

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURLAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Hybrid Word Sense Disambiguation Approach for Afaan Oromo Words

By

YEHUWALASHET BEKELE TESEMA

JUNE 2016

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURLAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Hybrid Word Sense Disambiguation Approach for 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

YEHUWALASHET BEKELE TESEMA

JUNE 2016

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURLAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Hybrid Word Sense Disambiguation Approach for Afaan Oromo Words

By

YEHUWALASHET BEKELE TESEMA

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Chair person,	_____	_____
_____	Advisor,	_____	_____
_____	Examiner,	_____	_____
_____	Examiner,	_____	_____

Dedication

This thesis work is dedicated to my father Bekele Tesema, my mother Abebu Gemachu,my sweetheart Firaol Gizachew, my brother Deselegn Bekele and all my brothers and sisters who were able to reap the fruit of their own.

Acknowledgment

First of all I would like to thank Jesus Christ who helped me to succeed in all my life long learn. Next, I would like to thank my advisor Ato Ermias Abebe for his valuable assistance in providing his genuine, professional advice and encouragement goes even beyond the accomplishment of this study. He initiated me to do by giving precious comments on necessary points. My thanks go to him again, since it is difficult to mention his contribution to my achievements in words, it is better to say my heart has recorded it forever. Next, Words cannot express my deepest thanks to my Father Mr. Bekele Tesema and mother Abebu Gemechu, my Sweetheart Firaol Gizachew and all my brothers and sisters for their moral and financial support I received from them throughout my life of learning. My honest thanks equally go to my all friends for their help in providing me concerning background information. I would like to thanks Addis Ababa University, Department of Linguistic, PhD Candidate Ato Birhanu Legese and all my colleagues who have helped me on my progress; Adisu Bole and Getachew Emiru.

Lastly but not least, I would like to thanks to ORTO and Oromia Folklore Bureau for their permission to data collection and professional contribution and AAU for financial support.

TABLE OF CONTENT

Acknowledgment i

Table of content ii

List of table……vi

List of figures……vii

List of appendices…… viii

List of acrimony and abbreviations……..... x

ABSTRACT…… xi

CHAPTER ONE.....1

1. Introduction 1

1.1. Background 2

1.2. Statement of the Problem.....4

1.3. Research Questions 4

1.4. Objective of the Study.....5

1.4.1. General Objective 5

1.4.2. Specific Objectives 5

1.5. Scope of Study.....5

1.6. Methodology.....6

1.6.1. Literature Review 6

1.6.2. Procedures 6

1.6.3. Tools for development of the system 6

1.6.4. Techniques 7

1.6.5. Data/Corpus Preparation 7

1.7. Significance 8

1.8. Organization of the thesis 8

CHAPTER TWO.....9

2. Literature Review 9

2.1. Introduction.....9

2.2. Basic Concepts of Word Sense Disambiguation 9

2.3. Approaches to Word Sense Disambiguation 10

2.3.1. Supervised Word Sense Disambiguation 10
2.3.2. Unsupervised Word Sense Disambiguation 11
2.3.3. Bootstrapping Machine Learning Approach 12
2.3.4. Corpus - Based Approaches 12
2.3.5. Knowledge - Based Approaches 12
2.3.6. Hybrid Approach 13
2.4. Context Window Size 14
2.5. Vector Space Model 14
2.6. Clustering Algorithms 15
2.6.1. Hierarchical Agglomerative Algorithms 15
A. Single Link Clustering 16
B. Complete Link Clustering 16
C. Average Link Clustering: 17
2.6.2. Partitional Algorithms 17
A. K-means 17
B. Expectation Maximization Algorithm 19
2.6.3. Divisive Clustering 20
2.7. Rule Based Approach 21
2.9. Related Works 22
2.9.1. Related Works for local languages 25
2.9.1.1. Works on WSD for Afaan Oromo 25
2.9.1.2. Works on WSD for Amharic 27
2.9.2. Critique 28
2.9.3. Summary of related work 28
CHAPTER THREE 30
3. The Methodology 30
3.1. Introduction 30
3.2. Method 30
3.2.1. Data Collection and Acquisition 30
3.2.2. Training and Testing 30
3.2.3. Implementation Tools 32
3.3. Corpus Preparation 32

3.3.1. Corpus Acquisition and Preparation Techniques 32
3.3.2. Corpus Preprocessing 33
A. Tokenization 33
B. Stopword Removal 33
C. Normalization 33
3.4. The WSD Algorithms 33
3.4.1. The Machine Learning Approach 35
3.4.1.1. Extracting the Context Words 35
3.4.1.2. Clustering the Contexts 35
3.4.1.3. Vector Space Model 36
3.4.1.4. The Clustering Algorithms 37
3.5. The Hybrid Approach 37
3.5.1. Constructing Rule for Extracting Contexts 38
3.5.1.1. Modifiers 38
3.6. Evaluation Method 39
CHAPTER FOUR 41
4.0. Experimentation Results, Discussions and Evaluations 41
4.1. Architecture of the Afaan Oromo WSD 41
4.2. Data requirements 42
4.3. Steps in preprocessing 42
4.4. Experiment on the Machine Learning Approach using different Context Window Sizes and clustering algorithms 44
4.4.1. EM algorithm 45
4.4.2. K-Means algorithm 46
4.4.3. Complete link clustering algorithm 47
4.4.4. Single Link Clustering Algorithm 48
4.4.5. Average Linkage clustering algorithm 49
4.4.1. Summary of the Machine learning clustering algorithms 50
4.5. Experiment on the Hybrid Approach 51
4.5.1. Experimentation on EM algorithm 53
4.5.2. Experimentation on K-Means algorithm 54
4.5.3. Experimentation on complete link clustering algorithm 55

4.5.4. Experimentation on single link clustering algorithm 56
4.5.5. Experimentation on Average link clustering algorithm 57
4.5.6. Summary of the Hybrid Approach clustering algorithms 58
4.6. Discussion 60
4.6.1. Comparision of machine learning and hybrid approaches 60
4.6.2. Walk-through Using an Example 63
A. Machine Learning 63
B. Hybrid Approach 64
CHAPTER FIVE 68
5. Conclusions and Recommendations 68
5.1. Conclusion 68
5.2. Recommendations 68
References 70

List of Tables

Table1.1. Summary of Word Sense Disambiguation Approaches 22
Table 2.1. Summary of related works for Afaan Oromo 28
Table 2.3: Summary of related works on WSD for Afaan Oromo.....29
Table 2.4: Summary of related works on WSD Ahmaric 29
Table 3.1. Sample of Ambiguous Words for test 31
Table 4.1.Window size experimentation for EM algorithm 45
Table 4.2. Window size experimentation for K-Means algorithm.....46
Table 4.3. Window size experimentation for Complete link algorithm.....47
Table 4.4. Window size experimentation for Single link algorithm 48
Table 4.5. Window size experimentation for Average link algorithm.....49
Table 4.6. Summary of best experiments.....50
Table 4.7. Summary of Machine Learning Clustering Algorithms.....50
Table 4.8.: Window size experimentation for EM algorithm 53
Table 4.9: Window size experimentation for K-Means algorithm.....54
Table 4.10. Window size experimentation for complete Link clustering algorithm……55
Table 4.11: Window size experimentation for single Link clustering algorithm……56
Table 4.12: Window size experimentation for Average Linkage clustering algorithm……57
Table 4.13. Summary of best Accuracy of the Hybrid Approach ……58
Table 4.14. Summary of Hybrid Approach Clustering Algorithm……58
Table 4.15. Comparison of machine learning and hybrid approaches ……60
Table 4.16. Evaluation of WSD……62
Table 4.17. Extracted Context Using Window Size……63
Table 4.18. Extracted Modifiers for ambiguity word *bahe*……64
Table 4.19. Summary of the Test……67

List of figures

Figure 1. The Surrounding Contexts 10
Figure 2. a) Single Link, b) Complete Link and c) Average Link.....17
Fig.3. Hierarchical Clustering and Divisive Clustering 21
Figure 4.1. WSD Architecture 41
Figure 4.2. User Interface of the WSD.....43
Figure 4.3. Expectation Maximization and K-Means Clustering for word bahe.....64
Figure 4.4. EM (left side) and K-Means (right side)…….....65

List of Appendices

Appendix A. Question for Collection of Ambiguous Words…….76

List of Acronyms and Abbreviations

AI: Artificial Intelligence

CSV: Comma-Separated Values

EM: Expect Maximization

HAC: Hierarchical Agglomerative Clustering

IR: Information Retrieval

KBD: Knowledge-Based Disambiguation

ML: Machine Learning

MT: Machine Translation

MAX: Maximum

MIN: Minimum

NLP: Natural Language Processing

ORTO: Oromia Radio and Television Organization (ORTO)

POS: Part-of-Speech

QA: Question Answering

VSM: Vector Space Model

WSD: Word Sense Disambiguation

ABSTRACT

Word Sense Disambiguation (WSD) is a technique in the field of NLP where the main task is to find the appropriate sense in which ambiguous word occurs in a particular context. In this thesis we have developed the prototype that offer the related meaning of the ambiguous word based on the underlying contexts. It is found to be a vital to help applications such as Machine Translation, Text Summarization, Question and Answering and Information Retrieval.

The problem this research attempts to address is, due to ambiguity of word in Afaan oromo there is a problem in, retrieving documents, text preprocessing, document translation and grammar analysis**.**

Thus objective of this thesis is to design and test a hybrid system which finds the meaning of words based on surrounding contexts combining unsupervised with rule based approach. Hence, this work presents a WSD strategy which combines unsupervised approach that exploits sense in a corpus and the manually crafted rule using hybrid method. The idea behind the approach is to overcome the problem of a bottleneck for the machine learning approaches, while hybrid method can improve the accuracy and suitable when there is scarcity of training data. A fundamental problem with corpus-based approach is sparseness of the training contexts for ambiguous word for assigning appropriate senses. This makes our approach suitable for disambiguation of languages when there is lack of resource and sense definitions. In this work, the meaning and context of a given word is captured using term co-occurrences within a defined window size of words. We have conducted experiments to define the optimal window size in this research. We conclude that the window size for extracting semantic contexts is window 1 and 2 words to the right and left of the ambiguous word achieved better result. The similar contexts of a given senses of ambiguous word are clustered using hierarchical and partitional clustering. Each cluster representing a unique sense. Of the test set, ambiguous words have two senses to five senses. The partitional Clustering (EM and K-means) has yields significantly higher accuracy as compared to hierarchical clustering for context clustering. The achieved result was encouraging; despite it is less resource requirement. The system yield accuracy of 76.05% for the unsupervised and 89.47% hybrid approach respectively. Yet; further experiments using different approaches that extend this work are needed for a better performance.

Key words: *Word Sense Disambiguation, Afaan Oromo, Ambiguous Word, Disambiguation, Rule Based, Hybrid*.

CHAPTER ONE

1. Introduction

In today's world, many users use web for searching and reading documents and texts. Sometimes it was observed that the result of search is not appropriate. The reason behind this is because of ambiguity in the query words. To solve the problem of ambiguity, word sense disambiguation is introduced. WSD is an important task in Natural Language Processing. Word Sense Disambiguation is the process where the system has to find the exact meaning of query word. For application like machine translation the query word should give proper meaning, and then only one can say that the output is similar to the expected output [13].

One of the first problems that any natural language processing (NLP) system encounters is lexical, syntactic or semantic ambiguity. The problem of resolving semantic ambiguity is generally known as word sense disambiguation (WSD) and has been proved to be more difficult than syntactic disambiguation [25]. Human language is ambiguous; so that many words can be interpreted in multiple ways depending on the context in which they occur, the identification of the specific meaning that a word assumes in context is only apparently simple. Unfortunately, the identification of the specific meaning that a word assumes in context is only apparently simple. The other problem of WSD is that words often have more than one meaning, sometimes fairly similar and sometimes completely different. The meaning of a word in a particular usage can only be determined by examining its context. This is, in general, a trivial task for the human language processing system. We immediately recognize that in the sentence. However, the task has is proved to be difficult for computer or some have believed that it is unsolvable [1].

The objective of this study is to investigate hybrid approach for word sense disambiguation to Afaan Oromo words. We propose a method for word sense disambiguation which makes use of corpus-based evidence combined with background knowledge. Hybrid approaches can combine advantages from both strategies, potentially yielding accurate and comprehensive systems. Linguistic knowledge is nowadays available in electronic resources suitable for practical use. However, the use of this information has been hampered by the limitations of the modeling techniques that have been explored: using such sources of knowledge is beyond the capabilities of such techniques.

A major difficult problem in natural language processing is the complexity of human language. With so many different meanings for any given word, it is no wonder that automatic language processing is difficult [25]. This research investigated the mentioned problems using Hybrid algorithm.

1.1. Background

In natural language processing, word sense disambiguation (WSD) is the problem of determining which sense (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people [14]. WSD is a natural classification problem, given a word and its possible senses, as defined by an occurrence of the word in context into one or more of its sense. The features of the context (such as neighboring words) provide the evidence.

Natural languages have an ambiguous words which need to be disambiguated so that the appropriate sense of a word in a given context can be identified. A word can take several senses depending on the context in which it appears. The same word can take different meanings depending on the context. Identifying the correct sense of polysemous words seems very easy at first, but it is too difficult. WSD is used to identify the appropriate sense of the target word in a given context. The surrounding words of the target word in a sentence provide the context for the target word and this context provides consistent clue as regards the appropriate sense of the target word [26].

Word sense disambiguation is the task of automatically determining the correct sense of a word within a text. In any other language for that matter, a word may be used in a variety of contexts with a variety of meanings. Each of these meanings is called a word sense.

Accurately determining the correct sense of a word has been the subject of a great deal of research for several decades now [6].

The potential applications of word sense disambiguation are numerous and wide ranging. For example: A search engine may use it to determine what a user wants to search for more accurately. For instance someone who searches for 'Java programming' is probably not looking for results on coffee. Also, a program that translates text in one language to another can find the correct translation of a homonym. For instance Translating Seenaa from Afaan Oromo to English would be either History or Get in. A program that converts speech to text can use it to determine

the correct spelling of a word, where multiple spellings of the word exist, or properly interpret easy to-miss words. And also program making a concept map from a transcript of a meeting can use it to identify central topics more clearly.

Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia. Currently, it is an official language of Oromia state (which is the largest Regional State among the current Federal States in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 40% of the total population [22].With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991.

In Afaan Oromo most words have many possible meanings. A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human [3]. The task of word sense disambiguation is to examine word tokens in context and specify which sense of each word is being used. WSD is a difficult problem. The most difficulties arise from the fact that the concept of a meaning is vague. Usually, there are no clear boundaries between one sense and the other [26].

Natural languages are ambiguous. A word can take several senses depending on the context in which it appears. Let's consider following two sentences:

a. **Mirga** oromoota eeguu.

b. **Mirga** qabadhuu deemi.

In the first sentence, the word '**Mirga'** means 'respect human right'. In the second sentence, it means 'go having right direction'. These words sharing same spelling and pronunciation but have different senses. These multiple senses can be totally unrelated to each other. Thus, the same word can take different meanings depending on the context. Identifying the correct sense of polysemous words seems very easy at first, but it's not so trivial, if we think computationally. It involves processing unstructured information to extract the underlying meaning. So, we don't know which sense of the word "**mirga**" is expected in the sentence without looking at the context and the domain.

1.2. Statement of the Problem

Afaan Oromo is one of the major languages that are widely spoken and used in Ethiopia. So far Tesfa K. [21] and Feyisa G. [22] attempted to address the problem of WSD in Afaan Oromo. In Afaan Oromo, like any other natural language, there are words that have the same form, but have more than one meaning. These words are homonyms which has multiple distinct meanings [27]. Therefore, it is difficult to understand the meaning of those words in a given context.

Hence to have a clear understanding of a word in the language, WSD for Afaan Oromo is needed. Now days, there is a rapid development of technology. Like other languages Afaan Oromo has also started to use technology for different purposes. Many different Afaan Oromo documents are stored various repositories.

Due to ambiguity of words in Afaan Oromo there is a problems in, retrieving Afaan Oromo documents, translation of Afaan Oromo documents to other languages, text processing, and speech processing and grammar analysis. To overcome these problems, word sense disambiguation for Afaan Oromo language is needed.

In Afaan Oromo 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 [28]. The absence of WSD in Afaan Oromo natural language processing system limits future efforts of making computer to understand Afaan Oromo Language.

Therefore, the major concern of this study is to investigate hybrid approach for Afaan Oromo WSD. More specifically, this study aims to build the hybrid WSD model for Afaan Oromo words using hybrid approach. So, in this study we attempted to design and test a prototype for Afaan Oromo word sense disambiguation system which enables Afaan Oromo native speakers to access and retrieve the vast online information sources using their own (native) language.

1.3. Research questions

At the end of this research the following question is answered.

1. What are the most important clustering algorithms to improve the performance of WSD for Afaan Oromo?

2. Which window sizes are effective for Afaan Oromo?

3. Which approach is the most important to improve the performance of WSD for Afaan Oromo?

1.4. Objective of the Study

1.4.1. General Objective

The general objective of this study is to design and test a hybrid word sense disambiguation for Afaan Oromo words.

1.4.2. Specific Objectives

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

- To review related works so as to have a conceptual understanding of the area and identify the state-of-the-art in WSD;
- To build dataset ;
- To select WSD algorithms for Afaan Oromo words;
- To build the hybrid WSD model for Afaan Oromo words;
- To train and test the WSD model using hybrid approach;
- To evaluate the performance of the model; and
- Draw conclusions and suggest some recommendations for future work

1.5.Scope of Study

To deal with ambiguity problems in Afaan Oromo language, hybrid WSD technique (combining unsupervised machine learning and rule based approach) is proposed.

In Afaan Oromo the problem of ambiguity definitely happen at any level like phonology, morphology, syntax, semantic, and discourse. This study is limited to explore an ambiguity at word level. The study does not include word sense at phrase, phoneme, syntactic, and discourse levels. Because of lack of annotated corpus, parser and dictionary for Afaan Oromo; the study is limited to the experimentation of twenty ambiguous words. We deal only with data in textual format without any kind of grammar and spelling correction.

1.6.Methodology

Research methodology is a process used to collect information and data for the purpose of making decision. Methodology can be qualitative or quantitative.

The general methodology of the research is quantitative experimental. The steps of procedure we followed to overcome this methodology are presented below.

1.6.1. 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. We were reviewed works on WSD for both local and foreign language detail to understand tools and techniques that can be applied in our work.

1.6.2. 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 continue 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 normalization were conducted. By using selected clustering algorithms, we generated models and evaluate 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 was selected and finally completes by comparison of the result obtained with unsupervised algorithms and rule base approach.

1.6.3. Tools for development of the system

To perform this experiment, the algorithms were implemented in Java; NetBeans 8.0.2 which run on the prepared corpus and the clustering were performed in Weka 3.7.9 tool. Java package is a general-purpose and open source programming language. Moreover, it is optimized for software quality, developer productivity, program portability, and component integration. Nowadays Java is commonly used around the world for Internet scripting, systems programming, user interfaces, product customization, numeric programming, and more. It is generally considered to be among the top four or five most widely-used programming languages in the world.

The Weka 3.7.9 is the other tool used to implement WSD specifically context clustering, using the hierarchical clustering (which are single link, complete link and average link) and partitional

clustering (EM, K-means) algorithms built into the tool. The tool has by default Euclidian distance which is used to measure the distance between clusters.

1.6.4. Techniques

As our research uses hybrid techniques, we choose clustering algorithms from Weka tools. The algorithms used for this research are; 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.

1.6.5. Data/Corpus Preparation

In this study, hybrid learning approach was selected to develop a model. In this approach a significant number of sense examples are required to make training possible for the algorithms, which is difficult to get for Afaan Oromo. For the reason that there is no corpus for Afaan Oromo language, so it developed through the study. The corpuses used in this study were collected from Department of Afaan Oromo, Institute of Oromia Radio and Television, Bulletins and Newspapers. These were selected as sample to be representative and balanced with respect to particular factors. A corpus is said to be "representative of a language variety" if the content of the corpus can be generalized to that variety [24]. Basically, if the content of the corpus, defined by specifications of linguistic phenomena examined or studied, reflects that of the larger population from which it is taken, then we can say that it "represents that language variety." These are considered as consisting different issues of the users like social, economic, technological and political issues. This will reduce the probability of making the corpus biased toward some specific words that do not appear in everyday life [18].

Feyisa G.[22] used a total of 1501 Afaan Oromo sense examples for selected seven ambiguous words namely ***sanyii, karaa, horii, sirna and qoqhii, ulfina, ifa***. In this study a set of 62986 lines of sentences and 412,725 words were used. We add thirteen ambiguous words on seven ambiguous words used by Feyisa. Totally twenty ambiguous words were used for this thesis.

Generally this study investigated hybrid approach for word sense disambiguation to Afaan Oromo words by increasing corpus size used by Feyisa and using different approaches which is hybrid approach (combining unsupervised machine learning and rule based approach).

1.7.Significance of the study

The results of this study is expected to produce experimental evidences that demonstrate different application areas of hybrid technique to word sense disambiguation of Afaan Oromo texts. It also contributes to future researches and development in the area of natural language processing specifically in machine translation, Information retrieval & hypertext navigation, speech processing, text processing, information retrieval, grammatical analysis, content and thematic analysis as those areas require word sense disambiguation as complement. It is expected that, the result of this study may used by different stakeholders like speakers of Afaan Oromo and new language learners to identify proper word senses.

1.8.Organization of the Thesis

The thesis is organized into five chapters comprising Introduction, Literature review, Methodology, Experimentation and Evaluation, and Conclusion & Recommendations. The first chapter gives the general introduction of the study. The second chapter presents review made on different literatures regarding Word Sense Disambiguation together with its approaches and different machine learning techniques. Chapter three discusses the methodology and presents the process of data preparation for the system to be developed for Afan Oromo. It presents the algorithms and rule preparation based on the analysis of rule categories as reviewed. The fourth chapter discusses the experimentation and discusses of the findings. Chapter sixth presents the conclusion and the recommendation.

CHAPTER TWO

2. Literature Review

2.1. Introduction

In this chapter, related literature review in the Word Sense Disambiguation (WSD) is presented. This part of study enlightens briefly on some of the work done by those researchers.

2.2. Basic Concepts of Word Sense Disambiguation

Word sense is a definition or meaning of a word [4]. In many natural languages, each word may have more than one meaning that is why a single word sometimes can have many senses. Disambiguation means to remove all ambiguity. WSD is a technique to use in finding the meaning of a word in a sentence. A word can have multiple meanings and humans decide the exact meaning of the word based upon context. Computers also can follow similar technique i.e. decide meaning of an ambiguous word in a sentence using context information. Likewise, it is the process of identification to decide the appropriate meaning of an ambiguous word in a particular context. People decide the meaning of a word based on the characteristic senses of a discussion or situation using their own merits. Machines have no ability to decide such an ambiguous situation unless some practices have planted into the machines' memory [25].

Human extracts Word sense by connecting it to that specific context. Machine disambiguates ambiguity of an individual word that can be used (in different contexts) to express multiple meanings. The context is the only real way to recognize the significance of the homonym term. Generally, context can use in the bag of words and relational information. Here the bag of words is the context that considered as words in some window surrounding the ambiguous word in terms of distance. The surrounding word of the ambiguous word decides correct choice of word. Assume, the ambiguity word afaan has surrounded with the following contexts. Therefore, these contexts give us the clue of information what the meaning of this ambiguous word is. It seems that these ambiguous words identify its senses with the help of its contexts as shown the Figure below:

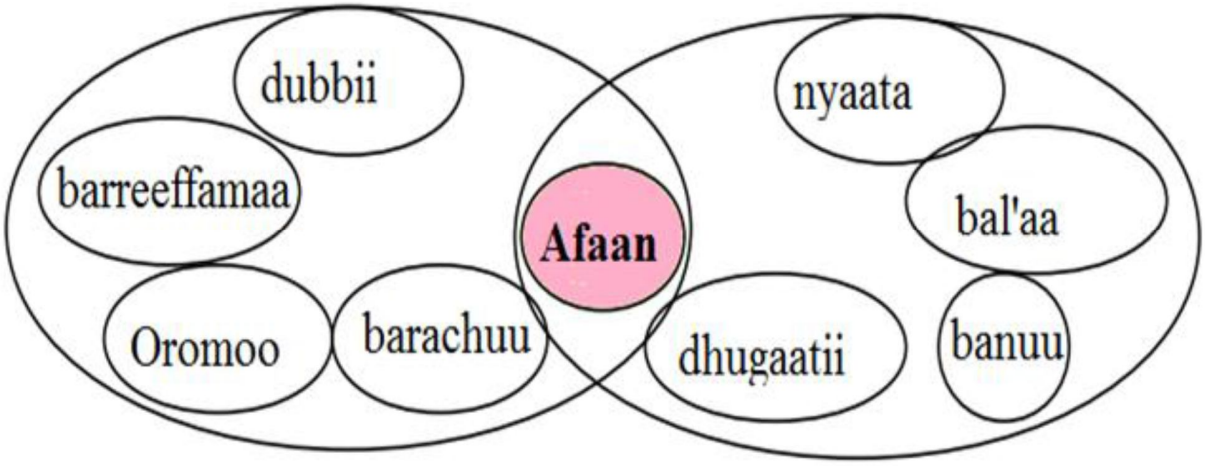


Figure 1. The Surrounding Contexts

2.3. Approaches to Word Sense Disambiguation

Regardless of the approach, WSD systems extract contextual information of an ambiguous word (in the corpus) and compare against the sense differentiation information stored in that word [40]. Many approaches have used through the evolution of WSD research.

Based on the level of supervision machine learning approaches categorized as; supervised machine learning, unsupervised machine learning and Bootstrapping machine learning approaches.

2.3.1. Supervised Word Sense Disambiguation

Supervised Word Sense Disambiguation use machine-learning techniques to learn a classifier from labeled training sets, that is, sets of examples encoded in terms of a number of features together with their appropriate sense label (or class [14]). The systems in the supervised learning approach category trained to develop a classifier that can used to assign a yet unseen example to one of a fixed number of senses. That means, there is trained corpus, where the system learns to classify and a test corpus that the system must annotate. So, supervised learning can considered as a classification task.

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

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

The main problem associated with supervised approach is the need for a large sense tagged training set. Despite the availability of large corpora in some language, manual sense tagging of a corpus is very difficult.

2.3.2. Unsupervised Word Sense Disambiguation

Unsupervised Word Sense Disambiguation methods are based on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word in context unlike supervised method [14]. Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck [15], that is, the lack of large-scale resources manually annotated with word senses. These approaches to WSD are based on the idea that the same sense of a word has similar neighboring words. They are able to induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters. They do not rely on labeled training text and, in their purest version, do not make use of any machine–readable resources like dictionaries, thesauri, ontology. However, the main disadvantage of fully unsupervised systems is that, as they do not exploit any dictionary, they cannot rely on a shared reference inventory of senses [17]. Most of the time, supervised approaches are superior to unsupervised in terms of accuracy of automatic disambiguation when used on the same type of texts that the systems trained on [2].

Nevertheless, there is another issue connected with the problem of the definition of a meaning, i.e., an issue of the creation of other resources used for automatic system performing WSD. This is especially evident in the creation of corpora that is manually annotated (tagged) with the senses ,which are used for training machine learning classifiers in a supervised setting. There are two important problems during manual sense tagging of a corpus: low inter annotator agreement (IA) and high cost of the annotation process. IA is a way of measuring how much an annotation assigned by one annotator differs from annotations assigned by another annotator. IA is used for the estimation of an upper bound on performance on automatic WSD but there is also another measure [19].

2.3.3. Bootstrapping Machine Learning Approach

The bootstrapping approach is situated between the supervised and unsupervised approach of WSD. The aim of bootstrapping is to build a sense classifier with little training data, and thus overcome the main problems of supervision: the data scarcity problem specially lack of annotated data. The bootstrapping methods use a small number of contexts labeled with senses having a high degree of confidence. This could be accomplished by hand tagging with senses the contexts of an ambiguous word w for which the sense of w is clear because some seed collocations [13] occur in these contexts. These labeled contexts are used as seeds to train an initial classifier. This is then used to extract a larger training set from the remaining untagged contexts. Repeating this process the number of training contexts grows and the number of untagged contexts reduces.

In addition to the above approaches, another criterion for categorization of approaches are use of knowledge (dictionary [25]).

2.3.4. Corpus - Based Approaches

A major challenge facing WSD research is the ability to acquire a large number of words with their different contexts. Corpus-based approaches came up with an alternate solution to the challenge by obtaining information necessary for WSD directly from textual data which is a corpus. A corpus provides a bank of samples, which enable the development of numerical language models, and thus the use of corpora goes hand-in-hand with empirical methods.

2.3.5. Knowledge - Based Approaches

Knowledge-based WSD is based on lexical resources like dictionaries, thesauri, and corpora where Machine-readable dictionaries (MRDs) are the primary source of acquisition of data. There are various Knowledge based approaches such as WSD uses selection Preferences, Lesk's algorithm, Walker's algorithm, WSD uses conceptual density and WSD uses Random Walk Algorithm. Rely on external knowledge resources (example, Word Net, Thesaurus), may use grammar rules for disambiguation and may use hand-coded rules for disambiguation.

Under this approach, disambiguation is carried out using information from an explicit lexicon or knowledge base. Since corpus based approaches require a considerable amount of work to create a classifier for each word in a language. Knowledge-based approaches use an explicit lexon like, MRD, thesauri, computational lexicons such as Word Net or (hand-crafted) knowledge bases as information source to resolve lexical ambiguities for many words.

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

Thesauri provide information about relationships among words, most notably synonymy [41]. Thesaurus based disambiguation makes use of the semantic categorization provided by a thesaurus or a dictionary with subject categories. The basic inference in thesaurus-based disambiguation is that semantic categories of the words in a context determine the semantic category of that context as a whole, and this category determines the correct senses that are used. Similar to machine readable dictionaries, a thesaurus is a resource for humans, so there is not enough information about word relations.

Computational Lexicons are a large electronic database containing useful lexical relations in linguistic Psycholinguistic and computational [3]. Lexicon like Word Net is used for sense evaluation and for similarity measure in WSD. For example [33] created a knowledge base from Word Net's hierarchy and apply a semantic similarity function to accomplish disambiguation, also for the purposes of information retrieval.

Finally, we can categorize WSD approaches as token based and type-based. Token based approaches associate a specific meaning with each occurrence of a word depending on the context in which it appears. In contrast, type-based disambiguation is based on the assumption that a word is consensually referred with the same sense within a single text. Consequently, these methods tend to infer a sense (called the predominant sense) for a word from the analysis of the entire text and possibly assign it to each occurrence within the text [25].

2.3.6. Hybrid Approach

The combinations of machine learning and rule based approaches have been used to make hybrid methods. A major weakness with unsupervised methods is the lack of ability to place a label on each discriminated word group. Combining an unsupervised method with a rule based method, particularly a machine readable dictionary, could overcome this weakness.

Since they obtain disambiguation information from both corpora and explicit knowledge bases, Hybrid Approaches do not fall into either knowledge or corpus-based. Hybrid systems aim to use

the strengths of the both conquering specific limitations associated with a particular approach, to improve WSD accuracy. They base both on a 'knowledge driven, corpus-supported' theme, utilizing as much information as possible from different sources. [11] used Bootstrapping approaches where initial data come from an explicit knowledge source which is then improved with information derived from corpora. He defines a small number of seed definitions for each of the senses of a word (the seeds can also be derived from dictionary definitions or lexicons such Word Net). Then the seed definitions are used to classify the obvious cases in a corpus.

Although above three approaches are very popular for WSD, they have weakness respectively. Semantic relations only cannot disambiguate any homonym word in any context. However, WordNet does not define the noun-verb relations. The supervised machine learning approach needs a great deal of training data that are expensive and sometimes hard to obtain. Therefore, a hybrid approach that uses both the machine and rule based methods to solve WSD problems. It is approaches that provides better results and make the results satisfy users' preferences.

2.4. Context Window Size

The context is the only means to identify the sense of a target word. All algorithms make use of the contexts to provide information for sense disambiguation. The context window size is defines the size of the window of context. A window size of N means that there will be a total of N words in the context window. If N is a (positive) even number, then there will be N word on the left and right side of the target word, where N is for example: 1, 2, 3, 4,5, etc. For example, if the window size is 5, then there will be 5 words on the left and right side of the target word. In order to disambiguate a given word, a wider context should be considered in the performance of the system to rise overall. However, a wider context implies more data and thus further features. The different methods and algorithms can benefit from the choice of the context size in a distinct way, which means that the optimal size of the context can depend on the used method and that the selection of the size of the context leads to a variation of the WSD output. [11] also claimed that the different types of ambiguity occurring in the data can be captured by a different size of the context.

2.5. Vector Space Model

Vector space Model is the most important model to obtain a vector that represents each word in the sentences. This vector is usually chosen to capture the contexts of the words can appear in. Each dimension corresponds to a separate word. If a word occurs with ambiguous word in the

corpus, its value in the vector is non-zero. All words in the corpus have its weight (the occurrence of word in the corpus); zero means the word simply doesn't exist in the corpus. Several different ways of computing these values as weights (term) have been used. One of the best known schemes is weighting (frequency occurrence of the word in the corpus [34]).
The dimensionality of the vector is the number of occurrence of word in the corpus. As all vectors under consideration, a cosine value of zero means that the ambiguous word and contexts are orthogonal and have no match (i.e. the word does not exist in the corpus). The context is formally a text that surrounds a language unit (e.g. a word) and helps to determine its interpretation. It represents the occurrences of target words as word vectors. From these vectors, context vectors are formed and meaning similarity is found that is a function of cosine between the contexts [34].

The computation of meaning similarity as operationalized by vector-based models has found widespread use in many tasks within NLP. The popularity of vector-based models lies in unsupervised nature and ease of computation. The model represents the meaning of each word as a point in a dimensional space, where each component corresponds to some co-occurring contextual [35]. The advantage of taking such a geometric approach is that the similarity of word meanings can be easily quantified frequency by measuring the cosine of the angle between them [36].

2.6. Clustering Algorithms

Clustering algorithms are categorized as hierarchical and partitional. The next section describes some common clustering algorithms. Here are general properties that characterize clustering algorithms [37].

Agglomerative vs Divisive: In agglomerative algorithms (bottom-up approach), each element is initially its own cluster and then the most similar clusters are iteratively merged until we are left with one large cluster containing all elements or until a stopping condition is met. Conversely, divisive algorithms (top-down approach) initially begin with a single all-encompassing cluster and iteratively split the clusters until each element belongs to its own cluster or until a stopping condition is met.

2.6.1. Hierarchical Agglomerative Algorithms

Hierarchical algorithms produce a nested partitioning of the data elements by merging or splitting clusters. Agglomerative algorithms iteratively merge clusters until an all-encompassing

cluster is formed [38], while divisive algorithms iteratively split clusters until each element belongs to its own cluster. The merge and split decisions are based on the similarity metric. The resulting decomposition (tree of clusters) is called a dendrogram.

The different versions of agglomerative clustering differ in how they compute cluster similarity. The most common versions of the agglomerative clustering algorithm are single-link, complete-link and average-link clustering. The complexity of these algorithms is O (n2logn [11]).

A. **Single Link Clustering**: The single link algorithm is a MIN version of the hierarchical agglomerative clustering method which is a bottom-up strategy: compare each point with each point.

Each context is placed in a separate cluster, and at each step merge the closest pair of clusters, until certain termination conditions are satisfied. For the single link, the distance of two clusters is defined as the minimum of the distance between any two points in the clusters. In single-link clustering the similarity between two clusters is the similarity between their most similar members (e.g. using the Euclidean distance [39]). It is capable of discovering clusters of varying shapes like the clusters of Figure 2.5 a. However, single-link is not practical because it suffers from the chaining effect [38]. For example, in Figure 2.5 a, single-link clustering generates an elongated cluster because of a bridge of elements connecting two clusters.

B. **Complete Link Clustering**: The complete linkage algorithm is the MAX version of the hierarchical agglomerative clustering method which is a bottom-up strategy: compare each point with each point. Each context is placed in a separate cluster, and at each step merge the farthest pair of clusters, until certain termination conditions are satisfied.

In complete-link clustering, the similarity between two clusters is the similarity between their maximum similar members (e.g. using the Euclidean distance [10]). Although complete-link clustering is not capable of discovering clusters like the two in Figure 2.5 b, it does not suffer from the chaining effect. Rather than producing straggly elongated clusters like single-link, complete-link generates compact clusters. Complete-link generates better clustering than single-link in many applications [40]. Figure 2.5a illustrates the cluster similarity of complete-link.

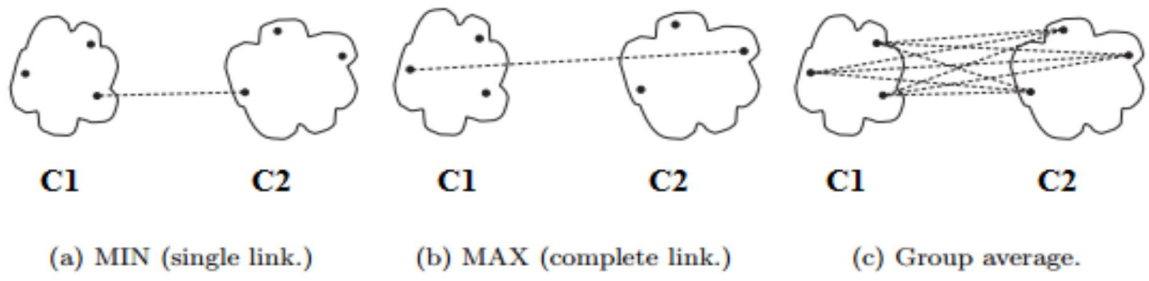


Figure 2. a) Single Link, b) Complete Link and c) Average Link

C. **Average Link Clustering:** Average-link clustering produces similar clusters to complete link clustering except that it is less susceptible to outliers [41]. It computes the similarity between two clusters, as the average similarity between all pairs of contexts across clusters (e.g. using the Euclidean distance). Figure 2.5 c shows merging decisions average linkage algorithms.

Generally, to use a hierarchical clustering procedure, need to express these distances in Weka tool. As an example, easily compute the Euclidean distance between cluster B and cluster C (generally referred to as d (B, C)) with regard to the two variables x and y by using the following formula, Weka by default has Euclidean distance.

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

2.6.2. Partitional Algorithms

Partitional algorithms do not produce a nested series of partitions. Instead, they generate a single partitioning, often of predefined size K, by optimizing some criterion. A combined search of all possible clustering's to find the optimal solution is clearly intractable. The algorithms are then typically run multiple times with different starting points. Partitional algorithms are not as versatile as hierarchical algorithms, but they often offer more efficient running time [41].

A. K-means

This algorithm has the objective of classifying a set of n contexts into k clusters, based on the closeness to the cluster centers. The closeness to cluster centers is measured by the use of a Euclidean distance algorithm. K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. A high degree of similarity among senses in clusters is obtained, while a high degree of dissimilarity among senses in different clusters achieved simultaneously [41].

Although the k-means algorithm often produces good results, it is not time-efficient and does not scale well. By saving distance information from iteration to the next, the actual number of distance calculations that must be made can be reduced. Some k-means variations examine ways to improve the chances of finding the global optimum. This often involves careful selection of initial clusters and means. Another variation is to allow clusters to be splited and merged. The variance within a cluster is examined and if it is too large, a cluster is split. Similarly, if the distance between two cluster centroids is less than a predefined threshold, they will be combined.
K-means clustering [32] is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results into a partitioning of the data space into Verona cells.
K-means [33] is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. Therefore, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

When no point is pending, the first step is completed and an early group age is done. At this point, we need to recalculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. Because of this loop, we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move any more. The algorithm is composed of the following steps:

a. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

b. Assign each object to the group that has the closest centroid.

c. When all objects have been assigned, recalculate the positions of the K centroids.

d. Repeat Steps b and c until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The most commonly used family of partitional algorithms is based on the K-means algorithm [41]. K-means clustering is often used on large data sets since its complexity is linear in K, the number of elements to be clustered. It creates a partitioning such that the intra-cluster similarity is high and the inter-cluster similarity is low. K-means using the concept of a centroid where a centroid represents the center of a cluster. A centroid is usually not an element of the cluster. Rather, a pseudo-element represents the center of all other elements. Often the mean of the feature vectors of the elements within a cluster is used as that cluster centroid. It is often difficult to define a centroid for categorical features.

K-means iteratively assigns each element to one of k clusters according to the centroid closest to it and recomputed the centroid of each cluster as the average of the cluster's elements. The following steps outline the algorithm for generating a set of k clusters:

1. Randomly select K elements as the initial centroids of the clusters;

2. Assign each element to a cluster according to the centroid closest to it;

3. Recomputed the centroid of each cluster as the average of the cluster's elements;

4. Repeat Steps 2-3 for T iterations or until a criterion converges, where T is a predetermined constant.

B. Expectation Maximization Algorithm

EM algorithm [42] is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. Expectation maximization (EM) algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM [43] iteration alternates between performing an expectation which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and maximization which computes parameters maximizing the expected log-likelihood found. These parameter-estimates are then used to determine the distribution of the latent variables.

Expectation maximization is a clustering algorithm that works based on partitioning methods. This algorithm is a memory efficient and easy to implement algorithm, with a profound probabilistic background. EM is widely used iterative algorithms for estimating model parameters [40].

Expectation maximization (EM) is a well-known algorithm used for clustering in the context of mixture models. This method estimates missing parameters of probabilistic models. Generally,

this is an optimization approach, which had given some initial approximation of the cluster parameters, iteratively performs two steps: first, the expectation step computes the values expected for the cluster probabilities, and second, the maximization step computes the distribution parameters and their likelihood given the data. It iterates until the parameters being optimized reach a fix point or until the log-likelihood function, which measures the quality of clustering, reaches its maximum [40].

There are two potential problems when using the EM algorithm. First, it is computationally expensive and convergence can be slow for problems with large numbers of model parameters. To solve the above problem we used small data set for this study. Second, if the likelihood function is very unbalanced it may always converge to a local maximum and not find the global maximum.

To simplify the discussion, we first briefly describe the EM algorithm. The algorithm is similar to the K-means procedure to that a set of parameters are re-computed until a desired convergence value is achieved. The parameters are re-computed until a desired convergence value is achieved. The finite mixture model assumes all attributes to be independent random variables [40].

A mixture is a set of N probability distributions where each distribution represents a cluster. An individual instance is assigned a probability that it would have a certain set of attribute values given it was a member of a specific cluster.

2.6.3. Divisive Clustering

Although it is not as common as agglomerative clustering. Divisive clustering algorithms start with a single cluster containing all elements. Considering all possible splits of the cluster into two clusters gives 2 (2n-1) – 1 possibilities. Using a splitting heuristic to iteratively split the largest cluster, Divisive clustering algorithms has worst case time complexity O (n2logn).

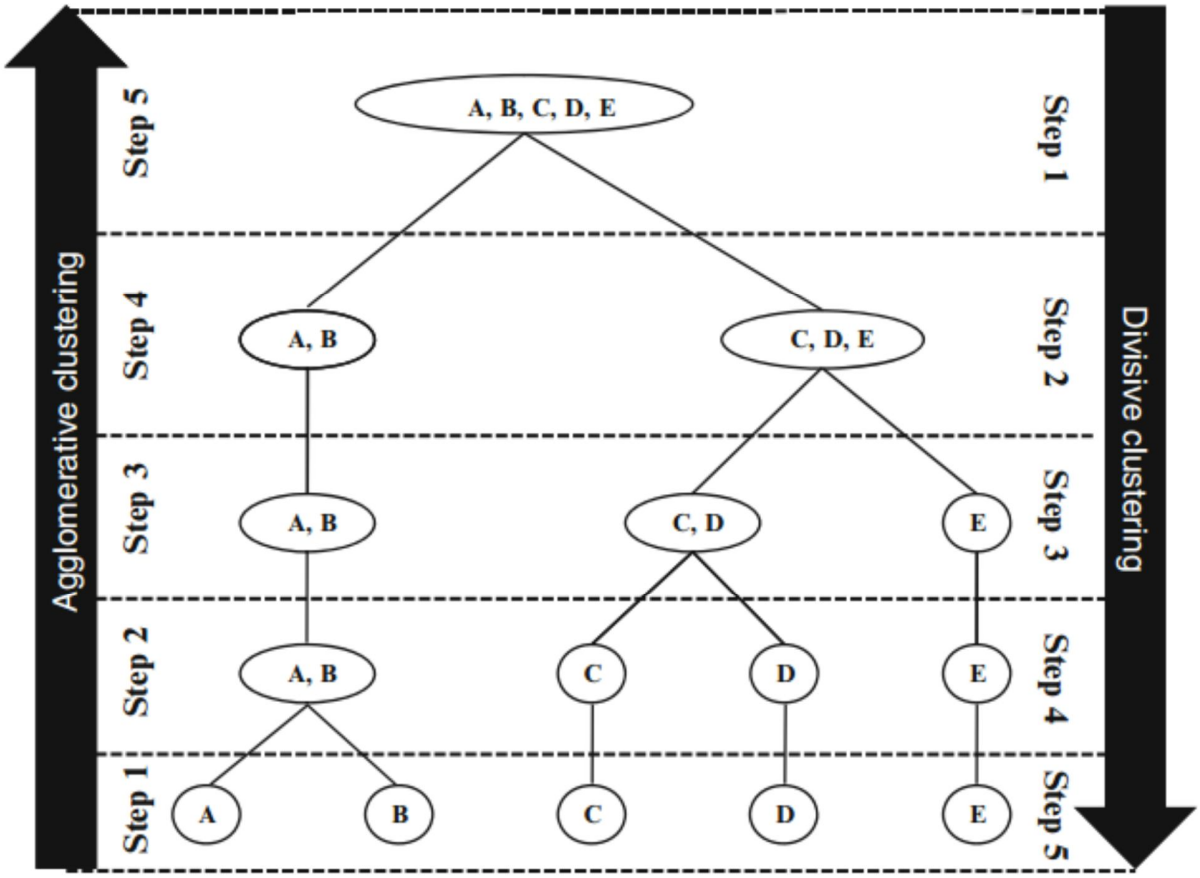


Fig.3. Hierarchical Clustering and Divisive Clustering

2.7. Rule Based Approach

The advantages of this approach are that it is easy to incorporate domain knowledge into the linguistic knowledge which provides highly accurate results. Furthermore, the linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system [36].

However, rule based is the most method used in NLP of disambiguation. When there is lack of available resources and their limitations, the rule based approach was used which is rely on hand-constructed linguistic rule in systems and resources. The use of rule based transformations are based on a core of solid linguistic knowledge [36].

Obviously, the manual creation of rule is an expensive and time consuming effort, which must be repeated every time the disambiguation scenario changes. Knowledge of the linguistic used for WSD are either lexical knowledge released to the public or world knowledge learned from a training corpus [37].

Algorithm Category	**Strengths**	**Weaknesses**
Knowledge Based	Accuracy	Rely on precompiled lexical knowledge resources
Supervised Approach	It needs pre-training (labeled data)	Dependent on pre-annotated corpora for training data
Unsupervised Approach	No pre-training necessary Works on multiple languages with no modification to the Algorithm	Merely discriminates between word senses; not disambiguate word senses
Rule Based	Accuracy	It takes time, needs expert and also it is very expensive
Hybrid Approach	Accuracy and it mix from others	-

Table1.1. Summary of Word Sense Disambiguation Approaches

2.9. Related Works

Alok [46] attempt to find meaning of words based on live contexts using supervised and unsupervised approaches. Unsupervised approaches use online dictionary for learning, and supervised approaches use manual learning sets. Hand tagged data are populated which might not be effective and sufficient for learning procedure. This limitation of information is the main flaw of the supervised approach.

They developed approach focuses to overcome the limitation using a learning set which is enriched in a dynamic way of maintaining new data. The trivial filtering method is utilized to

achieve appropriate training data. They introduce a mixed methodology having “Modified Lesk” approach and “Bag-of-Words” having enriched bags using learning methods. The approach establishes the superiority over individual “Modified Lesk” and “Bag-of-Words” approaches based on experimentation

[47] explored Word Sense Disambiguation using supervised, unsupervised, and knowledge-based approaches for WSD. The finding showed that supervised methods, undoubtedly perform better than the other approaches. However, relying on the availability of large training corpora for different domains, languages, and tasks is not a realistic assumption. Their finding also furnishes an idea of a few of the WSD algorithms and their performances, which compares and assesses the need of the word sense disambiguation.

According to [48] unsupervised WSD does not require much time in creating high quality resources and perform great in terms of accuracy. The other reason not to go for supervised approach is that, even if all possible resources are available to build a great supervised approach, it cannot be implemented in other languages easily. The resources have to be replicated for all possible languages. Another disadvantage of using the supervised approaches is, by using fixed sense repositories, the fixed number of senses present in that repository. It cannot discover new senses of words, which are not present in the sense corpus. Hence, only considering the accuracy of the approach is not a good idea, but considering its versatility and portability to other languages and domains is also equally important. These are the reasons that many unsupervised approaches were being tried by many researchers in WSD.

Manning and Schütze [50] describe word sense disambiguation difficulties that have several connotations or feelings. For such phrases provided of circumstance, there was hence ambiguity about how they are to be constructed. Writer clarify using instance of indecisiveness. They consider the word lender and two of the feelings that may found in Webster's New College Dictionary. Author describes several important methodological issues come up in the context of word sense disambiguation.

[49] state word sense disambiguation through a distributional semantic Model. They investigate that given a word and its context, Lesk algorithm exploits the idea of maximum number of shared words (maximum overlaps) between the context of a word and each definition of its senses (gloss) in order to select the proper meaning. The main finding of their approach relies on the use of a word similarity function defined on a distributional semantic space to compute the

gloss-context overlap. And conclude that an unsupervised WSD approach which selects the best sense according to the distributional similarity with respect to the context. In particular, the similarity is computed representing both the gloss and the context as vectors in a geometric space generated by a distributional semantic model based on LSA.

[51] stated that lexical meaning of words using vector spaces and linear algebra. He argued that the meanings of words will be represented using vectors, as part of a high-dimensional "semantic space". The fine-grained structure of this space is provided by considering the contexts in which words occur in large corpora of text. Words can easily be compared for similarity in the vector space, using any of the standard similarity or distance measures available from linear algebra, for example the cosine of the angle between two vectors. And he said that the hypothesis underlying distributional models of word meanings is the so-called distributional hypothesis: the idea that "Words that occur in similar contexts tend to have similar meanings". [51]Now the basis vectors correspond to whole sequences of grammatical relations, relating the target word and context word. Which paths to choose is a parameter of the approach, with the idea that some paths will be more informative than others.

[52]stated WSD in Nepali using Nepali WordNet. These approaches are overlap based approach and conceptual distance and semantic graph based approach which falls under Knowledge based approach. Conceptual distance and semantic graph distance are used as a measure to score WSD algorithm. They address that Word sense disambiguation (WSD) resolves the ambiguity by finding out which particular sense of a word is appropriate in a given context. WSD is of critical importance in the areas of machine translation, information retrieval, speech processing. The finding shows that the performance of overlap based approach is less than the combination of conceptual distance and semantic graph method. It is expected because overlap based approach suffer from sparse overlap. Nevertheless, especially in the case of nouns, the overlap based approach presented here gives better performance than the overlap based approach with machine readable dictionaries because not only the gloss and examples of the target and context synsets are taken but also the gloss and examples from their hypernyms have been taken into consideration. The adjective accuracy is more with the second method as the semantic graph distance score has been taken into account.

2.9.1. Related Works for local languages

Getachew *M*. [3] in his study, he has been identified the most well-known disambiguation problems is POS tagging. This means POS tagging assigns whether a given word is used as a noun, adjective and verb. In his work considered that Part-of-speech (POS) tagging is the act of assigning each word in sentences a tag that describes how that word is used in the sentences.

Daniel B. [2] in his study, development that of a corpus-based CLIR system which makes use of word based query translation for Afaan Oromo-English language. He was study that, Cross Language Information Retrieval (CLIR) is to provide users with access to information that is in a different language from their queries especially from Afaan Oromo-English language.

One of the study by Debela T.[18] focused on stemming for Afaan Oromo. This work attempts to make the search engines more effective in information retrieval was the usage of word stemming. This work takes as input a word and removes its affixes according to a rule based algorithm. It is the grammar checker for Afaan Oromo. Their work was focus on spelling checking for Afaan Oromo.

2.9.1.1.Works on WSD for Afaan Oromo

Tesfa K. [21] tried to solve the problem of word sense disambiguation (WSD) for Afaan Oromo language using a corpus based approach. He applied Naïve Baye's theory to find the prior probability and likelihood ratio of the sense in the given context for his experimentation. The system uses information gathered from training corpus to assign senses to unseen examples. The corpus he used contains 1240 sentences, and he evaluated for 5 Afaan Oromo ambiguous words namely ***sirna, karaa, sanyii, qophii*** **and** ***horii.***

By using these words he conducted two experiments. The first experiment was conducted to evaluate the performance of the algorithm; using 10-fold cross-validation. In this technique, first the total data set is divided into 10 mutually disjoint folds approximately of equal size using stratified sampling mechanism. Second, the training set and testing set was identified and separated from the total data set. In order to check the result using the developed system, he removed manually tagged sense examples from test set. Before doing the actual experiment, 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. 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.

Fayisa G. [22] presented a corpus based approach to disambiguation is employed where unsupervised machine learning techniques are applied to a corpus of Afaan Oromo language. He 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.

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 matrix. Out of five clustering 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%.

2.9.1.2. **Works on WSD for Amharic**

Solomon M. [19] used corpus based approach to 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. He used five Amharic words as test cases and he translated word senses from a total of 1045 English sense examples collected from British National Corpus (BNC) by using dictionary. 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 evaluated the classifiers for the five ambiguous words and achieved accuracy within the range of 70% to 83% which is very encouraging. 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. Finally he recommended that further experiments for other ambiguous words and using different approaches needs to be conducted.

Solomon A. [20] 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.[19]. 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.

2.9.2. **Critique**

It is obvious that the problems of word sense disambiguation can be solved through different approaches and techniques. As listed above different works has been done to solve the problems of word sense disambiguation for Afaan Oromo.

The corpora used by Tesfa [21] and Feyisa [22] contain 1240 and 1500 sentences respectively and they evaluated for 5 and 7 Afaan Oromo ambiguous words respectively. Feyisa added only 500 sentences and 2 words on Tesfa's corpus, which is a small size. The size of the corpus is an important factor in the quality of the disambiguated senses, with the general message that more data is better data to get better accuracy [29]. However, both researchers used small size of corpus.

Both researchers used corpus based approach for Afaan Oromo word sense disambiguation, but accuracy of word sense disambiguation is better if hybrid approach is used [23]. According to [23] Rule based transformation is based on solid core of linguistic knowledge. The rules are hand crafted rules, so it is more accurate. To overcome these gaps we proposed hybrid approach for Afaan Oromo word sense disambiguation.

2.9.3. **Summary of related work**

No	**Title**	**Author**	**Method/approach**	**Result**
1	Parts of Speech Tagging for Afaan Oromo	Getachew Mamo	Hidden Markov Model (HMM)	Bigram and unigram algorithms show an accuracy of 91.97% and 87.5% correctly tagged words in average respectively.
2	Afaan Oromo-English Cross-Lingual Information Retrieval (CLIR)	Daniel Bekele	A Corpus Based Approach	A maximum average precision of 0.468 and 0.316 for Afaan Oromo and English was obtained respectively
	Designing a Rule Based Stemmer for Afaan Oromo Text	Debela Tesfaye	Rule based	Overall accuracy of about 94.84% and error rate is about 4.27%.

Table 2.1. Summary of related works for Afaan Oromo

No	Title	Author	Method/approach	Result
1	Word Sense Disambiguation For Afaan Oromo Language	Tesfa K.	Corpus Based	He achieved 79% accuracy
2	Unsupervised Corpus Based Approach for Word Sense Disambiguation to Afaan Oromo Words	Feyisa Gemechu	A Corpus Based /Unsupervised	Average Accuracy obtained by different clustering algorithms is, with K Means 81.9%,with EM 78.9% and with Complete Linkage 76.1%.

Table 2.3: Summary of related works on WSD for Afaan Oromo

No	Title	Author	Method/approach	Result
1	Word Sense Disambiguation for Amharic Text	Solomon Mamo	corpus based/supervised	Achieved accuracy within the range of 70% to 83%
2	Unsupervised Machine Learning Approach for Word Sense disambiguation To Amharic Words	Solomon A.	A Corpus Based/Unsupervised	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.

Table 2.4: Summary of related works on WSD Ahmaric

CHAPTER THREE

3. The Methodology

3.1. Introduction

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

3.2.Method

There are many algorithms to disambiguate multi sense words. This study relies on the patterns learned from the corpus (unsupervised approach) in combination with the rule crafted manually. A fundamental problem with such corpus-based approach is scarcity of the training contexts, for ambiguous word has to be assigned their appropriate senses by rule. The idea of this research is to combine both the rule based and unsupervised machine learning approaches into a hybrid approach which relies on the patterns learned from the corpus in combination with the rule based approach. The reason why the hybrid approach was used is that it depends on the available and more reliable linguistic knowledge besides the semantic techniques and training methods which consider the role of the words in its context.

3.2.1. Data Collection and Acquisition

As there are no standardized corpora in the area of word sense disambiguation for Afaan Oromo, we have compiled our own data collected from different sources. The sources include newspapers, Bible, news (Oromia Radio and Television Organization (ORTO)), magazines, historical documents and bulletins. The reason to widen the coverage of the source is to make it relatively balanced, representative of a language variety and avoid bias.

3.2.2. Training and Testing

The WSD system was trained on unannotated corpus, which constitutes ambiguous words. The work was tested by the most frequently 20 ambiguous words which are publicly used in the society. In these work thirteen ambiguous words were added on seven ambiguous words used by Feyisa [22].

The lists of ambiguous words are presented in table 3.1.

No	Ambiguous Words	Possible Senses	Defined Number of Senses
1.	Sanyii	Seed / Type	2
2.	Karaa	Road / Way	2
3.	Ulfina	Weight / Respection	2
4.	Ifa	Light / Clear	2
5.	Qophii	Program / Preparation	2
6.	Sirna	Event / Systems	2
7.	Horii	Money / Cattle	2
8.	Afaan	Language / Mouth	2
9.	Bahe	Freedom / Highland / Cloth / Witness / Dead(pass)	5
10.	Boqote	Break (rest) / Died	2
11.	Darbe	Cross/ Pass from class to class / Died / broadcast	3
12.	Diige	Fence / Absence on Meeting / cancel to start new	3
13.	Dubbatate	Struggle / Wedding	2
14.	Tume	Make / Hit / Contraceptive	3
15.	Haare	Sad / Burn	2
16.	Ija	Eye/ Tree Fruit / Vengeance/ Wide-eyed/ a little	5
17.	Ji'a	Stars / Month	2
18.	Dhahe	Follow / Hit /Fail	2
19.	Mirga	Direction / Human right / Brave	3
20.	Waraabuu	Hyena / Fetch / Record	3

Table 3.1. Sample of Ambiguous Words for test

3.2.3. Implementation Tools

In building the WSD model, the study was used hybrid approach which is found in the existing implementation Java NetBeans 8.0.2 and weka 3.7.9 package. Java NetBeans 8.0.2 and weka 3.7.9 tools were selected due to the familiarity of the researcher to the tool and because of its accessibility, processing capability and language independent features.

In our work Weka 3.7.9 tool also is used to cluster, based on the hierarchical clustering, EM and K-means algorithms built into the package. The package has by default Euclidian distance which is used to measure the distance between clusters and Java NetBeans 8.0.2 programming language is used for preprocessing activities like tokenization, stop word removal and normalization purpose to make sense example sentences ready for experimentation.

3.3. Corpus Preparation

The size of the corpus is an important factor in the quality of the disambiguated senses, with the general message that more data is better data [61]. However, it is very difficult to obtain data for under resourced language like Afaan Oromo.

3.3.1. Corpus Acquisition and Preparation Techniques

Afaan Oromo has been one of the under-resourced languages in terms of electronic resources. The natural language ambiguities become challenging problems due to scarcity of resources in the language [64]. The lack of resource has led the researcher to explore the use of unannotated raw corpora to perform hybrid disambiguation. It should be noted that unsupervised disambiguation cannot actually label specific terms as a referring to a specific concept that would require more information than is available. Unsupervised disambiguation can achieve clusters the instances of a word into distinct categories without giving categories labels from a dictionary.

In this work corpus from various sources was used.

The corpus was collected from newspapers (Bariisaa, Kallacha Oromiya and Oromiya. Bariisaa is a weekly newspaper, whereas the rest two come out once in two weeks), linguistics, ORTO news, government and official websites.

Moreover, Oromia Radio and Television Organization release daily news through radio and television broadcast and on its official website. To reduce the data sparsity the researcher used data from these sources as well.

See: http://www.gada.com
See: http://www.orto.gov.et
See: http://www.wikipedia/oromo.com

3.3.2. Corpus Preprocessing

A. Tokenization

The corpus which is a set of sentences first tokenized into words. Since, Afaan Oromo uses Latin alphabet the sentences can split using similar word boundary detection techniques like the use white space in English. In this work tokenization is needed for some purposes that are to select contexts, stop word removal and for further steps. For instance if we have the sentence *akkanatti hin haftu ni dabarti* in our corpus. We have tokenized into set of words on the white space, like *akkanatti, hin, haftu, ni* and *dabarti.*

B. Stopword Removal

After tokenization take place, we have removed Afaan Oromo stop words, hence it has no effect on meaning of the words. In this work, Stop word removal is used to remove stop words from the selected contexts because the absence or presence of these words has no contribution to identify appropriate sense. Not all tokenized words are necessary for this work hence, one word carry the meaning than other words and other words that have no own meaning. For instance, words such as (‘*as’*, ‘*achi’*), conjunctions (“*’fi’*, ‘*akkasumas’*, *‘kana malees’*). Since stop words do not have significant discriminating powers in the meaning of ambiguous words; we filtered stop-words list to ensure only content bearing words are included.

C. Normalization

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

3.4. TheWSD Algorithms

In our approach, two important features need to be extracted: the first one is determining all possible contexts (the candidate sense words) of the ambiguous words and the other one is to group these various contexts (senses) of the word, each group representing a specific sense of the ambiguous word. To this end, we developed two kinds of approaches towards the word sense disambiguation. The first approach is completely machine learning in its nature. The machine

learning approach extracts the two important features (the various contexts of the ambiguous words and their clustering) automatically without supplying linguistic rule (Section 3.4.1.2). The second approach combined machine learning algorithm and rule based (background linguistic). Such algorithm is called hybrid approach, which is a mixture of rule based (background knowledge) and corpus-based approaches applied. In this latter approach we combined the rule learned from a linguistic feature of Afaan Oromo with the semantic feature learned from corpus using machine learning approaches. Additionally, the manually crafted linguistic rule to extract the various contexts assumed by the ambiguous word used. We then relied on the machine learning algorithm to cluster these contexts as described in (Section 3.4.1.4). Hence, this latter approach also extracts all possible contexts assumed by the ambiguous word employing knowledge-based approaches, which make use of linguistic knowledge, whereas the corpus-based approaches use information acquired from the corpus; and the hybrid approach merges characteristics from both approaches.

The machine learning approach followed towards the WSD is basically unsupervised and mainly used for clustering the possible contexts for giving words that express the same meaning. Hence we didn't provide explicit sense labels for each group as the machine learning approach is unsupervised. Yet, small list of ambiguous words is required to test the algorithm.

3.4.1. The Machine Learning Approach

As discussed in the above (Section 3.4), two important features need to be extracted. We have described the procedures followed to extract the features in the subsequent Sections.

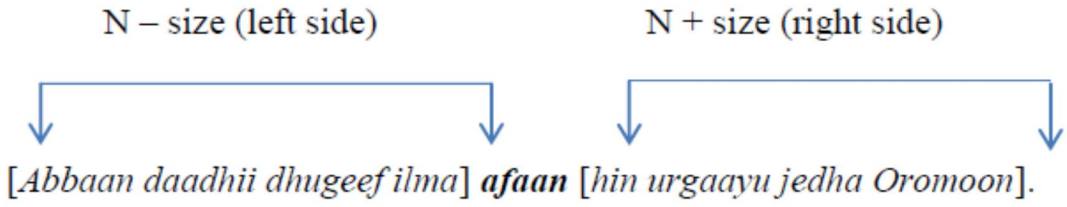
3.4.1.1. Extracting the Context Words

In the machine learning, the surrounding contexts actually selected by sliding window sizes. This machine learning approach extracts semantic features employing a vector space model. The vector space model relied on the intuition that meanings depend to some extent on a word's neighborhood. Words occurring in similar contexts tend to have similar meanings. This idea, known as the 'distributional hypothesis', has been proposed by various scholars. It implies that word meanings are context sensitive. A word's meaning cannot be fully grasped unless one takes the context into account. Meaning and context can be captured in terms of (more or less direct) neighborhood, i.e. words co-occurring within a defined window.

Most WSD algorithms make use of the contexts to provide information for sense disambiguation. For instance, [66] claimed that the different types of ambiguity occurring in the data can be captured by a different window size of the context. The reason why the researcher uses the

algorithm is that in the case of context window size, there is a probability to varying the size of the window if the surrounding context is not enough to predict the sense [64]. Window size avoids bias since it used varying sizes of context in the corpus. The context window size defines the size of the window of context. A window size of N means that there will be a total of N words in the context window. If N is a (positive) number, then there will be N word on the left and right side of the target word. For example, if the window size is 5, then there will be 5 words on the left side of the target word and 5 on the right. In order to disambiguate a given word, a wider context should be considered in the performance of the system to rise overall. However, a wider context implies more data and thus further features, which as a whole closes the circle of an endless loop over the trade-off between the amount and informative of the used data.

As already mentioned in this approach, we followed automatic procedure to extract all possible contexts assumed by the ambiguous word. To this end, we used the sliding window of words to extract words appearing n-words (n represent the size of window of words) to the left and the right of the ambiguous word. We have tried several window sizes in this research to propose the optimal window size for Afaan Oromo. Basically, the words that occur in similar contexts tend to have similar meanings. This is the main idea of the study, which is context based meaning of words. We have determined a set of contexts which are the most frequent words in the corpus with target words, by determining a window size contexts to the left and to the right.



3.4.1.2. Clustering the Contexts

For each context extracted in Section 3.4.1.1 above, construct a vector space matrix from co-occurrences. After the co-occurrence matrix, the cosine similarity was computed based on the angle between vectors of the contexts.

Finally, these cosine values, which are computed, are clustered by Weka 3.7.9 package. The cosine value, which is in .CSV (Comma Separated Value) file format, was entered into the Weka tool. The Weka tool clustered the contexts based on cosine value. Each cluster represents a

unique sense. To this end, we used hierarchical clustering, K-means and EM algorithms to merge the contexts based on the nature of the clustering algorithms using cosine similarity. The vector space model and the clustering algorithms are described in Section 3.4.1.3 and 3.4.1.4 respectively.

3.4.1.3. Vector Space Model

The vector space model is one of the algorithms used in the study. In the word senses, the words that occur in similar contexts tend to have similar meanings; this is the justification for applying the VSM to measuring word meaning similarity. For the given pair of contexts the algorithms extract contexts co-occurring with contexts in a predefined window. While in the experiments the researcher interested to identify the optimal size tried on several words in the sentences containing the target word, for instance, three, four, five, six, terms preceding and following the target word. After extracting the term co-occurrences the researcher provides weight based on their frequency. We did the same for the other pair of the target words. In order to identify the similarity between the target words the cosine value of the vectors computed. The cosine value of the vectors is then considered as the semantic similarity between the terms [23].

The context can be represented by a vector in which the contexts are derived from the occurrences of the ambiguous word in various contexts, such as windows of words. To illustrate this with a simple example: consider a co-occurrence matrix populated by simple frequency counting: if the context i co-occurs 5 times with context j in the corpus, we have used 5 in the fij in the contexts to compute the cosine similarity by co-occurrence matrix. The co-occurrences are normally counted within a context window spanning some number of words. The vector representation does not consider the ordering of the words in the corpus. It uses the angle between vectors instead of distance. The representation for a word is a point in a high-dimensional space. The dimensions stand for context items (for example, co-occurring words), and the coordinates depend on the co-occurrence counts. The co-occurrence of context with context words based on the frequently co-occur [67].

$$sim_{COS}(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|} = \frac{\sum_{k=1}^{n} x_i y_i}{\sqrt{\sum_{k=1}^{n} x_i^2}\sqrt{\sum_{k=1}^{n} y_i^2}}$$

3.4.1.4. The Clustering Algorithms

As already mentioned, the clustered contexts of the ambiguous words using their similarity values produced (in Section 3.4.1.2).

The clustering algorithms used in this study are hierarchical agglomerative clustering, which include single link, complete link, average link and partitional clustering include EM and K-means clustering (in Section 2.6). These clustering algorithms have their own unique nature of clustering. The hierarchical clustering begins by assuming that each context of a target word forms its own cluster (and therefore represents a unique sense). Then, it merges the contexts that have the minimum dissimilarity between them (and are therefore most alike). The partitional clustering algorithms have applied clustering as its own nature [68]. Likewise, the minimum specified cutoff the number of clusters is taken. In this case, the minimum specified cutoff of the number of clusters is two (vertical distance of dendrogram) since one ambiguity word has at least two senses.

In this method each occurrence of a context in a corpus is represented as a vector. The vector cosine is then clustered into groups, each identifying a sense of the target word. It is based on the idea of a vector space whose dimensions are words. A word in a context can be represented as a vector whose component counts the number of times that word co-occurs with word within a fixed context (a sentence or a larger context). The sense group can be performed by grouping the context vectors of a target word using a clustering algorithm. This algorithm group context of target word which groups the occurrences of an ambiguous word into closest clusters of senses, based on the contextual similarity (vector space model) between occurrences of contexts. All the senses in a cluster are contextually similar to each other, making it more likely that the given target word has been used with the same meaning in all of those senses [30].

3.5. The Hybrid Approach

According to [66] WSD heavily relies on knowledge of machine learning and linguistics. In fact, the skeletal procedure of any WSD system can be given a set of words (a sentence or a bag of words), a technique is applied, which makes use of one or more sources of knowledge to associate the most appropriate senses in context. In addition to the issue of machine learning, it needs a mechanism to determine the meaning of a word as it is used in the current context. Of course, human speakers exploit all sorts of contextual clues that are unavailable to computer systems, such as background knowledge and visual information of the situation.

The hybrid approach in this work, constitute the hand crafted rule to extract the contexts of the ambiguous word followed by the machine learning approach to cluster the contexts as described in the following Sections.

3.5.1. Constructing Rule for Extracting Contexts

The linguistic knowledge of the language plays an important role to create the rule. The knowledge required for the NLP can be obtained in different ways. In this study, the way of the word meaning obtained from Afaan Oromo was used as a rule. However, an effort has been to develop the rule of the language because of the linguistic knowledge plays a great role in developing an efficient system [69].

The most common way of representing the language is a rule [63]. The rule captures generalizations to identify word meaning. The rule underlies many linguistic theories, which in turn provide the bases for many natural language understanding systems [57]. The rule and contextual information were the basis of the linguistic properties of Afaan Oromo word meaning categories. In case of this study, the modifier is the way of deciding word meaning which the rule was developed. The modifiers where the word or phrase which provide the information about the word and give more description in words it precede. The modifiers can be single words or phrases, which establish understanding for the reader and important that modifiers refer clearly to the words they modify.

The meaning of words fundamentally based on the words preceded by the word it modifies [57]. The words are described (modified) by the Noun or Verb preceding them. However, since there is no annotated corpus of this language we didn't identify whether the modifiers are Noun or Verb. The word may appear in the middle or the end of a sentence, but it always becomes before the word it modifies. As an example, "*Seenaan kaleessa daara bahe*." [Seenaa got cloth yesterday]. From the construction of this sentence, the word "*bahe*" is modified by the Noun "*daara*". According to the rule in Afaan Oromo, is that the Noun and Verb always mandatory to go before the word its meaning decided [13].

3.5.1.1. Modifiers

In Afaan Oromo modifiers have a great role to decide a word's meaning. In Afaan Oromo modifiers can happen before the target word (the word it modifies or describes). The sentences would be pretty boring without modifiers to provide excitement and intrigue. A modifier adds detail or limits or changes the meaning of another word or phrase. Possible to identify a modifier

by its function in the sentence is it provides information, adding detail or describing something else. Modifiers usually have to accompany the thing they are modifying or go as close to it as possible [58].

In Afaan Oromo, the words preceding a specific word are more likely to influence the meaning of a word.



For example, [*Inni qabsoo hadhaan bilisa bahe*]. He got freedom by strong struggle

In this example, the core of the sentence is "[*Inni qabsoo hadhaan bilisa bahe*] he got freedom by strong struggle". The word "[*bilisa] freedom*" is a modifier; it gives extra information that is part of the sentence. In this case, it is a verb modifier, because it is modifying the verb "*bahe*". A modifier should be placed next to the word it describes.

3.6. Evaluation Method

To our knowledge, there were no previous standard Afaan Oromo word sense disambiguation dataset for evaluation as presented in Section 4.5. For this reason we did not evaluate against the other systems. In this work, the evaluations were undertaken on the basis of precision and recall. Precision is defined as the percentage of correctly disambiguated words out the total of disambiguated words. Recall is defined as the percentage of correctly disambiguated words out of the total number of ambiguous words [57].

$$\text{Precision}(\%) = \frac{\#\text{ correctly disambiguated words}}{\#\text{ disambiguated words}}$$

$$\text{Recall}(\%) = \frac{\#\text{ correctly disambiguated words}}{\#\text{ total number of ambiguous words}}$$

On the other hand, the clustering algorithms were evaluated comparing the result produced by the clustering algorithm with the manually grouped similar contexts of the ambiguous words in the test set by experts. The evaluation constitutes the following two points:

1. To evaluate how much the produced clusters are comply with the clusters prepared by human experts as a benchmark.

In order to achieve this we used the following criteria:

- How many of the clustered contexts are correct, i.e. to evaluate if all the similar contexts of the ambiguous words are placed in the same group.

2. Given the number of senses assumed by the ambiguous words in the test, judge the system on the basis of the number of senses identified by the system.

Similarly, in order to achieve these following steps performed:

a. Start with a small list of ambiguous words in the test with known number of senses N.

b. Run the algorithm on the test to identify the possible senses based on it's the number of clusters of the context as extracted from the big corpus

c. Count the number of clusters

d. Compare it against the already prepared sense clusters by experts.

CHAPTER FOUR

4.0. Experimentation Results, Discussions and Evaluations

4.1. Architecture of the Afaan Oromo WSD

In order to develop disambiguation model for Afaan Oromo, we followed a three steps process which involves: (a) Text preprocessing which take inputs and corpus, tokenize to remove stop words and perform normalization. Preprocessing is done offline which means it is completed in advance.

(b) Extracting context terms providing clue about the senses of the ambiguous term using two techniques (window size and rules),

(c) Clustering to group similar context terms of the given ambiguous terms, the number of clusters representing the number of senses encoded by the ambiguous term. In order to cluster similar context terms we computed the degree of similarity using the vectors constructed from co-occurrence information. The architecture (pipeline) of the system with the underlying steps is presented in figure 4.1 below:

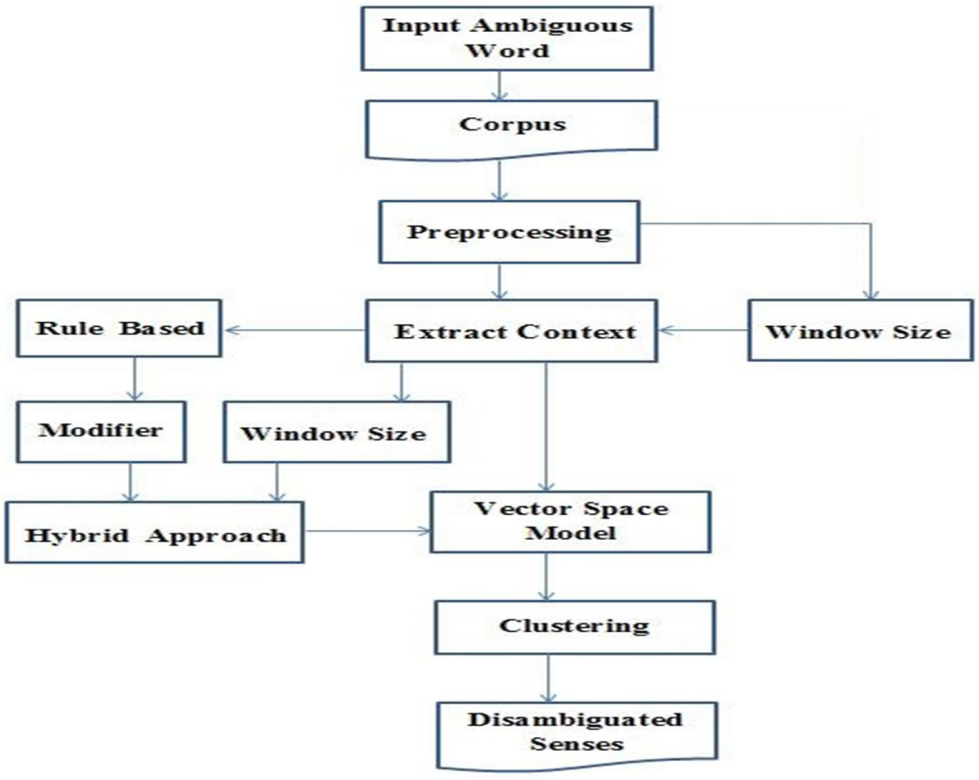


Figure 4.1. WSD Architecture

4.2. Data requirements

In this work, two types of data were used: (a) a small list of ambiguous words to test the algorithms. Twenty (20) highly frequent ambiguous words were selected from the language using the procedure described in Section 3.2.2, (b) a corpus containing thousands of sentences to extract contexts and its vectors to represent the group of contexts. Hence, the compiled corpus for the experiment is a free text. It is composed of a set of 62986 lines of sentences and 412,725 words. This corpus is prepared for the purpose of this study as there is no standard corpus for Afaan Oromo language.

4.3. Steps in Preprocessing

The 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 word list available for Afaan Oromo.

To get stop word list we have done an experiment which was implemented in Java programming language. The steps are presented below.

Step 1: Tokenization

In this step, the corpuses were first tokenized into words. Afaan Oromo uses white space as word separator. In this work tokenization is needed for some purposes that are to select contexts and stop word removal. For instance if we have the sentence *akkanatti hin haftu ni dabarti* in our corpus, the tokenized output would be like*:*

Akkanatti

hin

haftu

ni

dabarti

Step 2: Normalization

In Afaan Oromo the case of a word has no effect on the meanings of the words. Therefore, case folding was easily implemented.

For example: ***Qabsoo*** *is* similar to ***qabsoo***.

Step 3: Stop word removal

After tokenization and normalization take place, we went to remove those words considered irrelevant to represent the document.

To determine the stop word lists, first we calculated the whole frequency of each word in the corpus and sorted them in decreasing order by frequency.

Many of the most frequent words found to be stop words (like articles, prepositions) and are thus removed[18].

After the pre-processing has been completed out of the system, the cleaned corpuses were given to the Afaan Oromo WSD interface that was prepared by the researchers. The following figure shows the user interface of the word sense disambiguation used during the experimentation.

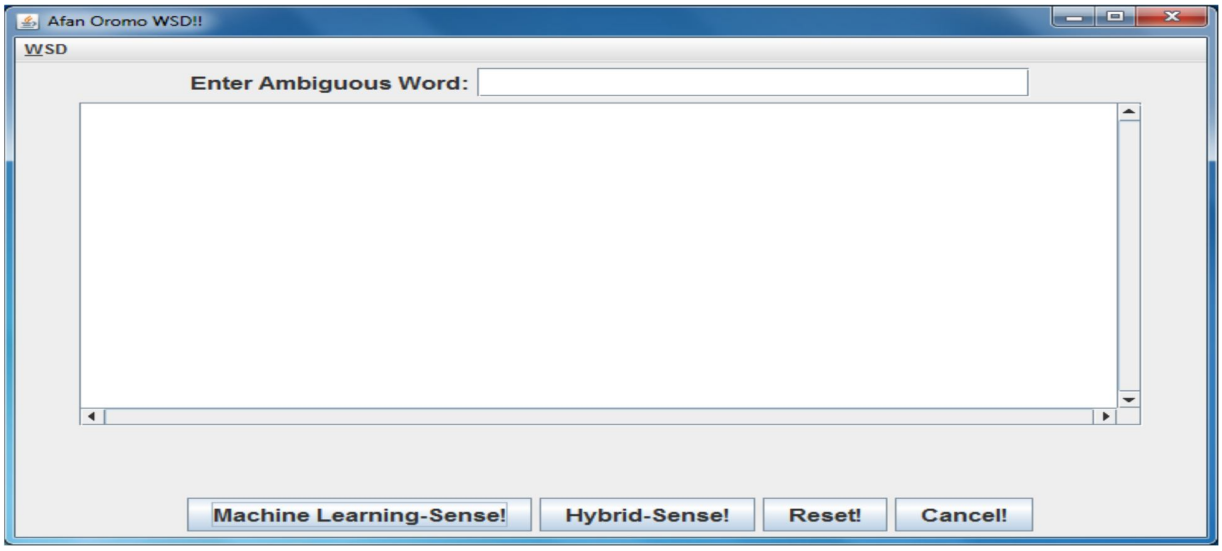


Figure 4.2. User Interface of the WSD

The interface was done by researchers on Java NetBeans 8.0.2. The interface works as follows:

First run the program from java and then the above interface is displayed.
After the interface is displayed, inter the ambiguous word in to the given text box. Then, select the type of WSD techniques (Machine learning sense or Hybrid sense).Then the system extract the context, by calculating cosine similarity and displaying on white space of the interface. These results are displayed in the format of CSV which are ready for clustering.
To disambiguate the given ambiguous word, it takes one ambiguous word at a time in the interface provided and produces the cosine similarity to cluster.

4.4. Experiment on the Machine Learning Approach using different Context Window Sizes and clustering algorithms

In the following section, we have presented the experimental procedure on different clustering algorithms and result on various windows size and suggested the optimal window size for Afaan Oromo.

As indicated in chapter three, the machine learning approach uses a window of *N* words surrounding the ambiguous word to extract contexts.
The extracted context words are then clustered together based on their similarity values using clustering algorithms and vector space model.
The system provides context terms to the left and right side of ambiguous word (excluding stop words) based on the provided window size. In this experiment, window sizes of up to ±10 words on both sides of ambiguous word have been tried.

In addition we used clustering algorithms to cluster contexts which are providing a useful way to discover semantically related senses. The clustering algorithms directly use the vector representations (cosine similarity measure) of the contexts of the ambiguous word (extracted using rules or the window of words) as input. The basic point of sense clustering is that related word senses must be associated and grouped in the same cluster. Hence, it identifies groups of senses, which are assumed to represent different meanings.

In this study the contexts of the ambiguous words are clustered using two clustering algorithms (EM and K-Means) and tree hierarchical clustering algorithms (Complete link, Single link and

Average Link). The number of clusters formed by the clustering algorithms depends upon the uniqueness and the dissimilarity present in the cosine similarity.

In total there are 20 ambiguous words to be discriminated, 12 words with 2 senses, 5 words with 3 senses, 1 word with 4 senses, and 2 words with 5 senses.

Five different configurations of clustering are run for each word by changing window size from 1-1 to 10-10 for each and every word. The results are presented different tables.

4.4.1. EM algorithm

Expectation maximization is clustering algorithm that works based on partitioning methods. This algorithm is a memory efficient and easy to implement algorithm, with a profound probabilistic background. As presented in table 4.1, EM algorithm performed best accuracy with window size one-one, which is 80.6 %.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	60	80	60	40	60	40	40	40
Afaan	50	50	50	50	50	50	50	50	50	50
Boqote	50	50	50	50	50	50	50	50	50	50
Darbe	100	75	50	50	50	50	50	50	50	50
Diige	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	50	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	33.3	33.3
Haare	100	50	50	50	50	50	50	50	50	50
Ija	75	75	60	60	75	40	60	40	40	40
Ji’a	100	50	50	50	50	50	50	50	50	50
Dhahe	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	33.33	66.67
Mirga	100	66.67	66.67	66.67	66.67	66.67	33.33	66.67	66.67	33.33
Waraabuu	100	100	66.67	66.67	66.67	33.33	66.67	66.67	66.67	33.33
Sanyii	100	50	50	50	100	50	50	50	50	50
Karaa	100	100	100	50	50	100	50	50	50	50
Ulfina	50	100	50	100	50	50	50	50	50	50
Ifa	100	100	50	50	50	50	100	50	100	50
Qophii	50	100	100	50	50	100	50	50	50	50
Sirna	50	100	100	50	50	50	50	50	50	100
Horii	100	50	50	50	100	50	50	50	50	50
Average	**80.6**	**78.2**	62.7	59.02	60.9	56.5	55.5	54.01	53.5	50.67

Table 4.1.Window size experimentation for EM algorithm

4.4.2. K-Means algorithm

This algorithm has the objective of classifying a set of n contexts into k clusters, based on the closeness to the cluster centers. The closeness to cluster centers is measured by the use of a Euclidean distance algorithm. According to the result found from this algorithms it performed best when compared to hierarchical clustering algorithms. It performed 78.1% with window size one-one and 75.2% with window size two-two.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	60	80	60	40	60	40	40	40
Afaan	50	50	50	50	50	50	50	50	50	50
Boqote	50	50	50	50	50	50	50	50	50	50
Darbe	100	75	50	50	50	50	50	50	50	50
Diige	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	50	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	33.33	66.67	33.3	33.3
Haare	100	50	50	50	50	50	50	50	50	50
Ija	75	75	60	60	75	40	60	40	40	40
Ji'a	100	50	50	50	50	50	50	50	50	50
Dhahe	66.67	66.67	66.67	33.33	66.67	66.67	66.67	66.67	33.33	66.67
Mirga	100	66.67	66.67	66.67	66.67	33.33	33.33	33.33	66.67	33.33
Waraabuu	100	66.67	33.33	66.67	66.67	33.33	66.67	66.67	33.33	33.33
Sanyii	100	100	50	50	100	50	50	50	50	50
Karaa	100	100	100	50	50	100	50	50	50	50
Ulfina	50	50	50	100	50	50	50	50	50	50
Ifa	50	100	50	50	50	50	100	50	100	50
Qophii	50	50	100	50	50	100	50	50	50	50
Sirna	50	100	100	50	50	50	50	50	50	100
Horii	100	50	50	50	100	50	50	50	50	50
Average	**78.1**	**75.2**	61.7	58.02	60.9	55.5	54.5	52.5	51.01	50.07

Table 4.2. Window size experimentation for K-means algorithm

4.4.3. Complete link clustering algorithm

The complete linkage algorithm is the MAX version of the hierarchical agglomerative clustering method which is a bottom-up strategy: compare each point with each point. Each context is placed in a separate cluster, and at each step merges the farthest pair of clusters, until certain termination conditions are satisfied. As presented in table 4.3 complete link algorithm perform better accuracy than single link and average link. The performance of complete link with window size one-one is 75.8%.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	60	80	60	40	60	40	40	40
Afaan	50	50	50	50	50	50	50	50	50	50
Boqote	50	50	50	50	50	50	50	50	50	50
Darbe	100	75	50	50	50	50	50	50	50	50
Diige	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	100	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	33.33	66.67	33.3	33.3
Haare	100	50	50	50	50	50	50	50	50	50
Ija	75	75	60	60	60	40	60	40	40	40
Ji’a	100	50	50	50	50	50	50	50	50	50
Dhahe	66.67	66.67	66.67	33.33	66.67	66.67	66.67	66.67	33.33	66.67
Mirga	100	66.67	66.67	66.67	66.67	33.33	33.33	33.33	66.67	33.33
Waraabuu	100	66.67	33.33	33.33	66.67	33.33	66.67	66.67	33.33	66.67
Sanyii	50	100	50	50	100	50	50	50	50	50
Karaa	100	50	100	50	50	100	50	50	50	50
Ulfina	50	100	100	100	50	50	50	50	50	50
Ifa	50	50	50	50	50	50	100	50	100	50
Qophii	100	100	100	50	50	100	50	50	50	50
Sirna	50	100	100	50	50	50	50	50	50	100
Horii	50	100	50	50	100	50	50	50	50	50
Average	**75.8**	**74.6**	64.2	56.35	60.65	55.5	54.5	52.5	51.01	51.73

Table 4.3: Window size experimentation for complete Link clustering algorithm

4.4.4. Single Link Clustering Algorithm

The single link algorithm is a MIN version of the hierarchical agglomerative clustering method which is a bottom-up strategy: compare each point with each point. Each context is placed in a separate cluster, and at each step merges the closest pair of clusters, until certain termination conditions are satisfied. As presented in table 4.4 the accuracy of single link were 74.6% and 73.6% with window size one-one and two-two respectively.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	60	80	40	40	60	60	40	40
Afaan	50	50	50	50	50	50	50	50	50	50
Boqote	50	50	50	50	50	50	50	50	50	50
Darbe	100	75	50	50	50	50	50	50	50	50
Diige	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	100	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	33.33	66.67	33.3	33.3
Haare	100	50	50	50	50	50	50	50	50	50
Ija	75	75	60	60	60	40	60	40	40	60
Ji'a	100	50	50	50	50	50	50	50	50	50
Dhahe	66.67	66.67	33.33	33.33	66.67	66.67	66.67	66.67	66.67	66.67
Mirga	100	33.33	66.67	66.67	66.67	33.33	33.33	33.33	66.67	33.33
Waraabuu	100	66.67	33.33	33.33	66.67	66.67	66.67	66.67	33.33	66.67
Sanyii	50	100	50	50	100	50	50	50	50	50
Karaa	100	50	100	50	50	100	50	50	50	50
Ulfina	50	100	100	100	50	50	50	50	50	50
Ifa	100	50	50	50	50	50	100	50	100	50
Qophii	50	50	100	50	50	100	50	50	50	50
Sirna	50	100	100	50	50	50	50	50	50	100
Horii	50	100	50	50	100	50	50	50	50	50
Average	**74.6**	**73.6**	62.6	56.35	59.65	57.16	54.5	53.5	52.7	52.73

Table 4.4.: Window size experimentation for single Link clustering algorithm

4.4.5. Average Linkage clustering algorithm

Average-link clustering produces similar clusters to complete link clustering except that it is less susceptible to outliers [41]. It computes the similarity between two clusters, as the average similarity between all pairs of contexts across clusters (e.g. using the Euclidean distance). As we can observed from table 4.5 it performed better accuracy with window size one-one which is 73.85%.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	60	80	40	40	60	60	40	40
Afaan	50	50	50	50	50	50	50	50	50	50
Boqote	50	50	50	50	50	50	50	50	50	50
Darbe	50	75	50	50	50	50	50	50	50	50
Diige	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	100	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	33.33	66.67	33.3	33.3
Haare	100	50	50	50	50	50	50	50	50	50
Ija	100	75	60	60	60	40	60	40	40	60
Ji’a	100	50	50	50	50	50	50	50	50	50
Dhahe	100	66.67	33.33	33.33	66.67	66.67	66.67	66.67	66.67	66.67
Mirga	100	66.67	66.67	66.67	66.67	33.33	33.33	33.33	66.67	33.33
Waraabuu	100	66.67	66.67	33.33	66.67	66.67	66.67	66.67	33.33	66.67
Sanyii	50	100	50	50	100	50	50	50	50	50
Karaa	100	50	100	50	50	100	50	50	50	50
Ulfina	50	100	100	100	100	50	50	50	50	50
Ifa	100	50	50	50	50	50	100	50	100	50
Qophii	50	50	100	50	50	100	50	50	50	100
Sirna	50	50	100	50	50	50	50	50	50	100
Horii	50	100	50	50	100	50	50	50	50	50
Average	**73.85**	**71.5**	64.2	56.35	62.15	57.16	54.5	53.5	52.7	54.3

Table 4.5. Window size experimentation for Average Linkage clustering algorithm

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

As can be seen from the above tables a narrow window of context, one and two words to either side was found to perform better than wider windows.

As can be seen from table, 4.1 and 4.2 window sizes one-one and two-two perform a better result with EM and K-means algorithms. Smaller window sizes (1-1 and 2-2) have yield significantly higher accuracy than other windows, which are 80.6% and 78.3% with EM ,78.1%and 75.2 %K-means,75.8 %and 74.6 %with complete link, 74.6 and 73.6 % with single link and 73.85% and 71.5 %with average link.

No	Clustering Algorithms	Window size	Accuracy (%)
1	EM	1-1	80.6 %
		2-2	78.3%
2	K-Means	1-1	78.1%
		2-2	75.2%
3	Complete Link	1-1	75.8 %
		2-2	74.6 %

Table 4.6. Summary of best experiments
Generally in this study best accuracy is achieved with EM (80.6%) algorithm.

4.4.1. Summary of the Machine learning clustering algorithms

Name of the Algorithms	**No of Clusters**	**Cluster Instances**	**No of Iterations**	**Within clusters sum of squared errors**	**Time taken to build model(in sec.)**	**Log likelihood**	**Uncluster ed Instances**
Single Link	28	0 - (50%) 1 - (50%)	-	-	0	-	-
Complete Link	28	0 - (70%) 1 - (30%)	-	-	0	-	-
Average Link	28	0 - (70%) 1 - (30%)	-	-	0	-	-
K-Means	30	0 - (40%) 1 - (60%)	2	7.66736	0	-	-
EM	33	0 - (70%) 1 - (30%)	-	-	0.4	45.37927	-

Table 4.7. Summary of Machine Learning Clustering Algorithms

As shown in table 4.7 the Weka experiment, the result of clustering algorithms varies with the distance they used.

As can be seen from table 4.6, out of the total clusters grouped by experts, the possible clusters are produced by different algorithms that correspond to different senses. An important issue in clustering is how to decide what is the best set of clusters for a given dataset, in terms of both the number of clusters and the membership of those clusters. For this reason we compare the number of clusters produced with the number of senses chosen by a group of human subjects (experts). However, counting the number of clusters only doesn't report about the accuracy of the system if one doesn't take the quality of the clusters into account. Hence, we have also evaluated how good the clusters are by comparing them against the clusters produced by experts. Put, differently, we measured the number of contexts assigned to correct and incorrect clusters (senses).

Based on the experiment, few of the clusters were wrongly clustered with different a context which has a different sense. For instance the pair of contexts *dhugaa* and *ragaa* are appearing in different clusters. The hierarchical clustering algorithms are not much satisfying result because of they are on a limited amount of information (single link), or they assume that all the cosine value in the cluster is very similar to each other at distance (complete link). The average link measuring the similarity of both these clustering algorithms, as the average of the pair wise similarity of the cosine value from each cluster.

4.5. Experiment on the Hybrid Approach

The second part of the experiment in this work is combining unsupervised machine learning and rule based approaches. Unlike unsupervised machine learning in Section 4.4, the hybrid approach relies on both machine learning and rule based algorhms. The rule based algorithms relies on, hand-constructed rules that are acquired from the structure of language rather than automatically trained from data. The rule based WSD benefit from various sources of linguistic knowledge [27]. This approach is used when there is a scarcity of data. The rule depend on domain knowledge to bridges the gap caused by data scarcity.

Unlike, unsupervised machine learning which uses window size only, the hybrid approach used rules to extract modifiers of the ambiguous word and consider them as contexts. These modifiers are therefore identified according to the developed rule planted.

The advantage of this approach is that it is easy to incorporate the linguistic knowledge [15]. In this research, we try to answer “Does the addition of linguistic knowledge, improve a word sense disambiguation performance for Afaan Oromo ambiguous word?”. As the conducted experiment shows the obtained result was more robust than unsupervised machine learning. In this case, with worldly knowledge and reasoning, it is necessary to assist disambiguation. According to [61] certain knowledge sources provide evidence as to the word sense used or disallowed by a particular context, but most of them use linguistic knowledge. According to [27] the hybrid defines a contextual representation is a characterization of the linguistic contexts in which a word appears.

However, in this work the rule based approach was not independently used to disambiguate, rather it is combined with unsupervised machine learning, just to mine the contexts of the ambiguous word. The modifiers (Section 3.5.1.1**.**) are identified accordingly the rule planted to it. Similar to unsupervised machine learning, we have used the same test set, window size and clustering algorithms in the hybrid approach.

Rule:

⇨ **IF** ambiguous words proceeded by Modifiers THEN collect the modifiers to disambiguate.

The following tables show the experiment taken place by hybrid approach.

4.5.1. Experimentation on EM algorithm

As presented on table 4.8 EM algorithm performed a better accuracy with window size one-one and two-two (90.35% and 86.28%) respectively. When we compared the performance of EM in hybrid approach and machine learning approach it results a better accuracy in hybrid approach.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	100	80	60	80	80	60	60	60	60	60
Afaan	100	100	100	100	50	50	50	50	50	50
Boqote	50	100	100	50	50	100	100	50	50	50
Darbe	100	75	100	50	100	100	50	50	50	50
Diige	100	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	50	100	100	100	50	50	50	50	50
Tume	100	66.67	100	66.67	66.67	100	66.67	66.67	66.67	66.67
Haare	100	50	50	50	50	100	50	50	50	50
Ija	100	75	75	75	75	75	60	60	60	60
Ji'a	100	50	50	50	50	50	100	50	50	50
Dhahe	66.67	100	66.67	66.67	66.67	100	66.67	66.67	66.67	66.67
Mirga	100	66.67	100	100	66.67	66.67	66.67	66.67	66.67	66.67
Waraabuu	100	100	66.67	100	66.67	66.67	66.67	66.67	66.67	66.67
Sanyii	100	100	50	100	100	100	50	50	50	50
Karaa	100	100	100	50	50	100	50	50	50	50
Ulfina	50	100	50	100	100	50	50	100	50	50
Ifa	100	100	100	50	50	100	100	50	100	50
Qophii	100	50	100	100	50	100	50	100	50	50
Sirna	50	100	100	100	50	100	100	50	100	100
Horii	100	50	100	50	100	50	50	100	50	100
Average	**90.35**	86.28	82.6	73.1	71.9	68.33	65.5	63.51	61.33	60.16

Table 4.8.: Window size experimentation for EM algorithm

4.5.2. Experimentation on K-Means algorithm

As shown on table 4.9 K-means performed a better result with accuracy of 86.6 % with window size one-one, which is appreciable when compared with its values on machine learning.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	100	60	80	80	60	60	60	60	60	60
Afaan	50	100	100	100	100	50	50	50	50	50
Boqote	100	50	100	100	50	100	50	50	50	50
Darbe	100	75	100	50	100	50	50	50	50	50
Diige	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	50	50	50	50	50	50	50	50	50
Tume	66.67	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Haare	100	50	100	50	50	50	50	50	50	50
Ija	75	75	60	60	75	75	60	60	60	75
Ji’a	100	50	100	50	100	50	50	50	50	50
Dhahe	66.67	66.67	100	66.67	66.67	66.67	66.67	66.67	100	66.67
Mirga	100	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Waraabuu	100	66.67	66.67	66.67	100	66.67	66.67	66.67	66.67	667.6
Sanyii	100	100	50	100	100	50	50	50	100	50
Karaa	100	100	100	50	100	100	50	50	50	50
Ulfina	100	50	50	100	50	50	100	100	50	50
Ifa	50	100	100	50	50	50	100	50	50	50
Qophii	100	50	100	100	50	100	50	100	50	50
Sirna	50	100	100	50	50	50	100	50	50	100
Horii	100	50	50	50	100	50	50	100	100	100
Average	**86.6**	83.55	81.21	70.5	72.5	66.08	65.33	62.67	62.2	60.6

Table 4.9: Window size experimentation for K-Means algorithm

4.5.3. Experimentation on complete link clustering algorithm

In complete link algorithm the distance between groups is 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. As can be seen from table 4.10 the excrement with complete link results the performance of 83.85% and 82.82.56% with window one-one and two-two respectively.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	60	80	60	60	60	60	40	40
Afaan	100	50	100	50	50	50	50	50	50	50
Boqote	50	50	100	50	100	50	50	50	50	50
Darbe	100	75	50	50	50	100	50	50	50	50
Diige	66.67	66.67	66.67	100	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	100	50	50	100	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	33.3
Haare	100	50	50	50	100	50	50	50	50	50
Ija	100	75	60	60	60	75	60	60	60	60
Ji’a	100	50	50	100	50	50	50	50	50	50
Dhahe	66.67	66.67	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Mirga	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Waraabuu	100	66.67	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Sanyii	50	100	50	100	100	50	100	50	50	50
Karaa	100	50	100	50	50	100	50	50	100	50
Ulfina	100	100	100	100	50	50	100	100	50	50
Ifa	100	50	100	50	50	50	100	50	100	50
Qophii	100	100	100	50	50	100	50	100	50	50
Sirna	50	100	100	50	50	50	100	50	100	100
Horii	50	100	100	100	100	50	50	50	50	50
Average	**83.85**	**82.56**	79.23	70.35	68.15	64.08	63.8	62.83	61.9	59.39

Table 4.10. Window size experimentation for complete Link clustering algorithm

4.5.4. Experimentation on single link clustering algorithm

Single linkage 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. As presented on table 4.11 the greatest performance of single is 81.6 and 80.26 with window one-one and two-two respectively.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	80	80	60	60	60	60	60	60
Afaan	100	100	50	100	50	50	100	50	50	50
Boqote	100	50	100	50	100	50	50	100	50	50
Darbe	100	75	50	100	50	50	50	100	100	50
Diige	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	100	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Haare	100	50	100	50	50	50	50	50	50	50
Ija	75	75	60	60	60	75	60	60	60	60
Ji'a	100	50	50	100	50	50	50	50	50	50
Dhahe	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Mirga	100	66.67	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Waraabuu	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Sanyii	50	100	50	50	100	50	50	50	50	50
Karaa	100	50	100	50	50	100	50	50	50	50
Ulfina	50	100	100	100	50	50	50	50	50	50
Ifa	100	50	50	100	100	50	100	50	100	50
Qophii	50	100	100	50	50	100	50	50	50	50
Sirna	50	100	100	50	50	50	50	50	50	100
Horii	50	100	50	100	100	50	50	50	50	50
Average	**81.6**	**80.26**	**76.1**	72.18	65.65	61.6	60.33	61.2	60.5	60.1

Table 4.11: Window size experimentation for single Link clustering algorithm

4.5.5. Experimentation on Average link clustering algorithm

Average linkage is clustering algorithm 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. According to our experimentation on average link clustering algorithm the best performances are resulted by window size one-one and two-two which are 80.6% and 79.1% respectively.

Ambiguous words	**Window size**									
	1-1	2-2	3-3	4-4	5-5	6-6	7-7	8-8	9-9	10-10
Bahe	80	60	80	80	60	60	60	60	60	60
Afaan	100	100	50	100	50	50	100	50	50	50
Boqote	100	50	100	50	100	50	50	100	50	50
Darbe	100	50	50	100	50	50	50	100	100	50
Diige	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Dubbatate	100	100	50	50	50	50	50	50	50	50
Tume	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Haare	100	50	100	50	50	50	50	50	50	50
Ija	75	75	60	60	60	75	60	60	60	60
Ji'a	100	50	50	100	50	50	50	50	50	50
Dhahe	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Mirga	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Waraabuu	100	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67
Sanyii	50	100	50	50	100	50	50	50	50	50
Karaa	100	50	100	50	50	100	50	50	50	50
Ulfina	50	100	100	100	50	50	50	50	50	50
Ifa	100	50	50	100	100	50	100	50	100	50
Qophii	50	100	100	50	50	100	50	50	50	50
Sirna	50	100	100	50	50	50	50	50	50	100
Horii	50	100	50	100	100	50	50	50	50	50
Average	**80.6**	**79.1**	**75.1**	72.18	65.65	61.6	60.33	61.2	60.5	60.1

Table 4.12: Window size experimentation for Average Linkage clustering algorithm

Generally this hybrid experiment, most of the tested ambiguous words have relatively higher performance when compared with machine learning.

As indicated in experiment results above, the hybrid approach achieves accuracy of; 90.35% and 86.28% with EM,86.6% and 83.55% with K-means, 83.85 % and 82.56% with Complete link, 81.6% and 80.26 % with Single link and 80.6 % and 79.1% with Average link by using window size of one-one and two-two respectively.

When we compare these results with results achieved by machine learning, hybrid approach yields a better accuracy. The reason behind the enhanced accuracy might be because the hybrid method brings mix the advantages of both method the Machine Learning and rule based approaches.

No	Clustering Algorithms	Window size	Accuracy (%)
1	EM	1-1	90.35 %
		2-2	86.28%
2	K-Means	1-1	86.6%
		2-2	83.55%
3	Complete Link	1-1	83.85 %
		2-2	82.56%

Table 4.13. Summary of best Accuracy of the Hybrid Approach

4.5.6. Summary of the Hybrid Approach clustering algorithms

Name of the Algorithms	**No of Clusters**	**Cluster Instances**	**No of Iterations**	**Within clusters sum of squared errors**	**Time taken to build model(sec.)**	**Log likelihood**	**Unclustered Instances**
Single Link	38	0: 50% 1: 50%	-	-	0	-	0
Complete Link	38	0: 50% 1: 50%	-	-	0	-	0
Average Link	38	0: 50% 1: 50%	-	-	0	-	0
K-Means	45	0: 75% 1: 25%	2	2.0	0	-	0
EM	49	0: 85% 1: 15%	-	-	0.04	50.19	0

Table 4.14. Summary of Hybrid Approach Clustering Algorithm

The way of using clustering algorithms in this hybrid approach is similar to the way used in unsupervised machine learning. Likewise, the manner of these algorithms used, determined and evaluated also similar but their performance is different. To evaluate the accuracy, we considered the basis of number of clusters produced and correctness of the produced clusters. Additionally, the time (rate of convergence) and the number of iterations were also the other evaluations (see table 4.14).

Based on the experiment on section 4.5, from the total five clusters, the EM and K-Means clusters which yield significantly higher accuracy than other clusters when compared with machine learning result (see 4.4). EM is the best-suited algorithm for the given datasets, since it depends on the probability distributions where each distribution represents a cluster. As presented in the Section 4.5.6, if the log likelihood probability is maximum it shows high accuracy. Similarly, K-Means clustering also showed better performance.

However, Single, complete and average link clusters did not perform as much as EM and K-Means clusters. Therefore, for WSD to Afaan Oromo, EM and K-means enhance the accuracy of sense clustering than hierarchical single, complete and average clustering algorithms.

4.6. Discussion

In this paper, we reported unsupervised machine learning and hybrid approach for Afaan Oromo WSD. We described the experiment we conducted to compare the performance of the two approaches. These approaches were evaluated with a benchmark test set (by experts), to facilitate the comparison with the results of both approaches. The accuracy results of our method, when compare machine learning and hybrid results (section 4.4 and section 4.5) are promising and proving method respectively [61].

In the machine learning experiment, the system yields a highest accuracy of 80.6% with window size one-one and EM clustering algorithm and in the hybrid experiment, the system yields an accuracy of 90.35 % with the same clustering algorithm and window size. Looking at the smallest dataset available to both machine learning and hybrid, they found an average figure of different yielded results. Initially, we evaluated our WSD method with all the 20 words. This lead, to a total of 20 words tested in this evaluation, and these ambiguous words have two senses to five senses. twelve terms have two senses (the terms with two senses are *afaan, boqote, dubbatate, haare, ji'a,* Sanyii ,Karaa ,Ulfina ,Ifa ,Qophii ,Sirna ,Horii), and five terms have three senses (the terms with three senses are *diige, tume, dhahe, mirga, waraabuu*) and two terms have five senses (the terms with five senses are *bahe, ija*) and the left one has four senses (the term with four senses is *darbe*) out of 20 ambiguous terms (see table 4.19)

4.6.1. Comparison of machine learning and hybrid approaches

No	Clustering Algorithms	Window size	**Machine learning**	**Hybrid approach**
1	EM	1-1	**80.6 %**	**90.35 %**
		2-2	**78.3%**	**86.28%**
2	K-Means	1-1	78.1%	86.6%
		2-2	75.2%	83.55%
3	Complete Link	1-1	75.8 %	83.85 %
		2-2	74.6 %	82.56%

Table 4.15. Comparison of machine learning and hybrid approaches

The hybrid approach shows promising result as compared to the performance of independent machine learning approaches to WSD task (see section 4.4 and 4.5 experimental results).

The rules were helpful to extract contexts overcoming the shortage of training data. This rule is beneficial for WSD, especially the integration of deep linguistic knowledge with algorithms markedly improves disambiguation accuracy. We argue that using hybrid approach to find senses, significantly increase disambiguation performance. This result also seems to support the result obtained by [71]. In this study we claim that using either a machine learning or hybrid approach (as discussed in Section 3.4 and Section 3.5 respectively) possible to find the sense of any ambiguous word in a given contexts but the hybrid approach was recommended when there are scarcity of training data and the accuracy needed [23].

Using a window size of ±2 words either side of the target word offered the accuracy of disambiguation than using the whole sentence. Therefore; smaller value of the window size, which leads to the proper choice of sense for the target word. Based on this result, we conclude that for Afaan Oromo window 0ne-one and two-two was recommended unlike Amharic language [20] which window size of 3 is recommended.

The conducted experiment shows that, the meaning of words are closely connected to the statistics of word usage, which are working with window size and vectors value derived from event frequencies; that is, we are dealing with Vector Space Model (cosine similarity) and clustering (Euclidean distance) [72]. By using cosine similarity we include important semantic information in the purely statistical process of selecting the appropriate sense for a particular word. This benefits both unsupervised, hybrid approaches to WSD by increasing the chances of matching a particular contexts.

As shown on table 4.15, the result obtained by unsupervised machine learning and hybrid approach were different as the semantic information extracted by the algorithms is distinct from the rule. However, the inclusion of unsupervised machine learning only as features does not always improve performance. This is [72] that machine learning algorithms were a useful information source for disambiguation but that it not as robust as a linguistic (modifiers in this case).

As the conducted experiment showed, each clusters have a context group, where the sense of these context groups are hopefully different. The underlying assumption is that the senses found in similar contexts are similar meanings.

Then, new occurrences of the context can be classified into the closest induced clusters (senses). All contexts of related senses are included in the clustering and thus performed over all the contexts in the sentences. The underlying hypothesis is that ambiguous word contexts clustering

captures the reflected unity among the contexts and each cluster reveal possible relationships existing among these contexts (see table 4.6 and table 4.13) [73].

The system evaluations of the system were undertaken on the basis of precision and recall (see section 3.6)

As the evaluation of the system indicates that, unsupervised machine learning usually 76.05% on the test, while the hybrid approach accuracy 89.47 (see table 4.16).

	Unsupervised Machine Learning		**Hybrid Approach**	
	Precision	Recall	Precision	Recall
Correctly Disambiguated Senses	76.05%	65%	89.47%	85%

Table 4.16. Evaluation of WSD

From Table 4.16, the hybrid approach WSD is an encouraging result than unsupervised machine learning approach. This is due to unsupervised approach is not as hybrid approach, especially hierarchical clustering result was noisy. As already discussed before, the obtained result in both approaches were different. Therefore, the linguistic knowledge (hybrid approach) the best approach to solve WSD than machine learning algorithms in Afaan Oromo [18] as shown in the experiment (Table 4.16). However, the overall system performance gained thus far is not surprising since this language was under resourced materials and tools.

From the finding of this experiment the addition of deep linguistic knowledge to a WSD system is a significant rise in disambiguation accuracy and coverage as compared with results discussed so far. It is especially interesting that using the preceding modifiers of the ambiguous word perform better result. We can conclude that modifiers contain a lot of valuable clues for disambiguation [74].

The WSD developed for Afaan Oromo has its own strength and weakness. As the results showed that, the experiment attempts to disambiguate any ambiguous words if it is in corpus rather than limiting itself to treating restricted ambiguous words. This is one of the strongest sides of this WSD. It is argued that this approach is more likely to assist the creation of practical systems.

However, there is some ambiguous word on which the performance of our approach is actually low (as shown in section 4.4 and Table 4.5).

The system reported that the vector space model was affected by the data scarcity. The frequencies of co-occurrence of most context words are zero due to the limited corpus size of the

language. The other weakness of the system is context clustering in hierarchical clustering which was a noisy and yielded low performance as compared to K-Means and EM clustering.

4.6.2. Walk-through Using an Example

A. Machine Learning

In this section, a walk-through of the experiment is described using an example in the following sections. As described in Section 4.4, the contexts are identified using window sizes. For instance, assuming the ambiguous word *bahe* the unsupervised machine learning WSD extracts the following contexts words.

No	Contexts
1	Bilisa
2	Ragaa
3	Dhugaa
4	Qabsoo
5	Tabba
6	Gaara
7	Uccuu
8	Duute
9	Daara
10	Lubbuu

Table 4.17. Extracted Context Using Window Size

As presented in section 4.4 and 4.5 EM and K-Means produce better result for word bahe with window size one-one and two-two, as compared with hierarchical clusters on both WSD approach (Machine learning and hybrid) . In case of the quality of the clusters, EM and K-Means are relatively good quality as compared with others.

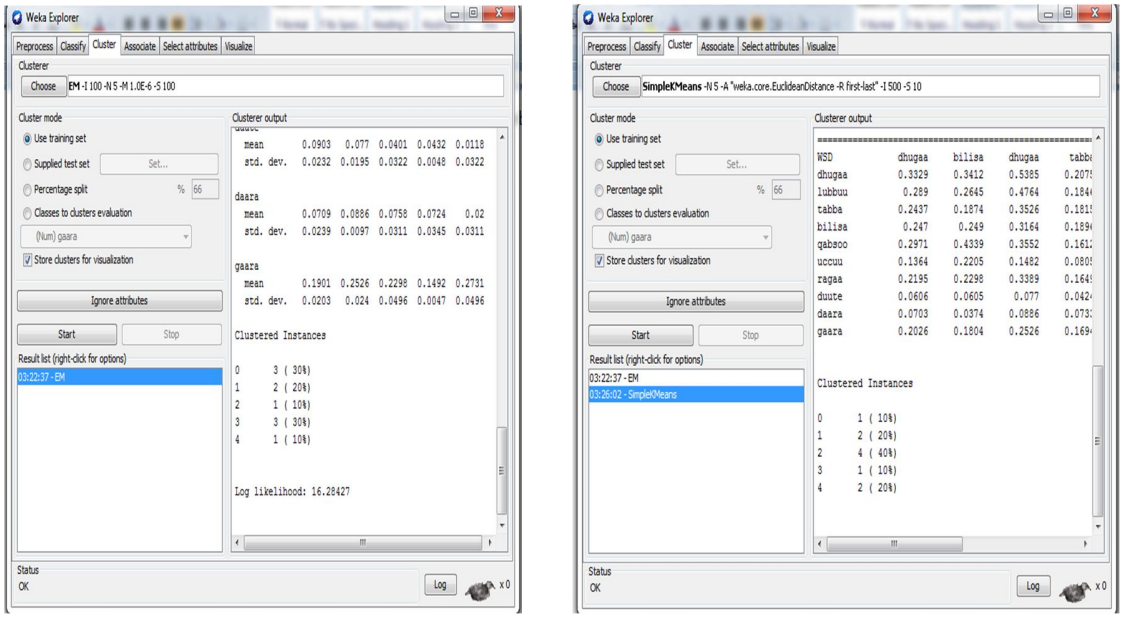


Figure 4.3. Expectation Maximization and K-Means Clustering for word bahe

B. Hybrid Approach

No	The Modifiers
1	Bilisa
2	Qabsoo
3	Tabba
4	Gaara
5	Dhugaa
6	Ragaa

Table 4.18. Extracted Modifiers for ambiguity word *bahe*

As can be seen from table 4.18, the modifiers identified accordingly the rule developed. Similar to machine learning, the identified modifiers with its cosine similarity in .CSV file loaded to Weka tool. The process of using clustering algorithms was similar to in unsupervised machine learning experiment. As the experiment shows that senses are similar between each other and are dissimilar to the senses belonging to different clusters. The clustering is which two or more senses are considered belong to the same cluster if it defines a concept common to all these senses.

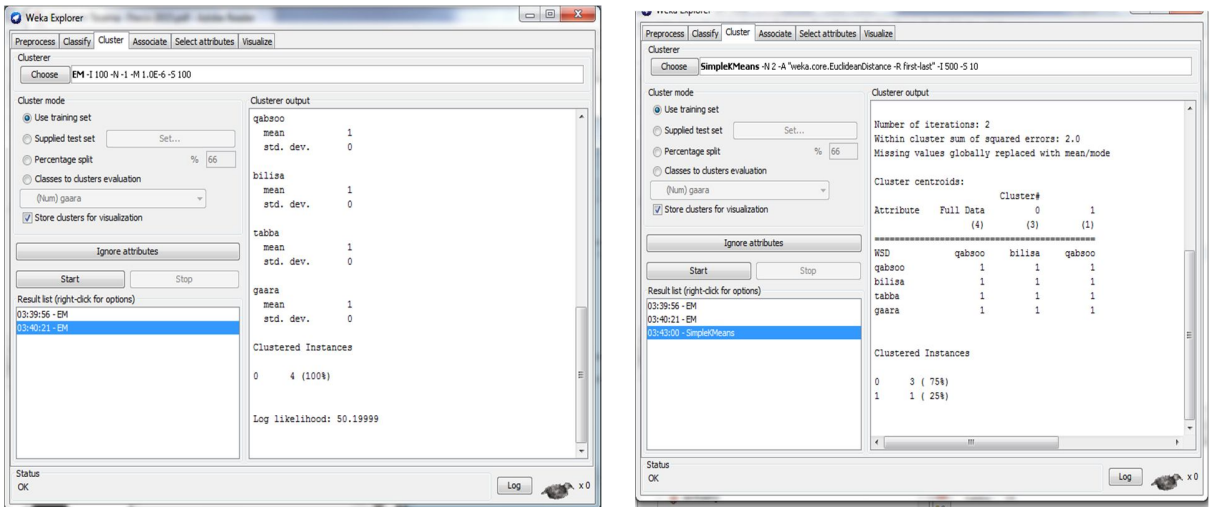


Figure 4.4. EM (left side) and K-Means (right side)

The other interesting way to examine the data in these clusters is inspect it visually. One of the options from this pop-up menu is visualizer cluster assignments. Windows pops up that the results and see them visually. This is shown in a chart how the clusters are grouped in terms of instances. Also, turn up the “Jitter” which is artificially scatter the plot points to allow to see more easily.

As a general, the WSD performance was promising result as it is hybrid approach which compiled unsupervised machine learning and rule based approaches. The hybrid approach was very important for WSD system especially for under resourced language like Afaan Oromo. In this work, it has essential role in this study in terms of accuracy and data scarcity. Similarly, it has a strict sense of well-formedness in mind and imposes linguistic constraints to satisfy well formedness. It relies on hand-constructed rules that are to be acquired from language specialists rather than automatically trained from data [75]. Based on the experiment, the possible senses of the test ambiguous word according to the given contexts has the following senses on table 4.19.

No	**Ambiguous Words**	**No of Senses**	List of sense	**Possible Senses**
1.	Sanyii	2	Sense1	Seed
			Sense2	Type
2	Karaa	2	Sense1	Road
			Sense2	Way
3	Ulfina	2	Sense1	Weight
			Sense2	Respection
4	Ifa	2	Sense1	Light
			Sense2	Clear
5	Qophii	2	Sense1	Program
			Sense2	Preparation
6	Sirna	2	Sense1	Event
			Sense2	Systems
7	Horii	2	Sense1	Money
			Sense2	Cattle
8	Afaan	2	Sense1	Language
			Sense2	Mouth
9	Bahe	5	Sense1	Freedom
			Sense2	Highland
			Sense3	Witness
			Sense4	Cloth
			Sense5	Dead(pass)
10	Boqote	2	Sense1	Break (rest)
			Sense2	Died
11	Darbe	4	Sense1	Cross
			Sense2	Pass from class to class
			Sense3	Died
			Sense4	broadcast
12	Diige	3	Sense1	Fence
			Sense2	Absence on Meeting
			Sense3	cancel to start new
13	Dubbatate	2	Sense1	Struggle
			Sense2	Wedding
14	Tume	3	Sense1	Make
			Sense2	Contraceptive
			Sense3	Hit
15	Haare	2	Sense1	Sad
			Sense2	Burn
16	Ija	5	Sense1	Eye
			Sense2	Vengeance
			Sense3	Wide-eyed
			Sense4	Tree Fruit
			Sense5	A little

17	Ji’a	2	Sense1	Stars
			Sense2	Month
18	Dhahe	3	Sense1	Follow
			Sense2	Hit
			Sense3	Fail
19	Mirga	3	Sense1	Direction
			Sense2	Human right
			Sense3	Brave
20	Waraabuu	3	Sense1	Hyena
			Sense2	Record
			Sense3	Fetch

Table 4.19. Summary of the Test

CHAPTER FIVE

5. Conclusions and Recommendations

5.1. Conclusions

The overall focus of this research was to investigate Word Sense Disambiguation which addresses the problem of deciding the correct sense of an ambiguous word based on its surrounding context's and the modifiers. In this study, a hybrid approach which combines unsupervised machine learning and rule based methods was used. To this end, we relied on several techniques which include: N- windows of words to extract context words, vector space model for measuring similarity between the context words and clustering algorithms (Single, Complete and Average clustering, EM and K-means clustering algorithms) to group related context words.

Based on the literature review, experiments and the results presented in previous chapters, we have forwarded the conclusions and recommendations as presented in the following sections:

- ❖ According to this study EM clustering algorithm is the most important clustering algorithm for Afaan Oromo WSD.

- ❖ Window size of one-one and two-two are enough for Afaan oromo word sense disambiguation.

- ❖ Hybrid approach is the best approach for Afaan Oromo word sense disambiguation. It yields the highest accuracy of 89.47%, which is better when compared with the accuracy of machine learning.

5.2. Recommendations

The following recommendations are forwarded based on our findings with regard to the developments of resources and future research directions to WSD for Afaan Oromo:

- ➢ Researches in WSD for other languages use knowledge resources like WordNet, lexicon, machine readable dictionaries and machine translation software. In this study, we faced a significant challenge as Afaan Oromo lacks those resources. Taking into account their contribution to WSD other researchers and institutions should develop these resources.

- Other developed languages do have standard sense annotated data for WSD research. The researcher could not find have such data for Afaan Oromo language. So, there need to be an initiative to prepare the data for WSD research.
- The size of the corpus used for this research was limited and doesn't have good quality. These limitations affected the accuracy of word sense disambiguation because if resources used are small, we may not be able to find all possible senses. Therefore, other works should be attempted with large and high quality corpus to minimize these problems.
- There are so many clustering algorithms in weka 3.7.9. But, due to time limitation, in this study only five clustering algorithms were experimented. 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.
- In addition to corpus based approach, there are also knowledge source based and hybrid approach (combination of knowledge source based and corpus based approach) which are used for WSD in another language and found a good result [77] like WordNet, annotate data. These approaches need to be investigated for Afan Oromo as well.

References

[1] Bar-Hillel, Yehoshua(1960)The present status of automatic translation of languages. Advances in Computers, ed. by Franz Alt et al. 91–163. New York: Academic Press.

[2] Daniel B.(2011) Afaan Oromo Information Retrieval (CLIR): A Corpus Based Approach, M.Sc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

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

[4] Ravi Mante, Mahesh Kshirsagar and Dr. Prashant Chatur(2014) A Review Of Literature On Word Sense Disambiguation, PhD, Government college of engineering Amravati, Maharashtra.

[5] Jurafsky, D., Martin, J.(2009) Speech and Language Processing (2nd Edition) Pearson Education.

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

[7] Gang Li, Guangzeng Kou, Ercui Zhou, and Ling Zhang(2009). Symmetric Trends: Optimal Local Context Window in Chinese Word Sense Disambiguation. In Proceedings of the 2009 Ninth International Conference on Hybrid Intelligent Systems – Vol. 03, Washington, DC, USA.IEEE Computer Society.

[8] Stevenson, Mark and Yorick Wilks, The interaction of knowledge sources in word sense disambiguation, Computational Linguistics, Vol.27, No.3, PP. 321–349, 2001.

[9] 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.

[10] King, B. (1967) Step-wise clustering procedures. Journal of the American Statistical Association, Vol.69.

[11] 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.

[12] 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.

[13] Ide, N., & Veronis, J. (1998) Introduction to the special issue on word sense disambiguation: the state of the art. Comput. Linguist, vol.24 (1).

[14] Roberto Navigli,(2009) Word Sense Disambiguation: A Survey, ACM Computing Surveys. USA.

[15] Gale, W., Church, K., & Yarowsky, D. (1993) A method for disambiguating word senses in a large corpus. Computers and the Humanities, vol.26 (5).

[16] 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.

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

[18]. 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.

[19] 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.

[20] Solomon A., "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.

[21] 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.

[22] Fayisa G., "Unsupervised Corpus Based Approach for Word Sense Disambiguation to Afaan Oromo words" ,M.sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[23] Abdel Monem, K. Shaalan, A. Rafea, H. Baraka(2008) Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation,Springer,vol.20(4).

[24] Geoffrey Leech (1991) The state of the art in corpus linguistics, in English Corpus Linguistics:Linguistic Studies in Honour of Jan Svartvik, London: Longman.

[25] Mohd Shahid Husain and Mohd Rizwan Beg (2013). Word Sense Ambiguity: A Survey, Lucknow, India, Volume 02–Issue 06, November 2013.

[26] Harris (1998) Distributional structure. Word, 10(23):146–162.

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

[28] Dorr, Bonnie, Machine translation divergences: A formal description and proposed solution, Computational Linguistics, Vol. 20, No. 4, pp., 1994, 597-633.

[29] Curran, James R. (2004) From Distributional to Semantic Similarity, Ph.D. thesis, University of Edinburgh.

[30**]** Nancy Ide and Jean Véronis, (2007)Word Sense Disambiguation Algorithms and Applications, Ney work, Spring, vol.33.

[31] Heine B., The Wata dialect of Oromo. (Language and dialect atlas of Kenya 4). Berlin:Dietrich Reimer Verlag, 1981.

[32] 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.

[33] Smeaton, A.F(1995) Linguistic Approaches to Text Management: An Appraisal of

[34] Marco Baroni and Gemma Boleda(2009) Distributional Semantics: Natural Language Processing.

[35] Scott McDonald(2000)Environmental Determinants of Lexical Processing Effort. Ph.D. thesis, University of Edinburgh.

[36] H. Schutze(1998)Automatic word sense discrimination. Computational Linguistics, vol.24(1).

[37] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999), Data clustering: a review. ACM Computing Surveys, Vol 31(3).

[38] Nagy, G(1968) State of the art in pattern recognition in Proceedings of IEEE.

[39] Sneath, P. H. A. and Sokal, R. R(1973) Numerical Taxonomy: The Principles and Practice of Numerical Classification., London, UK: Freeman.

[40] Jain, A. K. and Dubes(1988), R. C., Algorithms for Clustering Data, Prentice-Hall.

[41] Han, J. and Kamber, M.(2001) Data Mining–Concepts and Techniques: Morgan Kaufmann.

[42] A. Dempster, N. Laird, and D. Rubin(1977) Maximum likelihood from incomplete data via the EM algorithm .Journal of the Royal Statistical Society, Vol. 39.

[43] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics and data analysis, vol.14.

[44] Amaal Saleh Hasan (2010) Word Sense Disambiguation and Semantics techniques, MSc thesis, Sultan Qaboos University.

[45] Manning, C. D. and Schütze, H. , Foundations of Statistical Natural Language Processing.1999: MIT Press.

[46] Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha(2013) A Hybrid Approach To Word Sense Disambiguation Combining Supervised And Unsupervised Learning, West Bengal, India.

[47] J. Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amjan Shaik, P. Pavan Kumar(2012) Word Sense Disambiguation: An Empirical Survey, Vol.2, Issue-2, Hyderabad, India.

[48] Mishra N., An Unsupervised Approach to Hindi Word Sense Disambiguation, Indian Institute of Information Technology, Allahabad. UP, India, 2009.

[49] PierpaoloBasile, Annalina Caputo and Giovanni Semeraro(2014). An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model, University of Bari Aldo Moro Via E. Orabona, 4, Bari – 70125, Italy.

[50] Manning, C. D. and Schütze, H. , *Foundations of Statistical Natural Language Processing.*1999: MIT Press.

[51] Stephen Clark (2014). Vector Space Models of Lexical Meaning, Handbook of Contemporary Semantics, 2nd ed, edited by Shalom Lappin and Chris Fox.

[52] Arindam Roy, Sunita Sarkar and BipulSyamPurkayastha(2014) Knowledge Based Approaches to Nepali Word Sense Disambiguation, Msc thesis, Assam University,Silchar, Vol. 3.

[53] Curran, James R. (2004) From Distributional to Semantic Similarity, Ph.D. thesis, University of Edinburgh.

[54] Atelach Alemu Argaw and Lars Asker(2010) An Amharic Stemmer: Reducing Words to their Citation Forms, M.S Thesis, Department of Computer and Systems Sciences, Stockholm University, Sweden.

[55] F1orster, R.(2006) Document clustering in large German corpora using natural language processing, Ph.D. dissertation, University of Zurich.

[56] Malmkjaar, Kirsten(1995)The linguistics Encyclopedia, New York: Routledge.

[57] Flickinger, D.(2015)Natural Language Engineering-Efficient Processing with HPSG: Methods, Systems, Evaluation. At: http://www.coli.uni-sb.de/nlesi/ Accessed at 3/15/2015.

[58] Diriba Megersa(2002)An Automatic Sentence Parser For Oromo Language Using Supervised Learning Technique, M.Sc. Thesis, Addis Ababa University, Ethiopia.

[59] Girma Debele(2014)Afaan Oromo News Text Summarizer, Master's thesis, Pohang University of Science and Technology, Pohang, Korea.

[60] Temesgen Negassa(1995) word formation: the structure of independent and dependent clauses in oromo, Addis Abeba, Bole.

[61] David Yarowsky(1994)Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, ACL '94, Stroudsburg, PA, USA. Association for Computational Linguistics.

[62] Stephen Clark (2014)Vector Space Models of Lexical Meaning, Handbook of Contemporary Semantics, 2nd ed, edited by Shalom Lappin and Chris Fox.

[63] Tullu Guya(2003) CaasLuga Afaan Oromoo: Jildii-1,Gumii Qormaata Afaan Oromootiin Komishinii "Aadaa fi Turizimii Oromiyaa", Finfinnee.

[64] Schutze, H. (1992)Dimensions of meaning. In Supercomputing '92: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing. IEEE Computer Society Press, Los Alamitos, CA.

[65] Jeff Mitchell and Mirella Lapata(2008) Vector-based models of semantic composition. In Proceedings of ACL-08: HLT, Columbus, OH, USA.

[66] Katrin Erk and Sebastian Padó(2008) A structured vector space model for word meaning in

context. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA.

[67] Joseph Reisinger and Raymond J Mooney(2010)Multi-prototype vector-space models of word meaning. In Proceedings of HLT-NAACL.

[68] Hearst, Marti(1991) "Noun Homograph Disambiguation Using Local Context in Large Text Corpora," in Using Corpora, University of Waterloo, Waterloo, Ontario.

[69] Dagan, Ido and Alon Itai(1994) Word Sense Disambiguation Using a Second Language Monolingual Corpus, Computational Linguistics, vol.20.

[70] Ted Pedersen and Rebecca Bruce(1997)Distinguishing word senses in untagged text: In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, New York, Routledge.

[71] Rada Mihalcea, Ted Pedersen(2005) Advances in Word Sense Disambiguation, Tutorial at ACL Conference.

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

[73] Fei Shao, Yanjiao Cao(2005) A New Real-time Clustering Algorithm, Department of Computer Science and Technology, Chongqing University of Technology Chongqing 400050, China.Linguistics:Linguistic Studies in Honour of Jan Svartvik, London, Longman.

[74] Tilahun Gamta(1995) Seera Afaan Oromoo, Finfinnee, Boolee Press.

[75] Debela Tesfaye(2011)A Rule-based Afaan Oromo Grammar Checker, Addis Ababa University, Addis Ababa, Ethiopia.

[76] Rui Xu and Donald Wunsch (2005) Survey of Clustering Algorithms and Transactions On Neural Networks, Cambridge University Press, Cambridge, UK. Vol. 16.

[77] George Miller. (1990) Wordnet: An on-line lexical database. International Journal of Lexicography, vol.3 (4).

Appendix A: Question for Collection of Ambiguous Words

The questions are prepared and converted to Afaan Oromo language to make suitable to collect the ambiguous words test from public.

Saala: _____ Umurii:_____

Sadarka Barumsa: kutaa 1-5:_____

kutaa 6-12:_____

Barataa Kolleejjii:_____

Barataa Yuuniversitii:_____

Barsiisaa:_____

Q/Bulaa:_____

Kan biraa:_____

1. Jechoota hiika lamaa ol qaban (ambiguous words) kan naannawaa keessanitti fayyadamtu naaf tarreessi?

Lakk.	Jechoota	Baay'ina Hiikaa	Hiika

Declaration

This study is my original work and has not been submitted as a partial requirement for a degree in any University and that all sources of material used for the study have been duly acknowledged.

_____

Yehuwalshet Bekele Tesema

June 2016

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

Ermias Abebe _____ _____