLP) community for applications in Information Retrieval, Machine Translation, Word Sense Disambiguation etc. One of the most successful to WSD is to make use of WordNet [1].

Wordnet is created by George Miller and his team at Princeton University. It is a large electronic database organized as a semantic network built on typical relations including synonymy, hyponymy, antonymy, and entailment. WordNet evolved into a system that reflects current psycholinguistic theories about how humans organize their lexical memories. WordNet contains only open class words (nouns, verbs, adjectives, and adverbs) and which does not contain closed class words such as pronouns, conjunctions, and prepositions. WordNet is organized semantically (as part-of-speech) [7].

Afaan Oromo WSD(AOWSD) systems is developed which include, Preprocessing component, Morphological Analysis component, Afaan Oromo WordNet Database and disambiguation components.

2. Statement of the problem

In Ethiopia there are many languages spoken among those language Afaan Oromo is one of the major languages. Currently, it is an official language of Oromia national regional state. It is spoken by more than **30** million Oromo's within Ethiopia [10]. In addition, the language is also spoken in Somalia, Kenya, Uganda, Tanzania, Djibouti and other countries where the language speakers exist. Previously there is a research attempt to handle word sense disambiguation by [2, 3]. The difficulty with word sense disambiguation is word senses. There are no exact ways of identifying where one sense of a word ends and the next begins [7]. As a result of this it becomes mandatory to develop WSD for Afaan Oromo to accurately disambiguate words and retrieve documents from on line repository.

A corpus based approach to disambiguation is employed by [3] were supervised machine learning techniques are applied to a corpus of Afaan Oromo language, to acquire disambiguation of information automatically. Manually annotated corpus data for five selected ambiguous words becomes preprocessed to make it ready for experimentation.

A hybrid approach is applied by [2] which find the meaning of words based on surrounding contexts combining unsupervised with rule based approach. Hence, the researcher's work presents a WSD strategy which combines unsupervised approach that exploits sense in a corpus and the manually crafted rule using hybrid method. The researcher applies this approach to only twenty selected ambiguous words.

Those researchers [2, 3] have the following limitations. The study conducted by [3] was limited to five ambiguous words and that of [2] is limited to twenty ambiguous words to experiment. Manly available sense-annotated corpora are insufficient to cover all the senses of each of the ambiguous words and corpus used as a source of information for disambiguation. A sample of words is selected from the corpus and the selected words are disambiguated in a short given context. It is assumed that the word to be disambiguated has a fixed set of senses in the sense inventory, where the sense inventory contains the mapping of words and their different senses. Afaan Oromo WSD developed by researchers requires manually labeled sense examples, which is time taking and exhaustive when the number of corpus size increased and cluster the contexts of an ambiguous word into a number of groups. Manual sense tagging is very difficult, time taking and limiting the number of sense tagged words to be used.

To handle the above problem knowledge-based approach to Afaan Oromo WSD technique is proposed. Knowledge-base uses structured data called a knowledge source. These methods rely on information from the knowledge source about a concept such as its definition or synonym rather than training instances in manually annotated or unannotated data.

Afaan Oromo WordNet is used as a source of information for disambiguation to disambiguate words in a sentence, which is called all words disambiguation. The previous researcher uses lexical-sample disambiguation because lexical sample methods can only disambiguate words in which there exists a set of training data i.e. ambiguous words may not be known ahead of time. All word disambiguation is advantageous than lexical samples. We determine the correct concept of ambiguous words by first identifying the ambiguous words semantic type, which is a broad categorization of a concept. After the semantic type of the ambiguous words is identified, the correct concept is identified based on its semantic type from Afaan Oromo WordNet. The development of Afaan Oromo WordNet is an important step for WSD, even for other application areas such as Information Retrieval, Machine Translation and so on. This is the research gap that motivates us to use Afaan Oromo WordNet for WSD [7].

Lastly, we proposed a knowledge-based Afaan Oromo WSD method that does not require sense tagged corpus which identifies senses of all words in sentences or not a small number of words. Therefore, the major concern of this research was to investigate knowledge-based approach for Afaan Oromo WSD, test the results in order to develop a bit further natural language understanding and compare the results with the previous researches [2, 3].

3. Objectives

3.1 General objective

The general objective of this research work is to design and develop a model for Afaan Oromo word sense disambiguation using WordNet.

3.2 Specific objectives

The specific objectives of this research work are:

- To review a literature on the techniques and approaches of WSD adopted for other languages using WordNet.
- To collect data from Afaan Oromo dictionary and other relevant sources for developing Afaan Oromo WordNet.
- To Identify Afaan Oromo ambiguous words and their contextual meaning
- To design architecture for Afaan Oromo Word Sense Disambiguation.
- To develop algorithm for Afaan Oromo Word Sense Disambiguation.
- To develop a prototype of the system.
- Evaluating the performance of the developed model.

4. Methodology of the study

4.1 Research design

This research is an experimental design that tests the implementation of the word sense disambiguation for Afaan Oromo language. It uses Wordnet developed for the languages to get target word and find sense for the target word.

4.2 Literature review

For the purpose of understanding different literatures, books and other scholarly published materials are reviewed. In addition to this, the researches reviews different material to avoid duplication of research and to go through different techniques and algorithms that are applied by different researchers to design word sense disambiguation system for different languages specifically giving attention to word sense disambiguation.

4.3 Data collection

We collect our data from various institutions using Afaan Oromo documents, libraries and other relevant sources like Afaan Oromo dictionary[11, 51] having ambiguous words. We use fifteen ambiguous words for training and thirty five for testing our model.

4.4 Development tools

Different tools and techniques will be used to achieve the goal of the research. The main parts of the system are Word Sense Disambiguation, Afaan Oromo WordNet, Morphological Analyzer and Preprocessing. We use Java programming, Python programming and SQL Server. A prototype is developed to test the performance of the system. AS a morphological Analyzer we use Michael Gassers' tool developed for Afaan Oromo, Amharic and Tigrinya.

4.5 Experimentation

After we develop the prototype we perform experiment to see the effectiveness of knowledge based word sense disambiguation. The experiment is performed using the input sentence fed to the model.

5. Scope and limitation

Word Sense Disambiguation is complex because there are no decisive ways of identifying where one sense of a word ends and the next begins. There are no publicly available linguistic resources for Afaan Oromo. Researches in WSD for other language use linguistic resources like WordNet, thesaurus and machine-readable dictionaries but for Afaan Oromo this resource is not developed. The scope of this study is limited to retrieving senses of ambiguous word from Afaan Oromo WordNet, Identifying the ambiguous words and its context in the given text and assigning the appropriate sense to the given word in the given context from Afaan Oromo WordNet, which is developed manually.

The limitation of this study is that the developed system does not perform grammar and spelling correction and do not works for words, which do not exist in Afaan Oromo WordNet developed.

6. Significance of this study

The result of this study is expected to produce experimental evidences for word sense disambiguation of Afaan Oromo texts. It also contributes for future researcher's and development in the area of natural language processing specifically in machine translation, Information retrieval, speech processing, text processing, information retrieval, content and thematic analysis. It is expected that, the result of this study may used by different stakeholders like speakers of Afaan Oromo and new language learners to identify proper word senses. The other beneficiaries from this output are those who do not properly identify meaning of polysemous words when the words come with different context.

7. Thesis organization

Our thesis organization is presented in a summarized form as follows. Chapter Two presents literature review. In this chapter we reviewed literatures related to word sense disambiguation to have clear understanding on what is word sense disambiguation mean. Additionally, approaches to word sense disambiguation is clearly indicated. We also present Afaan Oromo Language and Afaan Oromo word ambiguity. Chapter three presents' works related to word sense disambiguation system. In this chapter works of previous researches done for local and foreign languages are presented. Chapter Four discusses the design (architecture) of Afaan Oromo knowledge base which is composed of Preprocessing, Morphological analysis, Afaan Oromo WordNet (**OROWORDNET**) developed from Afaan Oromo Dictionary and disambiguation component. The Fifth Chapter discusses the implementation of the system, development of prototype and experiment of the proposed WSD using a corpus prepared for the proposed system. Finally, chapter six deals with the conclusion and recommendations drawn from the findings of our study.

CHAPTER TWO: LITERATURE REVIEW

2. Introduction

This chapter focuses on literature in the field of word sense disambiguation (WSD). Thus, overview of word sense disambiguation and discussion on major approaches that have been employed for WSD research with special focus on knowledge-based approach, which is used in this study. The discussion of different approaches and algorithms would help to understand the central problem in WSD research and facilitates the comparison of existing approaches to the specific solutions that are employed in this study. Lastly, Afaan Oromo word ambiguity is presented shortly.

2.1 Word Sense Disambiguation

Word sense disambiguation (WSD) is a subfield within computational linguistics, which is also referred to as natural language processing (NLP),where computer systems are designed to identify the correct meaning (or sense) of a word in a given context[6, 7]. For example, the word "**Mirga**" is ambiguous. As a noun it can be used to mean "right or privilege". Those ambiguous words come with different sentence having different sense. As a result, word sense disambiguation technique is used in finding the meaning of a word in a various sentence [2]. For sense classification there should be approaches to differentiate those senses. Those, approaches are classified into knowledge based and corpus based [3]. According to[12] a corpus based approaches which is called as supervision machine learning approaches are categorized as; supervised machine learning, unsupervised machine learning and Bootstrapping machine learning approaches.

In knowledge based approach disambiguation is carried out using information contained in man-made lexical resources, like WordNet. The lexicon may be a machine readable dictionary, thesaurus or it may be hand-crafted.

For our research purpose knowledge based approach to WSD is applied to Afaan Oromo language. Knowledge based approach uses external lexical resources like WordNet to disambiguate words. We have developed a WSD tool using knowledge-based approach with Afaan Oromo WordNet that we develop to disambiguate words. WordNet is built from co-occurrence and collocation and it includes synset or synonyms, which belong to either noun, verb, adjective, or adverb [14, 9].

The knowledge-based approach does not rely on sense-annotated corpora. It uses information contained in large lexical resources, such as WordNet. The Lesk [15] algorithm is a classic example of the knowledge-based approach. The algorithm counts the number of words that are in both the neighborhood of the ambiguous word and in the definition of each sense in a dictionary. It then chooses the sense with the larger number of words. Even though it is simple, its approach is very sensitive to the exact wording of the definitions, so the absence of a certain word can radically change the results. This is a significant limitation as dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions. These kinds of limitation are common to most dictionary-based methods, as they have not realized the potential of combining the relatively limited information in such definitions with the abundant co-occurrence information extractable from text corpora.

According to [16], word sense disambiguation involves two steps. The first thing 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 dictionary, from the synonyms in a thesaurus, or from the translations in a translation dictionary. In the second phase to involves a means to assign the appropriate sense to each occurrence of a word in a context. All disambiguation work involves matching the context of an instance of the word to be disambiguated either by using information from external knowledge sources or with contexts of previously disambiguated instances of the word. To do this it becomes necessary to perform preprocessing or knowledge-extraction procedures representing the information as context features. The computer also 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.

Words do not have well-defined boundaries between their word of senses, and our task is to determine which meaning of the word is indented in a given context. This is first problem encountered by any natural language processing system, which is referred to as lexical semantic ambiguity.

We can distinguish two variants of the generic WSD task [13]:

1. Lexical sample (or targeted WSD) - where a system is required to disambiguate a restricted set of target words usually occurring one per sentence. Supervised systems are typically employed in this setting, as they can be trained using a number of hand-labeled instances (training set) and then applied to classify a set of unlabeled examples (test set).

2. All-words WSD- where systems are expected to disambiguate all open-class words in a text (i.e., nouns, verbs, adjectives, and adverbs). This task requires wide-coverage systems. Consequently, purely supervised systems can potentially suffer from the problem of data sparseness, as it is unlikely that a training set of adequate size is available which covers the full lexicon of the language of interest. This thesis focuses on all words disambiguation tasks but also highlights target word tasks.

2.2 Knowledge Sources for WSD

Knowledge is a fundamental component for word sense disambiguation. Knowledge sources provide data which are essential to associate senses with their appropriate words. They can vary from corpora of texts, either unannotated (unlabeled) or annotated with word senses, to machine-readable dictionaries, thesauri, glossaries, ontology's, WorldNet's, etc. [7, 9].

2.2.1 Lexical Knowledge

Work on WSD reached a turning point in the late 80s when large scale lexical resources, such as dictionaries, thesauri, became widely available. Attempts were made to automatically extract knowledge from these kinds of resources towards disambiguation of words [17]. In the mid 80s, work began on the construction of large scale knowledge bases by hand like WordNet. Fundamentally there exists two approaches to the construction of this semantic lexicons; the

enumerative approaches, wherein senses are explicitly provided, and the generative approach, in which semantic information associated with given words is underspecified, and generation rules are used to derive precise sense information [17].
WordNet is like a dictionary in that it stores words and meanings. In WordNet the concepts are defined as synonymy sets called synsets linked to each other through semantic relations like (hyperonymy, hyponymy, meronymy, antonymy, and so on). Each sense of a word is linked to a synset. In this research, we use Afaan Oromo WordNet that we develop to disambiguate words [4].

WordNet

WordNet [18] is a machine readable dictionary. Unlike most dictionaries, WordNet contains only open-class words (nouns, verbs, adjectives, and adverbs). WordNet does not contain closed-class words such as pronouns, conjunctions, and prepositions. WordNet groups sets of synonymous word senses into synonym sets or synsets. A word sense is a particular meaning of a word. For example, the word "Mirga" has several meanings; as a noun, it can refer to "wanna seeran namaaf kenname tokko". A synset contains one or more synonymous word senses. The synset is the basic organizational unit in WordNet. Each synset has a gloss (definition) associated with it.

Words with only one sense are said to be monosemous. Thus, according to wordnet principle those kinds of words exist only in one synset. In WordNet, each word occurs in as many synsets as it has senses. By definition, each synset in which a word appears is a different sense of that word. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is like a dictionary in that it stores words arranged semantically instead of alphabetically [18].

Structure of wordnet makes a wordnet a useful tool for computational linguistics and natural language processing. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms-strings of letters-but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus do not follow any explicit pattern other than meaning similarity [19].

WordNet is a lexical database of the English language. It was created in the cognitive science laboratory of Princeton University under the direction of psychology professor Miller starting in 1985 and has been directed in recent years by Christiane Fellbaum [21]. This English WordNet encourages us to develop Afaan Oromo WordNet which can be used for WSD. Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between Synsets.

The central object in WordNet is a synset, a set of synonyms. WordNet organizes the lexical information in terms of word meanings and it can be termed as lexicon based on psycholinguistic principles. Each word may have one or more senses and these are classified as Homonynms, Monosemous and Polysemous. Homonynms- A case of homonymy is one of an ambiguous word, where different cases are related to each other in any way. Words that are identical in sound and spelling are called homonyms. Monosemous-word with only one sense are said to be monosemous. When a word or phrase has several meanings, you can describe that word as polysemous [1].

Parts of Speech in WordNet

WordNet stores information about words that belong to four Part-Of-Speech: nouns, verbs, adjectives and adverbs [16]. These are arranged in their respective synsets. Prepositions and conjunctions don't belong to any synset.

Nouns in WordNet: Noun words have various relations defined in WordNet for the Noun Part of speech. These relations are Hypernymy and Hyponymy, Meronymy and Holonymy and Antonymy. Hypernymy and Hyponymy: These are two most common relations for nouns. They are semantic relationships that connect two synsets if the entity referred to by one is a kind of or is a specific example of entity referred to by other. Specifically, if synset A is kind of B synset, then B is a hyponym of A, and A is the Hypernym of B. The number of hypernym links is equal to the number of hyponym links since for every hypernym link there is a corresponding hyponym link. Meronymy and Holonymy: These are also semantic relationships that connect two synsets if A is part of B conversely B is a holonymy of A [16].

Verbs in WordNet: Verb words have various relations defined in WordNet for the Verb Part of speech. These relations are Troponymy, Antonymy, Entailment, and Cause. These Troponym and Antonymy are analogous to the noun hypernymy and hyponymy respectively. Synset A is the hypernym of B, if B is one way to A; A is then the troponym of B. Antonymy: Like nouns, verbs are also related through the relationship of antonymy that links two verbs that are opposite to each other in the meaning. This is a lexical relationship and does not belong to the other words in the synsets that both belong to. Entailment and Cause: Other relations defined for verbs include those of entailment and cause, both of which are semantic relations. A synset A is related to synset B through the entailment relationship if A entails B [16].

Adjectives and Adverbs in WordNet: Adjectives and Adverb words have various semantic relations defined in WordNet are Similar-to and Also-see.

Similar-to: It is defined for Adjectives. This semantic relationship links two adjective synsets that are similar in meaning, however not close enough to be put together in the same synset. Also-see: This relation is common to both adjective and verbs. All links of this type of adjective are semantic in nature but they are not lexical relations

2.2.2 Learned World Knowledge

World knowledge is very difficult or trivial to be verbalized completely. Because of this, there should be a smart strategy to automatically acquire world knowledge from the context of training corpora on demand by machine learning techniques [22]. The frequently used types of contextual features for learning are listed below.

Indicative Words surround the target and can serve as the indicator of target senses. In general, the closer to the target word, the more indicative to the sense. There are several ways, like fixed-size window, to extract candidate words. Domain-specific Knowledge, like selectional restrictions, is about the semantic restrictions on the use of each sense of the target word.

However, domain-specific knowledge can only acquire from training corpora, and can only be attached to WSD by empirical methods, rather than by symbolic reasoning. There are no significant distinctions between lexical knowledge and learned world knowledge. If the latter is general enough, it can be released in the form of lexical knowledge for public use. Usually, unsupervised approaches and knowledge-based approaches use lexical knowledge only, while supervised approaches employ learned world knowledge for WSD [22]. For our study, we use Lexical Knowledge as knowledge source.

2.3 Approaches to WSD

Context is the only means to identify the meaning of a polysemous word in a sentence. Getting the correct meaning of a word in context for computer is not as simple as for human being because computer lacks knowledge for word sensing based on a context surrounding the target words. As a result of this there should be approaches to provide clues or indicators for a computer to differentiate word senses for a target word. The knowledge source that we use to differentiate word sense depends on approaches that we follow. Methods that depend primarily on dictionaries, thesauri, lexical knowledge bases, and WordNet without using any corpus evidence are termed as dictionary-based or knowledge-based approaches [7]. Methods that rely on external information that uses sense-tagged corpora to train the sense model, which makes it possible to link contextual features (world knowledge) to word sense is known us supervised approaches. Theoretically, it should outperform approaches because more sense tagged information is fed into the system [22]. Methods that depend on external information and work directly from raw unannotated corpora are termed unsupervised methods (adopting terminology from machine learning). Included in this category are methods that use word-aligned corpora to gather cross-linguistic evidence for sense discrimination [7].

2.3.1 Knowledge-based WSD

The knowledge-based approach to natural language processing (NLP) concerns itself with methods for acquiring and representing knowledge for intelligent information access, automatic document classification, machine translation and for applying the knowledge to solve well-known problems in NLP such as ambiguity resolution [23]. Knowledge-based methods depend on information that can be extracted from a knowledge source, such as a dictionary, thesaurus or lexical database. Knowledge-based methods represent a distinct category in word sense disambiguation (WSD). The performance of such knowledge intensive methods is usually exceeded by their corpus-based alternatives, but they have the advantage of a larger coverage. Knowledge based methods for WSD are usually applicable to all words in unrestricted text, as opposed to corpus-based techniques, which are applicable only to those words for which annotated corpora are available [7].

The knowledge-based method relies on knowledge resources like WordNet as it is organized into synonym sets representing lexical concepts. WordNet also organizes words into a conceptual structure by representing a number of semantic relationships (hyponymy, hypernymy, meronymy, etc.) among synsets. Thus it uses these organized concepts to disambiguate contexts in a sentence. This approach usually picks the sense whose definition is most similar to the context of the ambiguous word, by means of textual overlap or other methods. Knowledge-based allows us to use grammar rules as well as hand coded rules for disambiguation. The overlap based approach strategized by knowledge based technique requires a Machine Readable Dictionary (MRD) [20]. The availability of massive lexicographic databases offers a promising route to overcoming the knowledge acquisition bottleneck [24].

There are many techniques to approach word sense disambiguation using knowledge base. Those techniques are: measures of semantic similarity, selection restriction, and heuristic based WSD and overlap based approach [7]. Knowledge sources used for WSD are either lexical knowledge which is released to the public, or world knowledge learned from a training corpus [22]. A large training corpus used in supervised approach is avoided by using knowledge based approaches. It is classified based on the function of the resources used as, Machine readable Dictionary, Thesauri, and Lexical Knowledge Bases.

Machine Readable Dictionaries (MRDs): It becomes a popular source of knowledge for natural language processing tasks. The machine read able dictionary provides a readymade source of information for word senses and it becomes a staple of WSD research. Since Lesk [15], the first WSD based on MRD and many researchers have used machine-readable dictionaries (MRDs) as a structured source of lexical knowledge to deal with WSD. However, MRDs contain inconsistencies and are created for human use, and not for machine exploitation. As stated by Agirre and Martinez [15], there are different types of information, which is useful for WSD. It is obtained from MRDs. This information includes part of speech, semantic word associations, syntactic cues, selection preferences, and frequency of senses.

Thesauri: It provide information about relationships among words like synonymy, antonymy and, possibly, further relations [9, 16]. Thesaurus based disambiguation makes use of the semantic categorization provided by a thesaurus or a dictionary with subject categories [16]. The most frequently used thesaurus in WSD is Roget's International thesaurus which was put into machine tractable form [25]. 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. This category then determines the correct senses that are used [16].

Below is brief description of four main types of knowledge-based methods for word sense disambiguation.

Lesk Algorithm for Word Sense Disambiguation

The Lesk algorithm [15] is one of the first algorithms developed for the semantic disambiguation of all words in unrestricted text. The only resource required by the algorithm is a set of dictionary entries, one for each possible word sense, and knowledge about the immediate context where the sense disambiguation is performed. The original Lesk algorithm [15] disambiguates words in short phrases. Given a word to disambiguate, the dictionary definition or gloss of each of its senses is compared to the glosses of every other word in the phrase. A word is assigned that sense whose gloss shares the largest number of words in common with the glosses of the other words. The algorithm begins anew for each word and does not utilize the senses it previously assigned. It is a simple knowledge-based approach, which relies on the calculation of the word overlap between the sense definitions of two or more target words. This approach is named gloss overlap or the Lesk algorithm after its author Lesk. The main idea behind the original definition of the algorithm is to disambiguate words by finding the overlap among their sense definitions.

Although traditionally considered a dictionary-based method, the idea behind the Lesk algorithm represents the starting seed for today’s corpus based algorithms [15].

Measures of Semantic Similarity for WSD

Words that share a common context are usually closely related in meaning, and therefore the appropriate senses can be selected by choosing those meanings found within the smallest semantic distance. Based on the size of the context they span, measures of the context are divided in to two main categories. These are local context and global context. In local context a given word does not take into account additional contextual information found outside a certain window size. On the other hand, a global context attempt to build threads of meaning throughout an entire text, with their scope extended beyond a small window centered on target words [7].

Among some of the known structures of meanings are Lexical chains. A lexical chain is a sequence of semantically related words, which creates a context and contributes to the continuity of meaning and coherence of a discourse. They are considered useful for various tasks in natural language processing, including text summarization, text categorization, and word sense disambiguation. Lexical chains are drawn independently of the grammatical structure of the text, and may span long distances in the text. However, solutions designed to increase the efficiency of the Lesk algorithm are equally applicable here, in which each ambiguous word in disambiguated individually, using a method similar in spirit with to the simplified Lesk algorithm. The application of measures of semantic similarity to the disambiguation of words in unrestricted text is not always a straightforward process [7].

Selectional Preferences (restrictions) for WSD

Selectional preferences try to capture the fact that linguistic elements prefer arguments of a certain semantic class. Selectional preferences capture information about the possible relations between word categories, and represent commonsense knowledge about classes of concepts. EAT-FOOD, DRINK-LIQUID, are examples of such semantic constraints, which can be used to rule out incorrect word meanings and select only those senses that are in harmony with common sense rules [15].

While selectional preferences are perceptive, and occur to us in a natural way, it is difficult to put them into practice to solve the problem of WSD. The main reason seems to be the circular relation between selectional preferences and WSD: learning accurate semantic constraints requires knowledge of the word senses involved in a candidate relation, and, vice versa, WSD can improve if large collections of selectional preferences are available[7,15].

Heuristics for Word Sense Disambiguation

Heuristic methods, consists simple rules that can reliably assign a sense to certain word categories, an easy and yet fairly precise way to predict word meanings is to rely on heuristics drawn from linguistic properties observed on large texts. One such heuristic, which is often used as a baseline in the evaluation of many WSD systems, is the most frequent sense heuristic. The other two heuristics are the tendency of a word to exhibit the same meaning in all its occurrences in a given discourse (one sense per discourse), in the same collocation (one sense per collocation) [7].

Gale *et al.* [26] introduced one Sense per Discourse heuristic. It states that a word tends to preserve its meaning across all its occurrences in a given discourse. This is a rather strong rule since it allows for the automatic disambiguation of all instances of a certain word, given that its meaning is identified in at least one such occurrence. The one-sense-per-collocation heuristic is similar in spirit to the one-sense per-discourse hypothesis, but it has a different scope. Yarowsky [27] introduced it, and it states that a word tends to preserve its meaning when used in the same collocation. In other words, nearby words provide strong and consistent clues to the sense of a target word. It was also observed that this effect is stronger for adjacent collocations, and becomes weaker as the distance between words increases. Yarowsky (25) used both one-sense-per-discourse (and one-sense per-collocation) in his iterative bootstrapping algorithm, which improved performance from 90.6% to 96.5%

2.3.2 Corpus-based WSD

In the last fifteen years, empirical and statistical approaches have had a significantly increased impact on NLP. The types of NLP problems initially addressed by statistical and machine-learning techniques are those of language- ambiguity resolution, in which the correct interpretation should be selected from among a set of alternatives in a particular context. These techniques are particularly adequate for NLP because they can be regarded as classification problems, which have been studied extensively in the ML community. Corpus-based approaches are those that build a classification model from examples. These methods involve two phases: learning and classification. The learning phase consists of learning a sense classification model from the training examples. The classification process consists of the application of this model to new examples in order to assign the output senses [28].

The three main approaches to WSD based on statistical methods are [9].

Supervised Corpus-Based Method

Supervised WSD uses machine-learning techniques for inducing a classifier from manually sense-annotated data sets. Usually, the classifier (often called *word expert*) is concerned with a single word and performs a classification task in order to assign the appropriate sense to each instance of that word. The training set used to learn the classifier typically contains a set of examples in which a given target word is manually tagged with a sense from the sense inventory of a reference dictionary [9].

Supervised methods are based on the assumption that the context can provide enough evidence on its own to disambiguate words hence; world knowledge and reasoning are deemed unnecessary. These supervised methods are subject to a new knowledge acquisition bottleneck since they rely on substantial amounts of manually sense tagged corpora for training, which are laborious and expensive to create.

Generally, supervised systems have obtained better results than unsupervised ones, a conclusion that is based on experimental work and international competitions. This approach uses semantically annotated corpora to train machine learning (ML) algorithms to decide which word sense to choose in which contexts. The words in such annotated corpora are tagged manually using semantic classes taken from a particular lexical semantic resource. Corpus-based methods are called "supervised" when they learn from previously sense-annotated data, and therefore they usually require a large amount of human intervention to annotate the training data. Although several attempts have been made to overcome the knowledge acquisition bottleneck (too many languages, too many words, too many senses, and too many examples per sense) it is still an open problem that poses serious challenges to the supervised learning approach for WSD.

In supervised approaches, a sense disambiguation system is learned from a representative set of labeled instances drawn from sense-annotated corpus. Input instances to these approaches are features encoded along with their appropriate labels. The output of the system is a classifier system capable of assigning labels to new feature encoded inputs.

According to [3] in supervised techniques words can be labeled with their senses. For example in the following two sentences **horii** is an ambiguous word and it is tagged with sense **beelada** (cattle) and **qarshii** (money) respectively.

- Tolaan horii<**beelada**> qale nyaate.
- Tolaan horii<**qarshii**> isa mana baanki kaa‟e.

Therefore supervised approaches can be seen as: – accept a corpus tagged with senses – define features that indicate one sense over another – learn a model that predicts the correct sense given the features

Unsupervised Corpus-Based Method

Unsupervised learning identifies patterns in a large sample of data, without the benefit of any manually labeled examples or external knowledge sources. These patterns are used to divide the data into clusters, where each member of a cluster has more in common with the other members of its own cluster than any other. If one may remove manual labels from supervised data and cluster, one may not discover the same classes as in supervised learning. In this way, supervised classification identifies features that trigger a sense tag and unsupervised clustering finds similarity between contexts. If sense tagged text is available, it can be used for evaluation. But these sense tags are not used for clustering or feature selection [2, 3, 16].

Unsupervised WSD approaches have a different aim than supervised and knowledge-based methods, that is, that of identifying sense clusters compared to that of assigning sense labels. However, sense discrimination and sense labeling are both sub problems of the word sense disambiguation task and are strictly related, to the point that the clusters produced can be used at a later stage to sense tag word occurrences [7].

2.3.3 Hybrid Approaches

These approaches are the hybrid between different methods like statistical based and rule based methods of machine learning approaches. It combines the advantages of corpus based and knowledge based methods to overcome the specific limitations associated with a particular approach and improve WSD accuracy. For example, Yarowsky [25] 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. Then the seed definitions are used to classify the obvious cases in a corpus. Luk's[12] system uses the textual definitions of senses from a machine readable dictionary to identify relations between senses. It then uses a corpus to calculate mutual information scores between the related senses in order to discover the most useful information. In this way, the amount of text needed in the training corpus is reduced.

2.4. Design requirements

In natural language processing like word sense disambiguation, machine translation and others the feature of the language is a determinant factor. Thus, designing Afaan Oromo WordNet based on features of the language composition becomes a crucial role for sense disambiguation. Four main elements are required in designing every word sense disambiguation system: the selection of word senses, the use of knowledge sources, the representation of context, and the selection of an automatic classification approach [13].

Knowledge sources

All disambiguation work involves matching the context of an instance of the word to be disambiguated either with information from external knowledge sources (like WordNet) or with contexts of previously disambiguated instances of the word from a training corpus using world knowledge [13]. Lexical knowledge is usually released with a dictionary where as world knowledge is too complex to be verbalized completely. Therefore, it is a good strategy to automatically acquire world knowledge from the context of training corpora on demand by machine learning techniques. For our study, we use Afaan Oromo WordNet as knowledge source that we have developed [20].

Selection of Word Senses

A *word sense* is a commonly accepted meaning of a word. For instance, consider the following two sentences:

- Tolaan mana baankiti ***horii*** baayyee qaba.
- Qonnaan bultoonni hedduun ***horii*** horsiisuun galii argatu.

The word "***horii***" is used in the above sentences with two different senses: **qarshii (money)** in the 1st sentence and **beelada (cattle)** in the 2nd sentence. The example makes it clear that determining the sense inventory of a word is a key problem in word sense disambiguation. A *sense inventory* partitions the range of meaning of a word into its senses. Word senses cannot be easily discretized, that is, reduced to a finite discrete set of entries, fact that the language is inherently subject to change and interpretation [13].

Representation of context

According to [3] for any word sense disambiguation there should be a standard approach to WSD which should consider the context of the ambiguous word and use the information from its neighboring or collocation words. This information is gathered from text representation of knowledge source which is an unstructured source of information. To make it a suitable input to WSD, it is usually transformed into a structured format. Therefore, preprocessing of the input sentence is usually performed, which typically includes normalization, tokenization and stop-word removal.

Choice of a Classification Approach

Three main approaches applied in the field of WSD are knowledge based approaches, corpus based approaches and hybrid approach. Corpus based approaches can be divided into two types, supervised and unsupervised learning approaches. Supervised learning approaches use information gathered from training on a corpus that has sense tagged for semantic disambiguation. Unsupervised leaning approaches determine the class membership of each object to be classified in a sample without using sense tagged training examples. Hybrid approach combines aspects of the above mentioned methodologies. Knowledge based approaches use WordNet. It relies on information provided by Afaan Oromo WordNet developed by the researcher [16]. For this study we use knowledge based approach is used t o disambiguate words.

2.5 Afaan Oromo Language and Afaan Oromo Word Ambiguity

2.5.1 Afaan Oromo Language

Afaan Oromo is a Cushitic language spoken by about 30 million people in Ethiopia, Kenya, Somalia and Egypt and is the 3rd largest language in Africa [10]. Currently, it is the official language of Oromia Regional State which is the largest regional state among the current Federal States in Ethiopia. It is used by Oromo people, who are the largest ethnic group in Ethiopia and account for more than 40% of the population [10]. Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region. Moreover, a number of literatures, newspapers, magazines, educational resources, official documents and religious writings are written and published in Afaan Oromo [30, 31]. With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo [31].

2.5.1.1 Afaan Oromo Alphabet and Writing System

The writing system of Afaan Oromo language is straightforward which is designed based on the Latin script. Thus, letters in the English language are also in Oromo except the way it is written. Afaan Oromo text is written from left to right and spaces between words use as demarcation [32].

2.5.1.2 Punctuation Marks In Afaan Oromo

Words in Afaan Oromo sentences are separated by white spaces the same way as it is used in English. Different Afaan Oromo punctuation marks follow the same punctuation pattern used in English and other languages that follow Latin writing system. For example, comma (,) is used to separate listing of ideas, concepts, names, items, etc and the full stop (.) in statement, the question mark (?) in interrogative and the exclamation mark (!) in command and exclamatory sentences mark the end of a sentence[33].

2.5.1.3 Consonant and Vowel Phonemes

Like most other Ethiopian languages, Afaan Oromo has a set of ejective consonants, that is, voiceless stops or affricates that are accompanied by glottalization and an explosive burst of air. Afaan Oromo has another glottalized phone that is more unusual, an implosive retroflex stop, "dh" in Afaan Oromo orthography, a sound that is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins [33]. Afaan Oromo has the typical Southern Cushitic set of five short (a, e, i, o, u) and five long vowels, indicated in the orthography by doubling the five vowel letters (aa, ee, ii, oo, uu). The difference in length of vowels results in change of meaning.

For Example:

Afaan Oromo	**English**
Hara	lake
Haaraa	new

Gemination (doubling a consonant) is also significant in Afaan Oromo. That is, consonant length can distinguish words from one another.

For Example:

Afaan Oromo	**English**
Badaa	bad
Baddaa	highland

In Afaan Oromo alphabet, a letter consists either of a single symbol or a digraph (ch, dh, ny, ph, sh). Gemination is not obligatorily marked for the digraphs [33].

2.5.1.4 Afaan Oromo Morphology

Like in a number of other African and Ethiopian languages, Afaan Oromo has a very complex and rich morphology [31]. It has the basic features of agglutinative languages involving very extensive inflectional and derivational morphological processes. In agglutinative languages like Afaan Oromo, most of the grammatical information is conveyed through affixes, (that is, prefixes and suffixes) attached to the root or stem of words. Although Afaan Oromo words have some prefixes and infixes, suffixes are the predominant morphological features in the language. Almost all Afaan Oromo nouns in a given text have person, number, gender and possession markers which are concatenated and affixed to a stem or singular noun form. In addition, Afaan Oromo noun plural markers or forms can have several alternatives. For instance, in comparison to the English noun plural marker, *s (-es)*, there are more than ten major and very common plural markers in Afaan Oromo including: ***-oota, -oolii, -wwan, -lee, -an, een, -eeyyii, -oo,*** etc.). As an example, the Afaan Oromo singular noun ***mana*** (house) can take the following different plural forms: ***manoota (mana* + *oota), manneen (mana* + *een), manawwan (mana* + *wwan)***. The construction and usages of such alternative affixes and attachments are governed by the morphological and syntactic rules of the language [3]. Afaan Oromo nouns have also a number of different cases and gender suffixes depending on the grammatical level and classification system used to analyze them. Frequent gender markers in Afaan Oromo include ***-eessa/-eettii, -a/-ttii*** or ***–aa/tuu***.

Example:

Afaan Oromo	**Construction**	**Gender**	**English**

Obboleessa	***obbol + eessa***	male	brother
Obboleettii	***obbol + eettii***	female	sister
beekaa	***beek + aa***	male	knowledgeable
beektuu	***beek + tuu***	female	knowledgeable

Likewise, Afaan Oromo adjectives have case, person, number, gender, and possession markers similar to Afaan Oromo nouns. Afaan Oromo verbs are also highly inflected for gender, person, number, tenses, voice, and transitivity. Furthermore, prepositions, postpositions and article markers are often indicated through affixes in Afaan Oromo [33]. The extensive inflectional and derivational features of Afaan Oromo are presenting various challenges for a number of NLP tasks in the language.

Usually, WSD systems do not consider morphological variations of the context words. While this might not have any serious consequences for the performance of the algorithms for English, however, this approach may not work well for morphologically rich languages like Afaan Oromo. In such languages, an ambiguous word might occur in several morphological forms and hence, without morphological analysis it would be impossible, even to identify these forms as ambiguous word forms, for assigning the correct sense [37]. A morphological-analyzer reduces the different forms of an ambiguous word into their root forms and plays an important role in this regard.

2.5.2 Afaan Oromo Word Ambiguity

Ambiguity can be referred as the ability of having more than one meaning or being understood in more than one way. Ambiguity can occur at various levels of language processing. Ambiguity could be Lexical, Syntactic, Semantic, Pragmatic etc. [35].

According to [3], identifies different types of ambiguity in Afaan Oromo based on Getahun's works for Amharic [34] such as Phonological, Lexical, Structural, Referential and Semantic ambiguity. Below can be the summary of the ambiguity identified and explained in the following sub sections that I adopt from [3].

2.5.2.1 Phonological Ambiguity

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

Karaa + itti du'e / karaatti du'e

In the above sentence, "+" sign shows the place where the pause is occurred. When the sentence is pronounced with pause, it means "*the way he was killed*" but the meaning differs if it is pronounced without pause. It will mean "*He died on the road*" [3].

2.5.2.2 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 [36]. There are different factors that can cause lexical ambiguity such as Categorical Ambiguity, Homonymy and others. There are different factors that can cause lexical ambiguity such as Categorical Ambiguity, Homonymy and others.

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:

Barsiisan kutaa ***seena*** jira.

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

I. The teacher is getting into the class room. [With nominal meaning]

II. The teacher is in the history room. [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:

Tolaan ***ulfina*** gudda qaba.

In the above example the word "***ulfina***" is an ambiguous word having the following two different senses:

I. Tolaa has a huge weight

II. Tolaa is a respected person

2.5.2.3 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:

Barsiisa seena Ferensay

The above sentence can have two different interpretations:

I. A French man who teaches History.

II. A person who teaches French History.

The structural organization of the constituent words in the above sentence is:

Barsiisa[N] seena[N] Ferensay[N]

2.5.2.4 Referential Ambiguity

Referential ambiguities happen when a word or phrases in the context of a particular sentence refer to two or more properties or things. Usually the context tells us which meaning is intended, but when it doesn't we may choose the wrong meaning. If we are not sure which reference is intended by the speaker, we will misunderstand the speaker's meaning, as a result we assign the wrong meaning to the word [3, 38]. For example, ***Tolaan nama gudda dha (tolaa is a big man)*** you will have to guess whether *gudda* (big) refers to his height (**dheera** dha), his weight (**furdaa** dha), social status (**kabajamaa** dha) or something else. As another example:

Gaadisaan gatii ebifaamef gamade.

The above sentence has two different meanings:

I. Gadisa was pleased because he graduated.

II. Somebody was pleased because Gaadisa graduated

III. Gadisa was pleased because he offered blessing.

Referential ambiguities are usually easy to spot and once recognized are easily avoided [38].

2.5.2.5 Semantic Ambiguity

Semantic ambiguity is the phenomenon when a word has multiple meanings. It is caused by polysemic and idiomatic constituents. The following sentence is an example of polysemic constituent which has multiple meanings.

Abaabon lalisee gudate jira.

The above sentence has two interpretations:

I. The flower has grown.

II. Lalise‟s(name of a person)flower has grown.

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: **Inni dhiiga kooti.** The literal meaning of the above example is "*that is my blood*" but the idiomatic expression refers to "that is my relative".

CHAPTER THREE: RELATED WORKS

3.1 Introduction

As stated in chapter two word sense disambiguation(WSD) is a task of natural language processing that identify which sense of a word (i.e. meaning) is activated by the use of the word in a particular context in a sentence, when a word has multiple meanings. Thus, WSD is used in order to increase the success rates of NLP applications like machine translation, information retrieval, natural language understanding, language study and etc.[39] As a result of this in the next section we briefly describe various works done by researchers applying different approaches to words for WSD like knowledge based, corpus based and others.

Research work on WSD was first formulated as a distinct computational task during the early days of machine translation in the late 1940s, making it one of the oldest problems in computational linguistics. As a solution to ambiguity some of the researchers follow the supervised approach in which labeled training set is utilized, some of them follow unsupervised approach, which attempts to disambiguate a word without previous training, or labeled corpora. In knowledge-based approach, the algorithm uses the underlying meaning of the text to disambiguate a word. The task of disambiguation system is to resolve the lexical ambiguity of a word in a given context. Lexical ambiguity can be resolved by lexical category disambiguation i.e., parts-of-speech tagging. As many words may belong to more than one lexical category part-of-speech tagging is the process of assigning a part-of-speech or lexical category such as a noun, verb, pronoun, preposition, adverb, adjective etc. to each word in a sentence. Lexical ambiguity refers to two different concepts “homonymy” and “polysemy”. The distinction between bank (“river edge”) and bank (“financial institution”) has been used as an example of homonym, and rust (verb) and rust (noun) for polysemy [1].

In this chapter we conduct survey of past research in the area of corpus-based and knowledge-based word sense disambiguation. Knowledge-based disambiguation is carried out by using information from an explicit lexicon or knowledge base. The lexicon may be a machine readable dictionary, thesaurus or hand-crafted. On the other hand corpus based approaches disambiguate words based on training data obtained from corpus, rather than taking it directly from an explicit knowledge source. [40, 41, 42] use WordNet as the knowledge-base to disambiguate word senses, and [43] uses Roget's International Thesaurus.

3.2 WSD for Afaan Oromo Language

Tesfa K. [3] used supervised approach to solve the problem of word sense disambiguation (WSD) for Afaan Oromo language. He applied Naïve Baye's theory to find the prior probability and likelihood ratio of the sense in the given context for his experimentation. The system uses information gathered from training corpus to assign senses to unseen examples. The corpus he used contains 1240 sentences, and he evaluated for 5 Afaan Oromo ambiguous words namely ***sirna, karaa, sanyii, qophii*** **and** ***horii.*** By using these words he conducted two experiments.

Experiment one

During this experiment he tries to evaluate the performance of his algorithm; using 10-fold cross-validation. In this technique, first the total data set is divided into 10 mutually disjoint folds approximately of equal size using stratified sampling mechanism. Second, the training set and testing set was identified and separated from the total data set. In order to check the result using the developed system, he removed manually tagged sense examples from test set. Before doing the actual experiment, pretest has been done by the researchers using sense examples in test set and comparing the result with manually tagged test set. The pre-test has been conducted iteratively to increase prototype's performance. The errors encountered during this experimentation have been corrected and the experiment has been done iteratively until the result is found to be satisfactory. Finally, the actual test was conducted using sense examples in test set. During this process nine fold were used for training the developed system whereas the remaining tenth fold was used for testing the system that was trained on the previous nine folds. The process was repeated ten times by taking other nine as training and tenth one as testing. After each training phase, the system was tested on average of 124 Afaan Oromo sentence. Each of the corresponding training set contains an average of 1116 sentences. The result on test data set was obtained by comparing the result returned by the system with the corresponding test set which was manually tagged.

The second experiment

During this phase of experiment he sought to investigate the effect of different context sizes on disambiguation accuracy for Afaan Oromo ambiguous word, and find out, if the standard two-word window applicable for other languages and especially English holds for Afaan Oromo.

For the first experiment, he got 76% precision, 88% recall, 81% F1- measure and 79% accuracy. During the second experiment he concluded from his experiment a four-word window on each side of the ambiguous word is enough for Afaan Oromo WSD.

A hybrid approach used by[2] relies on the patterns learned from the corpus in combination with the rule based approach to solve the problem of word sense disambiguation (WSD). She applies machine learning because it becomes used to extract various contexts of the ambiguous words and their clustering and secondly she used a combined machine learning algorithm and rule based to selected Afaan Oromo ambiguous word like **Sanyii,Karaa,Ulfina, Ifa, Qophii, Sirna, Horii, Afaan, Bahe, Boqote, Darbe, Diige, Dubbatate, Tume, Haare, Ija, Ji'a, Dhahe, Mirga, Waraabuu.**

She has determined a set of contexts which are the most frequent words in the corpus with target words, by determining window size contexts to the left and to the right. After the context of ambiguous word determined vector space matrix from co occurrences is constructed. As a result of co-occurrence matrix, the cosine similarity was computed based on the angle between vectors of the contexts. The cosine values, which are computed, are clustered by Weka 3.7.9 package.

Yehuwalashet conduct two experiments as explained below using 20 ambiguous words to be discriminated, 12 words with 2 senses, 5 words with 3 senses, 1 word with 4 senses, and 2 words with 5 senses.

Experiment One

She conducts her first experiment on the Machine Learning Approach using different Context Window Sizes and clustering algorithms. She used window sizes of ±10 and different clustering algorithm like two clustering algorithms (EM and K-Means) and three hierarchical clustering algorithms (Complete link, Single link and Average Link). Those clustering algorithms directly use the vector representations (cosine similarity measure) of the contexts of the ambiguous word (extracted using rules or the window of words) as input. According to her experiment window sizes one-one and two-two perform a better result with EM and K-means algorithms. Smaller window sizes (1-1 and 2-2) have yield significantly higher accuracy than other windows, which are 80.6% and 78.3% with EM ,78.1% and 75.2 % K_means, 75.8 % and 74.6 % with complete link, 74.6 and 73.6 % with single link and 73.85% and 71.5 % with average link. Generally with this experiment the best accuracy is achieved by expectation maximization (EM) which results 80.6%.

Experiment Two

For second experiment she uses a Hybrid Approach which combines unsupervised machine learning and rule based approaches. In addition to the window size in unsupervised machine learning a hybrid approach adds modifiers using rules which are planted to the developed system for experimenting.

Unlike, unsupervised machine learning which uses window size only, the hybrid approach used rules to extract modifiers of the ambiguous word and consider them as contexts. These modifiers are therefore identified according to the developed rule planted. Similar to unsupervised machine learning, she has used the same test set, window size and clustering algorithms in the hybrid approach.

At the end of the experiment using the hybrid, most of the tested ambiguous words have relatively higher performance when compared with machine learning. As indicated in experiment results the hybrid approach achieves accuracy of; 90.35% and 86.28% with EM,86.6% and 83.55% with K-means, 83.85 % and 82.56% with Complete link, 81.6% and 80.26 % with Single link and 80.6 % and 79.1% with Average link by using window size of one-one and two-two respectively.

From the experiment she concluded that the results achieved by hybrid approach yields a better accuracy. The reason behind this enhanced accuracy might be because the hybrid method brings advantages of both methods the Machine Learning and rule based approaches.

3.3 WSD for Amharic

A Knowledge based approach used by [1] to solve the problem of word sense disambiguation (WSD) for Amharic language using Amharic wordnet that he develop manually which contains 10,000 synsets and 2000 words. To test WSD system he prepares a test sentence of 200 random sentences containing the ambiguous words from the knowledge base created.

The first experiment was conducted to measure to what extent morphological analyzer in the Amharic WordNet will affect the accuracy of WSD. As a result he conducted the experiment on Amharic WordNet with and without morphological analyzer since knowledge-based methods use information from an external knowledge source like Amharic WordNet that he develops. The second experiment is investigating the effect of different windows context sizes on disambiguation accuracy for Amharic to point out the optimal window size. He tested on window size of variant data sets starting from 1-left and 1-right to 5-left and 5- right window sizes. The experiment that he conducts can be explained below:

Experiment one: The Effect of Morphological Analyzer on the Accuracy of the WSD
In the first experiment he tries to conduct the experiment by using morphological analyzer and without using morphological analyzer to disambiguate sense of ambiguous word. As a result of this experiment when morphological analyzer used the accuracy becomes better than that of the experiment without morphological analyzer. This accuracy difference comes because morphological analyzer reduces various forms of word into their common root or stem word. As result 80% accuracy becomes achieved using morphological analyzer on the other hand 57.5% accuracy is achieved without morphological analyzer.

Experiment two: Determining Optimal Context Window
On the second experiment context windows of various size starting from 1-left and 1-right to 5-left and 5- right becomes used to select best window size for knowledge based approaches. As a result of his experiment using various windows size, for the knowledge based Amharic word sense disambiguation the maximum accuracy of 86.5 % on two-two word window size becomes achieved.

3.4 WSD for English Language

Word sense disambiguation by [46] was implemented for English language using a knowledge-based approach. The authors propose a robust knowledge-based solution to the word sense disambiguation problem for English language. The solution to sense ambiguities are based on both knowledge referred by the context of the sentence and the grammatical knowledge of the natural language. Two phase word-sense disambiguation solution are applied. In the first phase all the possible knowledge objects corresponding to each term in the given sentence becomes identified. During this phase morphology rules are taught to the system to convert words into their base forms, which happen at the first step of parsing a sentence. The second phase is responsible for resolving the ambiguity among all possibilities to correctly identify the intended meaning. Because of the ambiguity reason the researcher classify ambiguities into two categories. One category may be resolved based on the grammar's requirement that a certain *pos* be at a specific place of the given sentence, and hence it becomes resolved during the parsing stage of a sentence. The other category is resolved during the understanding of the thought, which uses the context information available from the rest of the sentence. The researcher uses POS and contextual information found in the sentence. Therefore, resolving an ambiguous word based on the word's POS is possible when the parse tree is unambiguous. However, problems may arise when multiple parse trees can be formed due to the absence of an optional term and the presence of a term with an ambiguous POS.

Unsupervised word sense disambiguation researched by [48] using WordNet relatives like synonyms, hypernyms and hyponyms used to disambiguate polysemous target noun. The researcher uses the context words surrounding the target noun. The sense of a word in a context is determined by selecting a substituent word from WordNet relatives of the word. The selection of substituent word is based on the co-occurrence frequency between the relative words surrounding the target word in a given context. Generally the proposed method disambiguates senses of words through the set of WordNet relatives of the target words and a raw corpus. Lastly, the proposed system is tested on 186 documents in Brown Corpus and achieved 52.34% of recall and the researchers do not consider a way to utilize the similarity between definitions of words in WordNet.

CHAPTER FOUR: DESIGN OF AFAAN OROMO WORD SENSE DISAMBIGUATION

For word sense disambiguation in any language design or architecture must be there based on the behavior of the languages. Thus, we proposed architecture for Afaan Oromo word sense disambiguation. In the subsequent section we clearly describe knowledge based WSD for Afaan Oromo language. So, this chapter focuses on architecture of Afaan Oromo WSD, Afaan Oromo WordNet, design requirement, and prototype. In addition to this, the detail description of components on the architecture and their algorithms are also presented.

4.1 Architecture of Afaan Oromo WSD

The proposed architecture for Afaan Oromo word sense disambiguation system is composed of the following essential components

- Preprocessing component,
- Morphological analysis component,
- Afaan Oromo WordNet (**OROWORDNET**) database,
- Disambiguation component.

The architecture shows the overall functionality of Afaan Oromo Word Sense Disambiguation system. The system takes Afaan Oromo sentence as an input and identifies the ambiguous words and its sense from **OROWORDNET**. The sentences are preprocessed to make suitable for further processing. Morphological analysis is important for morphologically complex languages like Afaan Oromo since it is difficult to store all possible words in WordNet database. As a result of this morphological analysis is used for reducing various forms of a word to a single root word. Morphological analysis produces root word and provides the root word-to-word sense disambiguation component particularly to the ambiguous word identification. The **OROWORDNET** designed contains Afaan Oromo words along with their different meanings, Synset and semantic relations within concepts. This component helps to implement the components of WSD. Word sense disambiguation component is responsible to identify the ambiguous word and to assign the appropriate sense to ambiguous word. To accomplish this, it incorporates various components such as Ambiguous Word Identification, Context Selection, Sense Selection and Sense retrieval components. Those components of Afaan Oromo WSD are explained in the next subsections.

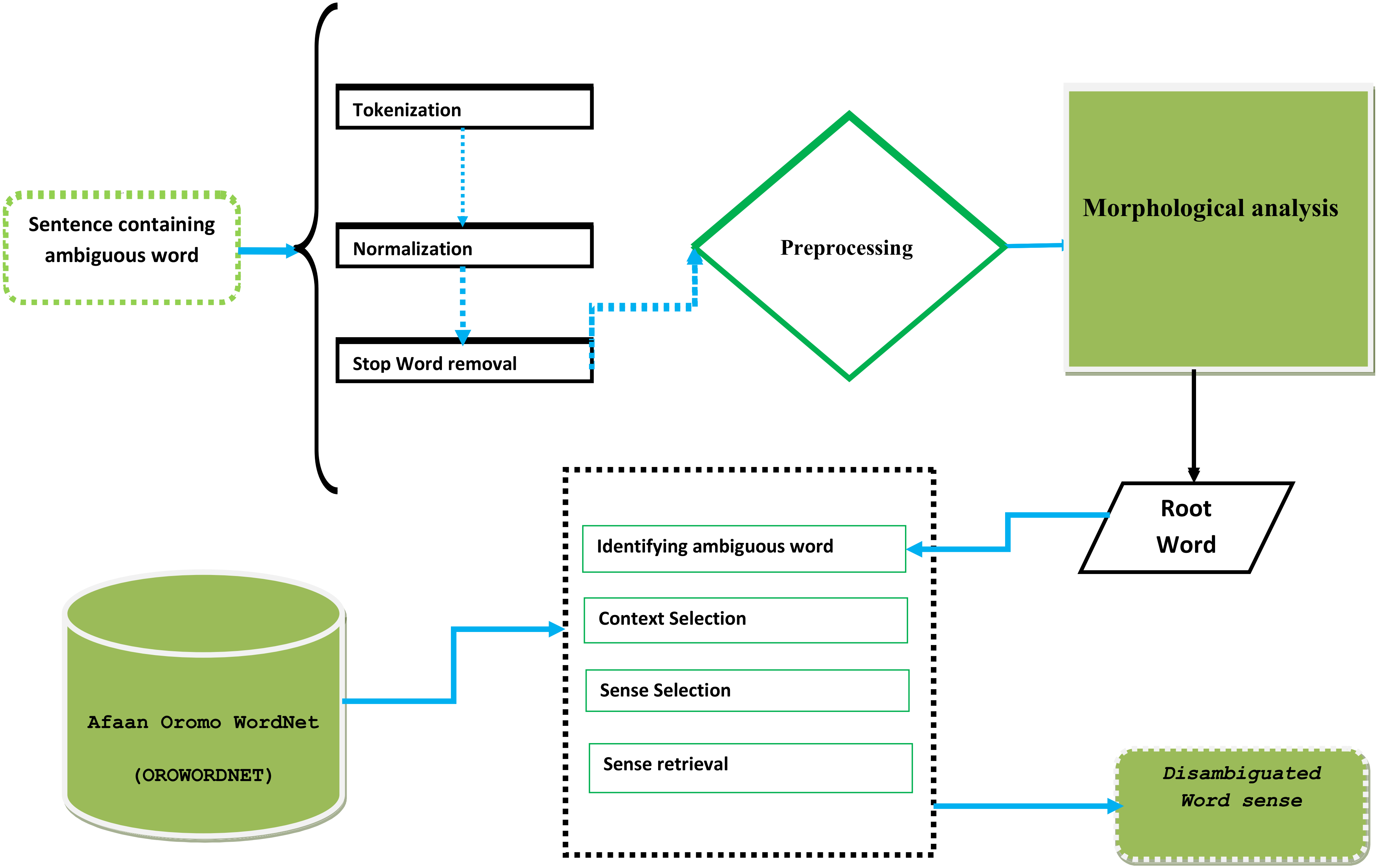


Figure 4.1a: Detailed Architecture of Afaan Oromo Word Sense Disambiguation (Adopted from [3, 15])

4.2 Preprocessing

We preprocessed our input sentence to make our model disambiguate ambiguous word for us. Thus the pre-processed sentence becomes ready for other components of our model. Our pre-processing consists of stages like tokenization, normalization and stop word removal.

4.2.1 Tokenization

Tokenization is a process which splits up the text into a set of tokens usually words, based on the boundaries of a written text [14]. Tokenizing of a given text depends on the characteristics of the language of the text in which it is written [14]. Word demarcation in Afaan Oromo is handled following space. Thus, Afaan Oromo tokenizer parses text into its constituent words usually by considering the space and punctuation mark. Punctuation mark usage in Afaan Oromo is similar to that of English which include semicolon (;), comma (,), full stop (.), question mark (?) and exclamation mark (!). These punctuation marks are removed from the text because they don't have any relevance in identifying the meaning of ambiguous words in WSD [3].

4.2.2 Stop Word Removal

Stop word removal is used to remove stop words from the input text [14]. Every language has its own list of stop words: words that have no significant discriminating powers in the meaning of ambiguous words like "**yommuu, booddee, isaa, keetii, saniif**". Stop words mainly consist of prepositions (**irra, irraa, itti, and jala**), conjunctions (**fi, garuu, immoo, yookin, moo, kanaaf, and kanaafu**) [3]. These words need to be removed during preprocessing phase. There are various techniques used to remove stop words. Among this IDF (inverse document frequency) value and dictionary lookup are the common one [15]. The IDF approach assumes words that appear in many documents as stop words. However, most of the existing stop words removal techniques are based on a dictionary lookup that contains a list of stop words [15]. This technique is much easier for well studied languages that have standard list of such words. As a result of this, dictionary lookup was employed for this study. For the purpose of this research work, list of around 100 stop words that is compiled from Afaan Oromo books during implementation of a stemmer by [44] is used. The algorithm is described in Figure 4.3.

```
1. Open stop word list
2. Read a sentence
   For each word in the sentence
           If a word is in stop word list then
           Remove a word
           End if
    End for
3. Return the remaining words
4. Stop processing
```

Figure 4.3: Algorithm for stop word remover

4.2.3 Normalization

In Afaan Oromo the some characters of the same words are sometimes represented in uppercase or lowercase in the sentence as well as in the user input and hence we have normalized them into lowercase. The purpose of normalization in our case is to make similar the words in different cases in our corpus. Given that to get at the meaning that underlies the words, it seems reasonable to normalize superficial variations by converting them to the same form. The most common types of normalization are case folding (converting all words to lower case). Case folding is easy in Afaan Oromo for example *Horii* similar to *horii*.

4.3Morphological Analysis

Morphological analysis is the process of segmenting words into morphemes or analyzing the process of word formation [50]. It is a primary step for various types of text analysis of any language. Morphological analyzer takes a word as an input and produces the root and its grammatical features as the output.

For Example:

Input: **horii**

Output :{ **POS**: verb, **root**: <hor>, singular}

Morphological analysis is a very significant step towards efficient natural language processing for highly inflectional languages like Afaan Oromo. Morphology is one of the complementary parts of the structural aspects of natural language expression. Afaan Oromo root words can generate hundreds of lexical forms of different meanings. The Afaan Oromo language makes use of prefixing, suffixing and infixing to create inflectional and derivational word forms. In morphologically complex language like Afaan Oromo, a morphological analysis will lead to significant improvements in WSD systems. In this thesis, we used Hornmorpho morphological analyzer developed by [50].

4.4 Afaan Oromo Wordnet

In Afaan Oromo WordNet, the words are grouped together according to their similarity of meanings like "**horii, qarshii**" to mean **money**. Afaan Oromo WordNet is a system for bringing together different lexical and semantic relations between the words. We follow the principle of English wordnet to develop Afaan Oromo wordnet for Word Sense Disambiguation system. The structure of WorldNet's becomes a useful tool for computational linguistics and natural language processing [21].

Structure of Afaan Oromo WordNet

In natural language processing system Synonyms are words that denote the same concept and are interchangeable in many contexts and are grouped into unordered sets (synsets) like "**bukkee, maddii**". Two words that can interchange in a context are synonymous for each other in that context. For each word there is a synonym set, or synset in Afaan Oromo WordNet, representing one lexical concept. This is done to remove ambiguity in case there exist a single word having multiple meanings. Synsets are the basic building blocks of WordNet. In our wordnet we use 5 hierarchies of POS" **gochima, maqaa, addeessa, maqdhala, dabal gochima**" i.e verb,noun,adjective,pronouns respectively. Every Synset is described by a brief gloss definition. Synsets in WordNet are connected by relations, which can be categorized into two kinds. The introduction of a "frequently used" or "highly expected" field in the synset structure of WordNets can scale-up the efficiency in determining winner sense of a polysemous word, as these highly related words will enrich the sense bag with more information, thereby enhancing the chances of appropriate overlap. WordNet defines the relations between synsets and relations between word senses. A relation between synsets is a semantic relation, and a relation between word senses is a lexical relation. The distinction between lexical relations and semantic relations is somewhat subtle. The difference is that lexical relations are relations between members of two different synsets, however semantic relations are relations between two whole synsets.

The following are semantic relations:

- Hypernymy(generalization) and Hyponymy(specialization)
    - Relation between word senses(synsets)
    - X is a hyponym of Y if X is a kind of Y
    - Hyponymy is transitive and asymmetrical
    - Hypernymy is inverse of Hyponymy
- Meronymy and Holonymy
    - Part-whole relation, branch is a part of tree
    - X is a meronymy of Y if X is a part of Y
    - Holonymy is the inverse relation of Meronymy

Lexical relations are:

- Antonymy

- Oppositeness in meaning
- Relation between word forms

Our Afaan Oromo database has seven basic tables that are Words, POS, Synset, Domain,Link type, lexicalRelations and Semantic Relations. This kind of Afaan Oromo WordNet Structure is adopted from English WordNet [21].

Afaan Oromo WordNet Database Schema

Our Afaan Oromo wordnet is developed based on the principle of coverage and minimality. Our schema contains SYNSET, WORD, POS, DOMAIN, LINKTYPE, LRELATION, SRELATION tables. WORD table maintain the unique words of Afaan Oromo language. We use this table to identify ambiguous words and get its definition through WID which has relation with SYNSET through the WID in SYNSET. SYNSET table is used to maintain the details of a synset (concept in a language). A synset (or concept) has a gloss and synonym word set. The purpose of this table is to maintain concepts which are used to describe a sense of words. LRELATION table is used to maintain the lexical relations with respect to the SYNSET. POS table used maintains the part of speeches such as Noun, Adjective, Adverb and Verb of the language. SRELATION is used to maintain the semantic relation like Hyponyms, Hypernym, Holonym, and Meronym between pair of synsets/concepts, which is a IS-A-KIND-OF/ PART-WHOLE type of a semantic relationship between synsets. DOMAIN table is used to maintain the source from which a concept or synset has been taken or belongs like finance concept, medical concept, agriculture concept, technology concept, language specific concept. Figure 4.2 shows the database schema of Afaan Oromo WordNet.

In our database the word “mirga” has a WID of 1002 having five senses. Since this word has five senses, the synset table holds gloss of those senses of the word under different SYNID with a gloss

2004. Maddii yookiin bukkee (cinaa) dhaqna namaa, Kan yammuu gara kaabaatti yookiin boroottii garagalan gara baha biiftuu oolu

2005. Bineensa akka arbaa, leencaa ajjeesanii faachii irraa fudhatamu

2006. Uumaman waan namni gochuu danda'uu

2007. Waan hojjetan irratti olaantumma namni sun qabu

Below is an example sentence containing ambiguous word. “**Tolaan karaa mirga koo dhaabate.”** from this sentence the ambiguous word is **mirga** and **karaa** is the context used to identify the sense of the ambiguous word, the synset of “**Mirga**” is indicated in the above gloss, and synset of “**karaa**” is “**lafa namni yookiin konkolaatan irra deemu, kan bakka tokko qabee bakka biraatti nama geessu**”,”**akkaataa, haala**”,”**roga dhimma nama ilaallatu**". In addition to sense overlap we use ontology to identify the context of the ambiguous word “**Mirga**” so that, the ontology of the word “**Karaa**” is **kallatti,cinaa,maddii.**

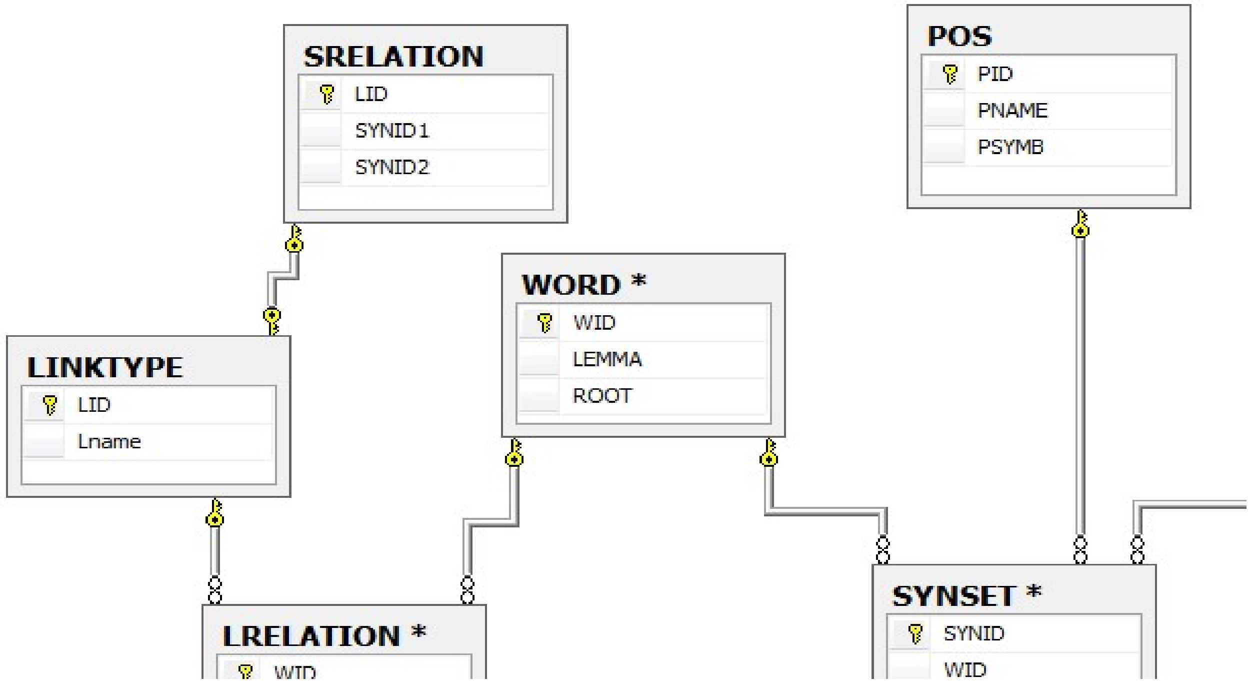


Figure 4.3 Afaan Oromo WordNet Database Schema Structure

4.5 Word Sense Disambiguation (WSD)

Word sense disambiguation is a main component of the Knowledge-Based Afaan Oromo Word Sense Disambiguation which contain Ambiguous Word Identifier, Context Selection, Sense Selection and Sense Retrieval components as discussed below:

4.5.1 Ambiguous Word Identifier

For each sentence as input it is disambiguated separately after preprocessed, starting with the first word of the sentence and working left to right. At each stage, the word being disambiguated is called the target word, and the surrounding words form the context window. Thus, the Ambiguous word Identifier is a component used to identify the ambiguous word from the input sentence using knowledge sources of Afaan Oromo WordNet Database. As a result Afaan Oromo WordNet identify the ambiguous word which contains more than one sense for a given words from the input sentence and check the existence of root word from the database. If words do not exist in database the word is discarded. For example, if the following sentence is the input sentence: "**Tolaan karaa mirga koo dhaabate.**" First, the input sentence is preprocessed. After morphological analysis, only four words will be left (i.e tol,karaa,mirg,dhaab) in the input sentence. Then each root word with respect to its sense is checked in the database. We found mirg in our case. So that, "mirg" is detected as ambiguous word in the input sentence and "mirg" is the root word for the word "mirga". As a result, "mirga" is ambiguous word and their sense is retrieved from Afaan Oromo WordNet database based on the context of the sentence. The algorithm is shown in Algorithm 4.5.1

```
Input: sentence
          Preprocess the sentence
          Perform morphological analysis of words
          Index=0
Read root words from Afaan Oromo WordNet
For words in array of buffer
          If root word does not exist in Afaan Oromo WordNet
                    Discard the word
                    Else add word to Array of buffers
                    Index++
          End IF
End For
Return Ambiguous word
Stop
Output: Ambiguous word
```

4.5.1: Ambiguous Word Identification Algorithm

4.5.2 Context Selection

Context selection is used to select the context (sense example) that also contains the ambiguous word. The disambiguation works involve matching the context of the word to be disambiguated with information from Afaan Oromo WordNet. For example the sentence "**Tolaan karaa mirga koo dhaabate.**" after morphological analysis, it becomes "**tol,karaa,mirg,dhaab**". Based on ambiguous word identifier the ambiguous word is **mirga** and the context are words surrounding the ambiguous word "**tol,karaa,dhaab**". The correct sense of a word is obtained from the context of the sentence. This component uses the words of the sentence itself as context, including ambiguous words and selects the context that contains ambiguous words from Afaan

Oromo WordNet. In other words, we can say that context uniquely identifies meaning of the sentence. Based on this interpretation, the ambiguity of word known as lexical ambiguity is disambiguated. Below is algorithm of context selection:

```
Input:  sentence

Preprocessing Afaan Oromo Sentence

Read root word w[i] from morphological analysis

Open Afaan Oromo WordNet

Find the root word from WordNet for Afaan Oromo

For each root word in the sentence

     If W[i] is ambiguous word

     Find the sense and ontology of words in Afaan Oromo
     WordNet
     Extract the sense of the ambiguous word
     Else
     Assign empty value to array buffer
 End for
     If end of Afaan Oromo WordNet not reached
     Read root word in Afaan Oromo WordNet
     Else
     Return sense and related words
      End if
Return sense and related words
Stop
Output: root word + sense of ambiguous word
```

4.5.2: **Context Selection Algorithm**

4.5.3 Sense Selection Component

Sense selection is a component used to identify the possible senses of ambiguous word in the given input sentence. A word sense is a commonly accepted meaning of a word. Our sense selection component is based on Lesk assumption [15]. We find the number of overlapping of the words from the set of words output by the context selection component with the sense of ambiguous word. As a result words having highest overlapping are selected as the senses of the ambiguous word. The senses of all words are search from Afaan Oromo WordNet. For example in the sentence "**Tolaan karaa mirga koo dhaabate.**" "**karaa**" is a context used to differenciate the meaning of the ambiguous word "**mirga**" for this sentence. So, the sense overlap of "**karaa**" and the ambiguous word "**Mirga**" are selected from Afaan Oromo WordNet. The algorithm for sense selection is adopted from [15]

```
Input:  root word
For every word w[i] in the sentence
    Let overlap= 0
    let BEST_SENSE = null
    Open Afaan Oromo WordNet
    Assign words in to array of Buffer
    Read root words from Afaan Oromo WordNet
 For every sense sense[s] of root word w[i]
  If word[i] is ambiguous word
      let maxoverlap = 0
    For every other word w[k] in the sentence, k != i
   overlap = overlap + number of words that occur In the gloss of both sense[j] and sentence
End for
    IF overlap> maxoverlap
    maxoverlap = overlap
    BEST_SENSE = w[i]
    End IF
     End for
     IF maxoverlap > 0
     Extract BEST_SENSE
          Else
     Output "Could not disambiguate w[i]"
    End If
End for
Stop
```

4.5.3: Context Selection Algorithm

CHAPTER FIVE: EXPERIMENT AND EVALUATION

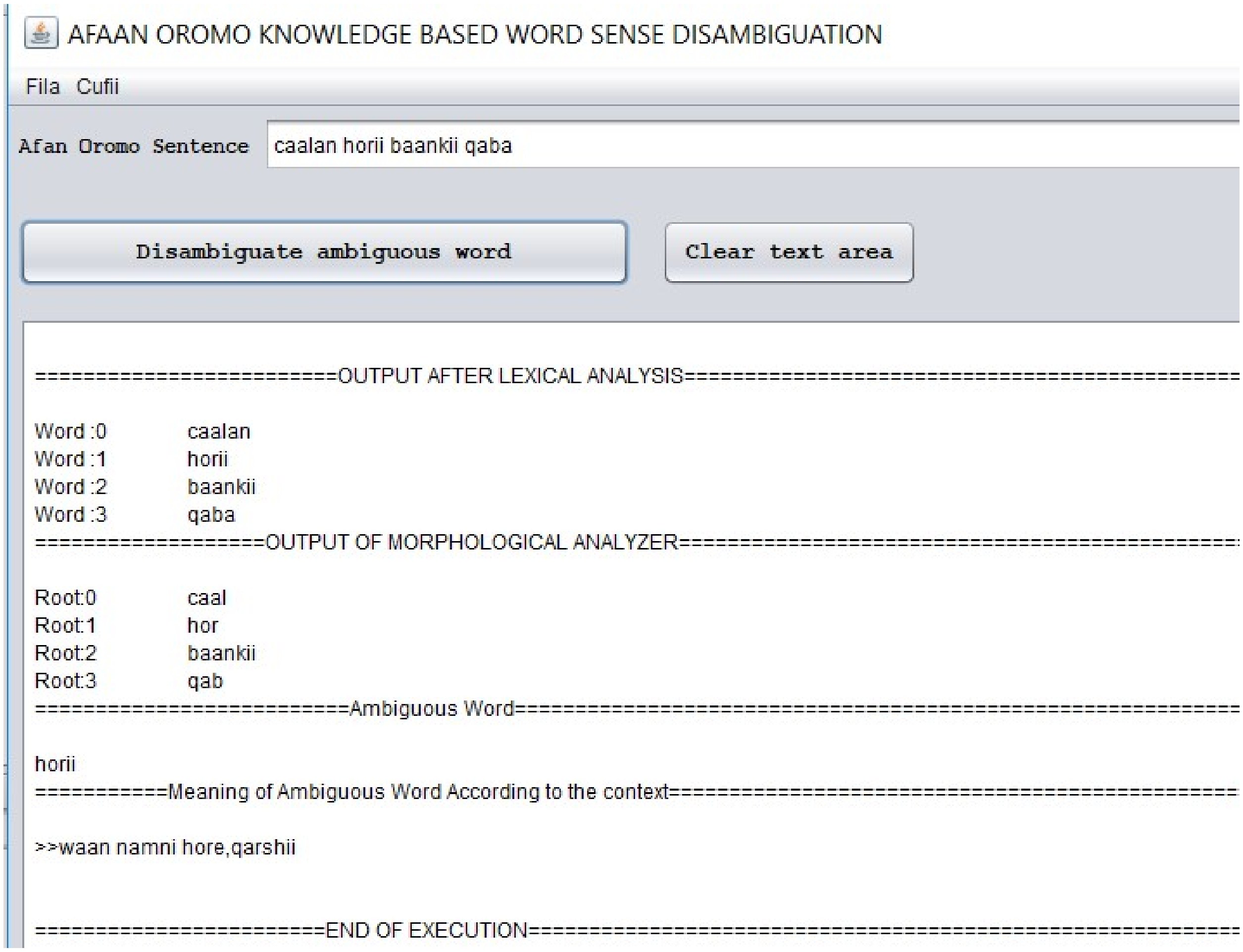
5.1 Introduction

Evaluation plays an important role to determine the accuracy of any system. For our study we select knowledge based word sense disambiguation as discussed in the previous chapter. We developed Afaan Oromo WordNet from scratch by collecting Afaan Oromo words from Afaan Oromo Dictionary. As we show the sample in the appendix our WordNet is developed from 100 ambiguous Afaan Oromo words and 267 Synsets. To develop the prototype we use python and java. We use python to extract root words of Afaan Oromo from Hornmorph as discussed previously. We use Java as it is dynamic in nature and can be run in any platform. Microsoft SQL server 2008 is used to develop Afaan Oromo WordNet. We perform an experiment to evaluate the performance of our system. However it is very difficult task since there is no standard rule while conducting the evaluation of WSD for all languages.

We use Hornmorph for morphological analysis. So, we conduct our experiment on Afaan Oromo ambiguous words using morphological analysis and without morphological analysis to see the effects of morphological analyzer for Afaan Oromo. Other researcher like [2, 3] conducts experiment by using various windows size of their sentence to disambiguate ambiguous word by employing different research methodology than that we are using for this research. So we also conduct our experiment using various window sizes. We collect fifty sentences from different Afaan Oromo documents and sites for our experiment, like news papers, bible and sites to test our system. Afaan Oromo Word sense disambiguation system takes the sentence as input and process each sentence one at a time. The meaning of the word and the target word is searched from Afaan Oromo WordNet that we developed. We have followed some set of procedures to conduct the experiment. The test environment, the set of activities defined under the procedures, and findings of the experiment are described in detail in the following sub sections.

5.2 The Prototype

Developing a prototype to demonstrate the usability of the proposed Afaan Oromo Word Sense Disambiguation is one of the objectives of this study. Hence, we developed the prototype of Afaan Oromo WSD using python and Java programming language. The main screen of the system is depicted in Figure 5.1 which shows the result of disambiguated sense of the ambiguous word for the given input sentence.


AFAAN OROMO KNOWLEDGE BASED WORD SENSE DISAMBIGUATION
Fila Cufii
Disambiguate ambiguous word
Clear text area

Afan Oromo Sentence: caalan horii baankii qaba

```
========================OUTPUT AFTER LEXICAL ANALYSIS============================================

Word :0        caalan
Word :1        horii
Word :2        baankii
Word :3        qaba
==================OUTPUT OF MORPHOLOGICAL ANALYZER================================================

Root:0         caal
Root:1         hor
Root:2         baankii
Root:3         qab
==========================Ambiguous Word=============================================================

horii
===========Meaning of Ambiguous Word According to the context==========================================

>>waan namni hore,qarshii


=======================END OF EXECUTION==============================================================
```

Figure 5.1: Screenshot of word sense disambiguation

5.3 Evaluation Metrics

For Afaan Oromo word sense disambiguation we use evaluation metrics for measuring the rate of disambiguation of our system, the most common evaluation techniques, which select a small sample of words and compare the results of the system with a human judge. We use the metrics such as precision P, recall R, F-measure and accuracy. The evaluation criteria were based on the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP counts the number of words that are recognized by WSD system and are found in the test data. TN counts the number of words that are not recognized by WSD system and are found in the test data. FP counts the number of words that are wrongly recognized by WSD system; however, they are not in the test. FN counts the number of words that are left unrecognized by WSD system; however, they are in the test data. Therefore, below is a formula that we apply to get precision, recall, F-measure and Accuracy

$$\text{Precision (P)} = \frac{TP}{TP+} \text{ ------------------------1}$$

$$\text{Recall (R)} = \frac{TP}{TP+} \text{ -------------------------2}$$

$$\text{F-Measure} = \frac{2*P*R}{P+R}, \text{ P+R} \neq 0 \text{ ------------------3}$$

$$\text{Accuracy} = \frac{TP+}{TP+FP+TN+FN} \text{ -----------------------4}$$

5.4 Test Results and Discussion

Word sense disambiguation in Afaan Oromo is conducted using different approaches. As a result, for each approach there should be a testing mechanism to verify the results of WSD. For our thesis we use knowledge bases to disambiguate ambiguous word which does not rely on manually or automatically generated data set as discussed in our previous chapters. In this study we conduct two experiments the first one is conducted by using morphological analysis i.e. experimenting the ambiguous word using morphological analyzer and without morphological analyzer to analyze the effect of Afaan Oromo WordNet on the accuracy of WSD. The second experiment is the effect of the context window size. We perform our experiment of context windows based on recommendation provided by different researchers on Afaan Oromo WSD like [2, 3]. Both researchers use different approaches of ours. Window size of 1-1 and 2-2 for word sense disambiguation is recommended by [2] and window size of 4-4 is recommended by [3]. Since our approach uses knowledge base we conduct the experiment using window size of 1-1 to 5-5 to the right and left of ambiguous word.

First Experiment: the consequence of Morphological analyzer on the accuracy of the WSD

For morphologically complex languages like Afaan Oromo morphological analysis improve word sense disambiguation as we discussed in the previous chapter. Thus, we use morphological analyzer to check whether the performance of WSD becomes improved or not. Linguistic resources and various WSD algorithms can be used for evaluation purpose. The linguistic resources are Afaan Oromo WordNet without morphological analyzer and Afaan Oromo WordNet with morphological analyzer. We used total number of 50 sentences to evaluate the performance of the system.

Experiment Without Morphological analyzer:

While we are conducting this experiment the result is obtained by comparing the instance of the input sentence with context build in to Afaan Oromo WordNet. As a result the system determines words that were correctly disambiguated and words that have wrong sense.

Experiment With Morphological analyzer:

During this experiment the importance of morphological analyzer is shown as morphological analyzer bring various forms of the some word to the some root. Ontology based related word and sense overlap is used to identify the context of the ambiguous word in the instance and ontology based related word is created manually due to lack of linguistic resources. We apply morphological analyzer WSD the accuracy of the system increases. As a result the system determines words that were correctly disambiguated and words that have wrong sense.

Afaan Oromo WordNet	Recall	Precision	F-Measures	Accuracy
Without morphological analyzer	62.5%	56.53%	58.35%	50.75%
With morphological analyzer	75.75%	69.78%	71.6%	63.95%

Table 5.1 ***Performance of WSD system with and without Morphological Analyzer***

As discussed previously morphological analyzer is very important for morphologically rich languages like Afaan Oromo to disambiguate ambiguous words. We verify that as shown in the above table increase of accuracy come because of the morphological analyzer. The context of word is found by determine the meaning of a word using domain based related words and the overlap of the sense of target word to each words, we achieved an accuracy of 63.95%.-Even though morphological analyzer reduces various forms of a word to its root forms the Hornmorph that we use for this research does not perform as expected for all words of Afaan Oromo. But we use these tools for morphological analyzer since there is no publically available resource as of my knowledge. The other problems encounter us during experiment is the tool we use does not contain all forms of the word. For example "**mirga**" to mean "right" is Afaan Oromo ambiguous word. According to Hornmorph its root form is “**mirg**” but if we want to find “**mirgaalee**” which is inflected form of the word “**mirga**” we cannot find from Hornmorph while performing the analysis. Lastly, the problem we encounter during experiment is that different ambiguous word can be stemmed to the some roots which impose a challenge to disambiguate the words. For example “**qaroo**” and “**qarree**” to mean *part of eye to see* and *sliding land* respectively can be stemmed to the some root word “**qar**”

Second experiment: By determining the context of window size

Window selection is the process of selecting words from the text containing the target word to the right and left of target word. These words are used for weighting the possible senses along with the knowledge data extracted from the knowledge base. In English, a standard two-word window on either side of the ambiguous word is found to be enough for disambiguation [52].

For Afaan Oromo WSD using supervised machine learning techniques by Tesfa K. [3] on five ambiguous word ***sanyii, karaa, horii, sirna and qoqhii***, he recommends four window size on both sides for the ambiguous word is found to be enough For Afaan Oromo. On the other hand Yewalashet B. [2] recommends window size of two using Hybrid Word Sense Disambiguation approach for Afaan Oromo Words **Sanyii , Karaa,Ulfina, Ifa,Qophii, Sirna,Horii,Afaan, Bahe, Boqote,Darbe, Diige, Dubbatate,Tume,Haare,Ija, Ji’a,Dhahe ,Mirga,Waraabuu.** As of my knowledge from reading various research paper this experiment is not conducted for Afaan Oromo using knowledge base like WordNet. So we conduct an experiment starting from windows size one to four for some ambiguous words and propose window size for knowledge bases.

Window size(N)	precision	Recall	F-Measure	Accuracy
1	73.5%	70.25%	73.28%	71.51%
2	72.51%	71.51%	76.51%	75.51%
3	68.23%	75.23%	74.46%	80.54%
4	63.7%	76.38%	75.54%	73.51%

Table 5.2: Experiment in different window sizes

By using Afaan Oromo WordNet for WSD that applies window of various sizes as shown above the maximum accuracy, precision and recall achieved 80.54 %, 68.23% and 76.38% on three-three word window size respectively. This shows Smaller window sizes lead to higher precision, while bigger window sizes lead to higher coverage at the cost of some precision. A greater window leads to better recall, though precision is decreased slightly. To get the sense of target polysemous word we define a window size around the target polysemous word and calculate the number of words in that window that overlap with each sense of the target polysemous word. Knowledge Based methods do not face the challenge of new knowledge acquisition since there is no training data required.

During the course of our research work we try to model word sense disambiguation to disambiguate the ambiguous word in a given sentences and check the results of our research by experimenting. As a result of this, we design Afaan Oromo WordNet. We use the designed Afaan Oromo WordNet and disambiguate ambiguous words when the word comes in different sentence. This model is generally used by any stake holders who want to know the meaning of polysemous word. Lastly, our work is used as input for various natural languages processing like information retrieval, information extraction and others when it is integrated.

CHAPTER SIX: CONCLUSION AND RECOMMENDATION

This research work is the first attempt to develop a word sense disambiguation system for Afaan Oromo Language using knowledge base. As there is no lexical knowledge available for Afaan Oromo we have constructed our WordNet by collecting resources from various Afaan Oromo Dictionaries. During the course of our study we reviewed works on WSD systems developed for local and non local languages to get clear information for our research work.

6.1 Conclusion

In Afaan Oromo there is much ambiguous word in which there meaning is changing with the context. This creates the user of the language to be confused about the meaning of those words. Our research work is based on a knowledge base as a source of information. As a result, we developed our WordNet manually from various Afaan Oromo documents like dictionary. We stored various ambiguous words in our database with their meaning. Based on those ambiguous words stored in our database we extract various sentences from newspapers and other documents by the help of language experts for testing our research work. A model of our word sense disambiguation contains: preprocessing, morphological analysis, Afaan Oromo WordNet database and word sense disambiguation phase to disambiguate ambiguous word in the sentence. Our model takes sentence as an input to process the sentence and show ambiguous word along with its meaning to the end user. During preprocessing stage, it segments the input sentence by using tokenization and removes stop words from the input sentence and print to text to be used by Hornmorph for morphological analysis. By considering the morphological variants of the language, morphological analyzer extracts the root word and print to text file to be used by the system for the next phase. After gathering information in the morphological analyzer step, the system uses the remaining words in the input sentence as context, which used ontology based related words and overlap features to identify the sense of ambiguous words. Disambiguation component is used to identify the ambiguous word and its sense based on information found in Afaan Oromo WordNet. Then, the system identify the context of ambiguous word using Ontology based related words and overlap of the sense of target word to each words and decides the most appropriate sense for a given ambiguous word in the input sentence. Two experiments are conducted. Those experiments are: first the use of Afaan Oromo wordnet with and without morphological analyzer and the second one is determining an optimal windows size for Afaan

Oromo WSD. The uses of morphological analyzer and without morphological analyzer have achieved an accuracy of 63.95% and 50.75% respectively. For the second experiment, there is no standard optimal context window size which refers to the number of surrounding words that is sufficient for extracting useful disambiguation. From the result of experiment three-three window on each side of the ambiguous word is enough for Afaan Oromo WSD.

During the course of our research work we try to model word sense disambiguation to disambiguate the ambiguous word in a given sentences and check the results of our research by experimenting. As a result of this, we developed Afaan Oromo WordNet. We used the developed Afaan Oromo WordNet and disambiguate ambiguous words when the word comes in different sentence. This model is generally used by any stake holders who want to know the meaning of polysemous word. Lastly, our work is used as input for various natural languages processing like information retrieval, information extraction and others when it is integrated.

6.2 Recommendations

The underlying hypothesis of the technique used in this thesis is that context based related words and sense overlap shows us about the intended meaning of a word. Thus for an accurate disambiguation, selecting the appropriate context is essential. Word sense disambiguation researches require variety of linguistic resources like thesaurus, WordNet and Machine Readable Dictionaries which is a challenge for Afaan Oromo. The other challenge we faced was lack of organized ambiguous word for evaluation and development of WordNet. Therefore, we forward the following recommendations for Afaan Oromo WSD texts:

- Researches in WSD for other language use linguistic resources like thesaurus and machine readable dictionaries. For Afaan Oromo those resources are not yet been developed. We recommend those resources to be included in the future work.
- We recommend well organized development of WordNet to be used as a knowledge base for WSD.
- The context with which the ambiguous word is expected to come with is developed manually to identify the meaning of ambiguous word in the sentences. Thus, we recommend the development of these resources.

- The system developed in this research work is just a prototype. Any interested body can do a project to make a full-fledged Afaan Oromo WSD that can be easily integrated into different Afaan Oromo NLP works such as machine translation, information retrieval, information extraction and speech.

REFERENCES

[1]. Segid Hassen Yesuf (2015). Amharic word sense disambiguation using wordnet: published Master’s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.

[2]. Yehuwalashet Bekele Tesema (2016).Hybrid Word Sense Disambiguation Approach for Afaan Oromo Words: published Master’s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.

[3].Tesfa Kebede Hundesa (2013). Word sense disambiguation for Afaan Oromo Language: published Master’s Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.

[4]. Satanjeev Banerjee(2002). Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet, Department of Computer Science, University of Minnesota, Duluth, Minnesota 55812 U.S.A.

[5]. Stefan Bordag . Sentence Co-occurrences as Small-world Graphs: A Solution to Automatic Lexical Disambiguation, University of Leipzig, Institute of Informatics

[6]. Mukti Desai (2013). Word Sense Disambiguation, Department of Computer Engineering Dwarkadas J. Sanghvi College of Engineering, Mumbai University, India.

[7]. Rada Mihalcea, E. Agirre and P. Edmonds Eds (2007). Word Sense Disambiguation Algorithms and Applications Text, Speech and Language Technology, Springer, VOLUME 33, Université de Provence and CNRS, France.

[8].Jason Michelizzi (2005). A Semantic Relatedness Applied to All Words Sense Disambiguation, Unpublished Master’s Thesis, Department of Computer Science, University Of Minnesota, Daluth, USA.

[9]. Roberto N. (2009).Word Sense Disambiguation: A Survey. ACM Computing Surveys, Vol 41 No. 2, Universit `a di Roma La Sapienza, Italy.

[10]. Omnigton “the online wncyclopedia of writing systems and language” accessed from http://www.omniglot.com/writing/oromo.htm, Sep 06, 2017.

[11]. Online Afaan Oromo Dictionary, accessed from http://oromodictionary.com/index.php , Sep 06, 2017.

[12]. Diana McCarthy (2009). Journal, Word Sense Disambiguation, University of Sussex

[13]. Navigli R., Word sense disambiguation: A survey, ACM Computing Surveys, Vol. 41, No. 2, Article 10, 2009.

[14]. Prity Bala (2013). Knowledge Based Approach for Word Sense Disambiguation using Hindi WordNet, In The International Journal Of Engineering And Science (IJES), Apaji Institute, Banasthali Vidyapith Newai, Rajesthan, India, pp. 36-41.

[15]. 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 5th SIGDOC.pp.24– 26.

[16]. Rohana Sharma (2008), Word Sense Disambiguation for Hindi Language, Thapar University, Patiala, India.

[17]. Nancy I., Jean V. (1998).Introduction to the special issue on word sense disambiguation: The state of the art,Volume 24, Number 1

[18]. Christiane Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[19]. Kerem Celik (2012). A Comprehensive Analysis of Using WordNet, Part-Of-Speech Tagging, And Word Sense Disambiguation in Text Categorization". Unpublished Master's Thesis, Department of Computer Science, Bogazici University, Turkey.

[20]. Prity Bala (2013). Knowledge Based Approach for Word Sense Disambiguation using Hindi WordNet, In The International Journal Of Engineering And Science (IJES), Apaji Institute, Banasthali Vidyapith Newai, Rajesthan, India, pp. 36-41.

[21]. G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244.

[22]. X. Zhou and H. Han. (2005). Survey of Word Sense Disambiguation Approaches, Proceedings of the 18th International FLAIRS Conference.

[23]. Kavi Mahesh and Sergei Nirenburg. Knowledge-Based Systems for Natural Language Processing, MCCS-96-296, New Mexico State University

[24]. Gale William, Ken Church & David Yarowsky (1992a). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Newark, U.S.A., 249–256.

[25]. 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.

[26]. Gale, William, Ken Church & David Yarowsky (1992b). One sense per discourse. Proceedings of the DARPA Speech and Natural Language Workshop, New York, U.S.A, 233–237.

[27]. Yarowsky, D. (1992): Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proc. of the 14th International Conference on Computational Linguistics (COLING-92), Nantes, France, pp. 454-460.

[28]. Gerard Escudero Bakx (2006). Machine Learning Techniques For Word Sense Disambiguation Ph.D. thesis, Department of Computer Science, Universitat Politµecnica de Catalunya.

[29]. Brown P., Pietra S., Pietra V. and Mercer R. (1991). Word sense disambiguation using statistical methods. Proc. of the 29th Meeting of the Association for Computational Linguistics (ACL- 91), Berkley, C.A. pp 264-270

[30]. Oromo language. Http://en.wikipedia.org/wiki/oromo_language; last accessed on NOV 10, 2017.

[31].Kula Kekeba Tune, Vasudeva Varma, Prasad Pingali, Evaluation of Oromo- English Cross-language Information Retrieval, ijcai 2007 workshop on clia, hyderabad, india, 2007.

[32]. Wakshum Mekonnen, Development of stemming algorithm for Oromo texts, Master"s Thesis, 2000.

[33]. Tesfaye Guta, Afaan Oromo search engine, Master"s Thesis, Addis Ababa University, Departement of Computer Science, 2010.

[34]. Getahun A., The analysis of ambiguity in Amharic, Journal of Ethiopian Studies, Volume 34#2, 2001.

[35]. Anjali M K, Babu Anto P(2014), Ambiguities in Natural Language processing, Departmentof Information Technology, Kannur University,Kerala, India.

[36]. Solomon Assemu, Unsupervised machine learning approach for word sense disambiguation to Amharic words, Master"s Thesis, Addis Ababa university school of Information Science, 2011.

[37]. Baskaran Sankaran, k. Vijay-Shanker, Influence of morphology in word sense disambiguation for Tamil, Anna University and University of Delaware Proceedings of International Conference on Natural Language Processing, 2003.

[38] A book called "critical thinking, fourth edition: an introduction to the basic skills" by William Hughes and Jonathan Lavery, 2004.

[39] N.Ide, J.Veronis, Word Sense Disambiguation, In Proceedings of the 19th International Conference on Computational Linguistic, 1-42, 1998.

[40]. Ellen M. Voorhees (1993). Using WordNet to Disambiguate Word Senses For Text Retrieval, Siemens Corporate Research, Inc. 755 College Road East Princeton, NJ 08540

[41] Agirre E., Rigau G (1996). Word sense disambiguation using conceptual density. Proc. Of COLING

[42] Richardson R. and Smeaton A. (1995). Using WordNet in a knowledge-based approach to information retrieval. Proc. of the BCS-IRSG Colloquium, Crewe.

[43] Yarowsky, D. (1992): Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proc. of the 14th International Conference on Computational Linguistics (COLING-92), Nantes, France, pp. 454-460.

[44]. Debela Tesfaye and Ermias Abebe, Designing a Stemmer for Afaan Oromo Text: Hybrid Approach, Master"s thesis, Addis Ababa University, Department of Information Science, 2010.

[45] Getachew M.(2011) Parts of Speech Tagging for Afaan Oromo, M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[46] W. Faris and K.H. Cheng (2013). A Knowledge-Based Approach to Word Sense Disambiguation Computer Science Department, University of Houston, Houston, Texas, USA.

[47] Hee-Cheol Seo, Hoojung Chung, Hae-Chang Rim, Sung Hyon Myaeng and Soo-Hong Kim (2004), unsupervised word sense disambiguation using WordNet relatives, Department of Computer Science and Engineering, Korea University, Published.

[48] Wanjiku NG'ANG'A (2005).Word Sense Disambiguation of Swahili: Extending Swahili Language Technology with Machine Learning, Faculty of Arts of the University of Helsinki, in auditorium XII, Unioninkatu 34, on the 18th of November.

[49] Vildan Ozdemir (2009). Word Sense Disambiguation for Turkish Lexical sample, Unpublished Master's Thesis, Department of Computer Engineering, Fatih University, Istanbul, Turkey.

[50] Gasser M. (2012). HornMorpho: A System for morphological processing of Amharic, Oromo,and Tigrinya. Conference on Human Language Technology for Development, Alexandria, Egypt.

[51] Galmee Jechoota Afaan Oromo. Wiirtu Qo'anno fi Qoranno Afaanota Itoophiyaa Yuniversiitii Finfinne .

[52] Kaplan A., An experimental study of ambiguity and context, Mechanical Translation, vol.2 no.2, 1955.

Appendix: 1 Sample Senses of Afaan Oromo Words

horii	1 uumama luka afurii kan akka loonii, hoolotaa 2 waan namni hore,qarshii
mirga	1 maddii yookiin bukkee(cinaa) dhaqna namaa, kan yammuu gara kaabaatti yookiin boroottii garagalan gara baha biiftuu oolu 2 bineensa akka arbaa, leencaa ajjeesanii faachii irraa fudhatamu 3 uumaman waan namni gochuu danda'uu 4 waan hojjetan irratti olaantumma namni sun qabu 5 waan tokko gochuuf dandeettii seeraan qaban
callaa	1 midhaan girdiin keessa baye 2 midhaan daakamee sirritti hinbullaayin 3 gosa dheedhii kan akka baaqelaa,atara 4 sharafa qarshii, saantima
caamsaa	1 yeroo aduun gar male ho'u,hongee 2 maqaa ji'aa kan eeblafi waxabajjii giddu oolu
sanyii	1 midhaan akka marguuf facaasan 2 waan gosa tokko ta'an, kan firooma wal irraa qaban
Karaa	1 lafa namni yookiin konkolaatan irra deemu, kan bakka tokko qabee bakka biraatti nama geessu 2 akkaataa, haala 3 roga dhimma nama ilaallatu
caffee	1 marga lafa jiidha qabutti margu 2 marga jiidha qabu 3 mana maree uummata oromo 4 mana maree bakka bu'oota uummataa
ifa	1 dukkana kan hin ta'iin

	2 dukkana kan dhabamsiisu, kan waan akka biiftu irraa bawuu
qophii	1 waan qophaa'ee dhiyaate 2 waan tokko raawwachuuf yookiin dhimma tokko baasuf wanti nama dandeessisu uumama, mija'aa 3 muka yookiin quba laaga of kaa'anii diddigaa baasa
cinaacha	1 lafee qaqal'aa qaama namaa yookiin horii bitaafi mirga irratti argamu 2 gar tokko, walakkaa
afaan	1 qaama funyaanii gadii kan midhaan ittiin nyaatan yookiin bishaan ittiin dhugan, akkasumas ittiin dubbatan 2 qooqa namni dubbatu
alkoolii	1 dhugaatii nama macheessu 2 dhangala'aa mana hakiimitti nafa madaaye jirbii cuubanii diban
darbu	1 bakka tokko keessaa deemanii bakka biraatti ceewuu, taruu, kutuu 2 nama caaluu 3 waa dhahuu
digirii	muka qal'aa babatteefii gindii ittin walitti qabsiisan, qicirtii 1 waraqataa barumsaa yuniversiitii nama fixeef akka ragaatti 2 kennamu, digriin eebifame 3 gosa meeshaa wayiin safaran, keessattu o'inaa,qorra
duree	1. tan dura taate 2. ijaarsa mana keessaa kan bakka ciisichaatiifi bakka taa'umsa addaan qoodu, cicha yookiin gorroo manaa 3. rifeensa sammuu namaa gubbaatti utuu hinhaadin dhiisan, guttiyyee,roggee
aarsuu	1 hamaa yookiin waan hintolle hojjetanii yookiin dubbatani nama dallansiisu 2 akka aarri keessaa yaa'uu gochuu

ala	1 keessa kan hin taane 2 naanno jiran keessa kan hin ta'iin, biyya keessa jiran kan hinta'iin
ashamaa	1 shurrubbaa furdatee dhahame 2 akkam bultan, ooltan,noora, ol seenaa
baabura	1 motora midhaan daaku 2 konkolaataa gommaa sibiilaa, kan sibiila irraa(hadiida irra) deemu
baasuu	1 keessaa gara alaatti yaasuu,fuudhuu 2 idaa kaffaluu, horii kennuu 3 weedduu,sirba,yookiin weellee sagaleen dhageessisuu
baqa	1 waan ta'e jalaa dheessa 2 dhadhaan yookiin wanti akkasii o'ee dhangala'atti geeddarama, 3. jiruun namatti toluun kan ka'e fuulli namaa cululuqa, fiila
bara	1 qoonqoon nama qabiisa, barbaachi nyaaata namatti dhaga'amina 2 waan nyaatan dhabanii rakkoo guddaan nama irra gayiisa,oongee,gadadoo 3 ji'a kudhalama, wagga,ogga 4 yeroo dheeraa,jabana 5 waan duraan hinbeekne beekumsa
bilbila	1 sibiila yoo urgufan yookiin waliin rukutan sagale kennu, kan morma farda,jabbii, daa'immaati hidhan,hashqura 2 shiboo alaalaa kan ittiin walitti dubbatan, silkii
boffee	1 furdaa keessi isaa jabeenya hinqabne 2 jirbii maagii ta'u, kan manatti fooyamu
bokkuu	1 seera gadaa keessatti ulee abbaan gadaa amma barri isaa raawwatutti qabatu 2 muka midhaan yookiin buna mooyyee keessatti ittiin tuman 3 dhiita yookiin waan dhiita fakkaatu

bulchaa	1 nama biyya bulchu 2 nyaata ganama nyaatan kan irbaata irraa(yeroo baay'ee) hafu

Appendix: 2 Sample Afaan Oromo Stop Words

akka hanga jechuun ol waan akkam
henna kan waggaa akkasumas hoggaa kanaaf
oliin woo akkum hogguu kanaafi yammuu
akkuma hoo kanaafuu osoo yemmuu mmo
illee kee otoo yeroo ani immoo
keenya otumallee ykn ani innaa keenyaa
otuu yommii booda inni keeti otuullee yommuu
booddee isaa keetii saniif yoo dura oliif

Appendix: 3 Sample Afaan Oromo ambiguous words and their roots

horii	hor	bara	bar
mirga	mirg	bulchaa	bulch
caamsaa	caams	bilbila	bilbil
ifa	if	baqa	baq
afaan	af	baasuu	baas
darbu	darb	ciraa	cir
aarsuu	aars	buufaa	buuf
citaa	cit	cufaa	cuf
daara	daar	eela	eel
fura	fur	gala	gal
guutuu	guut	haqa	haq
hartuu	har	hiddaa	hidd
kooraa	koor	koree	kor
kottee	kott	kutaa	kut

kuusaa	kuus	laga	lag