ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

UNSUPERVISED CORPUS BASED APPROACH FOR WORD SENSE DISAMBIGUATION TO AFAAN OROMO WORDS

By

FEYISA GEMECHU SHOGA

JUNE 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

UNSUPERVISED CORPUS BASED APPROACH FOR WORD SENSE DISAMBIGUATION TO AFAAN OROMO WORDS

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

FEYISA GEMECHU SHOGA

JUNE 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

UNSUPERVISED CORPUS BASED APPROACH FOR WORD SENSE DISAMBIGUATION TO AFAAN OROMO WORDS

BY
FEYISA GEMECHU SHOGA

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chair person,	_____	_____
Ermias Abebe	Advisor,	_____	_____
Dr. Solomon Tefera	Examiner,	_____	_____
Dr. Martha Yifru	Examiner,	_____	_____

Dedication

In loving memory of my most beloved and most cherished to **my Grandmother**, **Darare**, who lived a life of dignity, courage, wisdom, patience and above all affection, and who will remain my personal hero and my inspiration forever. May Allah bless her soul, Amen.

Acknowledgment

First of all, I would like to say "***Alhamdulillah Rabilalemin***". Second of all, I would like to sincerely thank my advisor, Ato Ermias Abebe without whom such type of work is difficult to do, for his supervision, advice, patience, punctuality and moral support throughout this thesis work.

I also want to thank Dr. Million Meshesha, Dr. Martha Yifru and Said Hassen for giving me variable ideas and guidance from the very beginning. I am also indebted to all IES Staff for the help they offered, understanding and supporting me during my stay in the office with them, especially Hailamariam Negussie, Genet Getaneh, Almaz Abuhay, Mahlet Teferawork and Daniel Mamo.

I would also like to give special thanks to my family. My family members are always there to support me in every situation. On top of all, gratitude goes to my lovely father, Ato Gemechu Shoga, who is my inspiration, my lovely mother, Shuna Gutema and her sister who is my ant Obse Gutema for remembering and motivating me. I deserve very grateful thanks for them because without them I would not be who I am today. My special thanks again goes to all my sisters and my brothers for their having being with me always ideally and above all, for being the greatest brothers and sisters.

I am grateful to my classmates who also, are now my friends. Their smiles, laughter and unreserved help, and most of all, solidarity through the hard times, made my stay at the school more fun and all the hard work bearable ideally supporting me. Finally, I would like to give my gratitude to people who are not mentioned in name but whose effort helped me much all along.

Last but not least, I would like to sincerely thank all of friends, colleagues and my students for their assistance, encouragement, and inspiration during this research. Though it is difficult to mention the name of persons who gave me their hand while doing this thesis, it is necessary to mention those who gave their precious time to read the thesis document, to share ideas, and gave me moral and material support. Tesfa Kebbede, Solomon Assemu, Tedros Gizaw, Duressa Deksiso, Meseret Gobena, Kibrom Haftu, Abraha Gebrekiros, Zemzem Adem, Ketsela Gelan, Tesfaye Basha and Jallanne Temesgen are few of them. I am very grateful to thank them for what they did.

Table of Contents **page**

LIST OF TABLE

LIST OF FIGURES

LIST OF APPENDICIES

LIST OF ACRONOMS AND ABBREVIATIONS

NLP Natural Language Processing

WSD Word sense Disambiguation

EM Expectation Maximization

MRD Machine-Readable Dictionaries

AI Artificial Intelligence

BNC British National Corpus

KNN k-Nearest Neighbor

NL Natural Language

IR Information Retrieval

CL Complete Linkage

SL Single Linkage

AV Average Linkage

LDOCE Longman Dictionary of Contemporary English

ABSTRACT

This thesis presents a research work on Word Sense Disambiguation for Afaan Oromo Language. A corpus based approach to disambiguation is employed where unsupervised machine learning techniques are applied to a corpus of Afaan Oromo language, to acquire disambiguation information automatically. We tested five clustering algorithms (simple k means, hierarchical agglomerative: Single, Average and complete link and Expectation Maximization algorithms) in the existing implementation of Weka 3.6.11 package. "Cluster via classification" evaluation mode was used to learn the selected algorithms in the preprocessed dataset.

Due to lack of sense annotated text to be able to do these types of studies; a total of 1500 Afaan Oromo sense examples were collected for selected seven ambiguous words namely ***sanyii, karaa, horii, sirna and qoqhii, ulfina, ifa.*** Different preprocessing activities like tokenization, stop word removal and stemming were applied on the sense example sentences to make it ready for experimentation. Hence, these sense examples were used as a corpus for disambiguation.

A standard approach to WSD is to consider the context of the ambiguous word and use the information from its neighboring or collocation words. The contextual features used in this thesis were co-occurrence feature which indicate word occurrence within some number of words to the left or right of the ambiguous word.

For the purpose of evaluating the system, a training dataset was applied using standard performance evaluation matrics. The achieved result was encouraging, because clustering algorithms were achieved better in terms of accuracy of supervised machine learning approaches on the some dataset similar. But, further experiments for other ambiguous words and using different approaches will be needed for a better natural language understanding of Afaan Oromo language.

Keywords: Natural Language Processing, Word Sense Disambiguation, Unsupervised Learning Method, Information Retrieval, Clustering algorithms.

1. CHAPTER ONE

1.1. Introduction/Background

Natural Language Processing is a computational technique for representing and analyzing naturally occurring texts at one or more levels of linguistic analysis. This is for the purpose of achieving human-like language processing for a range of applications. The goal of NLP is to design and build software that will analyze, understand, and generate languages that humans are using naturally in terms of technology. So to achieve this, natural language systems need to acquire extensive knowledge about the world which is not easy to acquire.

So, one of the extensive knowledge is lexical semantics which begins with recognition that a word is a conventional association between a lexicalized concept and utterance that plays a syntactic role [30]. This lexical semantics is understood clearly by proper disambiguation of words. The ambiguity of words can be achieved by one of the fields of natural language processing application known as Word Sense Disambiguation's (WSD) systems. Word sense disambiguation is the task of identifying the correct meaning of an ambiguous word, which has more than one sense, in a given context. When an ambiguous word is used in a sentence, humans are able to select the correct sense of that word without considering alternative senses [4]. However in any application, a computer cannot be able to identify and resolve the usage of ambiguous words in natural language processing (NLP).

This ambiguity is inherent to human language. In particular, word sense ambiguity is predominant in all natural languages; with a large number of words in any given language carrying more than one meaning. For humans, resolving ambiguity is a routine task that hardly requires conscious effort. In addition to having a deep understanding of language and its use, humans possess a broad and conscious understanding of the real world, and this equips them with the knowledge that is relevant to make sense disambiguation decisions effortlessly, in most cases. However, successful solutions for automatic resolution of ambiguity in natural language often require large amounts of annotated data/knowledge resources, to achieve good levels of accuracy. These issues are clearly reflected in the performance of current word-sense disambiguation systems. When given a large amount of training data for a particular word with reasonably clear sense distinctions, existing systems perform fairly well [2].

Many application of NLP, such as machine translation, information retrieval, information extraction, and question answering require semantic analysis, where WSD plays a crucial role.

As a result WSD has been an interesting concern since the earliest day of computer treatment of language in the 1950's [3]. Sense disambiguation is an intermediate task which is not an end by itself, but rather is necessary at one level or another to accomplish most Natural Language processing tasks. The word sense ambiguity is a hard problem for the developers of Natural Language processing (NLP) systems. Some words, often, have different meaning in various contexts. When language is capable of being understood in more than one way by a reasonable person, ambiguity exists based on the nature of words in context [1]. The reason behind this is that many words have more than one meaning depending on the context of use in sentences. The process by which the most appropriate meaning of an occurrence of an ambiguous word is determined is known as Word Sense Disambiguation (WSD), and remains an open problem in NLP.

However, since the 1950s, many approaches have been proposed for identifying senses of words in a context. The five main approaches applied in the area of WSD field are AI-based, methods based on the context window of the target word, knowledge-based approaches, corpus based approaches and hybrid approach. Knowledge based approach uses information provided by Machine Readable Dictionaries (MRD), Corpus based approach uses information gathered from training corpus and Hybrid approach combines aspects of the two methodologies [5].

Among more than 80 languages spoken in Ethiopia, Afaan Oromo is the one with the largest native speakers which is grouped under Cushitic family of languages [10]. Afaan Oromo uses Latin based script called “Qubee” and it has 26 basic characters. It is the official language of Oromia regional state of Ethiopia and is also the academic language for primary schools of the region [12]. Oromo language, literature and folklore are delivered as fields of study in many universities located in Ethiopia and other countries [11]. Afaan Oromo is rich with ambiguity in semantics which can benefits from WSD research and development.

1.2. Statement of the Problem

Afaan Oromo is one of the major languages that are widely spoken in Ethiopia. This language has also a number of ambiguous words like any other language. Therefore, it is difficult to understand the meaning of those words in a given context.

Hence, to have a clear understanding of ambiguous words in the language, WSD for Afaan Oromo language is also needed to be developed. Because, now a days as the development of technology is increasing rapidly, like any other language Afaan Oromo language has also started to use the technology for different purposes. Many different Afaan Oromo documents are storing in different sites as there is a problem to information retrieval, there is also a need for machine translation as we need to translate Afaan Oromo document to other language and many other tasks like text processing, speech processing and grammar analysis in which Afaan Oromo language is facing a problem with sense of words. To overcome this problem, the discussion of word sense disambiguation (WSD) for Afaan Oromo language is necessary.

A single word can have many senses and each of those senses can be mapped into many target language words. So, selecting the correct meaning of a word which is the most suitable for the context is a challenging problem [6]. The absence of WSD in Afaan Oromo Natural Language Processing system limits, future efforts of making computer to understand Afaan Oromo Language.

Generally, Afaan Oromo is associated with a different meaning, and the problem that remains is that of associating the context of an ambiguous Afaan Oromo word with one of the meanings. The distinction between different senses for a word is sometimes unclear even for human judges. Ambiguity is one of these problems which have a great relevance to Computational Linguistics. This is, even though people are unaware of the ambiguities in the language they use. The reason behind this is that, they are very good at resolving them using context and their knowledge of the world. But computer system which people are using today for different purpose does not have this knowledge, and consequently do not do a good job of making use of the context. Something is ambiguous when it can be understood in two or more possible ways or when it has more than one meaning. So, today this is a big problem issue in other natural language processing

application like information retrieval, text processing, speech recognition and machine translating for different stakeholders like learners, speakers, government, writers and researchers.

1.3. Research questions

At the end of this research the following question must be answered.

1) What are the most important clustering algorithms that improve the performance of unsupervised Afaan Oromo WSD system?
2) How effective is the unsupervised approach on the performance of unsupervised Afaan Oromo WSD systems?
3) Which window size is effective to identify the meaning of ambiguous words in the sentence contextually for unsupervised Afaan Oromo WSD?
4) What is the effect of the stemming on the performance of unsupervised Afaan Oromo WSD system?

1.4. Objective of the study

1.4.1. General objective

The general objective of this research was to design and test a prototype of unsupervised word sense disambiguation system for Afaan Oromo language.

1.4.2. Specific objective

The specific objectives of this research are to:

- Conduct literature review and related works to understand the technical approaches that we used for studying word sense disambiguation.
- Study the general grammatical structure of Afaan Oromo sentence, ambiguous words, and phrases due to their use for the development of corpus.
- Develop dataset and test with different clustering algorithms.
- Compare some selected unsupervised algorithms to see their effects on corpus developed from Afaan Oromo words.
- Train WSD model using the selected unsupervised machine learning algorithms.
- Evaluate the performance of the model.
- Draw conclusion and suggest some recommendations for future improvement of the systems.

1.5. Methodology

The general methodology of the research is Quantitative Experimental out of which we followed the following methods to overcome this methodology.

1.5.1. Literature Review

Literature review was done on different areas that are considered to be relevant for this study. Extensive literature review is conducted on Word Sense Disambiguation in order to obtain an in depth understanding of the area and to find useful approaches for the Afaan Oromo word sense Disambiguation. This research work is on a design and development of prototype for word sense disambiguation using corpus based approach. So both local and foreign language works on WSD were reviewed in detail to understand tools and techniques that can be applied in our work.

1.5.2. Procedures

Procedure shows the intended methodology of the major works and the flow of activities. Regarding our work the researchers started by determining the ambiguous Afaan Oromo words. Then, we continued with the action of developing the corpus for these words as well as creating a training dataset for the experiments. Then different preprocessing activities like stop word removal, tokenization and stemming were conducted. By using selected clustering algorithms, we generated models and evaluated the performance of the system. The analysis and interpretation of the generated result was followed with the comparison of each algorithm.

The best performing algorithm is selected and finally completes by comparison of the result obtained with supervised algoreithms.

1.5.3. Tools for development of system

Implementation of unsupervised algorithms in Weka 3.6.11 package was used to build and test the models. Weka 3.6.11 machine learning tool was selected due to the familiarity of the researcher to the tool and because of its accessibility, processing capability and language independent features.

1.5.4. Techniques

As our research uses unsupervised techniques, we choose clustering algorithms from Weka tools. These algorithms are like simple k-means algorithms, which represent simple, hard and flat clustering methods; agglomerative such as single, average and complete link algorithms for representative family of hierarchical clustering algorithms and the Expectation Maximization

algorithms also known as the EM which is probabilistic clustering algorithms to generates hierarchical clustering where clusters are described probabilistically.

1.5.5. Data Collection

For our work, we need to collect some ambiguous words and the sense of these words which is difficult to get for Afaan Oromo language. Because there is no standardized Afaan Oromo sense annotated data set, as for other languages like English. Afaan Oromo documents and information on Afaan Oromo was studied and ambiguous words were collected from different libraries and institutions. Data necessary for the purpose of experimentation and for preparation of corpus are acquired after assessing the different options that are available and also appropriate data are collected to analyze the different characteristics of Afaan Oromo word sense disambiguation using corpus based approach at different context. Based on this, previously a research was done on supervised WSD for Afaan Oromo [53]. We used the corpus and ambiguous words selected by [53] and make an arrangement that fits this research direction by adding some other corpus. Tesfa [53] developed Afaan Oromo corpus from the scratch. The corpus he developed contains 1240 sentences and he selected 5 Afaan Oromo ambiguous words namely **sirna, karaa, sanyii, qophii and horii** into which we have added two more words **Ifa** and **Ulfina** to increase the size of corpus**.** Out of 1240 corpus used by Tesfa We used 1000 corpus for our works. Here is the summary of the senses for each ambiguous word.

No.	**Ambiguous words**	**Sense of the words**	**English meaning**
1	Sanyii	Ija midhaani ykn biqiltu	Seed
		Gosa	Type/Kind
2	Horii	Qarshii	Money
		Beelada	Cattle
3	Karaa	Daandi	Road
		Akkaata ykn kallatti	Way/via
4	Sirna	Qophii	Event
		Seera	Procedure/system
5	Qophii	Haala Mijeessu	Preparation
		Saganta	Program/event
6	Ifa	Addenna/calanqisa wantootaa	Light
		Hubannaa	Clear
7	Ulfina	Ba’aa	Weight
		Kabaja	Respection

Table 1.1: Summary of senses for each ambiguous word

1.6. Scope and limitations of the study

Corpus based approaches use supervised, unsupervised and bootstrapping machine learning techniques for WSD. Due to time constraint to train, test and analyze all the results in this research, only five clustering algorithm were used to build and evaluate the WSD model. Because of unavailability of sense annotated data and linguistic resources; the study was limited to the experimentation of seven ambiguous words. We dealt only with data in textual format. The system deals with semantic level analysis predicting without any kind of grammar and spelling correction.

1.7. Significance of the study

The significance of the study was improving the performance of different NLP applications in the area of Word sense disambiguation. In the field of learning and education of Afaan Oromo word sense, it is used to identify the actual meaning of a semantic ambiguous word in its textual context. It also contributes to future research area in Natural language processing. The result of this research can also be used by different stakeholders like speakers to identify the sense of the words easly, learners, government, writers and researchers of the languages in the area of machine translation, information retrieval, text processing, and speech processing and grammar analysis and etc. We discussed the detail in chapter two.

1.8. Organization of the Research

This document contains a total of six chapters. The second chapter presents general concepts on word sense disambiguation (WSD) and related works. In chapter three, the brief history of Afaan Oromo with their writing systems, parts of speech, ambiguities grammers and dialects were described. The fourth chapter presents the data requirement, Preparation and system architecture design with data analysis. Chapter five presents the experimental results by incorporating discussion about the results. Finally, the conclusion and future works are presented in chapter six.

2. CHAPTER TWO: Literature Review

2.1. Word Sense Disambiguation

This chapter introduces all about the description of word sense disambiguation (WSD), different approaches used for word sense disambiguation (WSD) and some of the related works previously done by using different technique for WSD for different languages.

Many words have more than one meaning in natural language, and each one of the meaning is determined by its context. The automated process of recognizing word senses in context is known as Word Sense Disambiguation (WSD). It is the process of selecting the appropriate meaning or sense for a given word in a document. Word Sense Disambiguation (WSD) refers to also a task that automatically assigns a sense, selected from a set of pre-defined word senses to an instance of a polysemous word in a particular context [13].

One of the problems with word sense disambiguation is deciding what the senses are, in cases where at least some senses are different. In other cases, however, the different senses can be closely related (one meaning being a metaphorical extension of another), and in such cases division of words into senses becomes much more difficult [14].

Word sense disambiguation involves the association of a given word in a text or discourse with a definition or meaning or sense which is distinguishable from other meanings potentially attributable to that word [15]. According to [15] task of word sense disambiguation has two steps: firstly the determination of all the different senses for every word relevant at least to the text or discourse under consideration which includes activities like a listing senses found in dictionary, grouping features, categories, or associated words i.e. synonyms, as in a thesaurus and an entering in a transfer dictionary which includes translations in another language. Secondly, a means to assign each occurrence of a word to the appropriate sense relying on the context of the word to be disambiguated and external knowledge sources, including lexical, encyclopedic.

Word sense disambiguation (WSD) can be viewed as a classification task [16]: the first is in which word senses are the classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources. Alternatively word sense disambiguation can be a process in which tasks such as part-of-speech tagging i.e., the assignment of parts of speech to target words in context, named entity resolution the classification of target textual items into

predefined categories and text categorization i.e., the assignment of predefined labels to target texts are applied.

[17] Described word sense disambiguation having two main tasks: The first is 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. The second is All-words WSD, where systems are expected to disambiguate all open-class words in a text i.e., nouns, verbs, adjectives, and adverbs. Here semi supervised systems and knowledge-lean systems can potentially used.

The algorithms used in WSD can be classified as knowledge based, corpus based and hybrid, AI based and Methods Based on the Context Window of the Target Word. Corpus based algorithm can be further classified as supervised learning and unsupervised learning [7]. In knowledge based approach disambiguation is carried out using information from an explicit lexicon or knowledge base. The lexicon may be a machine readable dictionary, thesaurus or it may be hand- crafted. Supervised learning can be viewed as a classification task while unsupervised learning can be viewed as a clustering task [8]. A research on Unsupervised Machine Learning Approach for Word Sense Disambiguation to Amharic Words was also conducted by Solomon Assemu [9]. He has introduced corpus based approach to word sense disambiguation that only requires information that can be automatically extracted from untagged text. The researcher use unsupervised techniques to address the problem of automatically deciding the correct sense of an ambiguous word based on its surrounding context.

2.2. Elements of word sense disambiguation

2.2.1. Selection of Word Senses

A word sense is a commonly accepted meaning of a word. For instance, consider the following two Afaan Oromo 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 as [51] indicated in his thesis.

2.2.2. External Knowledge Sources

External Knowledge Sources for word sense disambiguation is seen as structured and unstructured resources.

2.2.2.1. Structured resources

Thesauri are a resource which provides information about relationships between words, like synonymy, antonymy representing opposite meanings.

Machine-readable dictionaries (MRDs) is other resource which has become a popular source of knowledge for natural language processing since the 1980s, when the first dictionaries were made available in electronic format: among these, we cite the Collins English Dictionary, the Oxford Advanced Learner's Dictionary of Current English, and the Oxford Dictionary of English [19].

Ontology is also other resource which is specifications of conceptualizations of specific domains of interest, usually including taxonomy and set of semantic relations. Ontology provides a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary [20].

Word-Net is other resource which can be considered as ontology. It encodes concepts in terms of sets of synonyms called synsets, that each word sense univocally identifies a single synset [22]. For each synset, Word-Net provides the information: The first is gloss which is a textual definition of the synset possibly with a set of usage. The second is lexical relations, which connect pairs of word senses. Lexical relations include Antonymy, Pertainymy and Nominalization. The third is semantic relations which is applied to synsets in their entirety i.e., to all members of a synset. It includes all such as: Hypernymy also called kind-of or is-a, Hyponymy and troponymy the inverse relations of hypernymy for nominal and verbal synsets, respectively, Meronymy also called part-of which holds for nominal synsets only, Holonymy which is the inverse of meronymy, Entailment, Similarity and Attribute.

2.2.2.2. Unstructured resources

Corpora are unstructured resources which are collections of texts used for learning language models. It can be sense-annotated and raw or unlabeled corpora which can be used in WSD, and they are most useful in supervised and unsupervised approaches, respectively [21].

Raw corpora such as Brown Corpus which include a million words balanced collection of texts published in the United States in 1961; British National Corpus (BNC) which include 100 million word collection of written and spoken samples of the English language often used to collect word frequencies and identify grammatical relations between words; Wall Street Journal corpus which include a collection of approximately 30 million words from WSJ, American National Corpus which includes 22 million words of written and spoken American English and Giga-word Corpus which include collection of 2 billion words of newspaper text.

The other corpora is sense-Annotated Corpora such as SemCor corpus is the largest and most used sense-tagged corpus, which includes 352 texts, tagged with around 234,000 sense annotations, MultiSemCor is an English-Italian parallel corpus annotated with senses from the English and Italian versions of Word-Net, the line-hard-serve corpus containing 4000 sense-tagged examples of these three words (noun, adjective, and verb, respectively), the interest corpus with 2369 sense-labeled examples of noun interest; the DSO corpus produced by the Defense Science Organization (DSO) of Singapore which includes 192,800 sense-tagged tokens of 191 words from the Brown and Wall Street Journal corpora and Open Mind Word Expert which include about 288 nouns [21].

Collocation resources is also another unstructured resources which register the tendency for words to occur regularly with others it include the Word Sketch Engine, JustTheWord, The British National Corpus collocations, the Collins Co-build Corpus Concordance.

2.2.3. Representation of Context

Text is an unstructured source of information. To make it a suitable input to an automatic method it is usually transformed into a structured format. So, preprocessing of the input text is usually performed in following steps: first tokenization, a normalization step which splits up the text into a set of tokens usually words. The next is part-of-speech tagging consisting in the assignment of a grammatical category to each word are tags for determiners like nouns, verbs, and adjectives. The next is lemmatization which is the reduction of morphological variants to their base form. The next is chunking which consists of dividing a text in syntactically correlated parts. And the last preprocess is parsing whose aim is to identify the syntactic structure of a sentence. The context of information which is performed in pre-processing activities can be represented with the following features [17]:

Local features - represent the local context of a word usage, that is, features of a small number of words surrounding the target word, including part-of-speech tags, word forms, and positions with respect to the target word.

Topical features – is in contrast to local features and define the general topic of a text or discourse, thus representing more general contexts (e.g., a window of words, a sentence, a phrase, a paragraph), usually as bags of words.

Syntactic features - representing syntactic cues and argument-head relations between the target word and other words within the same sentence note that these words might be outside the local context.

Semantic features - representing semantic information, such as previously established senses of words in context, domain indicators.

Based on this set of features, each word occurrence usually within a sentence can be converted to a feature vector.

2.3. Applications and use of WSD

Unfortunately, to date explicit WSD has not yet demonstrated real benefits in human language technology applications. Word sense disambiguation can be befits for the following applications in technology.

Information Retrieval (IR) - Ambiguity has to be resolved in some queries for information retrieval. For instance, given the query depression should the system return documents about illness, weather systems, or economics.

Information Extraction (IE) - WSD is required for the accurate analysis of text in many applications. For instance, an intelligence gathering system might require the flagging of say, all the references to illegal drugs, rather than medical drugs.

Machine translation (MT) - WSD is required for lexical choice in MT for words that have different translations for different senses and that are potentially ambiguous within a given domain (since non-domain senses could be removed during lexicon development). In MT, the senses are often represented directly as words in the target language. However, most MT models do not use explicit WSD. The lexicon is pre-disambiguated for a given domain, hand-crafted rules are devised, or WSD is folded into a statistical translation model [23].

Text mining - WSD is required for the accurate analysis of text in many applications. For instance, an intelligence gathering system might require the flagging of, say, all the references to illegal drugs, rather than medical drugs.

Content Analysis - The analysis of the general content of a text in terms of its ideas, themes, etc., can certainly benefit from the application of sense disambiguation.

Word Processing - Word processing is a relevant application of natural language processing, whose importance has been recognized for a long time as stated by [24]

Lexicography - Modern lexicography is corpus-based, thus WSD and lexicography can work in a loop, with WSD providing rough empirical sense groupings and statistically significant contextual indicators of sense to lexicographers, who provide better sense inventories and sense-annotated corpora to WSD.

The Semantic Web - it inherently needs domain-oriented and unrestricted sense disambiguation to deal with the semantics of (Web) documents, and enable interoperability between systems, ontology, and users.

2.4. Approach and technique of word sense disambiguation

Approach	Method	Details/summary
AI-Based	Symbolic	Different forms of logical Inference
	Connectionist	Spreading activation Models
Methods Based on the Context Window of the Target Word	Micro-context	words of context to the entire sentence in which the target word appears
	Topical context	substantive words which co-occur with a given sense of a word
Corpus Based	Supervised	Collection of labeled texts for classification
	Unsupervised	Collection of unlabeled texts for clustering
	Bootstrapping/semi-supervised	Combination of supervised and unsupervised with some resources
Knowledge Based	Machine-Readable Dictionary	Words derived from corpora
	Thesauri	provides information about relationships between words
	Computational lexicons	construction of semantic lexicons
Hybrid Based		Combination of corpus based and some knowledge based

Table 2.1: Summary of Approach to word sense disambiguation

2.4.1. AI-Based Approach

This method was begun in the early 1960's to attack the problem of language understanding [15]. So as a result, the problem of WSD in natural language was tried to solve by this approach. WSD in AI work was typically accomplished in the context of larger systems intended for full language understanding. Such systems were almost always grounded in some theory of human language understanding which they attempted to model and often involved the use of detailed knowledge about syntax and semantics to perform their task, which was exploited for WSD. There are two methods discussed under this approach as follows.

Symbolic method

In this methods, semantic networks is used which were developed in the late 1950's and were immediately applied to the problem of representing word meanings like in machine translation to derive the representation of sentences in an Interlingua comprised of fundamental language concepts [25]. Here she has developed a set of 100 concept types, in terms of which her group built a 15,000 entry concept dictionary, where concept types are organized in a lattice network with inheritance of properties from super concepts to sub concepts. Other researchers [28] built a network that includes links among words (tokens) and concepts (types), in which links are labeled with various semantic relations or simply indicate associations between words. The other is the use of frames which contained information about words and their roles and relations to other words in individual sentences. Specifically the use of a combination of a semantic network which consists of nodes representing noun senses and represented by verb senses and the other is case frames which impose IS-A and PART-OF relations on the network [29]. He shows that only homonyms can be fairly accurately disambiguated.

Connectionist method

This idea is realized in spreading activation models where concepts in a semantic network are activated upon use, and activation spreads to connected nodes.

Generally, disambiguation procedures embedded in AI-based approaches are most usually tested on only a very small test set in a limited context using a single sentence, making it impossible to determine their effectiveness on real texts. It was theoretically interesting but not at all practical for language understanding in any but extremely limited domains. So, the other approach is needed for word sense disambiguation.

2.4.2. Methods Based on the Context Window of the Target Word

This approach also used to solve the problems of natural language in case of word sense disambiguation. So here, the set of words to the left and right of the target word in the context, called window, is used for disambiguation. Several researchers recognized the importance of window and carried out experiments to determine the optimal window size [15]. Context is the only means to identify the meaning of a polysemous word. Therefore, all work on sense disambiguation relies on the context of the target word to provide information to be used for its disambiguation. For data-driven methods, context also provides the prior knowledge with which current context is compared to achieve disambiguation. Broadly speaking, context is used in two ways. One is bag of words approach in which context is considered as words in some window surrounding the target word, regarded as a group without consideration for their relationships to the target in terms of distance, grammatical relations, etc. The second is relational information in which context is considered in terms of some relation to the target, including distance from the target, syntactic relations, selectional preferences, orthographic properties, phrasal collocation, semantic categories and etc. In this approach two methods are used widely to sense selection, these are:-

Micro-context

Most disambiguation work uses the local context of a word occurrence as a primary information source for WSD. Local or “micro” context is generally considered to be some small window of words surrounding a word occurrence in a text or discourse, from a few words of context to the entire sentence in which the target word appears. Context is very often regarded as all words or characters falling within some window of the target, with no regard for distance, syntactic, or other relations

Topical context

Topical context is also the other methods which include substantive words which co-occur with a given sense of a word, usually within a window of several sentences. Unlike micro-context, which has played a role in disambiguation work since the early 1950's, topical context has been less consistently used. Methods relying on topical context exploit redundancy in a text--that is, the repeated use of words which are semantically related throughout a text on a given topic. Work involving topical context typically uses the bag of words approach, in which words in the context are regarded as an unordered set.

The use of topical context has been discussed in the field of word sense disambiguation (WSD) method. Yarowsky [40] used a 100-word window, both to derive classes of related words and as context surrounding the polysemous target, in his experiments using Roget's Thesaurus. Leacock et al. [47] have similarly explored topical context for WSD.

2.4.3. Corpus based approach

Corpus based methods grew in importance after the public availability of large scale digital corpora. Corpus is the data for the lexical sample task is typically a large number of naturally occurring sentences containing a given target word, each of which has been tagged with a pointer to a sense entry from the sense inventory. It is a collection of preferably published texts used for linguistics purposes. Texts should be selected across a variety of domain to cover different word senses since domain usually restricts words to one sense only. Corpora provide vast volume of information regarding language usage, therefore they are especially well suited for statistical or empirical methods. It gathers knowledge from corpora collection of edited and published texts used for linguistics purpose. Word sense disambiguation can be thought as consisting of two stages. First step is sense discrimination where the occurrences of a word are mapped into a number of classes depending on the sense they belong to. The second step, sense labeling, then assigns a sense to each class, hence to each word in that class. Schutze's [26] method is interesting in that it deals solely with the first step, eliminating the need to refer to any outside knowledge base such as MRD's, thus it can be considered as a purely corpus-based method. In corpus based approach there are many different methods such as supervised, semi-supervised and unsupervised approach.

2.4.3.1. Supervised method

Regarding word sense disambiguation (WSD), one of the most successful approaches is supervised learning approaches, in which statistical or ML classification models are induced from semantically annotated. This method is the way disambiguation methods utilize corpora forms as a classification which requires manually disambiguated words. This supervised machine learning algorithms use semantically annotated corpora to induce classification models for deciding the appropriate word sense for each particular context. Though, compilation of corpora for training and testing such systems requires a large human effort since all the words in these annotated corpora have to be manually tagged by lexicographers with semantic classes taken from a particular lexical semantic resource like Word-Net [30].

This method differs from unsupervised learning methods based on whether the correct senses of the target words in the training corpus are given or not. In supervised learning, training sentences are partitioned according to the given sense tags. Generally, supervised systems have obtained better results than the unsupervised ones, as shown by experimental work and international evaluation exercises such as Senseval tasks.

Even if this method obtained good results, it suffers from the lack of widely available semantically tagged corpora, from which to construct broad-coverage systems. This is known as the knowledge acquisition bottleneck. And the lack of annotated corpora is even worse for languages other than English. So, due to this obstacle, the use of unsupervised or statistical techniques for WSD was tried to avoid the manual annotation of a training corpus.

2.4.3.1.1. Main Approaches to Supervised WSD

Depending on the induction principle they use for acquiring their classification models supervised methods can be categorized or seen as the following approaches [51].

Probabilistic Methods – is statistical methods usually estimate a set of probabilistic parameters that express the conditional or joint probability distributions of categories and contexts which can be described by features. These parameters can be then used to assign to new examples with each particular category that maximizes the conditional probability of a category for the given observed context features. Naive-Bayes algorithm is a type this methods.

Methods Based on the Similarity of the Examples – The methods in this family perform disambiguation by taking into account a similarity metric. This can be done by comparing new examples to a set of learned vector prototypes one for each word sense and assigning the sense of the most similar prototype, or by searching in a stored base of annotated examples for the most similar examples and assigning the most frequent sense among them. The most widely used representative of this family of algorithms is the k-Nearest Neighbor (KNN) algorithm. In this algorithm the classification of a new example is performed by searching the set of the k most similar examples or nearest neighbors among a pre-stored set of labeled examples, and performing an average of their senses in order to make the prediction.

Methods Based on Discriminating Rules – in this category algorithm such as decision lists is can be considered as weighted if-then-else rules where the exceptional conditions appear at the beginning of the list high weights, the general conditions appear at the bottom low weights, and the last condition of the list is a "default" accepting all remaining case according to [27], while

decision trees is a way to represent classification rules underlying data by an n-ary branching tree structure that recursively partitions the training set. The two algorithms use selective rules associated with each word sense.

Methods Based on Rule Combination - combination refers to a set of homogeneous classification rules that are learned and combined by a single learning algorithm. The AdaBoost learning algorithm is one of the most successful approaches which is used to linearly combine many simple and not necessarily very accurate classification rules called weak rules or weak hypotheses into a strong classifier with an arbitrarily low error rate on the training set.

Linear Classifiers and Kernel-Based Approaches - Linear classifiers have been very popular in the field of information retrieval (IR), since they have been used successfully as simple and efficient models for text categorization. It is a hyper-plane in an n-dimensional feature space that can be represented with a weight vector w and a bias b indicating the distance of the hyper-plane to the origin. It includes on-line learning algorithms such as Perceptron, Widrow-Hoff, Winnow, Exponentiated-Gradient and Sleeping Experts.

2.4.3.1.2. Learning Algorithms under supervised WSD

Naive Bayes (NB) – Naive Bayes is the simplest representative of probabilistic learning methods which classifies a new example by assigning the sense that maximizes the conditional probability of the sense given the observed sequence of features of that example. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems and named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. It provides a useful perspective for understanding and evaluating many learning algorithms. Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents.

Exemplar-Based Learning – this algorithm is called Memory-based, Exemplar-based, Instance-based, or Case based learning because, the training step reduces to store all the examples in memory and sometimes it is also called Lazy learning because the generalization is postponed until each new example is being classified. But, the most widely used representative of this family of algorithms is the K-Nearest Neighbour (KNN) algorithm. In this algorithm the

classification of a new example is performed by searching the set of the k most similar examples (or nearest neighbours) among a pre-stored set of labeled examples, and performing an "average" of their senses in order to make the prediction. A very important issue in this technique is the definition of an appropriate similarity or distance metric for the task, which should take into account the relative importance of each attribute and be efficiently computable. The combination scheme for deciding the resulting sense among the k nearest neighbours also leads to several alternative algorithms. Escudero et al. [37] focused on certain contradictory results in the literature regarding the comparison of Naive Bayes and kNN methods for WSD.

Decision Lists (DL) – A decision list was described by [27] as an ordered set of rules for categorizing test instances (in the case of WSD, for assigning the appropriate sense to a target word). It is a simple learning algorithm that can be applied in this domain, which acquires a list of ordered classification rules of the form: "if-then-else" which is "if (**feature**=**value**) then **class**". When classifying a new example **x**, the list of rules is checked in order and the first rule that matches the example is applied. The exceptional conditions appear at the beginning of the list (high weights), the general conditions appear at the bottom (low weights), and the last condition of the list is a default accepting all remaining cases. Weights are calculated with a scoring function describing the association between the condition and the particular class, and they are estimated from the training corpus. When classifying a new example, each rule in the list is tested sequentially and the class of the first rule whose condition matches with the example is assigned as the result.

Decision tree is a predictive modeling technique used in classification and prediction tasks. It is a classifier expressed as a recursive partition of the instance space that is used in data mining to classify objects into values of the dependent variable based on the values of independent variables. There are two main types of decision trees. These are classification trees and regression trees. Classification trees are decision trees used to predict categorical variables, because they place instances in categories or classes. This can provide the confidence to correctly classify the data in which it reports the class probability, which is the confidence that a record is in a given class. And the second one is regression trees, which is a decision trees used to predict continuous variables which are not nominal. It estimates the value of a target variable that takes on numeric value [35].

The structure of decision tree is a tree like structure, where each internal node represents a test on an attribute, each branch characterizes an outcome of the test, and leaf nodes at the end represent classes in which the data is assigned.

AdaBoost (AB) is a general method for obtaining a highly accurate classification rule by combining many weak classifiers, each of which may be only moderately accurate. The main idea of the AdaBoost algorithm is to linearly combine many simple and not necessarily very accurate classification rules (called weak rules or weak hypotheses) into a strong classifier with an arbitrarily low error rate on the training set. AdaBoost has been successfully applied to many practical problems, including several NLP tasks [36]**.**

2.4.3.2. Semi supervised method

To overcome the knowledge acquisition bottleneck problem suffered by supervised methods, these methods make use of a small annotated corpus as seed data in a bootstrapping process [5]. Semi-supervised learning first starts with a supervised learner trained on available data. In a second step, data are added from automatically annotated sources. Semi-supervised approaches, especially when they do not optimize for individual words, often result in no or minimal improvements. This either means that the quality of the data is not good enough to be used for the given purpose, or that the approach in employing the data is not optimal. If the former is true, future approaches to semi-supervised learning must concentrate on distinguishing reliably from unreliably annotated examples.

The semi-supervised or minimally supervised methods are gaining popularity because of their ability to get by with only a small amount of annotated reference data while often outperforming totally unsupervised methods on large data sets. There are a host of diverse methods and approaches, which learn important characteristics from auxiliary data and cluster or annotate data using the acquired information.

2.4.3.3. Unsupervised method

Supervised learning requires many training sentences for each word. Bearing in mind that even in English, for which the most extensive research has been carried out historically, the sense tagged corpora are rather limited. It is a crying necessity to make better use of untagged corpora to be able to perform word sense disambiguation for any word in a running text. This is known as the knowledge acquisition bottleneck. And the lack of annotated corpora is even worse for

languages other than English. So, due to this obstacle, the use of unsupervised or statistical techniques for WSD was tried to avoid the manual annotation of a training corpus.

2.4.3.3.1. Main Approaches to Unsupervised WSD

Type-Based Discrimination – Type-based methods identify sets or clusters of words that are deemed to be related by virtue of their use in similar contexts. These methods often rely on measuring similarity between word co-occurrence vectors, and produce sets of word types. Note that the resulting clusters do not include any information regarding the individual occurrences of each word, which is why they are known as type-based methods. It creates a representation of the different words in a corpus that attempts to capture their contextual similarity, often in a high dimensional feature space. These representations are usually based on counts of word co-occurrences or measures of association between words. Given such information about a word, it is possible to identify other words that have a similar profile and are there by presumed to have occurred in related contexts and have similar meanings.

Token-Based Discrimination – Token-based methods cluster all of the contexts in which a given target word occur based on the similarity of those contexts. Its goal is to cluster the contexts in which a given target word occurs, such that the resulting clusters will be made up of contexts that use the target word in the same sense. Each context in which the target word occurs is a member of one of the resulting clusters.

2.4.3.3.2. Learning Algorithms under unsupervised WSD

Partitioning algorithm – when we say partitioning algorithm, given 'D' a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions, where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are similar, whereas the objects of different clusters are dissimilar in terms of the data set attributes. This includes all the following techniques like.

K-means - K-means algorithm is one of the partition clustering algorithms which can be described as given a set of initial clusters K (k-stands for numbers of clusters), assign each point to one of them and then each cluster center is replaced by the mean point on the respective cluster. A point is assigned to the cluster which is close in to the point [52]. In K-Means, the centroids are computed as the arithmetic mean of the cluster all points of a cluster. The distances are computed according to a given distance measure, that is Euclidean distance. Although K-

means has the great advantage of being easy to implement, but, it has two big drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success [52]. The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity. The k-means algorithm works as follows:-

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.

It then computes the new mean for each cluster. This process iterates until the criterion function converges. The square-error criterion is used, defined

$$E = \sum_{i=1}^{k} * \sum_{p \in ci}^{n} |\mathrm{p} - \mathrm{mi}|^2$$

Figure 2.1: k-means clustering formula

Where 'E' is the sum of the square error for all objects in the data set; 'p' is the point in space representing a given object; and 'mi' is the mean of cluster 'Ci' both p and mi are multidimensional.

In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed.

Expectation maximization – Expectation-Maximization (EM) algorithm is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It can be viewed as an extension of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. In other words, there are no strict boundaries between clusters. Therefore, new means are computed based on weighted measures. EM starts with an initial estimate or guess of the parameters of the mixture model collectively referred to as the parameter vector. It iteratively rescores the objects against the mixture density produced by the parameter vector. The rescored

objects are then used to update the parameter estimates. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster.

The algorithm works as follows:-

Make an initial guess of the parameter vector: This involves randomly selecting k objects to represent the cluster means or centers (as in k-means partitioning), as well as making guesses for the additional parameters.

Iteratively refine the parameters or clusters based on the following two steps:

(a) Expectation Step:

Assign each object **xi** to cluster Ck with the probability

$$\mathrm{p}(xi \in ck) \quad = p(c1|xi) = \frac{p(ck)p(xi|ck)}{p(xi)}$$

Figure 2.2: Expectation formula

Where p (**xi**jCk) = N(mk, Ek(**xi**)) follows the normal i.e., Gaussian distribution around mean, 'mk', with expectation, 'Ek'.

In other words, this step calculates the probability of cluster membership of object **xi**, for each of the clusters. These probabilities are the "expected" cluster memberships for object **xi**.

(b) Maximization Step:

Use the probability estimates from above to re-estimate (or refine) the model parameters. For example,

$$mk = \frac{1}{n}\sum_{n=1}^{n}\frac{\mathrm{xiP(xi} \in \mathrm{Ck)}}{\sum P(xi \in Cj)}$$

Figure 2.3: Maximization steps formula

This step is the "maximization" of the likelihood of the distributions given the data.

Hierarchical algorithm - A hierarchical clustering method works by grouping data objects into a tree of clusters unlike partitioning algorithm. These clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up merging or top-down splitting fashion.

In general, there are two types of hierarchical clustering methods:-

Agglomerative hierarchical clustering: This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category and they differ only in their definition of inter-cluster similarity. These includes:-

Single linkage clustering

One of the simplest agglomerative hierarchical clustering methods is single linkage, also known as the nearest neighbor technique. The defining feature of the method is that distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered. The minimum value of these distances is said to be the distance between clusters **A** and **B.** In other words, the distance between two clusters is given by the value of the shortest link between the clusters. At each stage of hierarchical clustering, the clusters **A** and **B**, for which **D (A, B)** is the minimum, are merged together.

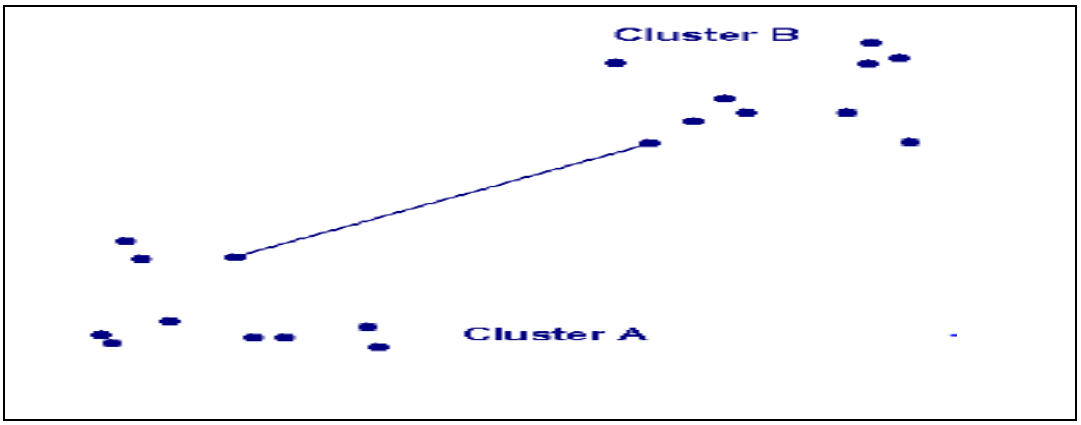

Figure 2.4: single linkage clustering

Complete linkage clustering

The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between groups is now defined as the distance between the most distant pair of objects, one from each group. Here the distance between every possible object pair (i,j) is computed, where object i is in cluster **A** and object j is in cluster **B** and the maximum value of these distances is said to be the distance between clusters **A** and **B.** In other words, the distance between two clusters is given by the value of the longest link between the clusters. At each stage

of hierarchical clustering, the clusters **A** and **B**, for which **D (A, B)** is the minimum, are merged together.

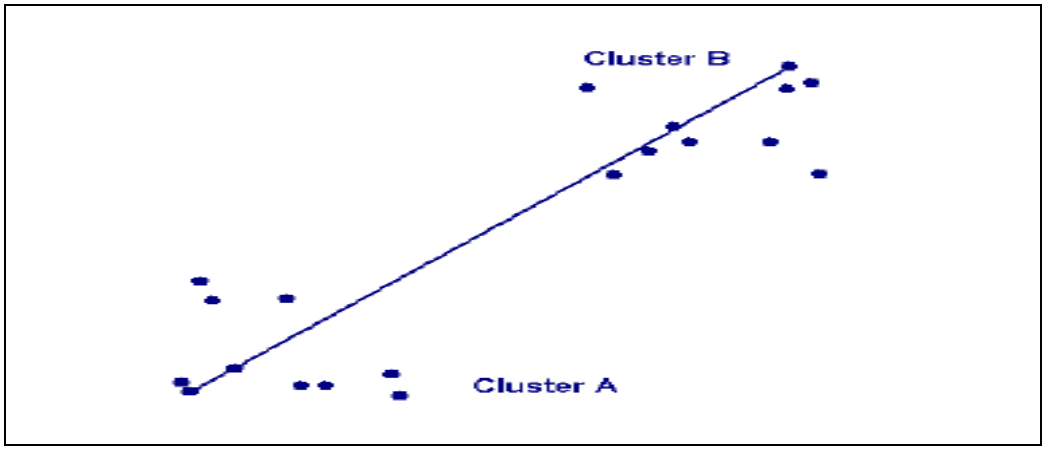

Figure 2.5: Complete linkage clustering

Average linkage clustering

The distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. In the average linkage method, **D(A, B)** is computed as **D(A, B) = TAB / (NA * NB)** Where **TAB** is the sum of all pairwise distances between cluster **A** and cluster **B. NA** and **NB** are the sizes of the clusters **A** and **B** respectively**.** At each stage of hierarchical clustering, the clusters **A** and **B**, for which **D (A, B)** is the minimum, are merged together.

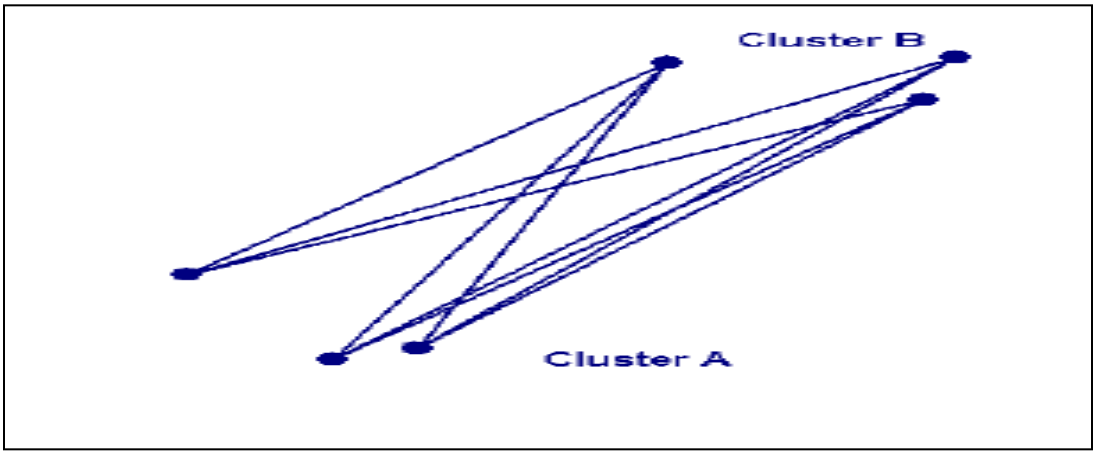

Figure 1.6: Average linkage clustering

Cobweb

Cobweb generates hierarchical clustering, where clusters are described probabilistically. The class attribute play is ignored using the ignore attributes panel in order to allow later classes to clusters evaluation. Doing this automatically through the Classes to clusters option does not make much sense for hierarchical clustering, because of the large number of clusters.

Divisive hierarchical clustering: This is top-down strategy and does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.

2.4.4. Knowledge based approach

In the above discussion of other WSD approach, AI-based was theoretically interesting but not at all practical for language understanding in any but extremely limited domains [15]. Work on WSD reached a turning point in the 1980's when large-scale lexical resources such as dictionaries, thesauri, and corpora became widely available as knowledge based approach [15].

The other is along with corpus-based methods which are applicable only to those words for which annotated corpora are available. So, as opposed to corpus-based techniques, Knowledge based methods for WSD are usually applicable to all words in unrestricted text. Knowledge based methods represent one of the main categories of algorithms developed for automatic sense tagging. 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 use external knowledge resources, which define explicit sense distinctions for assigning the correct sense of a word in context. It represents a distinct category in word sense disambiguation (WSD).

2.4.4.1. Lexical resource of knowledge based approach

Knowledge-based WSD is based on lexical resources like dictionaries, thesauri corpora and Machine-readable dictionaries as stated by Roy A. et.al [31]. These are defined as follows:

Machine-readable dictionaries (MRDs) became a popular source of knowledge for language processing tasks which was contributed significantly to lexical semantic studies but**,** the automatic extraction of large knowledge bases was not fully achieved.

Thesauri are another resource which provides information about relationships between words, like synonymy, antonymy (representing opposite meanings) [18]. Typically, each occurrence of the same word under different categories of the thesaurus represents different senses of that word; i.e., the categories correspond roughly to word senses [40]. Like machine-readable dictionaries, a thesaurus is a resource created for humans and is therefore not a source of perfect information about word relations.

A computational lexicon is also the other resource of knowledge based approach which was begun to construct large scale knowledge bases by hand unlike Machine-readable dictionaries does not automatically extract large knowledge bases. Actually there are two fundamental approaches to the construction of semantic lexicons:

The enumerative lexicons is where in senses are explicitly provided. The most examples of enumerative lexicon of computational lexicon are Word-Net [46]. It is the best known and the most utilized resource for word sense disambiguation in English and some of other language like Amharic, etc. Word-Net combines the features of many of the other semantic relations resources commonly exploited in disambiguation work such as synonymous, hyponomy/ hyperonymy, antonymy, meronymy, etc.

The other is generative lexicons in which semantic information associated with given words is underspecified, and generation rules are used to derive precise sense information [45]. Most WSD work to date has relied upon enumerative sense distinctions as found in dictionaries. However, there has been recent work on WSD which has exploited generative lexicons [41], like polysemy opposed to homonymy are not enumerated but rather are generated from rules which capture regularities in sense creation, as for metonymy, meronymy, etc also.

2.4.4.2. Methods under knowledge based approach

Knowledge based approach can be classified into four main types of methods [51].

2.4.4.2.1. Lesk algorithm

Lesk Algorithm [59] which 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. That is why this method is traditionally considered as dictionary-based method. The idea behind the Lesk algorithm represents the

starting seed for today's corpus based algorithms. The main idea behind the original definition of the algorithm is to disambiguate words by finding the overlap among their sense definitions.

Namely, given two words, W_1 and W_2, each with NW_1 and NW_2 senses defined in a dictionary, for each possible sense pair W_1 i and W_2 j, i = 1..NW_1, j = 1..NW_2, we first determine the overlap of the corresponding definitions by counting the number of words they have in common. Next, the sense pair with the highest overlap is selected, and therefore a sense is assigned to each word in the initial word pair. Previously the Lesk algorithm was evaluated on a sample of ambiguous word pairs manually annotated with respect to the Oxford Advanced Learner's Dictionary and he got a precision of 50–70% by Lesk [59].

But, this original definition of the lesk algorithm has some problems, though different variation version of lesk algorithm was developed. These are:-

Simulated Annealing - one of the notorious problems with the original Lesk algorithm is the fact that it leads to a combinatorial explosion when applied to the disambiguation of more than two words. So, the possible solution to this problem is to use simulated annealing which is another version of lesk algorithm, as proposed by Cowie et al. [48]. They defined a function E that reflects the combination of word senses in a given text, and whose minimum should correspond to the correct choice of word senses. For a given combination of senses, all corresponding definitions from a dictionary are collected, and each word appearing at least once in these definitions receives a score equal to its number of occurrences. Adding all these scores together gives the redundancy of the text. The E function is then defined as the inverse of redundancy, and the goal is to find a combination of senses that minimizes this function. To this end, an initial combination of senses is determined i.e. pick the most frequent sense for each word, and then several itera-replaced with a different sense, and the new selection is considered as correct only if it reduces the value of the E function. They said [48] the iterations stop when example sentences using this optimized Lesk algorithm led to 47% disambiguation precision at sense level and 72% at homograph level. This method was also evaluated by Stevenson and Wilks [39] and a similar average precision was observed during their experiments (65.24%) on a corpus annotated with senses from the Longman Dictionary of Contemporary English (LDOCE).

Simplified Lesk Algorithm is another version of the Lesk algorithm, which also attempts to solve the combinatorial explosion of word sense combinations, is a simplified variation that runs

a separate disambiguation process for each ambiguous word in the input text. In this simplified algorithm, the correct meaning of each word in a text is determined individually by finding the sense that leads to the highest overlap between its dictionary definition and the current context. Rather than seeking to simultaneously determine the meanings of all words in a given text, this approach tackles each word individually, regardless of the meaning of the other words occurring in the same context. A comparative evaluation performed by Vasilescu et al. [50] has shown that the simplified Lesk algorithm can significantly outperform the original definition of the algorithm, both in terms of precision and efficiency. By evaluating the disambiguation algorithms on the Senseval-2 English all words data, they measured a 58% precision using the simplified Lesk algorithm compared to only 42% under the original algorithm.

Augmented Semantic Spaces is another variation of the Lesk algorithm which is also called the adapted Lesk algorithm and it was introduced by [32], in which definitions of related words are used in addition to the definitions of the word itself to determine the most likely sense for a word in a given context. Banerjee and Pedersen [32] employ a function similar to the one defined by Cowie et al. [48] to determine a score for each possible combination of senses in a text, and attempt to identify the sense configuration that leads to the highest score. Unlike the original Lesk algorithm which considers strictly the definition of a word meaning as a source of contextual information for a given sense, Banerjee and Pedersen [32] extend this algorithm to related concepts and their definitions. Based on the Word-Net hierarchy, the adapted Lesk algorithm takes into account hypernyms, hyponyms, holonyms, meronyms, troponyms, attribute relations, and their associated definitions to build an enlarged context for a given word meaning. Hence, they attempt to enlarge the dictionary-context of a word sense by taking into account definitions of semantically related concepts. In comparative evaluations performed on the Senseval-2 English noun data set, they show that the adapted Lesk algorithm on a set of 4,320 ambiguous instances doubles the precision to 32%.

2.4.4.2.2. Semantic Similarity

As the natural property of human language and at the same time one of the most powerful constraints used in automatic word sense disambiguation is words must be related in meaning for the discourse to be coherent. 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 [44].

These methods target the local context of a given word, and do not take into account additional contextual information found outside a certain window size. There are other methods that rely on a global context which attempt to build threads of meaning throughout an entire text, with their scope extended beyond a small window centered on target words.

Measures of Semantic Similarity

There are a number of similarity measures that were developed to quantify the degree to which two words are semantically related. Most such measures rely on semantic networks and follow the original methodology proposed by Rada et al. [44] for computing metrics on semantic nets. A comprehensive survey of semantic similarity measures is reported by Budanitsky and Hirst [42], and a software tool that computes similarity metrics on Word-Net is made available by Patwardhan et al. [43]. There are some of the similarity measures proved to work well on the Word-Net by different researchers which are assumed as input a pair of concepts and return a value indicating their semantic relatedness.

One of the similarity measurements is done by Leacock et al. [47] which determine the minimum length of a connecting path between synsets including the input words. They developed the equation which calculates the similarity as following;

$$similarity(C1, C2) \quad = -\log\left(\frac{\text{path}(C1, C2)}{2D}\right)$$

Figure 2.7: Semantic similarity measures formula

Where Path (C_1, C_2) represents the length of the path connecting the two concepts (i.e., the number of arcs in the semantic network that are traversed going from C_1 to C_2), and D is the overall depth of the taxonomy.

The other similarity measurement is developed by Hirst and St-Onge which integrate into their similarity measure the direction of the links that form the connecting path. The equation is as follows;

$$\text{Similarity } (C_1, C_2) = \text{C- path } (C_1, C_2) - k*d$$

Figure 2.8: Semantic similarity measures formula

Where *C* and *k* are constants, *Path* is defined similarly as above, and *d* represents the number of changes of direction.

Other researchers [49] use an equation which introduces a formula to measure the semantic similarity between independent hierarchies, including hierarchies for different parts of speech. All previously mentioned measures are applicable only to concepts that are explicitly connected through arcs in the semantic network. Mihalcea and Moldovan [49] created virtual paths between different hierarchies through the gloss definitions found in Word-Net. But, this equation was developed to work well for the disambiguation of nouns and verbs connected by a syntactic relation.

They developed the following equation;

$$\text{Similarity (C1, C2)} = \frac{\sum_{k=1}^{|CD_{12}|} \text{Wk}}{\log(descendant(C2))}$$

Figure 2.9: Semantic similarity measures formula

Where $|CD_{12}|$ is the number of common words to the definitions in the hierarchy of C_1 and $C2$, descendants (C_2) is the number of concepts in the hierarchy of C_2, and W_k is a weight associated with each concept and is determined as the depth of the concept within the semantic hierarchy.

Using Semantic Similarity within a Local Context

In the above points discussed about the measures of semantic similarity, its application to the disambiguation of words in unrestricted text is not always a straightforward process. Because a text usually involves more than two ambiguous words, and therefore we typically deal with sets of ambiguous words in which the distance of a word to all the other words in the context influences its meaning in the given text. For this reason local context is used to limit the number of words in the set of ambiguous words.

2.4.4.2.3. Selectional Preferences

Selectional preferences capture information about the possible relations between word categories, and represent commonsense knowledge about classes of concepts. This selectional preference is used for word sense disambiguation as a way of constraining the possible meanings of a word in a given context. But, 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 is that the WSD can improve if large collections of selectional preferences are available.

The application of word-to-word, word-to-class, and class-to-class selectional preferences to WSD was evaluated by Agirre and Martínez [38]. While the results they obtain on a subset of Semcor nouns do not exceed the most-frequent-sense baseline, they observed, however, that class-to-class models lead to significantly better disambiguation results compared to word-to-word or word-to-class selectional preferences. For instance, on a set of 8 nouns, the most-frequent-sense baseline leads to 69% precision and 100% coverage, the word-to-word selectional preferences give 95.9% precision and 26% coverage, word-to-class preferences decrease the precision to 66.9% and increase the coverage to 86.7%, and finally the class- to-class preferences have a precision of 66.6% and a coverage of 97.3%.

Selectional preferences were also evaluated by Stevenson and Wilks [39], who implemented them as features in their larger WSD system. In their work, selectional preferences are derived using the LDOCE semantic codes, a custom-built hierarchy over these codes and grammatical relations such as subject-verb, verb-object, and noun-modifier identified using a shallow syntactic analyzer. They evaluated the individual contribution of each knowledge source in their WSD system, and found that selectional preferences alone could lead to a disambiguation precision of 44.8%.

The use of selectional preferences for WSD is an appealing method, in particular when these preferences can be learned without making use of sense-tagged data. McCarthy D. and Carroll J. [34] are automatically acquired selectional preferences for use in an unsupervised WSD system. They achieved 52.3% precision at a recall of only 20% on the Senseval-2 all-words corpus (58.5% precision on the nouns only), which, incidentally, reveals, the sparse applicability of selectional prefe0rences. The performance of WSD methods based on selectional preferences is however usually exceeded by the simple most-frequent-sense baseline (e.g., the all-words

baseline in Senseval-2 was 57%), suggesting that more work needs to be done for learning accurate selectional preferences [33].

2.4.5. Hybrid approach

The last approach to word sense disambiguation is hybrid approaches which obtain disambiguation information from both corpus-based and explicit knowledge-bases. These systems aim to use the strengths of both approaches to overcome the specific limitations associated with a particular approach and improve WSD accuracy. Sense this approach depends on both a knowledge-driven and corpus-supported theme; it is utilizing as much information as possible from different sources. Yarowsky [5] used bootstrapping approaches where initial data comes from an explicit knowledge source which is then improved with information derived from corpora. He used 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.

2.5. Related work

As briefly discussed above word sense disambiguation (WSD) is one of the most significant and widely studied Natural Language Processing tasks, which is used in order to increase the success rates of NLP applications like machine translation, information retrieval, natural language understanding and language study. So, this issue can be solved in different approaches such as AI-based, corpus based, knowledge based and hybrid based are some of them. By applying all these approach different researchers can be solve the problem of word sense disambiguation for different languages. Some of the work done for different languages is discussed below:

2.5.1. WSD for Amharic

Amharic language is one of the Semitic family languages in which there is a difficulties to identify the sense of words in a given context of sentences. Because it has many ambiguous words due to knowledge acquisition bottleneck. This problem was tried to be solved by different researchers.

Teshome K. [4] was tried to solve the problem of word sense disambiguation (WSD) for Amharic language. [4] Has studied how linguistic disambiguation can improve the effectiveness of an Amharic document query retrieval algorithm. During his study, he developed Amharic disambiguation algorithm based on the principles of semantic vectors analysis and implemented

in Java. He used the Ethiopian Penal Code which is composed of 865 Articles as a corpus. The disambiguation algorithm was then used to develop a document search engine. He developed his own algorithm based on distributional hypothesis stating that words with similar meanings tend to occur in similar contexts. For disambiguation of a given word, he computed the context vector of each occurrence of the words. The context vector was derived from the sum of the thesaurus vectors of the context words. He constructed the thesaurus by associating each word with its nearest neighbors. For evaluating WSD, he used pseudo words which are artificial words rather than real sense tagged words reasoning that it is costly to prepare sense annotation data. He compared his algorithm with Lucene algorithm and reported that the algorithm is superior over the Lucene"s one.

Solomon M. [54] is also the other researchers who tried to solve the problem of word sense disambiguation (WSD) for Amharic language. [54] Used corpus based approach of supervised machine learning methods. For his study, he used monolingual corpora of English language to acquire sense examples and the sense examples are translated back to Amharic to overcome the problem of knowledge acquisition bottleneck. Totally he used five Amharic words as corpus in which he translated these words from a total of 1045 English sense examples collected from British National Corpus (BNC) by using dictionary. Depending on Naive-Bayes classifier experiment is employed from Weka 3.62 package in both the training and testing phases to perform the supervised learning on the preprocessed dataset using 10-fold cross-validation. Solomon have evaluated the classifiers for the five ambiguous words and achieved accuracy within the range of 70% to 83% which is very encouraging but further experiments for other ambiguous words and using different approaches needs to be conducted. Finally he concluded that Naive Bayes methods achieve higher accuracy on the task of WSD for selected ambiguous word, provided that the quality of the labeled data is good.

The other researcher of WSD for Amharic language is Solomon A. [9]. He used a corpus based approach to word sense disambiguation that only requires information that can be automatically extracted from untagged text. Unsupervised machine learning technique was applied to address the problem of automatically deciding the correct sense of an ambiguous word. He used corpus of Amharic sentences, based on five selected ambiguous words, to acquire disambiguation information automatically. A total of 1045 English sense examples for the five ambiguous words

were collected from British National Corpus (BNC) as previous work by Solomon M.[ ]. The sense examples were translated to Amharic using the Amharic-English dictionary which is one approach of tackling knowledge acquisition bottleneck. He tested five clustering algorithms such as simple k means, single link, average link, complete link and expectation maximization algorithms, in the existing implementation of Weka 3.6.4 package. He achieved accuracy within the range of 65.1 to 79.4 % for simple k means, 67.9 to 76.9 for EM and 54.4 to 71.1 for complete link clustering algorithms for the five ambiguous words. Based on the selected algorithms, he concluded that simple k means and EM clustering algorithms achieved higher accuracy on the task of WSD for selected ambiguous word, provided with balanced sense distribution in corpus.

2.5.2. WSD for Hindi

Hindi is a national language of Indian, spoken by million people and ranking 4th by majority spoken in the world. Word sense disambiguation is an important concept that is to be suitable for computer to interpret a word in its proper sense according to its context. Some researchers have also tried to solve the problem of word sense disambiguation for hind language.

One of these researches is Sharma R. [56]. He used knowledge based approach to solve the problems. By applying Hindi Word-Net developed at IIT, Bombay containing different words and their sets of synonyms called synsets. He attempted to solve the ambiguity by making the comparisons between the different senses of the word in the sentence with the words present in the synset form of the Word-Net and the information related to these words in the form of parts-of-speech. Finally, he said that the best approach is knowledge based approaches and among knowledge based approach lesk algorithm he used also lesk algorithm as an example to show its applicability for word sense disambiguation (WSD) of Hindi language. For this Lesk algorithm he took an example paragraph and created its context bag and then extracted the semantic bag for the word to be disambiguated and had done the overlap between both bags corresponding to each sense of the word and then the appropriate sense of the word is found out. And also he said that the approach that he followed can successfully able to resolve the synonymy, antonymy, hypernymy, hyponymy, meronymy and holonymy relations for the different categories of part-of-speech.

The other researchers are Mishra.M.et.al [55]. They used an unsupervised word sense disambiguation algorithm for Hindi languages. Their algorithm learns a decision list using untagged instances and they said some seed instances are provided manually. The evaluation they had done has been made on 20 ambiguous words with multiple senses as defined in Hindi Word-Net. They used a total of 1856 training instances and a total of 1641 test instances for the experiments. They also applied stemming and stop words removal from the context. They measured the performance in terms of accuracy and the accuracy was also measured in terms of correctly classified instances. They made two test runs for their performance measurement and both the test runs have been made for 20 target words. The first test-run observes the results with and without stop word. Though an average accuracy of 86.3% was achieved with stop wards and 87.2% was achieved without stop wards. The second test run is to assess the impact of stop word removal and stemming. Though an average accuracy of 89.5% was achieved with stop ward removal and 91.8% achieved stemming. Finally, they said that the experimental investigation suggests that stop word removal and stemming improves performance of the algorithm for Hindi word sense disambiguation.

2.5.3. WSD for Arabic

Arabic language is also belongs to the Semitic group of languages. Like other language in Arabic language, the main cause of word ambiguity is the lack of diacritics of the most digital documents so the same word can occurs with different senses. So, to solve the problem of word sense disambiguation in Arabic language different researchers used different approaches.

One of the researchers is Elmougy.S.et.al [57]. They used the rooting algorithm with Naïve Bayes Classifier to solve the ambiguity of non diacritics words in Arabic language. They used Al-Shalabi, Kanaan, and Al-Serhan algorithm for root extraction. This algorithm extracts word roots by assigning weights and ranks to the letters that constitute a word. Weights are real numbers in the range 0 to 5. The mapping of weights to letters was determined by extensive experimentation with Arabic text. They applied stop word removal and stemming process. To prove the algorithm efficiency, they implemented it by using Microsoft C# programming language. The written code is running over Microsoft Framework 1.1 and their database is implemented using Microsoft SQL Server 2000. They collected the training set from the World Wide Web (WWW). For each ambiguous word, they collected ten training sample and ten

testing sample for the testing phase. Each document in this training set has one or more occurrence of a given ambiguous word. These documents are presented to the training module with selected sense of the ambiguous word. In the training phase the proposed algorithm has been trained with more than 1600 roots collected from all the training documents where each root has a reference to a set of ambiguous words sense. Finally, their result shows that combing Al-Shalabi, Kanaan, and Al-Serhan rooting algorithm with Naïve Bayes (NB) Classifier decreases the dimensionality of the training documents vector and enhances the accuracy by 16% and decrease the error rate ratio by 17%.

The other researchers are Zouaghi.A.et.al [58]. They used hybrid approach for Word Sense Disambiguation of Arabic Language (called WSD-AL), that combines unsupervised and knowledge-based methods. They applied some pre-processing steps to texts containing the ambiguous words in the corpus of 1500 texts which was extracted from the web, and the salient words that affect the meaning of these words are extracted. After that a Context Matching algorithm is used, it returns a semantic coherence score corresponding to the context of use that is semantically closest to the original sentence. The Contexts of use are generated using the glosses of the ambiguous word and the corpus. In their system they were depended on extraction of signatures, rooting and applying the exact string matching algorithm for the words of the glosses. They applied the Context-Matching algorithm, which measures the similarity between the contexts of use corresponding to the glosses of the word to be disambiguated and the original sentence. For a sample of 10 ambiguous Arabic words, that was chosen by their number of senses out of contexts. Finally, they said the proposed system achieved a precision of 79% and recall of 65%, using roots and signatures identifying each sense which is satisfactory.

2.5.4. WSD for Afaan Oromo

Afaan Oromo is a Cushitic language which is currently the official language of Oromia Regional State and it is the largest regional state among the current Federal States in Ethiopia which is used by Oromo people. Like any other language in Afaan Oromo there is also a big problem of word sense disambiguation. Because, the same words occur with different meanings which is an important concept that is to be suitable for computer to interpret a word in its proper sense according to its context. So, to overcome these problem researchers have attempted to solve it.

Tesfa K. [53] tried to solve the problem of word sense disambiguation (WSD) for Afaan Oromo language. [53] He used a corpus based approach to disambiguation where supervised machine learning techniques are applied to a corpus of Afaan Oromo language, to acquire disambiguation information automatically. He applied Naïve Baye‟s theorem to find the prior probability and likelihood ratio of the sense in the given context for his experimentation. Tesfa also used Java programming language to develop the prototype. Based on the analysis of the language he has done, he developed a WSD algorithm and he tried to cross check his systems with the linguists and an iterative improvement has been made on the system. The system uses information gathered from training corpus to assign senses to unseen examples. Hence, he developed Afaan Oromo corpus from the scratch. The corpus contains 1240 sentences and he evaluated for 5 Afaan Oromo ambiguous words namely ***sirna, karaa, sanyii, qophii*** **and** ***horii.***

By using these words he done two experiments, the first experiment was conducted to evaluate the performance of algorithm. To evaluate the performance of the algorithm, 10-fold cross-validation evaluation technique is used in our experiment. 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, pre-test 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 sought to investigate the effect of different context sizes on disambiguation accuracy for Afaan Oromo ambiguous word, and to find out, if the standard two-word window applicable for other languages and

especially English holds for Afaan Oromo. Finally, for the first experiment he has achieved 79% accuracy and for the second experiment he has found that four-word window on each side of the ambiguous word is enough for Afaan Oromo WSD.

2.5.5. Summary and critique

Obviously the problem in word sense disambiguation can be solved through different approach and techniques. Based on this different work was done by different scholars by using different techniques and methods. As listed above we tried to present a work on four languages, Amharic, Hindi, Arabic and Afaan Oromo by focusing on techniques they used to solve ambiguity and the different experimental result they achieved through different techniques and also the improvement made. In general, for each language different approach was used and different result was achieved.

Different types of algorithms were also applied for disambiguation such as classification (Naïve bayes), clustering (EM, single k means etc) and also knowledge based methods. The best result was obtained with the work done by clustering algorithm. Like that taking what we obtained from the review of the related work we propose a solution to WSD for Afaan Oromo language by clustering algorithm unlike classification which was done previously.

3. CHAPTER THREE: Afaan Oromo Language

3.1. Background of Afaan Oromo Language

The Oromo people are the largest ethnic group in Ethiopia and account for more than 40% of the population. They can be found all over Ethiopia and particularly in Wollega, Shoa, Illubabour, Jimma, Arsi, Bale, Hararghe, Wollo, Borana, and the southwestern part of Gojjam [60]. The official language of these people is called Afaan Oromo.

Until the 1970s, Afaan Oromo was written with either the Ge'ez script or the Latin alphabet. Afaan Oromo was traditionally an oral language only, and has only an official writing system for two decades. Books published on Afaan Oromo in the 19th and 20th centuries used a mixture of Ge'ez (Amharic) script. During the era of 1974 up to 1991, Oromo was written predominantly in Ge'ez script, though the script lacked important consonant and vowel sounds.

In 1991, an alphabet using Latin characters known as **Qubee** was officially adopted for writing Afaan Oromo. Soon after, Oromo was allowed to be used as the medium of instruction in elementary schools throughout Oromia and in print and broadcast media. The adoption of a single writing system allowed a certain amount of standardization of the language, and more texts were written in Oromo between 1991 and 1997 than that it had been in the previous 100 years ago [61]. However, spelling of certain words still varies by dialect and personal knowledge also.

Afaan Oromo is a Cushitic language widely used as both written and spoken by about 30 million people in Ethiopia and Kenya, Somalia which is family of Afro Asiatic languages. It is third largest language in Africa following Kiswahili and Hausa; 4th largest language, if Arabic is counted as an African language [61][62][63]. Besides being a working language of Regional Government of Oromiya, 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 [65][66]. It has also become a language of public media like television and radio programs on such broadcasters as Ethiopian Television (ETV), Oromiya Television (TV Oromiya), Ethiopian Radio, and Radio Fana.

3.2.Dialects and varieties in Afaan Oromo

According to Williams [68] language varies over time, across national and geographical boundaries, by gender, across age groups and by socioeconomic status. When variation occurs within a given language, we call the different versions of the same language dialects. Even if the definition of dialect is difficult, for the sake of discussion, we can construe a dialect as a more or less identifiable regional or social variety of the language-distinguishable in terms of vocabulary, syntax and sometimes pronunciation. Generally, we may see the dialects of language in two ways: one is as every language that is spoken over any significant area is spoken in somewhat different forms in different places; these are its regional dialects. The other is, even in a single community, the language may be spoken differently by members of different social groups; these different forms are social dialects or sociolects."

Afaan Oromo is a language with different dialects classified by different scholars. Gragg [69] said that "Afan Oromo spoken in Ethiopia might be classified into four dialect areas, namely: Western (Wallagga, Iluu Abbaa Bor, Jimma), Central (Shawaa), Eastern (Hararge) and Southern (Arsii-Baale, Gujii and Boorana). The Baate and Raayyaa of Wollo and Tigray, respectively, however, have not been included in this classification." In Kenya, Heine [70] recognizes two major dialect areas, 'Central Afan Oromo' and 'Tana Afan Oromo'. The other classification by Lloret [71] divides Afaan Oromo into various dialects as Western and Eastern Afaan Oromo groups with the former encompassing Raayyaa, Baate, Macca and Tuulama, and the latter including Harar, Arsii, Boorana, Gabra, Orma and Waata. Now a day's Afaan Oromo has four major varieties i.e Borana-Arsi-Guji Oromo, Eastern Oromo, Orma (Oromo in Kenya) and West Central Oromo. These four varieties depend on geographical area. Even if there is strong similarities among these four varieties, but there is also difference between them [86].

3.3. Alphabets in Afaan Oromo languages

Afaan Oromo is a language that is spoken in a way it is written. Additionally it can be characterized by the sound that it is the same in every word in contrast to English in which the same letter may sound differently in different words. Since 1991 Afaan Oromo uses Latin alphabets (Roman alphabet) but with some modifications on sound of consonant and vowels [72]. It has 28 letters called qubee'. However, later on a new letter 'Z' was included in the alphabet as there are words which require the letter. For example: Ziqaya (gold), Zeeytuuna

(guava), Azoole (river in Arsii), Zeekkara (Opera), Zalmaaxaya (mess), Waziiza (fire place or fire work) and Zawii (insanity) are Afaan oromo words written using Z‘. Additionally ‘P‘and ‘V‘are also added. ‘P‘ and ‘V‘ letters are not Afaan Oromo letters because there is no Oromo word written by use of either of them. But they are included by considering the fact of handling borrowed terms from other languages like English. For example: ‘Police‘, ‘Piano‘, ‘Television‘, ‘video‘, and etc. To sum up currently there are 31 **Qubee** letters of Afaan Oromo including ‘Z‘, ‘P‘, and ‘V‘[73].

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

Table 3.1: Upper Case and lower case alphabets of Afaan Oromo

Vowels

Afaan Oromo vowels are similar to that of English but sound differently. There are five vowels **a, e, o, u** and **i**. These vowels pronounced in sharp and clear fashion which means, each

and every word is pronounced briefly. These vowels are classified as short (a, e, i, o, u) and 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. Example:

Afaan Oromo	**English**
Hara	Lake
Haaraa	New
Boru	Tomorrow
Booruu	Dirty

Consonants

Most Afaan Oromo constants do not differ greatly from other languages, but there are some exceptions and few special combinations. It includes:

i. The single consonant like "g, h, k, etc" and

ii. The combinations consonant like NY and DH.

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

Afaan Oromo	**English**
Badaa	Bad
Baddaa	Highland
Hatuu	to steal
Hattuu	Thief

In Afaan Oromo alphabet, gemination is not obligatorily marked for the digraphs like (DH, SH, NY, CH, PH, TS) [65].

3.4. Afaan Oromo Grammars

Afaan Oromo is a grammatically complex language with its own morphology, syntax and semantics. Afaan Oromo also have its own grammar called **Seer-luga**. It includes all the following points.

Like a number of other African languages, Afaan Oromo has a very complex and rich morphology [64]. But, 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 [74]. A morphological-analyzer reduces the different forms of an ambiguous word into their root forms and plays an important role in this regard.

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.

3.4.1. Parts of speech (POS)

In Afaan Oromo seven parts of speech are recognized as discussed in this research.

3.4.1.1. Afaan Oromo Noun

Almost all Afaan Oromo nouns in a given text have number, gender and definiteness 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 [64].

Gender

Afaan Oromo recognizes two genders, masculine and feminine. There is no neutral gender as in this language [73]. Nouns may ends with *-eessa* (male) and *-eettii* (female), as do adjectives when they are used as nouns: *obboleessa* 'brother', *obboleettii* 'sister', *dureessa* 'the rich one (male.)', *dureettii* 'the poor one (female). Grammatical gender normally agrees with biological gender for people and animals; thus nouns such as *abbaa* 'father', *ilma* 'son', and *sangaa* 'ox' are masculine, while nouns such as *haadha* 'mother' and *intala* 'girl, daughter' are feminine.

However, most names for animals do not specify biological gender [73]. Names of astronomical bodies are feminine: *aduu* 'sun', *urjii* 'star'. The gender of other inanimate nouns varies somewhat among dialects.

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

Numbers

There are singular and plural numbers in Afaan Oromo. But nouns that refer to multiple entities are not obligatorily plural. Some singular noun may refer multiple entities: *–nama* –man or –people, *–nama shan* –five men or –five people [73]. When it is important to make the plurality of a referent clear, the plural form of a noun is used. Noun plurals are formed through the addition of suffixes. The most common plural suffix is *-oota*; a final vowel is dropped before the suffix, and in the western dialects, the suffix becomes *-ota* following a syllable with a long vowel: *mana* 'house', *manoota* 'houses', *hiriyaa* 'friend', *hiriyoota* 'friends', *barsiisaa* 'teacher', *barsiiso(o)ta* 'teachers'. Among the other common plural suffixes are *-(w)wan*, *-een*, and *-(a)an*; the latter two may cause a preceding consonant to be doubled: *waggaa* 'year', *waggaawwan* 'years', *laga* 'river', *laggeen* 'rivers', *ilma* 'son', *ilmaan* 'sons'[73].

Definiteness

There is no indefinite article (such as a, an, some) in Afaan Oromo. The definiteness article '_the' in English is *'(t)icha'* for masculine nouns (the ch is geminated though this is not normally indicated in writing) and '*-(t)ittii'* for feminine nouns in Afaan Oromo. Vowel endings of nouns are dropped before these suffixes: *karaa* 'road', *karicha* 'the road', *nama* 'man', *namicha/namticha* 'the man', *haroo* 'lake', *harittii* 'the lake'. Note that for animate nouns that can

take either gender, the definite suffix may indicate the intended gender: *qaalluu* 'priest', *qaallicha* 'the priest (male)', *qallittii* 'the priestess (female)'. The definite suffixes appear to be used less often than the in English, and they seem not to co-occur with the plural suffixes [73].

3.4.1.2. Afaan Oromo Pronouns

Afaan Oromo pronouns includes personal pronouns (persons speaking, the persons spoken to, or the persons or things spoken about) listed in the table below and others like indefinite pronouns, relative pronouns connect parts of sentences and reciprocal or reflexive pronouns.

English	Base	Subject	Dative	Instrume	Locative	Ablative	Possessive
I	ana, na	ani, an	naa, naaf,	Naan	Natty	narraa	koo, kiyya
You	Si, Isin	Ati,Isini	sii,siif, sitti	Siin, nuun	Sitti, nu,tti	Sirraa, isinirra	Kee, Keessan
He	Isa	Inni	isaa,atti	Isaatii	Isatti	isarraa	(i)saa
She	isii,ishii ,ishe	isiin, etc.	ishii, ishiif, ishiitti, etc.	ishiin, etc.	ishiitti, etc.	ishiirraa , etc.	(i)sii, (i)shii
We	Nu	nuti, nu'i, nuy, nu	nuu, nuuf,	Nuun	nu, tti	Nurraa	keenna, keenya [teenna,teenya(f.)]
They	Isaan	Isaani	isaanii, isaaniif,	Isaanii, Tiin	Isaanitti	Isaanirraa	(i)saani

Table 3.2: Examples of pronouns in Afaan Oromo

3.4.1.3. Afaan Oromo Adjectives

Afaan Oromo Adjectives are words that describe or modify another person or thing in a sentence [73].

bifoota —Colors	Hamma guddina-Sizes	Boca-shapes	Dhandhama-Tastes
Baluu--blue	Guddaa--big	Geengoo-circular	Hadhaawaa--bitter
Daalacha--gray	Gabaabaa--short	Sirrii-straight	Asheeta--fresh
Diimaa--red	Xinnaa--small	rolarfee-square	Sogidaawa--salty
Adii--white	Dheeraa--tall	rolsadee- triangular	oganaawaa--sour

Table 3.3: Examples of Adjective in Afaan Oromo

3.4.1.4. Afaan Oromo Adverbs

Afaan Oromo adverbs are words that modify can modify verbs, adjectives (including numbers), clauses, sentences and other adverbs [73]. The four categories of adverbs in Afaan Oromo are adverb of time, adverbs of place, adverbs of manner and adverbs of frequency [73].

Adverbs of time			
Afaan Oromo	**English**	**Afaan Oromo**	**English**
Kaleessa	Yesterday	**Adverbs of manner**	
harr'a	Today	Baayyee	Very
Bor	Tomorrow	Dafee	Fast
Amma	Now	Cimaa	Hard
Adverbs of place		Suuta	Slowly
Carefully	Here	Qalbiidhan	Carefully
Achi	There	**Adverbs of frequency**	
gara sana	over there	yeroo hunda	Always
Eessayyu	Nowhere	gaaffii gaaf	Sometimes
Fagoo	Away	darbee darbee	Rarely
Ala	Out	Yoomiyyuu	Never

Table 3.4: Examples of Adverbs in Afaan Oromo

3.4.1.5. Afaan Oromo verbs

An Oromo verb consists minimally of a stem, representing the lexical meaning of the verb, and a suffix, representing tense or aspect and subject agreement.

Afaan Oromo	**English**
Beekuu	Understanding
Deemuu	Walking
Fiiguu	Running
Rafuu	Sleeping
Nyaachuu	Eating
Dubbisuu	Reading

Table 3.5: Examples of verbs in Afaan Oromo

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 [64]. The extensive inflectional and derivational features of Afaan Oromo are presenting various challenges for text processing and information retrieval tasks in the language. In information retrieval, the abundance of different verbs forms and lexical variability may result in a greater likelihood of mismatch between the forms of a keyword in a query and its variant forms found in the document index databases. Stemming, a technique that is used to bring morphological variants of a word into the root or stem word, plays an important role in this regard.

3.4.1.6. Afaan Oromo Preposition

There are many prepositions in Afaan Oromo such as links, nouns, pronouns and phrases to other words in a sentence. The word or phrase that the preposition introduces is called the object of the preposition [73].

Oromo Prepositions	**English prepositions**
Akka	As
Itti	At
Garuu	But
Dhaan	By
ta'uyyuu	Despite
Utuu	During
Irraa	From
Ala	Out
Fi	Plus

Table 3.6: Example of prepositions in Afaan Oromo

3.4.2. Afaan Oromo Writing System and Punctuation marks

With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991. There are about twenty-six consonants and ten vowels (five short and five long) in the Afaan Oromo language [65]. The Oromo writing system is a modification to Latin writing system. Thus, the language shares a lot of features with English writing with some modification. The writing system of the language is known as “Qubee Afaan

Oromo" is straight forward which is designed based on the Latin script. Thus letters in 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 [67]. This means 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 [75].

3.4.3. Syntax in Afaan Oromo

Syntax is the study of sentence structure. It attempts to describe what is grammatical in a particular language in terms of rules. When sound arrange in a meaningful sense they transfer the message of the speaker they should obey rule and regulation of the specific language. Afaan Oromo is one of the languages that have its own writing system and syntaxes which are used across all Afaan Oromo dialects. It follows Subject-Object-Verb (SOV) format. But because it is a declined language (nouns change depending on their role in the sentence), word order can be flexible, though verbs always come after their subjects and objects. Typically, indirect objects follow direct objects. Examples:

Afaan Oromo syntax	**English meaning**
Ani kubbaa siif nan ha'e I ball to you threw	"I threw the ball to you
Isheen Ameerikaa irraa dhufte She America from came	"She came from America"

There are some modifiers in syntax of Afaan Oromo i.e. adjectives come after the nouns they modify, adverbs that modify adjectives go before the adjective, adverbs that modify verbs and adverbial clauses, and relative clauses tend to go at the beginning of the sentence before the subject. Examples:

Afaan Oromo syntax	**English meaning**
Biirii dooqeen kee kutaa koo keessa jira pen blue your room my in is	Your blue pen is in my room
Eessa akka ta'e ani hin beeku Where that it is I do not know	I do not know where it is
Hagam manni postaa fagaata? How post office is far?	How far is the post office
Edana maal gotta? Tonight what will you do?	What are you doing tonight?

3.5. Ambiguities in Afaan Oromo

There are different types of ambiguities in natural language processing. These are Phonological, Lexical, Structural, Referential and Semantic ambiguity. Here, there are also different types of ambiguities in Afaan Oromo. We now summarize each type of ambiguity with example.

3.5.1. 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".

3.4.2. 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:

- A French man who teaches History.
- A person who teaches French History.

3.4.3. Referential ambiguity

Referential ambiguity arises when a word or phrase in the context of a particular sentence refers to two or more properties or things. Usually the context tells us which meaning is intended, but when it does not 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 and we assign the wrong meaning to the words. 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*),
- Weight (*furdaa dha*),
- Social status (*kabajamaa dha*) and so on.

Example2: "*Gaadisaan gatii ebifaamef gamade*." The sentence has two different meanings:

- Gadisa was pleased because he graduated.
- Somebody was pleased because Gaadisa graduated
- Gadisa was pleased because he offered blessing.

3.4.4. 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:

- The flower has grown.
- 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 in Afaan Oromo: "*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".

3.4.5. Lexical ambiguity

The lexical ambiguity of a word or phrase pertains to its having more than one meaning in the language to which the word belongs. "Meaning" here refers to whatever should be captured by a good dictionary. For instance, the word "bank" has several distinct lexical definitions, including "financial institution" and "edge of a river". Another example is as in "apothecary". One could say "I bought herbs from the apothecary". This could mean one actually spoke to the apothecary (pharmacist) or went to the apothecary (pharmacy).

The context in which an ambiguous word is used often makes it evident which of the meanings is intended. If, for instance, someone says "I buried $100 in the bank", most people would not think someone used a shovel to dig in the mud. However, some linguistic contexts do not provide sufficient information to disambiguate a used word. For example, Lexical ambiguity can be addressed by algorithmic methods that automatically associate the appropriate meaning with a word in context, a task referred to as word sense disambiguation.

The use of multi-defined words requires the author or speaker to clarify their context, and sometimes elaborate on their specific intended meaning (in which case, a less ambiguous term should have been used). The goal of clear concise communication is that the receiver(s) have no misunderstanding about what was meant to be conveyed. An exception to this could include a politician whose "weasel words" and obfuscation are necessary to gain support from multiple constituents with mutually exclusive conflicting desires from their candidate of choice. Ambiguity is a powerful tool of political science.

More problematic are words whose senses express closely related concepts. "Good", for example, can mean "useful" or "functional" (*That's a good hammer*), "exemplary" (*She's a good student*), "pleasing" (*This is good soup*), "moral" (*a good person* versus *the lesson to be learned from a story*), "righteous", etc. "I have a good daughter" is not clear about which sense is intended. The various ways to apply prefixes and suffixes can also create ambiguity ("unlockable" can mean "capable of being unlocked" or "impossible to lock").

This lexical ambiguity has been the focus for the study of context effects in many natural languages processing like word recognition, information extraction, speech recognition and information rwetrival etc.

Simply lexical ambiguity refers to a case in which either a lexical unit belongs to different parts-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 parts-of-speech category [23]. There are different factors that can cause lexical ambiguity such as Categorical Ambiguity, Homonymy and others.

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:

- The teacher is getting into the class room. [With nominal meaning]
- The teacher is in the history room. [With verbal meaning]

When we say **homonyms** they 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:

- Tolaa has a huge weight
- Tolaa is a respected person

4. CHAPTER FOUR: System Architecture

4.1. Introduction

This chapter is devoted to describe the design and data preparation of WSD for Afaan Oromo language. It mainly focuses on architecture of Afaan Oromo WSD, data requirement, corpus preparation, selected clustering algorithms for testing experiment and Performance evaluation technique. In addition to this, the detail description of components on the architecture and document preprocessing are also presented.

4.2. Architecture of Afaan Oromo WSD system

Figure 4.1 below presents the general architecture of the proposed Afaan Oromo WSD systems. Generally, it contains different steps as firstly preparing corpus of sentences which contains ambiguous words. The second activity is preprocessing which includes tokenizing the process of splitting up the text into a set of tokens usually words based on the boundaries of a written text, removing stop words from them which is words that have no significant discriminating powers in the meaning of ambiguous words and finally stemming which is reducing morphological variety of words into their root or stem. The third activity is formatting and preparing the dataset which is suitable for the weka tools to experiment. Then, finally model developing and disambiguating is done by applying different clustering algorithms. While we are applying clustering algorithms we can develop models and disambiguate the sentence into their senses and perform the evaluation of the models.

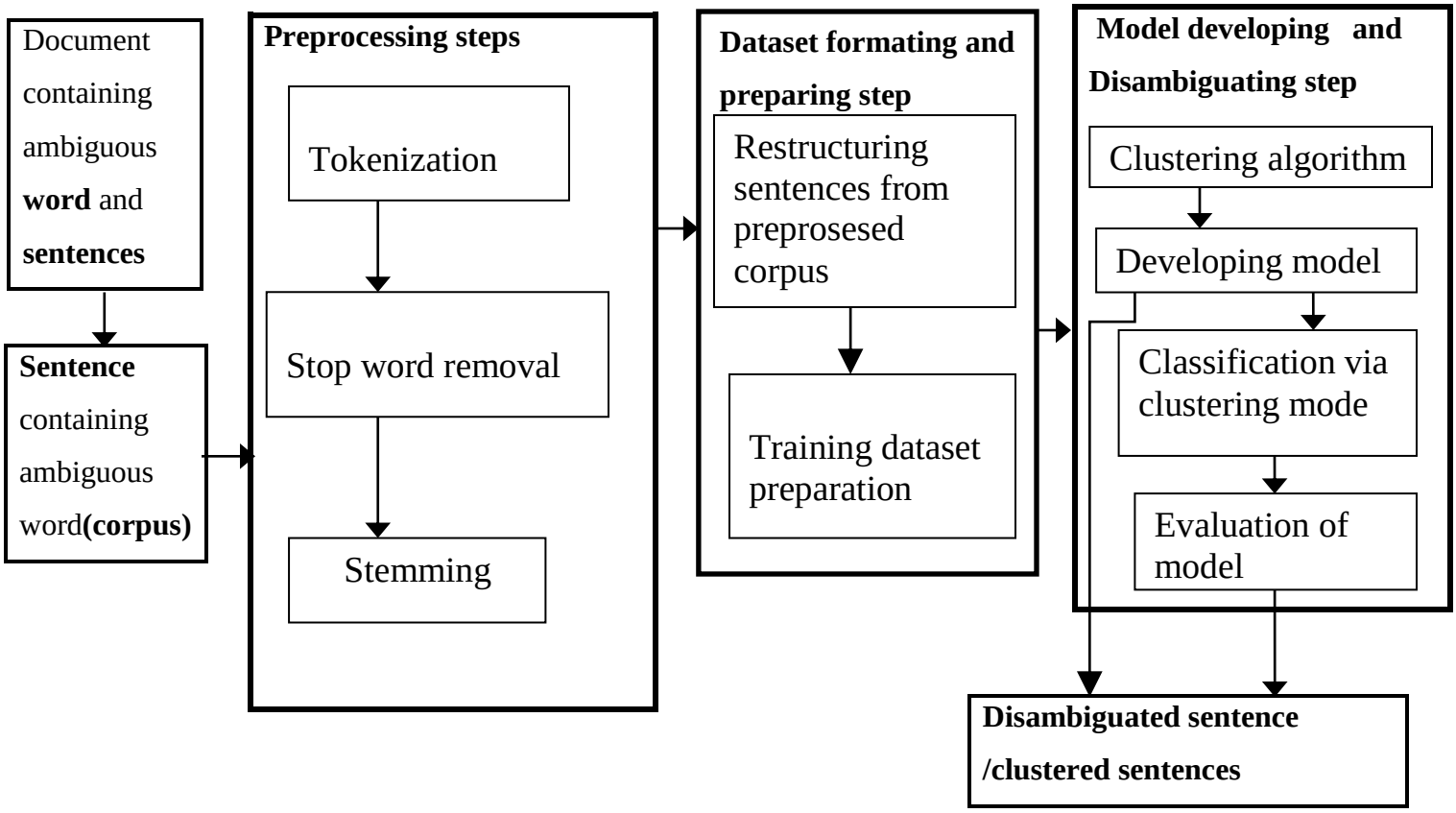

Figure 4.1: Architecture for Unsupervised Afaan Oromo Word Sense Disambiguation System

4.3. Data requirements

For different experiments especially in the linguistic area, data is the main resource that we have to gather from different sources and prepare them as corpus. The corpus is a fundamental tool for any type of research on language. Regarding our study, Afaan Oromo corpus was used which contains seven ambiguous words.

As we discussed in the literature review, supervised approach for Afaan Oromo WSD was done by Tesfa [53] in which he prepared the corpus from five Afaan Oromo ambiguous words. In addition to the five ambiguous words used by Tesfa [53], we used two more ambiguous words to develop the corpus. Tesfa [53] used a total of 1240 sentences with each ambiguous word having 100 and more than (for some words) sense sentences and he constructed these sentences from the scratch. Likewise we have constructed 501 sentences in addition to 1000 sentences adopted from Tesfa and totally 1501 sentences for seven ambiguous words each of which have two senses were used for this research.

Generally, for our work Afaan Oromo words are gathered from different dictionaries. We collected seven ambiguous words as data requirements (see table 4.1) which are balanced and distributed means each sense have some what the same size of sentences, because accuracy of machine learning algorithms degrade significantly when the training and testing samples have different distributions for the senses [77]. Therefore, we prepared 1501 sentences for seven ambiguous words each of which having more than and 100 sentences for each context.

No.	Ambiguous word	Sense of ambiguous words	Count	Total
1	Sanyii	Seed	100	200
		Type	100	
2	Karaa	Road	102	205
		Way	103	
3	Ulfina	Weight	110	221
		Respection	111	
4	Ifa	Light	120	235
		Clear	115	
5	Qophii	Program	101	201
		Preparation	100	
6	Sirna	Event	100	210
		Systems	110	
7	Horii	Money	115	229
		Cattle	114	

Table 4.1: Data requirement

4.4. Corpus preparation

A corpus is a collection of texts used for linguistic analyses, usually stored in an electronic database so that the data can be accessed easily by means of a computer [21]. It is the data for the lexical sample tasks. Typically it is a large number of naturally occurring sentences containing a given target word, each of which has been tagged with a pointer to a sense entry from the sense inventory. Corpus usually consists of many numbers of words that authentic (naturally occurring) either spoken or written [78]. According to [79], corpus is expected to have the following features:

Sampling and representativeness - many natural languages have large number of words and it is difficult to prepare the corpus that constitutes all the words in the language. Sample words are taken and used which can be representative of the other words. The sample has to represent variety of the words and their morphological and structural variation.

Machine readable - nowadays many corpora are also expected to be machine readable even though it is not always true.

A corpus may exist in two different forms: unannotated and annotated corpus [21]. Annotated corpus is a collection of texts that contains grammatical or linguistic information [21]. Whereas unannotated corpus is a collection of text without linguistic information [21]. Annotated corpus can be used for various purposes. In linguistics, properly annotated (tagged) corpus can be used to study linguistic features such as morphology and phonology of a language. It can also be used for part of speech taggers and WSD.

The corpus will be provided to the system as training data so that the system can learn/adapt some pattern from the corpus for each word or sentence. The size of the corpus affects the learning tendency of the system. Larger size of corpus provides greater learning tendency for the system. As a result, accuracy of the system will be better to automatically assign a meaning to ambiguous word. However, there is no such large size corpus which is already prepared for Afaan Oromo language for disambiguation purpose.

Totally, preparation of this large size corpus is expensive and time consuming task. As a result of this, we created the Afaan Oromo corpus manually and 1000 from [53]. First we had to select and collect ambiguous words which have two or more senses contextually from an Afaan Oromo dictionary. In determining the words to be used in WSD, the most common words and their senses were chosen with care. 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 first sentence and ***beelada*** **(cattle)** in the second sentence. 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.

Then the ambiguous word was searched in the document and sentences including this word were examined and used as corpus which was collected from different magazines, bulletins and news papers which have ambiguous words. As it is discussed in [80] bulletins, magazines and newspapers contain many social, economical, technological and political content. Hence, they are good source for collecting representative corpus for natural language processing. We also collected additional sentences from different sources such as Afaan Oromo books, Internet and others. In the case of our works, ambiguous words in the corpus have been prepared with their word senses with the help of Afaan Oromo dictionary. For instance:-

- ✓ *Kitaba bade **soquu** deemte.*
- ✓ *Lafa irra marga **soquu** deeme.*

The Word "***soquu***" in the above two sentence is annotated with sense (***barbadu*)** and (***qulquleessu*)** respectively.

During annotation of an ambiguous word, we ignored tagging of the ambiguous word with full definition. Rather a word is described using a single word or statement found in a dictionary. Those words are brief statements which are a unit of language that native speakers can easily identify them". For example, the meaning or definition of the word "*soquu*" in the first sentence is: "***wanta bade tokko barbadaani* argachu (looking carefully in order to find something missed or lost**)" which is to mean ***barbaadu* (find)**" using a single word. The meaning (definition) of the word "***soquu*"** in the second sentence is: "***waan xuraa'e tokko qulqulessu* (cleaning or removing dirt or unwanted matter from the surface of something by rubbing it hard)"** which is to mean ***qulqullessu* (scour)** using a single word. Therefore the above two sentences are annotated in the corpus using a single word i.e. "*barbaadu*" and "*qulqullessu*" respectively as follows:

- ✓ *Kitaba bade **soquu (barbadu)** deemte.*
- ✓ *Lafa irra marga **soquu (qulqullessu)** deeme.*

4.5. Document preprocessing

4.5.1. Tokenization

Tokenization is step which splits up the text into a set of tokens usually words based on the boundaries of a written text. This process detects the boundaries of a written text. Tokenizing of a given text depends on the characteristics of language of the text in which it is written.

In Afaan Oromo the word demarcation is white space. Thus, Afaan Oromo tokenizer splits text into its constituent words usually by considering white spaces and punctuation marks. Punctuation mark usage in Afaan Oromo is similar to that of English which includes semicolon (;), comma (,), full stop (.), question mark (?) and exclamation mark (!). These punctuation marks are removed from the text because they do not have any relevance in identifying the meaning of ambiguous words in WSD.

In this research tokenization is needed for some purposes that stop word removal and stemming is performed at word level.

For example: *Magaalaa Adaamaa fi Shashamannetti qophiin spoortii yeroo lamaf qophaa'e.* The tokens will be *Magaala, Adaama, fi, Shaashamannee, Qophiin, spoortii, yeroo, lamaaf, qophaa'e*. So we can define token as an instance of a sequence of characters.

1. *Read corpus(file)*
2. *Define word delimeter to space*
3. *Read each sentence from corpus(file)*
   While file delimiter
   If there is word delimeter
   Open new file
   Put each terms as separate tokens in new file
   End if
4. *Close file*

Figure 4.2: Algorithms for tokenization

4.5.2. Stop Word Removal

Stop word removal step is used to remove stop words from the input text. Stop words are words that have no significant discriminating powers in the meaning of ambiguous words. Stop words mainly consist of prepositions, conjunctions, articles, and particles. There is no standardized stop lists available for Afaan Oromo.

There are various techniques used to remove stop words. Among this IDF (inverse document frequency) value and dictionary lookup are the common one. 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. This technique is much easier for well studied languages that have standard list of such words.

A research was conducted by Mishra [55] for Hindi language by using unsupervised machine learning algorithms. As he reported, Stop word removal is needed that it performs better.

For the purpose of our work, list of around 100 stop words like for example yookin, immoo, akkasumas, Fi, booda , immoo, boooddee, moo, eegana , illee, eegasii, akka, erga, jechuu, eega and etc. that is compiled from Afaan Oromo books during implementation of a stemmer by Debela T. [81] is used. These words need to be removed during preprocessing phase. As a result of this, dictionary lookup was employed for this study.

```
1. Open stop word list
2. Read a sentence from corpus
      For each word in the sentence
          If word is in stop word list
            Open new file
              Then Remove word
              Put the left words in new file
        End if
      End for
3. Return the remaining words
4. Stop processing
```

Figure 4.3: Algorithm for stop word removal

4.5.3. Stemming

Stemming is a step that reduces morphological variants of words into root (base) by removing affixes or to a common form stem by stripping of its affixes [85]. In morphologically complex languages like Afaan Oromo, a stemmer will lead to significant improvements in WSD systems [64]. According to [64], morphologies of a word, especially suffixes, can be composed of attached, derivational, and inflectional suffixes. Afaan Oromo attached suffixes are particles or postpositions. Derivational suffixes are mainly used for the formation of new words in the language from stem or base form of a word. Inflectional suffixes of a word may indicate tense, case, plurality (number), and gender differences.

The most common order/sequence of Afaan Oromo suffixes (within a given word) is [64] : <stem> <derivational suffixes> <inflectional suffixes> <attached suffixes>. Thus, Afaan Oromo stemmer is expected to remove (from the right end of a given word) first all the possible attached suffixes, then inflectional suffixes and finally derivational suffixes step by step. For example, the word *barattootarratti* (on the students) is composed of *itti*, *irra* (attached suffixes), *oota* (inflectional suffix), *at* (derivational suffix), and *bar* (the stem). Therefore firs *-tti*, then *-rra*, then *-oota* and finally *-at* is removed to get the root "*bar-*". Affixes that are formed out of this sequence can also be removed.

The two way of reducing morphological variety of the word were applied in this work:

The first is stemming which reduces morphologies of words by stripping off its endings. For Afaan Oromo language this was developed by Debela Tesfaye [81]. Debela's algorithm is rule based Afaan Oromo Stemmer. It is porter stemmer based algorithm for Afaan Oromo text document which is good in information retrieval. Debela identified six stemmer rule clusters depending on the Afaan Oromo language grammatical rule such as measure (the number of vowel consonant sequence), ending of the remaining stem with specific character, ending of the remaining stem with consonant, ending of the remaining stem with short or long vowel, matching of the ending of the word with one of the suffixes and [81]. Such type of stemming is used in information retrieval and web search (IR). It is a procedure that reduces all words with the same stem to a common form by stripping off the affixes of the words rather than finding only the root of the words. These affixes are Prefixes which precede the stem, suffixes which follow the stem, circumfixes do both, and infixes are inserted inside the stem [81].

1. *Read the corpus*
2. *Read the next word to be stemmed*
3. *Open stop word file*
   Read a word from the file until match occurs or end of file reached
   If word exists in the stop word list
   Go to 5
4. *If word matches with one of the rules*
   Remove the suffix and do necessary adjustment
   Go back to 3
   Else
   Go to 6
5. *Return the word and record it in stem dictionary*
6. *If end of file not reached*
   Go to 1
   Else
   Stop processing
7. *If there is no applicable condition and action exist*
   Remove vowel and return the result

Figure 4.4: Stemming algorithm adopted from [81]

The second is lemmatization which reduces morphology of the words into similar root which is used in natural language (NL). This is adjusting stemmer developed by Debela [81]. That means we modified the stemmer to work to reduces affixes and to find root of the words. The following is sample example of the algorithms.

```
1. Read corpus, stop word list, word tokenize
2. While not end of corpus file do
        For each word the token
            If word starts with prefix
              If word not in stop word list then
                 Remove prefix
              End If
            End If
            If word ends with suffix
                If word not in stop word list then
                   Remove suffix
                End If
            End if
        End for
3. IF there is no applicable condition and action exist
        Return the word as it is
4. End while
5. Close files
```

Figure 4.5: Lemmatization algorithm modified from (81)

4.6. Dataset preparation and description

Dataset is a collection of many data which is stored in an organized format to make suitable for Weka tools. To prepare the dataset we restructured the preprocessed corpus into the format of sentences. Then, we prepared the datasets from preprocessed and restructured sentences of corpus in an organized way for the seven ambiguous words. As standard approach to WSD, to consider the context of the ambiguous word we use the information from its neighboring or collocation words. We prepared our dataset based on the context or neighbor words to the ambiguous words in sentences. This information is gathered from text representation of knowledge source (i.e. corpus) which is an unstructured source of information. To make it

suitable for input to WSD, it is usually transformed into a structured format. The dataset can be created after all the preprocessing activity of the input text is usually performed, which includes tokenizing, stop-word removing and stemming. So, once all the necessary preprocessing tasks were done on the corpus, preparation of dataset for training and evaluating the selected algorithms is followed.

For clustering there is no need to split the data into training and test sets for evaluation because of unsupervised nature of clustering algorithms [83]. Since, train-test split is used to avoid overfitting in machine learning, in unsupervised clustering we cannot need to evaluate and we cannot overfit in this way because in clustering there is no learning labels means something that is already labeled, but we want to discover some new structure in our data.

Any dataset may contain different components and characters such as the attribute, value and data type of the value. Attributes it indicate the components which are stored as head of the value. Value is the components to describe the attributes. The value may be exist in different formats or data types such as nominal, numerical and etc. of which we used nominal data for our studies.

There are three different groups of attributes in our datasets such as the **contexts** or the word surrounding the ambiguous word, the **ambiguous word** itself and the **meaning** of the ambiguous words. Context attributes were assigned as the left and right context of the ambiguous words in the sentences Rcontext (k) and Lcontext (k) which refer to eight words to the left and right from ambiguous words means the words that surrounds the ambiguous word to the right and left respectively, where k ϵ (1, 2, …, 8), the 'ambiguous_word' attribute holds the ambiguous word and Meaning attribute takes the senses of the ambiguous word, but the word classes are not practically used for experimentation or clustering senses rather, they were used for evaluation of clustering assignments. In the datasets all the sentences may not be equal in length, so the k^{th} left or right word from the ambiguous word may not exist. If it does not exist an empty value will be assigned to indicate that there is no context. Only if the sentences in the corpus are long it may constitute a maximum of eight words to the left and the right of the ambiguous word.

Generally, we used 8 words to the left and the right of the ambiguous word as contexts surrounding it and totally we have 18 attributes. The 8 sorrounding words are selected without any porpuse because it does not have any matter rather it is the matter of the instances the whole dataset holds. The following table is the format of our dataset for one sentence taken from the corpus which contains the ambiguous word "*sanyii*" with its context "seed". The first row is attribute in which LContext refers to left context, RContext refers to right context, ambigues_word holds words with two contexts and Meanig holds context of the ambiguous words which was used to show cluster assignment. This sample for instance contains seven left contexts and four right contexts surrounding the ambiguous word. In the Left context 8 is empty while Right context 5, 6, 7 and 8 are empty. The empty context indicates, all the sentences have no equal lenth and some sentences may start from the ambiguous words as well some of them may end with ambiguous word. So, because of this some context may be empty to the left and right of ambiguous words.

LContext_8	LContext_7	LContext_6	LContext_5	LContext_4	LContext_3	LContext_2	LContext_1	**Ambigues_word**	Rcontext_1	Rcontext_2	Rcontext_3	Rcontext_4	Rcontext_5	Rcontext_6	Rcontext_7	Rcontext_8	**Meaning(class)**
	kurmana	Bara	hojii	Misoma	Bosona	Deggera	Hoji	**Sanyii**	Biqiltuu	Qopheesa	Raawata	Hektara					<seed>

Table 4.2: Summary of Dataset format

4.7. The clustering algorithms

About five approaches have been applied in the field of WSD. These are AI-based, context window of the target word, knowledge based, corpus based and hybrid approaches. Three of them are the main approaches: Knowledge based approach which uses wordNet and Machine Readable Dictionaries (MRD) which depends on information provided by MRD. The other is corpus based approach which can be divided into three types, supervised, semisupervised and unsupervised learning approaches. Supervised learning approaches use information gathered from training on a corpus that has sense tagged for semantic disambiguation while unsupervised leaning approaches determine the class membership of each object to be classified in a sample without using sense tagged training examples. The last is hybrid approach which combines

aspects of supervised and unsupervised approaches. For our study we used unsupervised corpus based learning approach which of its selected clustering algorithm is discussed below.

As previously discussed in chapter two 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. Unlike Supervised classification which identifies features that trigger a sense tag, unsupervised clustering finds similarity between contexts. There are many clustering algorithm which can be grouped as hierarchical, partitional and probabilistic in Weka tool. For our work we selected five algorithms for experiments which found in weka tool.

Among the clustering algorithms, one is hierarchical clustering including agglomerative cluster such as **single linkage** which can measure clusters by distance between the cluster or the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered, **average linkage** in which the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group and **complete linkage** in which distance between groups is now defined as the distance between the most distant pair of objects one from each cluster or in other words the distance between two clusters is given by the value of the longest link between the clusters are selected.

The other is the probabilistic clustering like **EM** (expectation maximization) which is iterative refinement algorithm that can be used for finding the parameter estimates. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. This algorithm is also selected for our experiments.

The third algorithm we used is partitioning cluster like **K-means** which can be described given a set of initial clusters K (k-stands for numbers of clusters). It assigns each point to one of the cluster and then each cluster center is replaced by the mean point on the respective cluster.

4.8. Performance evaluation techniques

There are different performance evaluation techniques these are general measures not particular to weka for clustering. To measure the rate of disambiguation of our system we used the most common evaluation techniques. These are confussion metrics like precision (P), recall (R), F-measure and accuracy.

To evaluate the clustering results, precision, recall, and F-measure were calculated over pairs of points. For each pair of points that share at least one cluster in the overlapping clustering results, these measures try to estimate whether the prediction of this pair as being in the same cluster was correct with respect to the underlying true categories in the data. Where TP, TN, FP and FN refer to true positives, true negatives, false positives and false negatives respectively and Pt and Nt refer to the total number of positive and negative examples in the test set respectively and P and R refer to precision and Recall respectively:-

Precision is calculated as the fraction of pairs correctly put in the same cluster,

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

Figure 4.6: Precision formula

Precision also referred to as Positive predictive value (PPV), is the proportion of the examples which truly have cluster X among all those which were classified as cluster X. This means truly clustered as X divided by total clustered as cluster X. In the matrix, this is the diagonal element divided by the sum over the relevant column.

Recall is the fraction of actual pairs that were identified and,

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

Figure 4.7: Recall formula

Recall is the TP rate (also referred to as sensitivity) what fraction of those that is actually positive. Recall is rate of the proportion of examples which were clustered as cluster X, among

all examples which truly have cluster X, i.e. how much part of the class was captured. This means dividing clustered as X by actual total in cluster X. In the confusion matrix, this is the diagonal element divided by the sum over the relevant row.

F-measure is the harmonic mean of precision and recall. Due to the performance trade-off between precision and recall, the F-measure, which is a harmonic mean between these two values, yields a single number by which performance can be measured. This provides a convenient way to compare the performance of two or more cluster on the same problem, ranking them in order of quality of prediction.

$$\mathrm{F-Measure} = \frac{2 * \mathrm{P} * \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Figure 4.8: F-Measure formula

The other evaluation metric used in this study is **accuracy** which is basically used to identify the performance of algorithms. It is the rate of correctly clustered documents to all of documents in the given test bed. Accuracy is the easiest and most common way of reporting the performance of machine learning methods and calculated as follows.

$$\mathrm{Accuracy\ (Acc)} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{Pt} + \mathrm{Nt}}$$

Equation 4.9: Accuracy formula

Accuracy is also a ratio of ((no. of correctly classified instances) / (total no. of instances)) *100). The accuracy value enables comparison of a classifier's performance against a given base line such as the majority classifier which acts as the lower bound for the performance of probabilistic classifiers.

The evaluation of every clustering algorithm essentially depends on the characteristics of the dataset and on the input parameters as incorrect input parameters may lead to clusters that deviate from those in the dataset.

Typical objective functions in clustering is to formalize the goal of attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents

from different clusters are dissimilar) to determine the correct number of groups from a dataset which can be measured in terms of accuracy. These can be viewed as internal and external criteria of evaluating clustering algorithm.

External criteria is for which a priori knowledge of dataset information is required, but it is hard to say if they can be used in real problems (usually, real problems do not have prior information of the dataset in question). It evaluates how well the clustering matches the gold standard classes. This implies that we evaluate the results of a clustering algorithm based on a prespecified structure, which is imposed on a dataset, i.e. external information that is not contained in the dataset [82].

Unlike external criteria internal criteria is do not require a priori information from dataset. We may evaluate the results of a clustering algorithm using information that involves the vectors of the datasets themselves. It is direct evaluation in the application of interest but it is expensive, especially if large user studies are necessary and it is especially for the quality of a clustering. Internal criteria can roughly be subdivided into two groups: the one that assesses the fit between the data and the expected structure and others that focus on the stability of the solution [82].

In general, clustering evaluation is usually defined by combining compactness which measures closeness of cluster elements by using variance and separability which indicates how distinct two clusters are and it computes the distance between two different clusters.

For our study an external evaluation is used which is based on a comparison of an algorithm's result to a gold standard means one with each other. The way Weka evaluates the clustering's depends on the cluster mode you may select, because based on the version of weka tool the evaluation mode for accuracy may differ from each other even if their result is the same. For this study, classification via clustering mode was selected in implementation of Weka 3.6.11 package in order to satisfy our evaluation method.

5. CHAPTER FIVE: Experimentation and Results

5.1. Introduction

Experimentation is the process of conducting tests to measure the performance of a given systems. In this work several unsupervised algorithms were tested for their performance in Afaan Oromo word sense disambiguation.

For our experiment seven ambiguous words namely *sanyii, qophii, horii, ifa, karaa, ulfina, and sirna* of each having two senses and whose distribution in the corpus is not skewed to a particular sense, i.e. both senses appear with comparable (equal) frequencies. The contextual features used in this experiment were the ones which indicate a word occurs within some number of words to the left or right of the ambiguous word.

Machine learning methods need data for training and testing, in order to evaluate the performance of the system during experiment. There are several ways of doing evaluation and the most common is to split data into two sets, training set and test set which is important in supervised classification methods. But, in case of unsupervised method there is no need to split the data into training and test sets for evaluation because of the algorithm do not need train-test split of data [84]. Since, train-test split is used to avoid overfitting in machine learning, in unsupervised clustering we cannot need to evaluate. And also we cannot overfit in this way because in clustering algorithms there are no learning labels. This means something that is already labeled, but we want to discover some new structure in our data.

For our experiment we used five words of the dataset used by Tesfa [53] and two of them from the researchers in order to allow comparison.

5.2. Experimental

For this study four sets of experiments were conducted using EM, single linkage, complete linkage, average linkage and simple K means clustering and also NaïveBayes algorithms of weka 6.4.11. The first experiment was conducted to evaluate the effect of stemming on the accuracy of the result on selected Afaan Oromo ambiguous words. The second is to determine the optimal context window on disambiguation accuracy for Afaan Oromo ambiguous words where the contextual information was obtained from 1-left and 1-right to 8-left and 8-right consequent surrounding words. The third is to evaluate the performance of supervised classification methods with NaïveBayes algorithm for seven ambiguous words. And the last is the comparison of

clustering algorithms that are tested in this study against the results obtained from supervised classification algorithms.

5.3. Evaluation measures

To measure the performance of disambiguation of our system, we used the confusion Matrix such as precision, recall, F1-measure and accuracy as evaluation confission Matrix.

All this measurement can be calculated from confusion Matrix of which the columns represent the predictions and the rows represent the actual class. An edge is denoted as:-True positive (TP), if it is a positive or negative link and predicted also as a positive or negative link, respectively. False positives (FP) are all predicted positive or negative links which are not correctly predicted, i.e., they are non-existent. True negatives (TN) denote correctly predicted non-existent edges. False negatives (FN) is falsely predicted non-existent edges are defined i.e., an edge is predicted to be non-existent but it is a positive or a negative link in the reference network. Totally, confussion Matrix have precision, recall, F-measure and accuracy.

5.4. Results of the Experiments

This section discusses experiment carried on the training data sets for clustering algorithms and on both training and test dataset of which test dataset is used to evaluate the performance of the system for supervised classification algorithms. It also describes the result of each experiment for different adjusted data by different algorithms.

5.4.1. Experiment Set I: WSD with stemming dataset

As discussed earlier in chapter four, stemming has been found to give a significant improvement on performance of WSD for morphologically complex languages [82]. In our work, we have tried to test with two types of stemming, stemming which is used in information retrieval and lemmatization which is used in natural language. When we say lemmatization, it is simply trying to remove only the affixes into their root word as stated in the (figure 4.5) while stemming is a procedure that reduces all words with their stem to a common form by stripping off its derivational and inflectional affixes. These affixes are Prefixes which precede the stem, suffixes which follow the stem, circumfixes do both, and infixes are inserted inside the stem [85]. Debela's stemmer algorithm indicated in (figure 4.4)

This experiment is performed using "classification via clustering" evaluation mode to test whether this applies to unsupervised WSD for Afaan Oromo words with the default algorithm. During the experiment it assigned classes to the clusters, based on the majority value of the class

attribute within each cluster. Then it computed the clustering error, based on this assignment and also showed the corresponding confusion matrix. From these the accuracy was used to measure how well it has been able to generalize from the training data to evaluate the model.

5.4.1.1. Result of WSD with lemmatization

Ambiguous words	Effect of lemmatization on Accuracy with k means							
	Before				After			
	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure	Accuracy
Sanyii	0.60	0.62	0.59	**60.6**	0.62	0.64	0.61	**62.2**
Karaa	0.75	0.82	0.74	**75.24**	0.69	0.70	0.68	**68.9**
Ulfina	0.59	0.67	0.51	**58.9**	0.59	0.59	0.59	**59.4**
Ifa	0.65	0.70	0.65	**65.4**	0.68	0.68	0.68	**68.2**
Qophii	0.63	0.63	0.63	**63.2**	0.65	0.65	0.65	**64.7**
Sirna	0.62	0.62	0.61	**61.6**	0.62	0.62	0.62	**62.1**
Horii	0.56	0.56	0.55	**55.70**	0.58	0.61	0.54	**57.9**
Average				**62.9**				**63.3**

Table 5.1: Effect of lemmatization with KMeans algorithms

Ambiguous words	Effect of lemmatization on Accuracy with EM							
	Before				After			
	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure	Accuracy
Sanyii	0.59	0.63	0.56	**59.4**	0.62	0.62	0.61	**61.7**
Karaa	0.752	0.81	0.73	**75.2**	0.74	0.82	0.72	**74.3**
Ulfina	0.53	0.51	0.39	**53.3**	0.57	0.58	0.53	**57.2**
Ifa	0.60	0.59	0.57	**59.8**	0.34	0.57	0.42	**59.8**
Qophii	0.66	0.77	0.62	**65.7**	0.64	0.73	0.60	**64.2**
Sirna	0.60	0.63	0.60	**60.1**	0.62	0.64	0.60	**62.1**
Horii	0.557	0.562	0.549	**55.7**	0.53	0.57	0.44	**52.6**
Average				**61.3**				**61.7**

Table 5.2: Effect of lemmatization with EM algorithms

Ambiguous words	Effect of lemmatization on Accuracy with single linkage							
	Before				After			
	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure	Accuracy
Sanyii	0.50	0.76	0.35	**50.0**	0.514	0.264	0.349	**51.1**
Karaa	1.0	0.507	0.673	**50.5**	0.507	0.257	0.341	**50.5**
Ulfina	0.53	0.28	0.37	**52.8**	0.531	0.282	0.368	**52.8**
Ifa	0.56	0.32	0.40	**55.9**	0.562	0.316	0.404	**55.9**
Qophii	0.505	0.255	0.339	**50.2**	0.502	0.751	0.342	**50.2**
Sirna	0.505	0.751	0.344	**50.5**	0.505	0.751	0.344	**50.5**
Horii	0.498	0.248	0.331	**49.6**	0.504	0.751	0.343	**50.4**
Average				**51.3**				**51.6**

Table 5.3: Effect of lemmatization with Single linkage algorithms

Ambiguous words	Effect of lemmatization on Accuracy with complete linkage							
	Before				After			
	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure	Accuracy
Sanyii	0.511	0.261	0.346	**50.0**	0.544	0.544	0.544	**54.4**
Karaa	0.67	0.705	0.656	**67.0**	0.529	0.529	0.528	**52.9**
Ulfina	0.567	0.565	0.565	**56.7**	0.528	0.535	0.527	**52.8**
Ifa	0.609	0.619	0.61	**60.9**	0.654	0.651	0.649	**65.4**
Qophii	0.637	0.648	0.63	**63.7**	0.507	0.751	0.347	**50.7**
Sirna	0.53	0.53	0.53	**53.0**	0.702	0.792	0.677	**70.2**
Horii	0.526	0.534	0.498	**52.6**	0.548	0.55	0.545	**54.8**
Average				**57.7**				**57.3**

Table 5.4: Effect of lemmatization with complete linkage algorithms

Ambiguous words	Effect of lemmatization on Accuracy with average linkage							
	Before				After			
	Recall	**Precision**	**F-measure**	**Accuracy**	**Recall**	**Precision**	**F-measure**	**Accuracy**
Sanyii	0.533	0.675	0.402	**53.3**	0.511	0.261	0.346	**50.6**
Karaa	0.51	0.26	0.344	**50.5**	0.519	0.754	0.37	**51.9**
Ulfina	0.531	0.282	0.368	**52.8**	0.531	0.282	0.368	**52.8**
Ifa	0.556	0.309	0.398	**55.3**	0.559	0.313	0.401	**55.3**
Qophii	0.507	0.751	0.347	**50.7**	0.507	0.751	0.347	**50.7**
Sirna	0.505	0.751	0.344	**50.5**	0.503	0.253	0.336	**50.5**
Horii	0.504	0.751	0.343	**50.4**	0.509	0.752	0.353	**50.9**
Average				**51.9**				**51.8**

Table 5.5: Effect of lemmatization with Average linkage algorithms

Stemming has been found to give a significant improvement on performance of WSD for morphologically complex languages as shown in the above experiment. As indicated in the above tables stemming improved the accuracy of ambiguous words in simple K-means, Expectation Maximization and single linkage clustering algorithms in (table 5.1, 5.2 and 5.3) respectively. But, for the two algorithms like complete linkage and average linkage clustering algorithms stemming does not improve the accuracy as presented in (table 5.4 and 5.5) respectively.

In the experiment using simple Kmeans algorithm, except the word ‘karaa’ all the other words accuracies were improved with stemming. In (table 5.2) for EM algorithm accuracy of four of them was improved while these of the three on left did not improve. Likewise in (table 5.3) except the word ‘Sanyii and Horii’ which improves the accuracy, the others remained the same with unstemmed one. On average the stemmed dataset performed better than unstemmed dataset. As stated earlier, WSD models determine the meaning of a word by learning the pattern of surrounding words. If stemming is done, the variant of a word is taken as the same pattern, which will improve the accuracy of the algorithms. For example, before stemming, surrounding words ‘sanyiicha’ and ‘sanyiilee’ would be assumed as different but, basically they are the variants of the same word ‘*sanyii’.* After stemming, these words are taken as the same pattern. Therefore, in subsequent experiments the stemmed dataset was used as it enhanced the performance of the models.

In general, the experiments are performed using default parameters both with unstemmed and stemmed datasets of which stemmed dataset perform optimal accuracy in average. The three algorithms simple Kmeans, EM and single linkage improve accuracy while complete linkage and average linkage does not improve.

5.4.1.2. Result of WSD with stemming

Ambiguous words	Effect of stemming on Accuracy with different algorithms									
	Before					After				
	KM	EM	SL	CL	AL	KM	EM	SL	CL	AL
Sanyii	60.6	59.4	50.0	50.0	53.3	59.6	61.1	51.6	54.4	51.6
Karaa	75.24	75.2	50.5	67.0	50.5	75.2	76.7	50	69.4	51.5
Ulfina	58.9	53.3	52.8	56.7	52.8	57.1	55.7	52.8	53.9	53.9
Ifa	65.4	59.8	55.9	60.9	55.3	65.9	57.4	56.4	50.8	55.9
Qophii	63.2	65.7	50.2	63.7	50.7	60.9	59.9	50.5	62.4	50.5
Sirna	61.6	60.1	50.5	53.0	50.5	63.1	62.1	50.5	50.5	50.5
Horii	55.70	55.7	49.6	52.6	50.4	59.4	57.9	50.5	52.9	50.9
Average	**62.9**	**61.3**	**51.4**	**57.7**	**51.9**	**63.0**	**61.5**	**51.8**	**56.3**	**52.1**

Table 5.6: Effect of stemming on ambiguous words

As presented in table 5.6, we also tried to evaluate with stemming which reduces all words with the same stem to a common form by stripping of its derivational and inflectional suffixes. It differs from lemmatization. Means unlike lemmatization which find the meaningful root of the word; this can make also the words meaningless which are suitable for searching. As we can see from the table above the accuracy of all words were improved with stemmed dataset than unstemmed dataset.

5.4.1.3. Summary of lemmatization and stemming

Ambiguous words	Effect of lemmatization and stemming on Accuracy with different algorithms									
	Lemmatization					Stemming				
	KM	EM	SL	CL	AL	KM	EM	SL	CL	AL
Sanyii	62.2	61.7	51.1	54.4	50.6	59.6	61.1	51.6	54.4	51.6
Karaa	68.9	74.3	50.5	52.9	51.9	**75.2**	76.7	50	69.4	51.5
Ulfina	59.4	57.2	52.8	52.8	52.8	57.1	55.7	52.8	53.9	53.9
Ifa	68.2	59.8	55.9	65.4	55.3	65.9	57.4	56.4	50.8	55.9
Qophii	64.7	64.2	50.2	50.7	50.7	60.9	59.9	50.5	62.4	50.5
Sirna	62.1	62.1	50.5	70.2	50.5	63.1	62.1	50.5	50.5	50.5
Horii	57.9	52.6	50.4	54.8	50.9	59.4	57.9	50.5	52.9	50.9
Average	**63.3**	**61.7**	**51.6**	**57.3**	**51.8**	**63.0**	**61.5**	**51.8**	**56.3**	**62.9**

Table 5.7: Summary of effect of lemmatization and stemming on accuracy of ambiguous words

The above (table 5.7) is the comparison of lemmatization and stemming. Even if lemmatization is better in accuracy with Kmeans, EM and complete linkage algorithm than stemmeing, in average stemming is better than lemmatization. In general, the average accuracy of the stemmed dataset is better than that of the unstemmed dataset. As a result, for our further experiments stemmed dataset with stemming used in IR is used.

5.4.2. Experiment Set II: Determining optimal context window

Windows size refers to the number of words needed to be considered, to the left and to the right of the ambiguous word, for the purpose of disambiguation.

For Amharic supervised WSD, three window size on both sides for the ambiguous word is found to be enough [54]. And also for Amharic unsupervised WSD Window size of 3-3 was considered to be effective for Simple K means and EM clustering algorithms [9]. On the other hand for agglomerative SL and CL clustering algorithms the average accuracy of two word window size was found to be effective [9].

For Afaan Oromo supervised WSD window size of four-four was considered to be effective [53], but no one has suggested so far the optimal window size for unsupervised WSD. So, to

determine optimal window size for Afaan Oromo unsupervised WSD, experiments were carried out eight times from one-one window to eight-eight window on both side of the ambiguous word.

5.4.2.1. Results of Window size experiments with 7 words

Window size	Ambiguous word and algorithm result for K means							
	Sanyii	**Karaa**	**Ulfina**	**Ifa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	67.8	**88.8**	77.7	80.4	73.6	81.8	67.8	76.8
2 – 2	68.9	88.3	74.4	77.1	73.2	80.8	79.2	77.4
3 – 3	72.2	88.3	**82.2**	77.1	78.2	84.8	72.8	79.3
4 – 4	**74.4**	88.3	**82.2**	76.5	**78.2**	**87.4**	**86.6**	**81.9**
5 – 5	71.7	87.9	81.7	79.9	72.8	67.7	80.7	77.5
6 – 6	71.7	87.9	81.7	**84.9**	72.8	80.8	78.2	79.7
7 – 7	71.7	88.3	76.1	84.3	78.2	83.3	78.2	80.0
8 – 8	68.9	79.6	73.9	70.9	61.9	66.7	69.3	70.2

Table 5.8: Window size experimentation for KMeans algorithm with 7 words

Window size	Ambiguous word and algorithm result for EM							
	Sanyii	**Karaa**	**Ulfina**	**Ifa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	62.8	88.8	71.7	**82.1**	73.8	68.7	67.8	73.7
2 – 2	65.0	88.8	71.1	79.3	72.7	68.7	**80.2**	75.1
3 – 3	**70.0**	87.4	**83.9**	74.3	**75.7**	**88.4**	79.2	**79.8**
4 – 4	67.2	**89.9**	81.1	73.2	**75.7**	71.7	79.2	76.9
5 – 5	62.2	87.9	72.8	76.5	72.8	77.3	66.8	73.8
6 – 6	65.6	87.9	69.4	73.2	72.8	67.7	67.8	72.1
7 – 7	63.3	76.2	77.2	72.6	71.8	67.7	68.2	71.0
8 – 8	60.6	85.9	60.0	63.7	71.8	70.5	70.2	68.9

Table 5.9: Window size experimentation for EM algorithm

Window size	Ambiguous word and algorithm result for single linkage							
	Sanyii	**Karaa**	**Ulfina**	**Ifa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	**61.7**	**77.7**	57.2	71.5	64.4	69.7	62.9	66.4
2 – 2	58.3	76.7	58.9	75.9	**73.8**	68.7	65.3	68.2
3 – 3	56.7	76.7	59.4	73.7	68.3	69.2	80.7	69.2
4 – 4	**61.7**	76.7	**61.1**	73.2	62.4	**72.2**	79.2	**69.5**
5 – 5	**61.7**	65.5	60.6	73.2	62.4	69.2	61.9	64.9
6 – 6	56.1	67.9	57.2	72.1	57.9	68.7	61.4	63.0
7 – 7	56.1	64.3	57.2	74.9	68.3	69.2	63.4	64.8
8 – 8	57.7	65.0	57.2	**76.5**	61.9	65.2	61.4	63.6

Table 5.10: Window size experimentation for Single Link clustering algorithm

Window size	Ambiguous word and algorithm result for complete linkage							
	Sanyii	**Karaa**	**Ulfina**	**Ifa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	64.4	**88.9**	65.6	72.1	72.3	69.2	63.4	70.8
2 – 2	66.7	88.8	63.3	74.3	72.3	66.2	62.4	70.6
3 – 3	64.4	88.8	84.4	74.9	70.8	**86.2**	62.9	**76.1**
4 – 4	66.7	81.6	72.2	72.1	69.8	69.2	63.4	70.7
5 – 5	66.1	79.1	62.8	74.9	68.8	68.7	63.4	69.1
6 – 6	64.4	79.6	**82.2**	83.3	71.8	68.7	62.4	73.2
7 – 7	**69.4**	69.4	78.9	85.5	68.8	66.4	60.5	71.3
8 – 8	66.7	73.3	74.4	**86.0**	**72.8**	65.3	61.7	71.5

Table 5.11: Window size experimentation for Complete Link clustering algorithm

Window size	Ambiguous word and algorithm result for Average linkage							
	Sanyii	**Karaa**	**Ulfina**	**Ifa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	58.3	**88.8**	**64.4**	71.5	72.3	69.2	63.4	69.7
2 – 2	63.3	**88.8**	58.3	75.9	72.3	**71.2**	62.9	**70.4**
3 – 3	**64.4**	81.6	63.3	68.7	67.3	70.1	62.9	68.3
4 – 4	60.0	81.6	58.3	73.2	67.3	**71.2**	**72.8**	69.2
5 – 5	61.1	81.6	60.6	68.7	59.9	69.2	60.4	65.9
6 – 6	58.3	81.6	57.8	68.7	**72.8**	69.2	66.8	67.9
7 – 7	60	72.8	61.1	75.4	57.9	56.3	56.9	62.9
8 – 8	63.3	72.8	56.1	**76.5**	64.4	57.2	54.6	63.6

Table 5.12: Window size experimentation for Average Linkage clustering algorithm

The above experiments was done by using five unsupervised algorithms K means, Expectation Maximization (EM), Single Linkage, Complete Linkage and Average Linkage respectively from one-one up to eight-eight to the left and to the right of the ambiguous word.

In terms of the algorithms selected with (K Means, EM, Complete Linkage and Average Linkage) (table 5.8, 5.9, 5.11 and 5.12 respectively) the word '*karaa'* achieved highest accuracy while with (Single Linkage) (table 5.10) the word '*Ifa*' got highest accuracy. The second highest accuracy is achieved by the word 'sirna' with (simple K means) (table 5.8), 'Ifa' with (EM, Complete link and Average link) (table 5.9, 5.11 and 5.12 respectively) and 'karaa' with (Single link) (table 5.10). The list accuracy is also scored for the word 'sanyii' with (simple K means and EM) (table 5.8 and 5.9 respectively), 'ulfina' with (Single link and Average accuracy) (table 5.10 and 5.12 respectively) and 'horii' with (Complite link) (table 5.11).

Accordingly, (table 5.8 and 5.10) shows that for K Means and Single linkage algorithms window size four-four produces the best result. On the other hand (table 5.9 and 5.11) shows that for EM and complete linkage algorithms window size of three-three gives the best results. While (table 5.12) indicates that for Average link algorithms window size two-two is enough.

When we see accuracy in terms of each window size, for the word window size of one-one and two-two ambiguous word '*karaa*' in all selected algorithm achieved highest accuracy. In three-

three window size '*karaa*' in (Kmeans, Complete linkage and Average linkage), '*Sirna*' in Expectation maximization and '*Horii*' in Single linkage were achieved highest score. In window size four-four the word '*karaa*' achieved highest accuracy in all algorithms. In window size five-five except the word '*ifa*' which achieved highest accuracy in Single linkage algorithms, the word '*karaa*' has highest accuracy in other algorithms. In window size six-six the word '*Ifa*' in Single and Complete linkage while '*karaa*' in the left others have highest accuracy. In window size seven-seven '*ifa*' in single linkage, '*karaa*' in Kmeans and Average link and '*Ulfina*' in EM and Complete linkage achieved highest accuracy. While in window size eight-eight the word '*karaa*' in (Kmeans and EM) and '*Ifa*' in the three left others achieve highest accuracy.

Totally in our study best accuracy is achieved with Kmeans, EM and complete link algorithms. For these algorithms window size four-four is enough in Kmeans and three-three is enough for EM and complete link algorithm.

Ambiguous Word	**Window size**		
	Three - Three (3 – 3)		**Four - Four (4-4)**
	Expectation Maximization	**Complete Link**	**K Means**
Sanyii	70	64.4	74.4
Karaa	87.4	88.8	88.3
Ulfina	83.9	84.4	82.2
Ifa	74.3	74.9	76.5
Qophii	75.7	70.8	78.2
Sirna	88.4	86.2	87.4
Horii	79.2	62.9	86.6

Table 5.13: Summary of Window size experimentation for clustering algorithm

As indicated in the above (table 5.13) the algorithms achieved average accuracy of 81.9% in simple k means, 79.8% in EM and 76.1% in Complete Link Algorithms.

5.4.2.2. Results of Window size experiments for 5 words with 1000 datasets

Window size	Ambiguous word and algorithm result for K means					
	Sanyii	**Karaa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	66.5	68.1	71.3	81.8	67.8	71.1
2 – 2	**78.1**	78.3	73.2	70.8	65.3	73.1
3 – 3	72.2	**88.3**	78.2	64.8	72.8	75.2
4 – 4	67.9	78.5	**80.2**	**83.4**	**86.6**	**79.3**
5 – 5	76.3	86.4	72.4	65.8	80.7	76.3
6 – 6	71.7	81.7	71.1	82.2	58.2	72.9
7 – 7	72.3	62.3	71.2	53.3	78.2	67.4
8 – 8	78.0	74.3	61.7	62.7	69.3	69.2

Table 5.14: Window size experimentation for KMeans algorithm with 1000 dataset

The above experiment was done by using simple KMeans unsupervised algorithm which performed better in previous experiment. KMeans algorithm as shown above (table 5.14) the word '*karaa'* achieved highest accuracy. The second highest accuracy is achieved by the word '*sanyii*' and the list accuracy is also scored for the word '*sirna*'. In this experiment KMeans algorithms shows, window size four-four produces the best result and indicates that window size four-four is enough to identify the meaning of ambiguous word.

5.4.2.3. Results of Window size experiments for 5 words with 1240 dataset

Window size	Ambiguous word and algorithm result for KMeans					
	Sanyii	**Karaa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	75.5	74.1	82.9	87.2	64.6	76.8
2 – 2	80.3	72.7	84.4	**89.4**	63.1	77.9
3 – 3	79.4	76.4	84.0	88.6	63.4	78.3
4 – 4	**84.3**	**79.9**	81.7	85.6	**70.9**	**80.5**
5 – 5	76.1	79.8	83.5	74.8	55.7	73.9
6 – 6	76.2	79.7	**86.1**	74.9	66.3	76.6
7 – 7	74.9	79.6	85.7	79.1	62.1	76.2
8 – 8	76.8	74.0	82.6	68.6	68.5	74.1

Table 5.15: Window size experimentation for KMeans algorithm 1240 datasets

The above experiments were done by using simple KMeans unsupervised clustering algorithm which was performed better in previous experiment. KMeans algorithm, as shown above (table 5.15) the word '*qophii'* achieved highest accuracy. The second highest accuracy is achieved by the word '*sirna*' and the list accuracy is also scored for the word '*horii*'. In this experiment KMeans algorithms shows, window size four-four produces the best result and indicates that window size four-four is enough to identify the meaning of ambiguous word.

5.4.3. Experiment Set III: Experiments with supervised approaches

Here for supervised classification technique, there are several ways of doing evaluation and the most common way is to split data into two sets, training set and test set. In this regard, different training and test data sets were prepared for each ambiguous word, where the contextual information was obtained from 1-left and 1-right to 8-left and 8-right consequent surrounding words as we used for clustering algorithms. The training set was used to train the system and the testing set was used to measure the performance of the developed system. During the preparation of the dataset, in order to evaluate the performance of the developed system, we remove manually tagged sense examples. Sense the dataset we are using is not sufficiently large, to evaluate the performance of the algorithm, 10-fold cross-validation evaluation technique is used for classification. We used the whole data without splitting into training and test datasets.

In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds," D1, D2, …, Dk, each of approximately equal size. Training and testing is performed k times. In iteration i, partition Di is reserved as the test set, and the remaining partitions are collectively used to train the system. That is, in the first iteration, subsets D2, … , Dk collectively serve as the training set which is tested on D1; the second iteration is trained on subsets D1, D3, … , Dk and tested on D2; and so on. Hence, different combinations of training and testing data are used, i.e. the data set is divided in different ways in such a way that different combination of sense examples are available for training and testing each time. For these experiments, 10-fold cross validation is used which divides the data set into ten sets, each set containing 10% of the total data.

5.4.3.1. Results of Window size experiments for NaïveBayes algorithm for 7 words

Window size	Effect of NaïveBayes classification algorithm for seven words							
	Sanyii	**Karaa**	**Ulfina**	**Ifa**	**Qophii**	**Sirna**	**Horii**	**Average**
1 – 1	87.4	88.3	78.7	85.8	84.1	78.4	81.2	83.4
2 – 2	85.4	89.8	81.4	87.2	86.1	84.2	84.1	85.5
3 – 3	88.4	92.2	80.1	85.8	**83.6**	**84.2**	88.2	86.1
4 – 4	**84.9**	91.2	**80.1**	85.8	**82.6**	83.7	**84.6**	84.7
5 – 5	84.9	91.2	80.1	**84.6**	83.6	83.2	85.6	84.7
6 – 6	85.4	91.2	79.6	**82.4**	83.1	84.6	85.6	84.6
7 – 7	85.4	92.3	80.1	81.6	82.1	84.2	86.1	84.5
8 – 8	85.4	**92.7**	79.1	81.2	82.1	83.7	86.6	84.4

Table 5.16: Window size experimentation with NaïveBayes classification algorithm

As shown in (Table 5.16) above for the ambiguous word *sanyii,* the maximum accuracy was achieved on four-four word window size. Whereas, for ambiguous word *karaa* the highest accuracy was attained on eight-eight word window size and for ambiguous word *ulfina* the highest accuracy was attained on four-four word window size. Among all, the maximum accuracy was achieved for ambiguous word *ifa* with five-five and six-six windows size. For the ambiguous word *qophii* maximum accuracy was achived at three-three and four-four window size. For the ambiguous word *sirna* maximum accuracy was achived at three-three window size. And finally for the ambiguous word *horii* maximum accuracy was achived at four-four window size.

Most of the maximum accuracy result is achieved with windows size between two-two windows size up to six-six windows size. Hence, the result agreed with the findings in other language that the nearest words surrounding the ambiguous word give more disambiguation information than words far from the ambiguous word [84]. In this experiment, since the average accuracy result for windows size three-three is larger than all the other windows, window size three-three was considered to be effective for Afaan Oromo Word Sense Disambiguation.

5.4.3.2. Results of Window size experiments for NaïveBayes algorithm for 5 words with 1000 datasets

Window size	Effect of NaïveBayes classification algorithm					
	Sanyii	Karaa	Qophii	Sirna	Horii	Average
1 – 1	87.4	87.4	84.2	77.3	72.5	81.8
2 – 2	88.4	89.9	85.7	75.4	84	84.7
3 – 3	88.4	91.4	74.7	85.9	84	84.9
4 – 4	85.9	80.5	83.7	84.4	87	84.3
5 – 5	84.9	91.4	84.2	73.9	86.5	84.2
6 – 6	84.9	91.9	84.2	78.5	84	84.7
7 – 7	85.3	92.4	84.2	84.4	77	84.7
8 – 8	85.4	82.4	84.2	83.5	85	84.1

Table 5.17: Experimentation with NaïveBayes classification algorithm for 5 words

As shown in (Table 5.17) above, since the average accuracy result for windows size three-three is larger than all the other windows, window size three-three was considered to be effective for Afaan Oromo Word Sense Disambiguation.

5.4.3.3. Results of Window size experiments for NaïveBayes algorithm for 5 words with 1240 datasets

Window size	Effect of NaïveBayes classification algorithm					
	Sanyii	Karaa	Qophii	Sirna	Horii	Average
1 – 1	93.7	90.1	80.9	82.7	74.9	84.5
2 – 2	93.7	81.3	81.7	88.1	79.3	84.8
3 – 3	94.2	94.2	80.5	73.1	86.2	85.6
4 – 4	93.3	83.0	90.9	87.2	71.8	85.2
5 – 5	92.4	92.2	81.9	86.4	72.2	85.0
6 – 6	91.3	92.2	80.5	76.4	84.9	85.1
7 – 7	92.0	83.4	80.1	76.8	88.8	84.2
8 – 8	92.0	93.4	80.1	82.8	78.8	85..4

Table 5.18: Experimentation with NaïveBayes classification algorithm for 5 words with 1240 dataset

As shown in (Table 5.18) above also the average accuracy result for windows size three-three is larger than all the other windows which indicates that window size three-three was considered to be effective for Afaan Oromo Word Sense Disambiguation.

5.4.4. Experiment Set III: Comparing unsupervised and supervised algorithms on Afaan Oromo WSD.

Based on the experiments conducted so far (I and II), stemmed dataset with three-three and four-four window size is selected for comparisons against the performance recorded with supervised algorithm on experiment (III) and the one reported by Tesfa [53]. Two comparisons were done based on the dataset size used by the researchers. The first is comparison by using seven ambiguous words of 1501 dataset size. In this by using an average accuracy achieved, in which supervised classification algorithms performs better than unsupervised clustering algorithms with best accuracy of 84.1% with NaiveBayes algorithms at three-three window size. On the other hand, unsupervised WSD clustering algorithms out of five algorithms selected for experiment three of them scored best accuracy. These are simple K Means with average accuracy of 81.9%, Expectation Maximization with average accuracy of 78.9% and Complete Linkage with average accuracy of 76.1%.

The experiments were performed for both unsupervised clustering algorithm (table 5.14 and 15) and supervised algorithm (table 5.17 and 18) based on the size of the datasets for 1000 and 1240 dataset size respectively. When we see the clustering algorithms with 1000 in (table 5.14) and 1240 in (table 5.15) it performed an accuracy of 79.3%, while the second dataset performed better with accuracy of 80.5%. In the case of supervised algorithms with 1000 in (table 5.17) and 1240 in (table 5.18), again the highest dataset performed better with little difference of 0.7%. In all cases supervised algorithm performed better than unsupervised algorithms. The greater the size of the dataset performs the better accuracy. The second is comparing the results of unsupervised algorithms with supervised algorithms reported by Tesfa [53]. The experiment performed by using the same dataset, with the same size in which unsupervised achieved 80.5% and supervised achieved 79% as reported by Tesfa. But, the same experiments with the same dataset were performed by the researcher using supervised algorithm in which 1000 and 1240 dataset accuracy (84.9% and 85.6%) respectively was achieved better than Tesfas result. The reason may be that Tesfa used the taraining and test dataset for the evaluation which we used the whole data.

6. CHAPTER SIX: Summary, Conclusion and Recommendation

6.1. Summary

The main objective of this paper is word sense disambiguation using corpus-based unsupervised machine learning methods for Afaan Oromo language which addresses the problem of automatically deciding the correct sense of an ambiguous word based on its surrounding context's. Many words have more than one meaning in natural language, and each one of the meaning is determined by its context. The automated process of recognizing word senses in context is known as Word Sense Disambiguation (WSD). It is the process of selecting the appropriate meaning or sense for a given word in a document. Word Sense Disambiguation (WSD) refers to also a task that automatically assigns a sense, selected from a set of pre-defined word senses to an instance of a polysemous word in a particular context [13].

One of the problems with word sense disambiguation is deciding what the senses are, in cases where at least some senses are different. In other cases, however, the different senses can be closely related (one meaning being a metaphorical extension of another), and in such cases division of words into senses becomes much more difficult [14]. This problem is solved through different approaches such as AI-Based approach, Methods Based on the Context Window of the Target Word, Corpus-based, Knowledge-based and Hybrid-based. AI-Based approach uses Different forms of logical Inference and Spreading activation Models. Methods based on the Context Window of the Target Word approach uses words of context to the entire sentence in which the target word appears and substantive words which co-occur with a given sense of a word. Knowledge based approaches use information provided by Machine Readable Dictionaries (MRD), Corpus based approaches use information gathered from training corpus and Hybrid approach combines aspects of the two methodologies, corpus based and knowledge based approaches [15].

Corpus based methods grew in importance after the public availability of large scale digital corpora. Corpus is the data for the lexical sample task is typically a large number of naturally occurring sentences containing a given target word, each of which has been tagged with a pointer to a sense entry from the sense inventory. This method is further divided as supervised, semi-supervised and unsupervised approach based on whether sense examples in a corpus is manually tagged with their sense, partially or not. The researchers used unsupervised methods

to solve the problem for Afaan Oromo WSD. This method avoids the problem of knowledge acquisition bottleneck, that is, lack of large-scale resources manually annotated with word senses as discussed in the literature review parts. Regarding this methods five selected algorithms were used; these are Simple k means, EM, single, average and complete link clustering algorithms. To solve the problem of natural language corpus is the main resources. There is no large size corpus which is already prepared for Afaan Oromo language for WSD purpose. So, we prepared Afaan Oromo corpus manually for this study by using 7 selected ambiguous words namely *sanyii, horii, ulfina, ifa, karaa, qophii and sirna* having two senses, out of which 5 of them was used by Tesfa [53] in supervised methods. Based on these 7 words, we extracted 1501 sentences from Afaan Oromo news paper and by creating those as our training set out of them 1000 were used by Tesfa [53].

The work flow of our study contains three main steps: the preprocessing steps, dataset preparation steps and disambiguating steps. The preprocessing steps contains tokenization of the sentences into words, removing stop words and applying the two stemming process. These stemming processes are lemmatization which is simply trying to remove only the affixes, the prefix and suffixes to find only the root of the words. And stemming that reduces all words with the same stem to a common form by stripping off the affixes (Prefixes, suffixes, circumfixes and infixes) of the words rather than finding only the root of the words [85]. The dataset preparation step is making the data suitable for the weka tools while the last disambiguating step is used to identify the sense of the ambiguous words contextually through clustering algorithms.

Generally, using these clustering algorithms two main experiment was conducted. The first is evaluation of the effect of stemmed data versus unstemmed both for lemmatization and stemming. Though, the stemmed dataset performed better than unstemmed dataset in both stemmers. But, when we compare the two stemmers, stemming is the one with better accuracy which we used for our second experiment further. This shows that stemming is good to improve the accuracy of the algorithms. The reason is that stemming brings variants of a word into their common stem. This minimizes the consideration of the variants of a word as different word by WSD model.

The second experiment is evaluated with the window size dataset of one-one to eight-eight. In window size experiment the best accuracy was achieved an average accuracy of 81.9 for simple

K Means, 79.8 for EM and 76.1 for complete linkage. While average accuracy of 70.4 for Average link and of 69.5 for single link was recorded. So, find that window size three-three is enough for EM and complete link algorithm while four-four is enough for simple K Means algorithm to identify the sense of the words.

At the end we compared the accuracy achieved in the supervised and unsupervised Afaan Oromo WSD with NaïveBayes algorithms for 7 words. In which unsupervised approach achieved the less performance with average accuracy of 81.9% for simple K Means, 79.8% for EM and 76.1% for complete linkage respectively. Supervised Afaan Oromo WSD algorithms achieved an accuracy of 84.1% with NaïveBayes algorithm. And also when compare it with the accuracy of 79% reported by Tesfa [53] for 5 words unsupervised approach shows better performance. But, supervised is better as we achieved better accuracy with the same dataset. Because, Tesfa used training and test dataset for evaluation unlike we used the whole dataset for evaluation.

6.2. Conclusion

Based on the literature review, experiments and the results presented in the previous chapter, the following conclusions may be drawn:

- Afaan Oromo is a grammatically complex language with its own morphology, syntax and semantics. It has also its own grammar called **Seer-luga** which includes all the points like Parts of speech (nouns, pronouns, adjectives, adverbs, verbs, preposition), writing systems, punctuation marks and syntax.
- The experiment of this study was performed by preprocessing the datasets. One of the preprocessing activities is stemming activities. So, firstly evaluation of the effect of stemmed dataset versus unstemmed dataset both for lemmatization and stemming were performed. Though, the stemmed dataset performed better than unstemmed dataset in both stemmers which shows stemming datasets for experiment is important. The reason is that stemming brings variants of a word into their common stem. This minimizes the consideration of the variants of a word as different word by WSD model. But, when we compare the two, stemming is the one with better accuracy than lemmatization.

- Regarding clustering algorithms used in this study, five algorithms such as Simple k means, EM, single, average and complete link were selected. These algorithms were thought as important for WSD out of which simple K Means algorithm is the best of all to identify the meaning of ambiguous word in a context.
- To solve the problem of word sense disambiguation area, one of the techniques is corpus based approach. This approach have supervised and unsupervised methods of which the researcher used unsupervised methods. Unsupervised method is important for WSD as we evaluated unsupervised WSD with seven ambiguous words. But, supervised method achieves better accuracy than unsupervised approaches.

- Windows size which refers to the number of words needed to be considered as consequent surrounding words, to the left and to the right of the ambiguous word, for the purpose of disambiguation. The nearest words surrounding the ambiguous word give more disambiguation information than words far from the ambiguous word [84]. Hence, depending on the experiment, window size four-four was considered to be effective for Afaan Oromo Word Sense Disambiguation for unsupervised methods and three-three was considered to be effective for Afaan Oromo Word Sense Disambiguation for supervised methods.

6.3. Recommendations

A number of additional tasks need to be done to develop fully functional WSD system for Afaan Oromo. Even though we achieved encouraging results the following improvement are recommended for an even better systems; further work could be done along the following points:

- Trying to continue the work with (AI-Based approach, Methods Based on the Context Window of the Target Word, Corpus-based, Knowledge-based and Hybrid-based) approaches to improve the system performance in addition to corpus based approach good. But, while applying these approaches stemming the dataset can performs good accuracy. As we have stated in our work out of two types of stemming we used, in which stemming performs better than lemmatization. We recommend other researchers that these approaches need to be investigated for Afaan Oromo languages by using stemming as well.
- Different researches for WSD in other languages use linguistic resources like Thesaurus, Word-Net, Machine Readable Dictionaries and machine translation software. Regarding this

knowledge sources adopted by WSD systems, in recent years, the results of many research efforts for the construction of online lexical knowledge repositories, ontologies and glossaries became available creating new opportunities for knowledge-based sense disambiguation methods. We recommend other researcher for the development of these resources to enhance WSD for Afaan Oromo languages.

- The main resource in linguistic area to solve the problem is the availability of corpus. For other languages standard sense annotated corpora is available for WSD research and also for testing a WSD system. We do not have such data for Afaan Oromo language which makes the study to be limited for seven ambiguous words. So, large size Afaan Oromo corpus from many different ambiguous words especially for unsupervised WSD research is necessary as it need large corpus. It is better if an interested body or corpus developer create a large standardized corpus from different words for Afaan Oromo language.
- In linguistic area large corpus is recommended to do experiments. Because, large data may not contain some missed points as it is a collection of huge data. So it increases the accuracy of the evaluation. So, by extending the experimentation done using both supervised and unsupervised WSD, for other researchers, including other more ambiguous words to increase the size of corpus in addition to those covered in the research is better as we obtained from our results. This means one of the cases for the improvement of supervised system with seven words of 1501 corpus with that of reported by Tesfa [53] with five words of 1240 corpus is the difference of corpus.
- There are so many clustering algorithms in weka 3.6.11. But, due to time limitation, in this study only five clustering algorithms were experimented that are implemented in Weka 3.6.11 package. Trying the experiment with other algorithms like Clustering by Committee (CBC), Growing Hierarchical Self-Organizing Map (GHSOM) and Graph-based algorithms may have different effects.

Reference

1. Rada Mihalcea, Ted Pedersen, Advances in Word Sense Disambiguation, Tutorial at ACL Conference, 2005.
2. Vickrey D. et al, Word-Sense Disambiguation for Machine Translation, Department of Computer Science, Stanford University, Stanford, The Association for Computational Linguistics, 2005, pp. 771–778.
3. Jason Michelizzi, "A semantic relatedness applied to all Words Sense Disambiguation", A thesis submitted to the faculty of the graduate school of the University of Minnesota, In partial fulfillment of the requirements for the Degree of Master of science, Minnesota, 2005.
4. Teshome Kassie, "Word Sense Disambiguation for Amharic Text Retrieval: A case study For Legal Documents". A Thesis Submitted to the School of Graduate Studies of the Addis Ababa University in Partial Fulfilment for the Degree of Master of Science in Computer science, AA, 2008.
5. Yarowsky, D., Unsupervised word sense disambiguation rivaling supervised methods, In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995. Cambridge, M.A.
6. Dorr, Bonnie, Machine translation divergences: A formal description and proposed solution, Computational Linguistics, Vol. 20, No. 4, pp., 1994, 597-633.
7. Litkowski, K. C., Computational lexicons and dictionaries, In Encyclopedia of Language and Linguistics (2nd ed.), K. R. Brown, Ed. Elsevier Publishers, Oxford, U.K., 753–61, 2005.
8. Doina Tatar, Gabriela Serban, A New Algorithm for Word Sense Disambiguation, Studia Universitatis "Babes-Bolyai", Seria Informatica, Volume-XLVI, 2001.
9. Solomon Assemu. "Unsupervised Machine Learning Approach for Word Sense disambiguation To Amharic Words". A Thesis Submitted to the School of Graduate studies of the Addis Ababa University in Partial Fulfillment for the Degree of Master of Science in Information Science, 2011.
10. C. D. Manning, P. Raghavan, and H. Schutze, An Introduction to Information Retrieval, Online Edition, Cambridge: Cambridge UP, 2009.

11. Simon Ager, Oromo Language, Online edition, 2012. [Online]. Available: www.sas.upenn.edu/African_Studies/Hornet/Afaan_Oromo_19777.html. [Accessed: 25-Feb-2012].
12. Gezehagn G., Afaan Oromo Text Retrieval System. A Thesis submitted in Partial fulfillment of the requirement for the degree of master of information science, AA, 2012.
13. Saleh A., Word Sense Disambiguation and Semantics techniques. Sultan Qaboos University, College Of Science, Computer Science Department Oman, 2009.
14. Palta E., Word Sense Disambiguation, in partial fulfillment of the requirements of the degree of Master of Technology, 2006-2007.
15. Ide N. and Véronis J., Word Sense Disambiguation: The State of the Art, Computational Linguistics, vol. 24, no. 1, 1998.
16. Manning, C. and Schutze, H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.
17. Navigli R., Word sense disambiguation: A survey, ACM Computing Surveys, Vol. 41, No. 2, Article 10, 2009.
18. KILGARRIFF, A, Word senses. In Word Sense Disambiguation: Algorithms and Applications, E. c and P. Edmonds, Eds. Springer, New York, NY, 29–46, 2006.
19. Soanes C. and Stevenson, A., Eds, Oxford Dictionary of English. Oxford University Press, Oxford, U.K, 2003.
20. T. R. Gruber, A translation approach to portable ontology. Knowledge acquisition, vol. 5, no. 2, pp. 199-220, 1993.
21. Miller G. et al, A semantic concordance. In Proceedings of the ARPA Workshop on Human Language Technology, pp. 303–308, 1993.
22. Magnini, B. AND Cavagli`A, G., Integrating subject field codes into Word-Net. In Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC, Athens, Greece), pp. 1413–1418, 2000.

23. Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer. 1991. A statistical approach to sense disambiguation in machine translation. In Fourth DARPA workshop on Speech and Natural Language**,** Pacific Grove, CA, February.
24. Church, K. W. and Rau, L. F., Commercial applications of natural language processing. Commun.ACM vol.38, No.11, PP. 71–79, 1995.
25. Masterman M., "Semantic message detection for machine translation, using an Interlingua." International conference on machine translation of languages and applied language analysis, her majesty's stationery office, London, 1962, PP. 437-475.
26. Schutze, H., "Automatic Word Sense Discrimination," Computational Linguistics, No. 24, pp.97-123, March 1998.
27. Rivest, R. L., Learning decision lists. Machine Learning, vol.2, No.3, PP. 229-246, 1987.
28. Richard H. "Interlingual machine translation." Computer Journal, Vol.1, No.3, PP.144-47, 1958.
29. Hayes, Philip J., on semantic nets, frames and associations, Proceedings of the 5th International, Joint Conference on Artificial Intelligence, Cambridge, Massachusetts, 1977, PP. 99-107.
30. George A. Miller et al, Word-Net: An on-line lexical database. International Journal of Lexicography, Vol. 3, No. 4, PP. 235–312, August. 1993.
31. Roy A., S. Sarkar and B. S. Purkayastha, "Knowledge Based Approaches to Nepali word Sense Disambiguation.", International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014.
32. Banerjee, Sid & Ted Pedersen, An adapted Lesk algorithm for word sense disambiguation using Word-Net, Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLING), Mexico City, Mexico, 2002, pp.136–145.
33. Massimiliano C. and Johnson M., "Explaining away ambiguity: Learning verb selectional preference with Bayesian networks", Proceedings of the International Conference on Computational Linguistics (COLING), Saarbrucken, Germany, 2000, pp.187–193.

34. McCarthy D. and Carroll J., Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences, Computational Linguistics, Vol. 29, No. 4, PP. 639-654, 2003.
35. Loh W., Classification and regression trees, Department of Statistics University of Wisconsin-Madison, Madison, WI, USA, Vol. 1, 2011.
36. Schapire, Robert E., The boosting approach to machine learning: An overview, Nonlinear Estimation and Classification, ed. by D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu. New York, U.S.A.: Springer, 2003.
37. Escudero G. et.al, Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited, PP. 1–10, 2000.
38. Agirre E. and Martinecz D., Knowledge Sources for Word Sense Disambiguation, Springer-Verlag Berlin Heidelberg, 2001.
39. Stevenson, Mark and Yorick Wilks, The interaction of knowledge sources in word sense disambiguation, Computational Linguistics, Vol.27, No.3, PP. 321–349, 2001.
40. Yarowsky D. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, PP.454–460, 1992.
41. Pustejovsky J., Generativity and Explanation in Semantics: A Reply to Fodor and Lepore, Linguistic Inquiry, Vol. 29, No. 2, PP. 289–311, spring 1998.
42. Budanitsky, Alex and Graeme Hirst, Semantic distance in Word-Net: An experimental, application-oriented evaluation of five measures, Proceedings of the NAACL Workshop on Word-Net and Other Lexical Resources, Pittsburgh, U.S.A., pp. 29–34, 2001.
43. Patwardhan et al., Word-Net::Similarity - Measuring the Relatedness of Concepts, American Association for Artificial Intelligence, 2003.
44. Rada et al., Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity, 1989.
45. Fellbaum, WordNet(s). In: Keith Brown, (Editor-in-Chief) Encyclopedia of Language & Linguistics, Second Edition, Vol. 13, pp. 665-670, 2006.
46. Miller et al., Introduction to Word-Net: An On-line Lexical Database, 1990.

47. Leacock, Claudia, Martin Chodorow and George Miller, Using corpus statistics and Word-Net relations for sense identification. Computational Linguistics, Vol.24, No.1, PP. 147–165, 1998.
48. Cowie, Jim, Joe A. Guthrie & Louise Guthrie, Lexical disambiguation using simulated annealing. Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, PP. 359–365, 1992.
49. Mihalcea R. & Moldovan D., Semantic indexing using Word-Net senses. Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong, 2000.
50. Vasilescu, Florentina, Philippe L. and Lapalme G., 2004. Evaluating variants of the Lesk approach for disambiguating words, Proceedings of the Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, PP. 633–636, 2004.
51. Agirre E. et al., Word Sense Disambiguation, Text, Speech and Language Technology, Agirre E. and Edmonds PH., Word Sense Disambiguation Algorithms and Applications, Springer, 2007, pp. 107-2007.
52. Pamulaparty L., Rao G., A Novel Approach to Perform Document Clustering Using Effectiveness and Efficiency of Simhash , International Journal of Engineering and Advanced Technology (IJEAT), Vol. 2, No. 3, February 2013.
53. Tesfa K., "Word Sense Disambiguation For Afaan Oromo Language", A thesis submitted to the school of graduate studies of Addis Ababa University in partial fulfillment for the degree of Masters of science in computer science, 2013.
54. Solomon M., "Word Sense Disambiguation for Amharic Text", A Machine Learning Approach, A thesis submitted to the school of graduate studies of Addis Ababa University in partial fulfillment of the requirements for the degree of Master of Science in information science, 2010.
55. Mishra N., An Unsupervised Approach to Hindi Word Sense Disambiguation, Indian Institute of Information Technology, Allahabad. UP, India, 2009.
56. Sharma R., Word sense disambiguation for Hindi language, thesis submitted in partial fulfillment of the requirements for the award of degree of Master of Engineering in computer science and Engineering, Patiala, 2008.
57. Elmougy S., Naïve Bayes Classifier for Arabic Word Sense Disambiguation, 2008.

58. Zouaghi A.et al, A Hybrid Approach for Arabic Word Sense Disambiguation, International Journal of Computer Processing of Oriental Languages, China, 2011.
59. Lesk, M., Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone in Proceedings of the SIGDOC Conference. 1986. Toronto, Ontario.
60. Etnologue, Show Language, *Online edition*, 2009. [Online]. Available: http://www.ethnologue.com/web.asp. [Accessed: 01-Mar-2012].
61. Language Materials Project, Afaan Oromo, 2010. [Online]. Available: http://www.lmp.ucla.edu/default.aspx?menu=001. [Accessed: 04-Mar-2012].
62. Oromo Language. http://en.wikipedia.org/wiki/Oromo_language; Last accessed on August 25, 2009.
63. Omnigton "the online wncyclopedia of writing systems and language" accessed from http://www.omniglot.com/writing/oromo.htm, May 10, 2013.
64. Kula Kekeba Tune, Vasudeva Varma and Prasad Pingali, "Evaluation of Oromo-English Cross-Language Information Retrieval", IJCAI 2007 Workshop on CLIA, Hyderabad, India, 2007.
65. Oromo Language: Encyclopedia, http://en.allexperts.com/e/o/or/oromo_language.htm, Last accessed on April 10, 2010.
66. C. Griefenew-Mewis, "A Grammatical Sketch of Written Oromo", Druckerei Franz Hansen, Bergisch Gladbach, Germany, 2001.
67. Wakshum Mekonnen, "Development of a Stemming Algorithm for Afaan Oromoo Text," Master's thesis, Addis Ababa University, School of Information Studies, 2000.
68. Williams D., The Teacher's Grammar Book. 2nd Ed. Mahwah: Lawrence Erlbaum Associates, Inc. p. 220, 2005.
69. Gragg G., 'Oromo of Wallegga'. In M. Bender (ed.). The non-Semitic languages of Ethiopia. East Lansing (MI) and Carbondale (IL): African Studies Center of the MSU and Southern Illinois University. pp: 166-95, 1976.
70. Heine B., The Wata dialect of Oromo. (Language and dialect atlas of Kenya 4). Berlin:Dietrich Reimer Verlag, 1981.

71. Lloret M., A Comparative Study of Consonant Assimilation in Some Oromo Dialects. Ms. A paper presented at the 3rd International Symposium on Cushitic and Omotic Languages, Berlin, 1994.
72. Simon Ager, Oromo Language, *Online edition*, 2012. [Online]. Available: www.sas.upenn.edu/African_Studies/Hornet/Afaan_Oromo_19777.html. [Accessed: 25-Feb-2012].
73. Mylanguage.org, Oromo Numbers, 2011. [Online]. Available: http://www.mylanguages.org/learn_oromo.php. [Accessed: 12-Jan-2012].
74. 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.
75. Tesfaye Guta, Afaan Oromo search engine, Master's Thesis, Addis Ababa University, Departement of Computer Science, 2010.
76. Getahun A., The analysis of ambiguity in Amharic, Journal of Ethiopian Studies, Vol. 34, No.c2, 2001.
77. Agirre, E. and Martinez, D., Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content. 2000.
78. http://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics, accessed April 20, 2013.
79. Tony McEnery, Andrew Wilson, Corpus Linguistic, Edinburgh University, published by Edinburgh University Press, 2001.
80. Getachew Mamo, Automatic Part Of Speech Tagging for Afaan Oromo Language, Master‟s Thesis, School of Graduate studies, Addis Ababa University, 2009.
81. 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.
82. L. JB. "Development of a stemming algorithm". Mechanical Translation and Computational Linguistics, 11: 22-31 (1968).

83. Taeho Jo1 and Malrey Lee, The Evaluation Measure of Text Clustering for the Variable Number of Clusters, Advanced Graduate Education Center of Jeonbuk for Electronics and Information, South korea, 1998.
84. Kaplan A., An experimental study of ambiguity and context, Mechanical Translation, vol.2 no.2, 1955.
85. Daniel Jurafsky & James H.Martin, Speech and Language Processing: An introduction to speech recognition, natural language processing, and computational linguistics, October 12, 2007.

Appendix A: Sample list of Afaan Oromo sense examples used in the corpus

- Dhaabbatichi hanga kurmaana 2ffaa bara 2005tti hojiiwwan misooma bosonaa deggeran; hojii sanyii<Ija_mukaa_ykn_midhaani> qopheessuu biqiltuu qopheessuu kan raawwate yoo ta'u lafa hektaara 5389 irrattis hojiin biqiltuu kunuunsuu raawwatamuusaa Dar.olaanaan dhaabbatichaa obbo Didhaa Dirribaa ibsaniiru.
- Dhaabbatichi hojii misooma sululaa hirphuuf biiroo qonnaatiif qarshii miliona 2 kan kenne yammuu ta'u, sanyii<Gosa> mukaa KG 2400 kan qarshii miliona 1.2 baasu funaanaa akka jiru ibsa kanarraa hubachuun danda'ameera.
- Qamooleen badhasa kanarratti hirmaatan keessaa qonnaan bultoota adda duree, waldalee hojii gamtaa, jiduugaleessaa qorannoo qonnaa fi qoratoota, waldalee garaagaraa sanyiiwwaan<Ija_mukaa_ykn_midhaani> adda addaa baayi'suurraatti boba'aan akka badhaafamanis dubbataniiru.
- Ejensichi qonnaan bulaa fi mootummaa fayyadamaa gochuuf hojii sanyii<Gosa> loonii fooyyessuuf oolan babal'isaa jira jedhan.
- Naayitiroojinii dhangala'aan sanyii<Gosa> kormaa yeroo dheeraaf kaa'uun yeroo barbaachisetti tajaajila mala namaan loon diqaalomsuu raawwachuuf teekinooloojii faayidaa gudda qabudha.
- Haaluma wal fakkaatuun Sanyiin<Gosa> kormaa bara 2004 hanga kurmaana lammaffaatti qonnaan bulaaf dhiyaate doozii 41,754 ture bara kana hanga ji'a jahaattii doozii 70,775 ga'eera Faayidaasaa hubachuurraa kan ka'es fedhii qonnaan bulaan sanyii loonsaa fooyyessuuf qabu daran dabalaa waan jiruuf giddu galoonni tajaajila kana kennaa jiran gahaadha nama hin jechisiisu.
- Biyya keenya keessatti namoota beekamoo fi sanyii<Gosa> moototaatti of hirkisuun aadaa barame dha.
- Teekinooloojii omisha duraa lafa qopheessu, lafa qotuu, biyyee bulleessuu, sanyii<Ija_mukaa_ykn_midhaani> facaasuu, biqiltuu midhaanii babbaquu, teekinoolojiiwwan jal'isii fi bishaan harkisu ta'a.
- Fakkeenyaaf, qotaa fooyya'aa, bulleessaa biyyee sanyii<Ija_mukaa_ykn_midhaani> facaastuu omisha booda midhaan dhahan, calleessan, qulqulleessan, geejiban, kuusanii fi kkf yoo ta'u, fakkeenyaaf dhahaa midhaan adda addaa, calleessaa boqqolloo, maashina midhaan adda addaa qulqulleessu, gaarii gaalaan harkifamuu fi gaariiwwan adda addaa, kuusaa adda addaa hojjachuun dhiyeessaa jiru.
- Hanqina Sanyii<Ija_mukaa_ykn_midhaani> filatamaa jiru furuuf hojii hojecha jiru.
- Hojii sanyii<Ija_mukaa_ykn_midhaani> baay'isuu fi qulqulleessuus hojjechaa jiru.
- Xaa'oo fi sanyii<Ija_mukaa_ykn_midhaani> filatamaa yeroon qonnaan bulaa biraan gahuufis ta'e omisha isaa gabaatti geeffachuu akka danda'u daandiin bu'uura misoomaa hundaa ol murteessaadhas jedhaniiru.
- Haaluma kanaan guutuu Oromiyaa aanaalee sagantaa fooyya'iinsa sanyii<Gosa> looniitiin hammataman 75 keessatti loowwan kuma 45 ol tibbana bifa duulaan guraandhala 30 irraa

eegalee diqaalomfamaa akka jiran Ejensichatti Abbaan Adeemsa Hojii Dhimmoota Kominikeeshinii Mootummaa obbo Mulaatuu Haayilee himaniiru.

- Kaayyoon bifa duulaatiin loowwan diqaalomsuu kunis, loowwan hedduu yeroo tokkoon sanyii kormaa qabsiisuun yeroo gabaabaa keessatti loon sanyii<Gosa> filatamoo horachuun oomishaa fi oomishtummaa misooma horii dabaluudhaan fedhii ummatni aannanii fi bu'aa horiirraa argatu akka guutamu taasisudha jedhaniiru.
- Sagantaa fooyya'iinsa sanyii<Gosa> looniitiin barana loowwan kuma 96 fi 800 diqaalomsuuf karoorfame keessaa hanga ammaatti loowwan kuma 50 ta'an diqaalomsamuusaanii obbo Mulaatuun eeranii, hojii kanarratti qonnaan bultoonni kuma 45 ol ni hiirmaatu jedhamee ni tilmaamamas jedhaniiru.
- Qonnaan bultoonni Godina Wallagga Bahaa Aanaa Bonayyaa Boshee sanyii<Ija_mukaa_ykn_midhaani> filatamaa boqqolloo baayisuurratti bobba'an bu'aa birrii miliyoona 4 fi kuma 700 caalu argatan.
- Qonnaan bultoonni aanaa kanaa 746 ta'an sanyii<Ija_mukaa_ykn_midhaani> filatamaa boqqolloo bara omishaa 2004 hanga 2005 tti lafa hektaara 289 irratti baayisaani jiru.
- Akka Gadaa Booranaatti immoo, hayyuun Sanyii<Gosa> sadii qaba; Isaanis: Hayyuu Adulaa, hayyuu Garbaa fi hayyuu Meendhichaati.

- Waldaan Aksiyoona Liiqii fi Qusannaa Oromiyaa imaammataa fi tarsiimoo mootummaan baase deeggaruun jiruu fi jireenya ummataa fooyyeessuuf waggaa 15 dura hundaa'uun tajaajila liqaa, qusannaa, inshuraansii xixiqqaa, daddabarsa horii<Maallaqaa_ykn_Qarshii> biyya keessaa fi tajaajila gorsaa maamiltootaaf kennaa jira.
- Bara 2004tti ummanni Leeccalloo, horii<Maallaqaa_ykn_Qarshii> fi humnaan walumatti birrii 477,343,858.10 caalu gumaachuun miira abbummaa mana barumsaatiif qabu mirkaneesseera.
- Magaalaa keenya caalmaatti guddina ishii saffisiisuuf horii<Maallaqaa_ykn_Qarshii> fi humna qabnu qindeessinee investimentii magaalichaa keessa yoo galle dha.
- Waldaalee kanaafis gurmii irraa jalqabee deeggarsi leenjii, liqii horii<Maallaqaa_ykn_Qarshii> fi iddooo hojii osoo walirraa hin-citin kennamaa tureera.
- Gama baajataatiin immoo, akka godina Shawaa Kaabaatti, baajanni gama mootummaatiin qabame horii<Maallaqaa_ykn_Qarshii> miliyoona 49 ta'a jedhame yaadama.
- gama hirmaannaa uummata godinichaatiin horii<Maallaqaa_ykn_Qarshii> walitti qabamuuf karoofamee, birriin miliyoonni 14 fi kumni 84 fi 804 walitti qabameera.
- Birrii misooma Jaarraa irraa qabame miliyoona 49 keessaa, gorsitootaa fi kontiraankiterootaaf, horii<Maallaqaa_ykn_Qarshii> birriin miliyoonni 32 kafalameeraaf.
- Horii<Maallaqaa_ykn_Qarshii> baankii kaa'ameera.
- Birriin baankii ta'e horii<Maallaqaa_ykn_Qarshii> hirmaannaa ummataatiin walitti qabame dha.

- Ummanni godina Shawaa Kaabaa, ummata misooma jaalatu waan ta'eef, kiisii isaaniitii horii<Maallaqaa_ykn_Qarshii> baasuu irra darbanii hojii tolaa humna isaaniitiin hojjachaa jiru.
- Hirmaannaan ummanni aanichaa taasisu kunis, hojii humnaatii fi buusii horii<Maallaqaa_ykn_Qarshii> ti.
- Hojiiwwaan daandii aanichaa hojjachuuf baajata hirmaannaa uummataatiin horii<Maallaqaa_ykn_Qarshii> kumni 800 walitti qabameera.
- Bara 2004tti hojii gama kanaan hojjeterraa galiin horii<Maallaqaa_ykn_Qarshii> birrii miiliyoona 13.7 ta'u galuu danda'eera.
- Akka ragaa Ejansii Misooma IMX Oromiyaatti, interpiraayizotni ijaaraman kunneen priojektiiwwan mootummaa baadiyyaa fi magaalaatti gaggeeffaman hunda irrati akka hirmaatan kan taasifame yoo ta'u, gama biraatiin ammoo, hojii dhabdoonni baadiyyaatti argaman hojii qonna midhaanii fi horii<beelladaa> furdisuu, akkasumas magaalaatti, hojiiwwan kobil istoonii, hojii mukaa fi sibiilaa keessatti bal'inaan hirmaataa jiru.
- Yuuniyeenichi, rakkoolee hawaasa naannoo hiikuudhaafis, kilinika fayyaa horii<beelladaa> fi keellaa fayyaa namaa tokko hojjechuun faayidaa irra oolcheera.
- Keellaa fayyaa horii<beelladaa> hojjechiise tajaajilaaf oolcheera.
- Buufati fayyaa lubbu horii<beelladaa> darbuu danda'u oolchuuf fayada.
- Oyiruu qonnaa keenya keessatti nyaata horii<beelladaa> dhaabuun faayidaa dachaa argataa jirra jedhaniiru.
- Hippoo horii<beelladaa> hin qabduu ni elmatti Deebiinsaa silmii ykn balqii jechuu dha.
- Tikseen horiin<beelladaa> dura bahee; horiin duuba gala.
- Marga gargaaramuun horii<beelladaa> furdisuun ni danda'ama.
- Hojiilee qabeenya uumamaa hojjechuun bakka turetti deebisuu fi sanyii horii<beelladaa> horsiisee bulaa fooyyessuu hojjetamaa jiru.
- Rakkoowwan bishaan dhugaatii namaa fi horii<beelladaa> furuudhaaf piroojektoota gurguddoo ta'an waliin ta'iinsa mootummaa fi mit-mootummaatiin hojjetamaa jiru.
- Waldoonni afur ta'nis hojii horii<beelladaa> furdisuu irratti bobba'aniiru.
- Har'a jireenyi keenya fooyya'uusaatiin badhaasaaf geenyeerra jedhanii, hojii misooma horii<beelladaa> fooyya'aa fi misooma qonnaa teekinooloojiin deeggarametti fayyadamuun jireenyi isaanii akka fooyya'es himaniiru.
- Ayyaana sadarkaa naannoo Oromiyaatti horsiisee bultoonni horii<beelladaa> 66 fi akka biyyaalessaatti ammoo, horsiisee bultoonni 220 hojii misooma horiitin badhafamaniiru.

- Haaluma kanaan bulchiinsa magaalaa sabbataatti ji'oottan jahan darban keessa hojiileen ibsaa, karaa<daandi> keessaa keessaa cirrachaa fi dhagaa koobiliitiin hojjechuu kiiloometira 53, diichiin lolaa kiiloometira 4, boononn bishaanii 6, riqichi guddana tokkoo fi kkf baasii qarshii miiliyoona 19.2n kan raawwataman ta'uu kanatiibaan bulchiisa magaalattii Obbo Yemaanee yiggazuu ibsaniiru.
- Tolaan meeshaa karaa<daandi> ittin hojetaan ergifate deebise.

- Bu'uuraaleen misoomaa kanneen hafan kan akka karaa<daandi>, ibsaa, bilbila, bishaan dhugaatii qulqulluu waliin gahuuf hojii hojjetamaa turee fi hojjetamaa jiruun ummanni naannoo keenyaa sadarkaa sadarkaan irraa fayyadamaa ta'aa jira.
- Godina Oromiyaa garaa garaa aanaalee sadii keessatti karaan<daandi> bonnaa fi ganna tajaajilu kiloomeetirri 197 ol bajata mootummaan rammadee fi hirmaanaa ummataa qarshii miliyoona 87 oliin hojjetamaa akka jiru Waajjirri Abbaa Taayitaa Daandiiwwan aanaalee kanneenni beeksisan.
- Godina Wallaggaa Lixaa aanaa Sayyoo Nooleetti karaan<daandi> bonaa fi ganna tajaajiluu kiloomeetirri 49 bajata mootummaan rammadee fi hirmaanaa ummataa qarshii miliyoona 22 oliin hojjettamaa akka jiru Abbaan Taayitaa Daandiiwwan aanichaa beeksise.
- Ogeessi Teekinishaanii karaa<daandi> aanichaa Obbo Masgabuu Abdiisaa akka jedhanitti karaan kun bara darbe keessa bajata mootummaa fi hirmaanaa ummataatiin akka eegalamu taasifameera.
- Karaan<daandi> kun yeroo tajaajila kennuu eegaluutti gandoota aanichaa 13 kan walqunnamsiisu ta'uus himaniiru.
- Haaluma walfakkaatuun, godina Harargee Bahaa aanaa Baabbilee fi Gursumitti karaan<daandi> kiloomeetirri 148 ol hirmaanaa ummataa fi bajata mootummaan ramade qarshii miliyoona 65 oliin hojjetamaa akka jiru Waajjirri Abbaa Taayitaa Daandiiwwan aanaalee kanneenii ibsan.
- Bahee hojiirra oolaa jiruun aanaalee kanneenitii karaa<daandi> bonaa fi ganna tajaajiluu danda'u kiloomeetirri 148 fi meetira 600 tahuu hojjetamaa jira.
- Aanaa Baabileetti karaan<daandi> kiloomeetira 57 ta'u hojjettamaa jiru keessaa daandiin kiloomeetira 19 tahuu xumuramuun tajaajila kennaa akka jiru itti gaafatamtuun Abbaa Taayitaa Daandiwwan aanichaa Aadde Immabeet Boggaalaa dubbataniiru.
- Pirojektiin karaa<daandi> Asfaaltii Baalee kiiloometira 132 haguugu birr.mili 300 fi mil 80 oliin ijaarame tibbana sirna ho'aan eebbifame.
- Ijaarsa pirojektii karaa<daandi> Asfaaltii Roobee Gindhiiriifis qohiin barbaachisu xumurame.
- Ijaarsi karaa<daandi> kunis deebii argachuu kan danda'e gaaffii maanguddoonni muummicha ministeeraa duraanii Mallas Zeenaawiitiif dhiyeessaniin.
- Muummichi ministeeraa duraaniis rakkoo daandii ummanni godinichaa ittiin dararamaa tureef deebii kennuuf karaan<daandi> Asfaaltii Baalee bara 1999 Abbaa Taayitaa Daandiiwwan Itoophiyaatiin hogganamee ijaarsi akka eegalu ta'e.
- karaan<daandi> kun bara 2004 xumuramee tajaajila kennuu eegale.
- Biyya guddachuu irratti argamtu tokkoof tajaajila hawaasummaa, dinagdee fi bulchiinsaa babal'isuuf misoomni karaa<daandi> ga'ee olaanaa qabaata.
- Keessumaa walitti dhufeenya hawaasa fi dinagdee jiraattota baadiyyaa fi magaalaa gama cimsuutiin karaan<daandi> jiraachuun murteessaa dha.
- Bu'aalee qonnaa fi industirii ummata bakkeen lamaan jiraatu qaqqabsiisuun kan danda'amu karaan<daandi> sadarkaa amansiisaa ta'een yoo jiraate dha.

- Tajaajilawwan fayyaa, barnootaa, bishaan dhugaatii, ekisteenshinii fi misooma adda addaa ummata magaalaa fi baadiyyaan ga'uufis karaan<daandi> ga'ee ol'aanaa taphata.
- Bakka karaan<daandi> hin jirettis bu'uuraalee misoomaa hawaasa biraan ga'uun kan yaadamu miti.
- Bara 1948titti maanguddoonni kiiloomeetira hedduu lafoo fi fardaan kutanii gara Finfinnee dhufuun karaan<daandi> godina Baalee fi giddu gala biyyaa walquunnamsiisu akka hojjetamuuf gaafachaa turaniiru.
- Haaluma kanaan manichi kan argamu magaalaa Dannabaa keessatti yoo ta'u ka'umsi caalbaasii manichaa qarshii 43,540 dha caalbaasiin kan gaggeefamu bakumaa mannichaatti sa'aatii 3:30 – 6:30tti ta'a mannicha namni kamillee karaa<akkaata_kallatti> bakka bu'aa isaa ykn qaaman dhihatee bitachuu nidanda'a.
- Himataan Dhaabbata Gargarsaa maallaqaa adaa fi himatamtooni isin jidduu kan jiru falmii siviilii ilaalchisee isin himatamtoonni mana murtii ol'aanaa Godina Shawaa Bahaatti himatamuu keessaan beekitanii himannaa isin irratti dhiyaatee karaa<akkaata_kallatti> kutaa ofisarraa seera garee hariroo hawaasatiin galmee Lakk. 33225 ta'e irraa gaafa 10/7/2005tiin dura dhiyaatanii ergaa fudhatanii booda deebii keessani bareeffamaan qopheefatanii beellama armaan olitti ibsameetti sa'aatii 4:00tti dhaaddacha siv. 3ffaa irratti akka dhiyattan yoo dhiyachuu baattan falmiin bakka isin hin jirretti kan ilaalamee murtiin kan laatamu ta'uu beeksifnaa. Mana Murtii Ol/Go/Shawaa Bahaa.
- Himataan Dhaabbataa Gargarsaa Maallaqaa adaa fi himatamtoonni isin jidduu kan jiru falmii siviilii ilaalchisee isin himatamtoonni mana murtii Ol'aanaa Godinaa Shawaa Bahaatti himatamuu keessaan beektaanii himanaa isin irratti dhiyaate karaa<akkaata_kallatti> kutaa ofisaraa seera garee hariroo hawaasatiin galmee Lakk. 33226 ta'e irraa gaafa 10/7/2005 tiin dura dhiyaattanii ergaa fudhatanii booda deebii keessaani barreefamaan qopheefatanii beellama armaan olitti ibsametti sa'aatii 5:00tti dhaaddacha siv. 3ffaa irratti akka dhiyattan yoo dhiyachuu baattaan falmiin bakka isin hin jirretti kan ilaalamee murtiin kan laatamu ta'uu beeksifna. Mana Murtii Ol/Go/Shawaa Bahaa.
- M/A/Mirgaa Obbo Waldahaannaa Aayimekuu Birzuu fi M/A/Idaa 2ffaa Obbo Tashoomaa Kaasahuun jidduu kan jiruu falmii raawwachiisaa murtii ilaalchisee isiin M/A/Idaa 2ffaa kun mana murtii ol'aanaa Godina shawaa Bahaatti himannoo raawwii waan isiin irratti dhiyaateef himannoo raawwii dhiyyate kan karaa<akkaata_kallatti> kutaa ofiisaraa seera siviilii itti Lakk. galmee 33456 ta'ee irraa ergaa fudhatani booda akkataa murtii isiin irraatti kennameen raawwatanii akka dhiyyaattan yoo raawwachuu baattani sababaa raawwachuu dhabdaaniif yoo qabatanii gaafa beellama 12/7/2005 tti ganama sa'aatii 3:30tti dhaddachaa 1ffaa irraatti dhiyyatanii akka ibsitan M/M/ajajeera. M/M/O/G/Shawaa Bahaa.
- Gama biraatiin Koreen kun kan ilaale, aadaa gaarii ummanni rakkoowwan mudatan karaa<akkaata_kallatti> nagaatiin hiikkachuuf tumsi inni yeroo garaa garaa taasisaa jiru ammas cimee akka itti fufuu fi jajjabaachuu qabas jedheera.

- Galmi kun ammoo milkaa'uu kan danda'u hawaasni faayidaa barnootaa karaa<akkaata_kallatti> guutu ta'een hubachuun qooda fudhannaan inni dhimma mana barumsaarratti qabu yoo dabalee dha.
- Yeroo ammaa karaa<akkaata_kallatti> marii sadarkaa raayyaa dubartootaatti, sagantaalee paakeejii fayyaa 16n keessaa tokko kan ta'e karoora maatiirratti mar'achuun, daa'imman meeqaafi akkamiin akka godhachuu qabanirratti ni mari'atu.
- bulchiinsi mootummaa fi Ummataas karaa<akkaata_kallatti> seeraatiin akka raawwatamu ciminaan kan qabsaa'ee fi biyya olaantumman seeraa itti mirkanaa'e uumuuf halkanii fi guyyaa taffaafachaa kan ture hogganaa bilchaataa ture.
- Namoonni dogongora isaanii yoo sirreeffatanii fi karaa<akkaata_kallatti> seeraafi nagaatiin yoo sochoo'an jedhee kan amanu hogganaa garaa bal'atu ture.
- Hogganoonni dhaabbiilee mormitootaa karaa<akkaata_kallatti> seera fi nagaatiin mormii isaanii akka tarkanfachiisan bilisummaa kan gonfachiisee fi ejjannoo dimokraatawaan paartiilee kanneen wajjin hojjachuuf fedhii agarsiise qabatamaan mirkaneessuuf hoggansa bilchaataa kenneera.
- Ejjannoon isaa kun ejjannoon siyaasaa kamiyyuu karaa<akkaata_kallatti> bilisaa fi seera qabeessa ta'een adeemuu qaba jedhee amanuudhaan biyya keenya keessatti yeroo jalqabaatiif siyaasaan mormii fi sirni paartii hedduu hojiiraa akka oolu gochuuf ejjannoo bilchaataa gootummaan tarkanfachisaa ture wajjin kan wal qabatae dha.
- Sochii kanaanis, magaalonni naannoo keenyaa haala mijaa'aa misoomni baadiyyaa si'ataa uumeefitti fayyadamuudhaan karaa<akkaata_kallatti> guddina misooma dinagdee ariifachiisaa itti mirkaneessuu danda'anii fi guddinni misooma dinagdee magaalotaa fi baadiyyaa akka wal deeggaraa deemu taasifameera.
- Kanaan ala, karaa<akkaata_kallatti> kamiiniyyuu aangoo argachuunis ta'e aangoo mirkaneeffachuun akka hin danda'amne heerri mootummaa keenya ifatti kaa'eera.
- Somaaliyaa keessatti yeroo ammaa kana humni garee Alqaayidaa keessaa tokko ta'e Alshabaab dadhabaa dhufuu fi biyyattiin nageenya argataa dhufteen ummannishee karaa<akkaata_kallatti> nagaatiin mari'achuu eegaluun Itoophiyaafis milkaa'ina ta'uutu himame.
- Qabsoo farra hiyyummaa gama hundaan jalqabame fiixa baasuuf ammas caalmaatti tarsiimoowwan kanaaf bahan karaa<akkaata_kallatti> guutuu ta'een hojiirra ooluu qabu.
- Qophii keenya kanaanis tarsiimoo misooma barnootaa karaa<akkaata_kallatti> fooyyee qabuun hojiirra oolchuun qabsoo hiyyummaa waliin godhamu ariifachiisaa warra jiran irratti xiyyeeffanneerra.
- Sagantaaleen kunneen karaa<akkaata_kallatti> guutuu ta'een hojiirra oolanii qulqullinni barnootaa mirkanaa'uu kan danda'u, sochii barsiisoonni, barattoonnii fi hoggansi barnootaa taasisuun qofa miti.
- Aanaan Digaluu fi Xijjoo aanaalee godina Arsii PMQB karaa<akkaata_kallatti> fooyyee qabuun itti mirkanaa'ee fi ijaarsi raayyaa misooma barnootaa cimaan itti uumamee keessaa tokko Obbo Qaasim Shankoo I/G/W/Barnootaa aanaa kanaati.

- Tarsiimoowwan kunniin karaa<akkaata_kallatti> guutuu ta'een hirmaachistummaa dargaggootaa fi dubartootaatiin akka mirkanaa'u taasisuuf liigonni amaanaa guddaatu itti kenname.
- Hojiiwwan mana mootummaa karaa<akkaata_kallatti> ummataaf ifa ta'een kan raawwatamu ta'uu qaba.
- Gaaffiiwwan armaan olitti kaafaman kunneen, yeroodhaaf deebii quubsaa ta'e haa dhabanii malee, sirna Gadaa keessatti namni miseensa Gadaa ta'e kamiiyyuu, beekumsaa fi yaadasaa karaa<akkaata_kallatti> dimookiraatawwaa ta'een waan ibsatuuf, seerri Gadaa Oromoo yeroodhaa gara yerootti namoota adda addaatiin waan fooyya'uuf, kalaqiinsi sirna Gadaa kalaqa nama dhuufaatiin himamuun ykn waamuun gonkumaa waan hin yaadamnee dha.

- Hojiilee hojjetamaa turan keessaa sirna<seera_ykn_aadaa> gibiraa fi taaksii ammayyeessuu, maddoota galii babal'isuu, kafaltoota gibiraa fi taaksii barsiisuun yakka gibiraa fi taaksiirratti raawwatamu ittisuudhaan galii mootummaa fi manneen qopheessaa guddinni dinagdee naannichaa maddisiisuu danda'u si'oominaa fi gahumsaan walitti qabuuf bal'inaan hojjetamaa tureera.
- Margaaret Taacher sirna<seera_ykn_aadaa> sochii dinagdee biyyattii bu'uurarraa jijjiiruun hojiin bu'uuraalee misoomaa mootummaan hojjetamaa ture irraa jalaan akka dhaabatu ta'eera.Sababoota kanaafis dhaabbileen fayyaa,barnootaa fi kanneen biroos bu'aadhaaf akka dhaabatan ta'e.
- Mohaammad mootummaan naannoo Oromiyaa sagantaa fooyya'iinsa sirna<seera_ykn_aadaa> haqaa hojiirra oolchuun olaantummaan seeraa fi bulchiinsi gaariin akka mirkanaa'uuf hojjechaa jira jedhaniiru.
- Daldalaan karaa haqa qabeessaan gabaa keessatti dorggomuun ofis fayyadee guddina dinagdee biyyatti keessatti qooda isaa akka bahuf mootummaan haala mijataa uumuuf imaammata gabaa bilisaa baasee hojiirra oolchuusaatiin sirna<seera_ykn_aadaa> daldalaarraa bu'aa guddaan argamuu danda'eera.
- Haata'u malee sirna<seera_ykn_aadaa> gabaa bilisaa kana karaa sirrii fi haala imaammatichi jedhuun hojiitti hiikuurratti rakkoo guddaatu mul'ata . Rakkoon kun ammoo walitti kuufamuun gabaa bilisaa dorgommii mijataa ta'e kana dadhabsiisuu bira darbee shammattootarratti dhiibbaa gochaa jira.
- Karoora waggoota shaniif (2003-2007) qophaa'e keessattis daa'imman fedhii addaa qaban qaama miidhamtoota ta'an bara barnoota darban keessatti carraa barnootaa dhabanii turan ilaalchii fi xiyyeeffannaan addaa itti kennamee sirna<seera_ykn_aadaa> barnoota hunda hammatootiin manneen barnootaa hunda keessatti hiriyootasaanii waliin akka baratan gochuuf karoorfameera.
- Hojimaata kanaanis sirna<seera_ykn_aadaa> barnootaa sanyiidhaan, qabeenyaan, saalaan, amantiidhaan, miidhama qaamaatii fi rakkoowwan birootiin osoo hin qoodin barattoota hunda simachuu fi keessummeessuu kan dandaeesisudha. Gama biraatiin barnoota hunda

hammatoon barattoonni hundi barachuuf manneen barnootaa idilee fi dhaabbilee leenjii ogummaa fi teekniikaa keessatti carraa akka argatan ta'a.

- Humna raawwachiisummaa jechuun hojjetoota Inistiitiyuutichaa fi mamiltoota rakkoo gama ilaalchaa,dandeetii fi hordooffii degarsaa qaban lenjiiwwan garaa garaatiin cimsuudhaan raayyaa jijjiiramaa tokko ta'uudhaan sirna<seera_ykn_aadaa> diriirsuun kan raawwatamu ta'a.
- Filannoo, qacarraa fi ramaddii hojii akkasumas haala sirna<seera_ykn_aadaa> hojii hunda irratti loogiin Qaama Miidhamummaa bu'uurefachuun qaama Miidhamtoota irratti taasifamu dhorkaa ta'uun isaa keewwata 27(a) jalatti ifatti tumamee jira.
- Himataan Dhaa/Gargaarsaa Maallaqaa addaa fi himatamtoonni 1ffaa Abarraash Tasfaayee 2ffaa Marii Tufaa 3ffaa Tasfaayee Waakkennee jidduu kan jiru falmii siviilii sirna<qophii_ykn_sagaanta> gabaabaan akka illaalamuuf dhiyaate himannaan isiin irratti waan dhiyaateef guyyaa waamichi isiinif darbee irraa eegalee guyyaa 10 keessatti ofirraa ittisuuf ykn falmachuuf hayyamsiifachuu kan qabdan ta'uu beeksisa, ta'uu yoo baate himataaf murtiin kan kennamu ta'uu manni murtii ajajeera.
- Ummanni Oromoo ilma fuudhaafi gaheefi intala heerumaaf geesse sirna<qophii_ykn_sagaanta> walitti fiduu mataasaa danda'e qaba.
- Adeemsa sirna<qophii_ykn_sagaanta> kanaa keessatti dursa mucaan fuudhaaf gahe mucayyoo fuuchuuf barbaadu yeroo bishaan buutu, yeroo qoraan cabsituufi erga adda addaa warrasheetii ergamtee deemtu ilaallachuun fedhiisaa warrasaa gurra buusa.
- namoonni sirna<qophii_ykn_sagaanta> sanarratti argaman hundi ka'uun ijaajjanii simatu.
- sirna<qophii_ykn_sagaanta> cufiinsa shaampiyoonaa Ispoortii yeroo sadaffaaf akka godina addaa Oromiyaa Naannawa Finfinneetti adeemsifamaa ture irratti itti aanaan bulchaa godinichaa obbo Admaasuu Tashoomaa haasawa taasisaniin kaayyoon shaampiyoonaa ispoorti kanaas ispoortessitoota ciccimoo dandeettifi gahumsa qaban baay'inaan horachuun dorgommiilee adda addaarraatti godinicha bakka bu'uu danda'an filachuuf akka ta'e himaniiru.
- Sana boodas, warra intalaatti jaarsa erguudhaan uumaa fi duudhaa jiruun sirna<qophii_ykn_sagaanta> gaa'elaa geggeessan.
- sirna<qophii_ykn_sagaanta> eebbaarratti haasa kan taasisan pirezidanti Alamaayyoon tajaajila fayyaa gahaa ta'e saffinaafi qulqullina qabu ummataan ga'uudhaan dhukkuboota daddarboo ittisuufi waldhaanuudhaan du'aafi dhibama sadarkaa sadarkaan xiqqeessuun hawaasa fayya buleessa ijaaruun omishtummaa ummataa guddisuuf hojjetamaa jira.
- sirna<qophii_ykn_sagaanta> kenniinsa boondii godina Addaa Naannawaa Finfinnee magaalaa Sabbataatti raawwatamerratti argamuun haasaa kan taasisan hogganaan biirichaa Obbo Siiraj Kadir akka jedhanitti biyyi keenya guddina saffisaa hawaasni sadarkaan irraa fayyadamu galmeesisuutti argamti.
- sirna<qophii_ykn_sagaanta> madaallii fi badhaasaa bara 2005 marsaa tokkoffaa sadarkaa sadarkaan gaggeeffamuuf haalli qabiinsa ragaa haala gaarii irra jiraachuu isaa ibsaniiru.

- Bara kanas sagantaa badhaasa gootota misoomaa marsaa 7ffaa haala ho'aa ta'een geggeessuuf qophii<Haala_mijeessu> barbaachisaan kan xumurameera.
- Ijaarsi pirojektii daandii asfaaltii Roobee-Gindhiiris yeroo amma hojjetamuuf qophiin<Haala_mijeessu> hundi xumuramuu beekameera.
- Bara baajeta 2005 dhufuttis hojii bal'aan gama kanaan qophiin<Haala_mijeessu> barbaachisu godhamaa jira.
- Tooftaa safarrii lafaa sadarkaa lammaffaatiinis naannolee garaa garaa keessatti bara kana lafa hektaara miliyoona 50 ta'u safaruun kenni waraqaa raga qabiyyee lafaa mirkaneessu akka raawwatu gochuuf qophiin<Haala_mijeessu> xumurameera jedhan.
- Keessumattuu murtii haqaafi qulqullina qabu kennisiisuu irratti dhiibbaa fidaa kan jiru ragaa sobaa irratti qorannoon kan gaggeeffame yoo ta'u, qulqullina qorannoo yakkaa fooyyessuuf giddu-galeessa qorannoo fooreensikii naannichatti hundeessuun qorannoo xumuramee humna namaa, meeshaafi baajanni barbaachisu adda bahee mootummaaf dhiyaachuuf qophii<Haala_mijeessu> irratti argama.
- Akkasumas, hanqinaalee ilaalchaa fi naamusaa waliin walqabatee jiru fooyyeessuuf hoggantoota qaamoolee haqaa fi nageenyaatiif leenjii kennuuf qophiin<Haala_mijeessu> xumurameera.
- Tajaajila bilbila baadiyyaa waliin gahuudhaaf, kanaan dura ragaaleen gandoota baadiyyaa bilbila qabaniifi hinqabne kan addaan bahee yoo ta'u, bara baajata kana keessatti gandoota baadiyyaa naannoo keenyaa hunda keessatti tajaajila waliin gahuuf qaamolee raawwachiiftuu sadarkaa godinaalee waliin kan irratti mari'atamee, karoorrii fi kallattiin raawwii kan gadi bu'ee ta'ee qophiin<Haala_mijeessu> gama Ethio-telecom'n jiru eegamaa jira.
- Dargommii tapha shaampiyoonaa isopoortii manneen barnoota 2ffaa naannoo Oromiyaa Guraandhala 3 -17 bara 2005 magaalota Adaamaafi Asallaa geggeessuuf qophiin<Haala_mijeessu> barbaachisu xumurameera.
- Haaluma kanaan milkaa'ina dorgommii kanaaf qophii<Haala_mijeessu> guutuun taasifameera.
- Dorgommichi jalqabaa hanaa xumuraatti naga-qabeessa akka ta'u gochuuf qophii<Haala_mijeessu> gahaan yoo taasiifameyyuu, hundaa ol hirmaannaan hawaasaafi ispoortessitootaa ga'ee bakka hin bu'amne qaba.
- qophiin<saganta> kun wanta dubbiftootaaf dhiyeessu ni qaba.
- Gareen kubbaa miillaa biyyaalessa Itoophiyaa qophii<saganta> tapha waancaa kubbaa miilaa Afrikaa 29ffaaf taasisaa jiruun tapha wiixata darbe garee biyyaaleessa Tuuniziyaa waliin taasiseen 1fi 1n walqixa xumuraniiru.
- kutaa jalqabaa qophii<saganta> keenyaa maxxansa darbeen ilaalleera.
- qophiin<saganta> kun qophii adda dilbata ganama dhiyaatu dha.
- Tumsa ummanni dhugeeffannaadhaan taasisuun, seenaa kana caaluyyuu hojjechuun ni danda'a kan jedhu ammoo dhaamsa qophii<saganta> keenya ittii goolabnu ta'a.
- qophii<saganta> barreeffama ejjannoo mootummaa naannoo Oromiyaa balballoomsuu ilaalchisee dhimmootni wayitaawoon, imaammataafi qabxiiwwan iftoomina barbaadanirratti

ejjennoo mootummaan naannichaa dhimmicharratti qabu ifa taasisuun ummanni hubannoo akka argatuufi miidiyaafis ka'umsa ta'ee haala itti fufinsa

- Garuu qophii<saganta>n qophaa'uufi guyyaa facaafatamu tokko.
- Cumboon kan qophaa'us yeroo mara osoo hin ta'in, yeroo qophii<saganta> aadaafi ayyaanonni garaa garaa kabajamanittifi keessummaan kabajaa kanneen akka Soddaa argamanitti nyaataf kan qophaa'udha.
- Waajjirichatti ogeessa qulqullina bunaa kan ta'an Obbo Immiruu Tesammaa akka dubbatanitti dhaabbii bara 2005/6f buufataalee biqiltuu mootummaa afuriifi kan dhuunfaa 720 irratti biqiltuuleen bunaa dhukkuba dandamachuu dandaa'an miliyoonni 3,740,050 kan qopheeffaman yemmuu ta'u, biqiltuulee kanneen keessaas miliyoonni 3,181,800 buufata dhuunfaarratti akka ta'eefi qophii<saganta> kanarrattis qonnaan bultoonni 720 hirmaachuusaanii beeksisaniiru.
- Ammas, kutaa lmaaffaa qophii<saganta> kanaa Ob. Tasammaa G/Madhin Bulchaa I/A fi Hogganaan Waajjira Qonnaa godina Arsii-Lixaa kan nutti himanii fi qaamolee abbootii gahee sadarkaan jiran irraa odeeffannoo arganne seentuu taasifachuudhaan itti seenna.

- Barumsicha dubbisuukee yommuu itti fuftu, yeroo baay'ee yaanni sun ifa<hubannaa> sii ta'aa deema.
- Beekumsa ulfina isaa isa fuula Kristos irraa mul'atu ifatti<hubannaa> baasuudhaaf garaa keenya keessatti nuuf ibse" jechuudhaan Yihowaa jajateera.
- Ibsuun, odeeffannoo dabalataa dhiheessanii dhimma sana caalaatti ifa<hubannaa> gochuu gaafata.
- Kanaaf wanta Waaqayyo dubbiisaa raawwachuuf godherratti amantii dhabuunsaanii, wanta garaasaanii keessa jiru ifa<hubannaa> godheera.
- Caqasawwan dubbiste ibsuufi caqasawwan kun murtoo tokkorra ga'uuf kan gargaaran akkamitti akka ta'e ifa<hubannaa> gochuu qabda.
- Barsiisaa gaariin, haasaan tokko kaayyoo ifa<hubannaa> ta'e qabaachuu akka qabu ni hubata.
- Qabxiiwwan ijoo ta'an akka gaariitti ifa<hubannaa> ta'anii dhihaachuufi kan yaadataman ta'uu qabu.
- Caqasoonni tokko tokko haalawwan dhimma sana duuba jiran ifa<hubannaa> gochuuf qofa kan galan ta'uu danda'u.
- Simbirroota qilleensa keessa balali'an, daraarota bakkee karaa isa dhiphoo, mana dhagaarratti ijaaramee fi fakkeenyawwan hedduu kanaa wajjin wal fakkaatanitti fayyadamuunsaa, barumsisaa xiyyeeffannaa namaa kan hawwatu, ifaafi<hubannaa> kan hin irraanfatamne akka ta'u godheera.
- Qabxiiwwan gurguddaaniifi sa'aatiin matadureewwan xixinnoo ibsuuf barbaachisan ifatti<hubannaa> kan kaa'aman si'a ta'u, murtoowwan kaan garuu siif dhiifamaniiru.
- Sagalee akka gaariitti dhaga'amuufi karaa ifa<hubannaa> ta'een dubbachuun kee, turjumaana keetiif barbaachisaadha.
- Mala wantoota walbira qabanii barsiisuutti fayyadamuun, barumsa Macaafa Qulqulluu barbaachisaa ta'e kana dhaggeeffattootaaf ifa<hubannaa> gochuuf gargaara.

- Garaagarummaa dhugumaan nama Waaqayyoo ta'uufi gocha fakkeessaa Fariisonni raawwatan gidduu jiru ifa<hubannaa> godheera.
- Ogummaafi kaayyoo ifa<hubannaa> ta'e sammuutti qabattee kana gochuu qabda.
- Yesus Caaffata Qulqullaa'oorraa seenaa isaan beekan tokko caqasuudhaan akka gaariitti erga isaanii ibsee booda, karaa isaan hin eegneen Waaqayyo namoota du'an akka kaasu ragaa ifa<hubannaa> ta'e dhiheessuudhaan deebii isaanii kenneera.
- Bu'aa gaarii argachuuf, wanta itti amantu ilaalchisee deebii salphaafi ifa<hubannaa> ta'e kennuudhaan alattis wanti si barbaachisu jiraachuu danda'a.
- Deebii akka deebistu yommuu gaafatamtu deebii salphaa, ifaafi<hubannaa> gabaabaa kenni.
- Dhaqxee isaa wajjin haasa'uu mannaa xalayaa barreessuufii kan barbaadde maaliif akka ta'e ifa<hubannaa> godhiif.
- Kaayyoon xalayichaa, keeyyata jalqabaarratti karaa salphaafi ifa<hubannaa> ta'een ibsameera.
- Bakka hundumaatti amalawwan Waaqayyoon gammachiisan kana guutummaatti kan calaqqiftu yoo ta'e guddinnikee ifatti <hubannaa> mul'ata.
- Fuula ifaadhaan <addenna> nagaa nama gaafachuun gaariidha.
- Mana barumsaa tajaajila barnootaa fi faayidaa beekumsaa argachaa yommuu adeemtu, mana keettis barata cimaa tahuuf halkan halkan yoo baadiyaa jiraatte ifa<addenna> gaasii, solaria fi kkf dubbisuu qabda.
- Kanaaf guddina gama kanaan gootutti gammadi, guddinni kees ifatti <addenna> dukana kessa bahee ni mul'ata.
- Naannoowwan tokko tokkotti wantoonni dubbisaaf sababii ta'an rakkoo ijaafi ifa<addenna> ga'aa argachuu dhabuu ta'uu danda'a.
- Haasaa kee karaa mi'aa qabuun dhiheessuun kee miira ho'aadhaan akka dubbattu argisiisuu irraa kan hafe fuula keerraa ifatu <addenna> mul'achuu qaba.
- Namni tokko gammachuu itti dhaga'amu fuula ifaadhaan <addenna> yommuu ibsu arguun nama gammachiisa.
- Akkkuma yaadatan tajaajila eeraman kamiyyuu namni barbaadu duraan dursa dirqama guutuu qabu akka guutu amanuun tajaajilamaa tokkoon tokkoof tarreeffama kana bakka ifaa <addenna> ta'e keenyeerra.
- Naannoo ifaan <addenna> elektiriikii hin jiretti gaasitti fayyadamuun ija waan nama dhukkubsuuf yoo danda'ame solaritti fayyadamuu qabna jedhe ogessi fayyaa.
- Daandiin qajeelotaa akka ifa<addenna> barii ganamaa, isa yeroo biiftuun baatee ol adeemtu, ittumaa ifaa<addenna> ol adeemuuti.
- Tajaajili mootumman ilma namaaf kennu hedduu fi tajaajila kannen kessaa akka dirqama lammummaatti namni kamiyyu bishaanii fi ifa<addenna> elektiriikii mootummaa irra qixa argachuu qaba.
- Waggoota sana boodas, daandiin qajeelotaa akka ifa<addenna> barii ganamaa, isa yeroo biiftuun baatee ol adeemtu ittuma ifaa<addenna> ol adeemuu ifuusaa itti fufeera.
- Bishaan, nyaataa, ifa<addenna> fi waantota dhala namaaf barbaachisu argachuu ykn sanuma dhabuudhaan adabamuun hidhamtootni yaroo qoratamanitti akka itti deebii kennan irratti hundaaya.
- Hidhamtootni mana dukkana kessatti hidhaman ifa<addenna> guyyaa arguu, mana fincaanii argaachuuf fi bakka bal'insa balbala duraatti gadi bayuuf carraan qaban daangeffamaadha.

- Akkasumas, hidhamtoonni haala hamaa keessati akka hidhaman himu, kunis, nyata gayaa fi ifa<addenna> aduu gayaa akka hin argatne, haala naannoo qulqullina hin qabne keessa akka jiraatan fi wal'aansa fayyaa gayaa argachuu akka hin dandeenye himatu.
- Sochiin dhala namaa baay'inaa fi amallan wantoota qilleensa marsaa lafaa keessatti ifa<addenna> biiftuu balaqqisiisuu fi xuuxuun dhiibbaa uuman akka uumaman godhee jira.
- Akkuma gaasonni oo'a harkisan mana magariisaa daaw'itee ifa<addenna> biiftuu irraa dhufu akka qilleensa marsaa lafaa keessa seenuu fi oo'a qilleensa marsaa lafaa keessaa bahu hambisuuf gargaaru.
- Humna ifaa<addenna> biiftuu irraa maddu keessaa 50% qilleensa marsaa lafaa keessa dabruun fuula lafaatiin fudhatama.
- kofa callaqqeettii ifa<addenna> aduu fi gaaddidduu muka meetirii 15 dheeratu gidduutti kan uumamu agarsiisa.
- Yeroo jalqabaaf Kaampaaniin Edisen jedhamu namoota bizinesii lama wajjin bara 1850tti waliin hojjachuuf walii galan, jarri lamaan kun qarshii barbaachisuu dhiyeessanii kaampaanii ifa<addenna> elektriikii kennu dhaabaniiru.
- Erga Kaampaaniin Edisen gadi lakkisee bayee booda, Teslaan namoota bizinesii lama wajjin bara 1886tti waliin hojjachuuf walii galan, kampaanii maqaan isaa Ifa<addenna> Elektirikii fi Warshaa Teslaa jedhamu dhaabuf murteessan.

- Obboo Bulchaa Dammaqsaa, ilaalcha fedhan haa qabaatan, akka ilma Oromoo tokkootti; akka hayyuu Oromoo hangafa tokkoottii; akka manguddoo keenyaatti; nama ulfina<kabajaa> guddaa argatanii dha.
- kan nama ta'e hundinuu ulfina<kabajaa><ba'aa> nameenyaa qaba.
- Horiin namaa ulfina<kabajaa> qaba.
- gaa'illi hunduma biratti ulfina<kabajaa> qabeessa haa ta'u, ciisichi gaa'elas hin xureeffamin jedha.
- Rabbi beeku fi isaa soodachu irraa kan ka'e hameenya irra fagachu, ulifina<kabajaa> rabbiif bulchu yooki kennuu fi gaaruuma namootatiif ta'u hojjechu jechu dha.
- Ulfina<kabajaa> f safuus waaqaaf kennuu qabana jedha.
- Oromoon lafaaf (dacheef) ulfina<kabajaa> guddaa qaba.
- Kunis nu warri rabbiin abdachuu irratti kanneen jalqabaa taane akka galata ulfina<kabajaa> isaaf taanuu.
- Namni tokko nama guutuu kan ta'u, yoo namaaf kabaja, waaqaaf ulfina<kabajaa> kenne, seera Gadaa yoo guutatee dha.
- Abbaan waaqaa gaditti kan ulfini kennamuufi dha.
- Innis hanga warri kan Waaqayyoo ta'an galata ulfina<kabajaa> isaatiif furamanitti qabdii dhaala keenna nuu mirkaneessuu dha.
- Abbaan ulfina<kabajaa>, rabbiin keenya isin uumee akka isin isa beektaniif gara isaa qajeeltan isin taasise.
- Xuriifi fincaan jettee osoo hin jibbin, harma hoosistee kan nama guddiste waan taateef seerii, haati ulfina<kabajaa> gudda qabdi jedha.
- Abbaan waraa tokko hawaasa keessatti fudhatamaa fi kabaja argachuuf dursee maatii ofii keessatti nama seera maatiif ulfina<kabajaa> kennu ta'uu qaba.

- akkasumas abdiin inni itti isin waame sun maal akka ta'e, badhaadhummaan dhaala qulqullootaa ulfina<kabajaa> qabeessi sun maal akka ta'e akka beettaniif ijji qalbii keessanii akka isinii banamu rabbi nan kadhadha.
- Kanaaf sababii ani isiniif rakkachaa jiruuf jettanii akka hin dhiphanne isiniin jedha kunis ulfinnuma keessaniifi.
- Abbaan akkuma badhaadhummaa ulfina<kabajaa> isaatti fi namummaa isaatiin namoota hundaan kabajama.
- Waldaan tokko tokko nama malamaltumma fi loogumma hojjatu akka nama waan gaarii tokko hojjatetti ulfina<kabajaa> an uf-dura isaan dhaabuudhaan waldicha akka hogganu godhu.
- abbaa keetii fi haadha teetiif ulfina<kabajaa> kennuun dirqmaa nama kamiituyyu.
- Kunis ulfinni<kabajaa> inni qabu darbee ilma isaatiin dhaalamuudhaan ilmi isaa akka nama hundaan jaalatamuu fi gochaa inni dalaguun bakka hundatti akka galateeffamuu taasifameera.
- Innis hanga warri kan Waaqayyoo ta'an galata ulfina<kabajaa> isaatiif furamanitti qabdii dhaala keenna nuu mirkaneessuu dha.
- Gariin dhaloota qubee jedhaa inn garuu habaqaala dhiigni goototaa biqilcheedha kan kabajaa Oromiyaa fi ulfina<kabajaa> Oromummaaf dhiigaa lubbuu ofii hin mararsiifanne.
- Sirna Gadaa keessatti, seeri yakka raawwatamuu malu hundaaf akka salphinaa fi ulfina<kabajaa> isaatti jira.
- Bokkuu jechuun ulee ulfina<kabajaa> kan warri
- Daa'imti keessan hamma daangaa dheerina ykn ulfina<ba'aa> teessoo konkolaatichaaf hayyamame bira gahutti yookii geessutti daa'ima kana teessoo konkolaataa fuullisaa fuulduraatti garagaluuf qabattoo qabxii 5 qabu keessa kaa'aa.
- Daa'imti keessan teesso olka'aa fayyadamuun dura yoo xiqaate paawundii 40 ulfina<ba'aa> qabaachuu qaba.
- Qulqullini buna dheedhii jechuun gudina, roga fi ulfina<ba'aa> firii bunaa jechuudha.
- Grichia jalqabaaf aannani sassabu yeroo jalqabu, aannani miseensa irra diyaatuuf ilaalun bitachuufii fi kafaltii ammoo qabiyee ykn ulfina<ba'aa> isaa irratti hunda'uudhan kafaluu ni danda'a.
- Yeroo annan bitan Kafaltiin isaa qabiyee maraan waan ta'eef qulqulinna annanii dabaluudhaan ulfina<ba'aa> aannani madalun garii dha.
- Dhadhaan dhangala'a gartokkee ulfina<ba'aa> qabu fi tilmaaman qabatii aannani % 80-85, bishaan % 15-16 fi dhangala'a qibatii hin ta'in % 2 of keessa kan qabu dha.
- Omisha baadu keessatti: aannani pasturized ta'e qabdu almuniyemmii /hin dammesofneetti danfisuu, raacatiin ulfina<ba'aa> aannani %2 itti dabaluuu, sukiin akka uumamu rennetii ulfina<ba'aa> annani sana % 1 itti dabaluuu, aannanichi hanga suuki godhatu tti akka ta'u gochuu.
- Yeroo dhalatan daa'imman ulfina<ba'aa> gad aanaa Kg 2.5 gad ta'an akkamitti sooruun danda'ama.
- Aannan harma garaacha daa'immatiif salphati kan akka bullaa'u ta'e umamaan qopha'e fi kan ittin ulfaanne<ba'aa> dha.
- Yeroo dhukkubsatan isa dur barame caalaa harma hoosiisun daa'imman dafani akka dandamatan fi ulfiina<ba'aa> qaama akka hin hirrisine garagaara.
- Yeroo dhalatan daa'imman ulfiina<ba'aa> baayye gadi aanaa qaban fi kan harma qabachuu hin dandeenye sooruuf.

- Daa'imman dhukkubsatani fayyan ulfiina<ba'aa> qaama hirrisan akka deebi'u fi guddina sirri akka itti fufu danda'an isa dur nyaatan caalaa akka nyaatan jajjabessuun barbaachisa dha..
- Daa'imman ulfina<ba'aa> dhaloota gad aanaa qaban, haatii yoo kan vaayirasii HIV Waliin jiraatu taate, daa'imman sirritti sooruuf deegarsa addaa isaan barbaachisa
- Aannan harma haadhaa daa'imman ulfina<ba'aa> qaama gad aanaa qabaniif soorata filmaata hinqabine dha.
- Daa'imman ulfiina<ba'aa> qaama gadi aanaa qaban tokko tokko harma hoodhuu waan hindandeenyeef aannan eelmamee kubbaayyaan dhuguu qabu.
- Annisa qulqulluun beeyladaaf ulfina<ba'aa> qaamaa dabaluuf gargaaru.
- Bishaan dabalatee waliigalati hooriidhaaf jireenya, guudina, ulfina<ba'aa> qama dabaludhaa fi sadarkaa oomishtummaa horii irraa barbaadamu dhugoomsuuf nyaata barbachisaan jechuudha.
- Daa'imman akka dafani dandamatan fi ulfiina<ba'aa> dhukkubsachu dura qabaacha turan akka deebifatan, yeroo dhukkubsatan isa barame caalaa hoosiisuun barbaachisa dha.
- Beeyladoonni kun osoo nyaata fooyya'aa hin argatiin gabaatti waan dhiyaataniif carraa
- uulfina<ba'aa> ,oomisha fooni fi uumurii ni xinesa.
- Beelladoonni nyaata qulqullina hin qabne yeroo nyaatan ulfina<ba'aa> qalamuu isaan dandeessisu jalaa tursa, foon beeyladoota kanaas qulqullina barbaadamaa ta'e hin qabaatu.
- Baayyinni nyaata goggogaa beeyladni tokko fudhachuu danda'u tilmaamaan dhibbantaa ulfina<ba'aa> qaama isaanii 2 – 3 kan ta'uu danda'u dha jedhamee ni yaadama.
- Haaluma kanaan hojii furdisuu keessatti oomisha jechuun dabaliinsa ulfina<ba'aa> qaamaa fi fooyya'iinsa haala qaamaa beeylada jechuudha.
- Kanaaf horiiwwan sagantaa furdisuu keessa galan nyaata qulqulluu jireenyaa fi ulfina<ba'aa> qaamaa akka dabalan isaan gargaaru barbaadu.
- Fedhiin annisaa beeyladoonni barbaadan annisaa qulqulluu lubbuu tursuuf gargaaru yookiin annisaa qulqulluu ulfina<ba'aa> qaamaa dabaluuf gargaaruu jedhamee ibsama.
- Annisaa qulqulluun lubbuu tursuuf gargaaru kan hundaa'u ulfina<ba'aa> qaamaa meetaboliikii beeyladaa irratti yoo ta'u kunis haala kanan gaditti ibsameen hojjetama.

Appendix B. Lists of Affixes removed from the token (Debela, 2010).

Suffix			
Se	aa	olee	oma
Ssi	dha	olii	fis
Nya	Nna	ota	siis
Tu	Tee	oolee	ooma
Chaaf	Uu	oota	siif
Tiif	suu	icha	fam
Chuu	Sa	ichi	ata

Appendix C: Lists of stop words removed from sentences adopted from (Debela, 2010).

akka	hanga	jechuun	ol	waan
akkam	henna	kan	oliif	waggaa
akkasumas	hoggaa	kanaaf	oliin	woo
akkum	hogguu	kanaafi	yammuu	akkuma
hoo	kanaafuu	osoo	yemmuu	ammo
illee	kee	otoo	yeroo	an
immoo	keenya	otumallee	ykn	ani
innaa	keenyaa	otuu	yommii	booda
inni	keeti	otuullee	yommuu	booddee
isaa	keetii	saniif	yoo	dura
isaan	koo	silaa	yookaan	eega
isee	kun	simmoo	yookiin	eegana
iseen	malee	sun	yookinimoo	eegasii
ishee	moo	ta`ullee	yoom	ennaa
isheen	nu	tahullee	garuu	erga
itumallee	nuti	tanaaf	Jechuu	fakkeenyaaf
ituu	nuyi	tanaafi	oggaa	fi
ituullee	odoo	tanaafuu	ut	fkn
Jechaan	ofii	tawullee		

Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university

_____
Feyisa Gemechu Shoga
June 2015

The thesis has been submitted for examination with our approval as university Advisors.

Ermias Abebe _____ _____