ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

USE OF PART OF SPEECH TAGGING FOR AFAAN OROMO WORD SENSE MODELING

By:
Lalise Daniel Beka

February, 2019

Addis Ababa Ethiopia

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

USE OF PART OF SPEECH TAGGING FOR AFAAN OROMO WORD SENSE MODELING

By:
Lalise Daniel Beka

A Thesis submitted to Addis Ababa University in partial fulfillment of the requirement for the Degree of Masters of Science in Information Science

February, 2019

Addis Ababa Ethiopia

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

USE OF PART OF SPEECH TAGGING FOR AFAAN OROMO WORD SENSE MODELING

By:
Lalise Daniel Beka

Name and Signature of the Examining Board for Approval

Name	Signature	Date
Solomon Teferra (PhD) , Advisor	_____	_____
_____, Examiner	_____	_____
_____, Examiner	_____	_____

Declaration

I declare that this research is my original work and has not been presented for a degree in any university.

Declared by:

Name: Lalise Daniel

Signature: _____

This research has been submitted for Examination with my approval as university advisor.

Name: Solomon Teferra (PhD), Advisor

Signature: _____

Date: _____

February, 2019

Addis Ababa, Ethiopia

Acknowledgement

First I would like to thank My God, Who was with me in all difficulties and helped me to achieve each success I have achieved.

My special gratitude goes to my advisor Dr. Solomon. I deeply appreciate all the advice, opportunities, support, and encouragement that he has given me during my thesis. I would like to express my heart-felt thanks to Dr. Martha Yifru for her understanding and advice. Her endless energy and dedication to her students inspired me every day. She is my model and the reason why I chose NLP research.

Thanks to my teacher, Ermias Abebe for his wonderful assistance. His advice made me more confident to complete my work efficiently and present it in a good manner.

Next I would like to express thanks for my Husband, Aboma Amsalu and my parents: my Mom, Genet Tesemma, and My Dad, Daniel Beka for their prayer, Love, encouragement, advice and care.

I would like to express my gratitude to Dano Endalu (Ass. Prof.) who helped me to communicate the linguists. I am very thankful for Linguists: Tujube Amansa (PHD candidate), Fatuma Jeldo (MA) and Eba Wecho (MA) who helped me by tagging Afaan Oromo words.

I would like to thank Getachew Emiru, Yehuwalashet Bekele and Feyisa Gemechu for supporting me by giving me documents and sharing their experience on their previous work. Finally I want to thank all my families, my friends and colleagues who helped me through all my works.

Abstract

Word sense induction (WSI) is the task of automatically discovering all senses of an ambiguous word in a corpus. Induced senses can lead researchers in machine translation and information retrieval to improved performance.

In this thesis we have investigated the application of POS tagging to increase the performance of Word Sense Disambiguation for Afaan Oromo by word sense modeling.

In order to conduct the study the untagged corpus was taken from yehuwalashet [1]. *We prepared annotated corpus by implementing POS tagging on the data. A total corpus of 424397 words for WSM and* 29845 *words for POS tagging with 20 ambiguous words were used to test the system. For POS tagging purpose NLTK and Python Programming were used and to run the WSM system Java Neatbean were used. Different preprocessing tasks such as Tokenization, stop word removal and normalization were applied on both unannotated and POS tagged annotated corpus to make them ready for the experiment.*

The experiments were done with two clustering algorithms: EM and K-means and one to three context window size. Experiment results show that using annotated corpus for both approach improved the performance of the system. ML approach with EM algorithm achieved 74.85% for annotated corpus and 70.35% for unannotated one. Hybrid approach with k-means algorithm scored 79.1% for annotated corpus and 74.85% for unannotated corpus. EM algorithm generated error results for hybrid approach. The result showed that using annotated corpus improves the WSM system of Afaan Oromo Words and hybrid approach of WSM system performed good using POS annotated corpus for Afaan Oromo words .

List Of tables

List of figures

List of abbreviation

BNG	-	British National Corpus
CSV	-	Comma Separated Value
EM	-	expect Maximization
IA	-	Inter Annotator
ML	-	Machine Learning
NLP	-	Natural Language Processing
NLTK	-	Natural Language Toolkit
NLU	-	Natural Language Understanding
POS	-	Part of Speech
VSM	-	Vector Space Model
WSD	-	Word Sense Disambiguation
WSI	-	Word Sense Induction
WSM	-	Word Sense Modeling

Contents

CHAPTER ONE

Introduction

1.1 Background

In Human language there is a problem of ambiguities of words. Which means, depending on the context in which they occur the words of a given language can have multiple meanings or they can be understood in different ways. Humans understand the specific meaning that words take on in a context even without thinking about the ambiguities of the words. But, computers need to process unstructured textual information and transform them into data structures which must be tested in order to determine the right meaning [1].

Word sense induction (WSI) is the task of automatically discovering all senses of an ambiguous word in a corpus. The inputs to WSI are instances of the ambiguous word with its surrounding context. The output is a grouping of these instances into clusters corresponding to the induced senses. WSI is generally conducted as an unsupervised learning task, relying on the assumption that the surrounding context of a word indicates its meaning [2].

WSI is related to but distinct from word sense disambiguation (WSD) [2]. Word Sense Disambiguation (WSD) is the task of automatically identifying the correct meaning of a word that has multiple meanings [3]. In WSD meanings are referred to as senses, or concepts, which are obtained from a sense-inventory. The ambiguous word is referred to as the target word and the context in which the target word is used is called an instance [3]. WSD seeks to assign a particular sense label to each target word instance, where the sense labels are known and usually drawn from an existing sense inventory like WordNet. Although extensive research has been devoted to WSD, WSI may be more useful for downstream tasks. WSD relies on sense inventories whose construction is time-intensive, expensive, and subject to poor inter-annotator agreement. Sense inventories also impose a fixed sense granularity for each ambiguous word, which may not match the ideal granularity for the task of interest. Finally, they may lack domain-specific senses and are difficult to adapt to low-resource domains or languages. In contrast, senses induced by WSI are more likely to represent the task and domain of interest. Researchers in machine translation and

information retrieval have found that predefined senses are often not well-suited for these tasks, while induced senses can lead to improved performance [2].

Researchers worked on Afaan Oromo Word Sense Disambiguation using different approaches of Word Sense Disambiguation. WSD system which is rely on unsupervised approach consists three modules, namely preprocessing module, sense cluster module and sense disambiguation module [4]. In this thesis we investigated application of syntactic feature (Part of Speech Tagging) for Afaan Oromo Word Sense Modeling (sense cluster).

1.2 Statement of the Problem

Afaan Oromo belongs to the Cushitic branch of Afro-asiatic language phylum. It is probably the third-most widely spoken Afro-asiatic language in the world, after Arabic and Hausa. Ethiopia, Afaan Oromo is used as a day to day means of communication by a great many people of various ethno national groups other than its native speakers [5].

In Afaan Oromo there are homonym words that have the same form but have different meanings. As a result it is difficult to understand the meaning of those words in a given context. Therefore Word Sense Clustering, which can be an input for Word Sense Disambiguation for Afaan Oromo is needed to have a clear understanding of word in the language. In addition, due to ambiguity of words in Afaan Oromo, retrieving Afaan Oromo documents, translation of Afaan Oromo documents to other languages, text processing, and speech processing and grammar analysis are not easy tasks. To solve these problems Word Sense Modeling (clustering) for Afaan Oromo is needed [1].

Some researchers have conducted research [3] [6] [1] on Afaan Oromo WSD. Tesfa [3] has used 1240 sense examples for selected five ambiguous words. Feyisa [6] has used 1500 Afaan Oromo examples for seven target words. Yehuwalashet [1] has used 62986 sentences for twenty ambiguous words. The researchers who have conducted research on this area did not use any POS tag information. However, other researcher [7], showed that use of POS tagging improves the performance of WSD. Therefore, absence of POS tag information in Afaan Oromo WSM system might have resulted with low performance. Because, Part of speech tagging features are useful in capturing the local context of the target word [8].

Part of speech tagging is the process of assigning a part of speech or other syntactic and or grammatical class marker to each word in a corpus [9]. Gaustad [9] presented an application

oriented evaluation of three Part of Speech taggers; Hidden Markov Model tagger, Memory-Based tagger and Transformation-based tagger in a word sense disambiguation (WSD) system. Following the intuition that high quality input is likely to influence the final results of a complex system, he tested whether the more accurate taggers also produce better results when integrated into the WSD system. He used a stand-alone evaluation of the POS taggers to assess which tagger is the most accurate. The results of the WSD task, computed on the training section of the Dutch Senseval-2 data, including the POS information from all three taggers show that the most accurate POS tags do indeed lead to the best results. Which means highly accurate input into a WSD system is producing better results than qualitatively lesser input [9].

Biruk [7] worked on the investigation of Application of POS tag information on the development of WSD prototype model. He did experiment on supervised unsupervised and semi supervised approach by using clustering and classification algorithms. He concluded that addition of POS tag information on the corpus used for semi-supervised machine learning algorithm has been found to yield better performance score unlike supervised and unsupervised machine learning methods.
Word Sense Clustering is generalized work than word sense disambiguation (WSD) and it can produce well clustered document inputs for WSD, information retrieval (IR), machine translation (MT) and other NLP tasks [2]. Based on the insights we get from above works, we have conducted different experiments to see the improvement on the performance of Afaan Oromo WSM by adding POS tag information.

1.3 Research questions

Based on the above justifications of the problem, the following research questions were answered at the end of this research.

- What happens on the performance of the WSM system of Afaan Oromo words if POS tag information is added?
- Which approach of WSM system performs good using POS annotated corpus for Afaan Oromo words.

1.4 Objective of the study

1.4.1 General Objective

The general Objective of study is to investigate the application of Part of Speech Tagging on Word Sense Modeling of Afaan Oromo Words.

1.4.2 Specific Objectives

To achieve the general objective, the following specific objectives are addressed:

- To review related works in order to have a conceptual understanding of the problem area of WSM and the methods that are used to solve the problem;
- To prepare dataset that are used as a source for the system;
- To train WSM model with machine learning and the hybrid approach using the prepared data set;
- To conduct experiment and test the performance of the system;
- To draw conclusions and suggest recommendations for future work.

1.5 Scope and limitation of the study

We decided to use machine learning (unsupervised method) and hybrid approach (unsupervised with rule based). Because, the aim of this thesis is to investigate the application of POS tagging on WSM system of Afaan Oromo Words built by [1]. As, lack of time and difficulty to get volunteers linguists we used 23033 manually tagged words by 3 linguists and 6812 words that was taken from [10] and using 35 identified tagset .using those tagged words we tagged 424,397 words with brill tagger for the purpose of training and testing the model. Since, there is scarcity of annotated corpus for the language the target words are limited to twenty words.

1.6 Significance of the study

The result of this study is expected to produce experimental evidence that demonstrate the application of Part of Speech Tagging on Word Sense Modeling System of Afaan Oromo words. So this research resulted into a method/knowledge of how to develop a well performing WSM and provides a usable Part of Speech tagged and sense Clustered corpus. For the academic community we could revile that Part of Speech tagging contributes for the betterment of Word Sense Modeling and Disambiguation system.

1.7 Methodology

Research methodology is a systematic way to solve a problem. It consists of procedures and techniques for conducting a study [11]. To achieve the objective of this research quantitative experimental methodology is applied. The tools and techniques that are employed in this study are described below.

1.7.1 Literature Review

Various literatures that are considered to be relevant for the study are reviewed to get better understanding of the area and to have detailed knowledge on the various techniques that are essential for Afaan Oromo Part of Speech tagging and Word Sense clustering systems.

1.7.2 Data/corpus preparation

Corpus is a collection of text, speech documents, etc .It can be a text ,speech or documents where each word in the sentence is attached with linguistic information [10] .The corpus we used in this study was collected from previous work [1]. It contains 64503 lines of sentences and 424,397 words. As yehuwalashet [1] stated the sentences were collected from Department of Afaan Oromo, Institute of Oromia Radio and Television, Bulletins and newspapers. From total corpus 23033 words were selected randomly to be tagged manually by linguists and 6812 words had taken from [10] . We used totally 29845 words manually tagged sentences to tag the total corpus using brill tagger. As shown in table 1.1 ambiguous words that were used in the study were twenty in number.

No	Ambiguous Words	Possible Senses	Defined Number of Senses
1	Sanyii	Seed / Type	2
2	Karaa	Road / Way	2
3	Ulfina	Weight / Respecting	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	3
19	Mirga	Direction / Human right / Brave	3
20	Waraabuu	Hyena / Fetch / Record	3

Table 1.1: list of Ambiguous words used for testing with their respective senses.

1.7.3 Tools and Techniques

To perform this experiment an open source Natural language Tooklit (NLTK) and python programming languages which was implemented on python 2.7.13 were used tagging the sentences. Since The WSM model for Afaan Oromo words was adopted from previous work, the algorithm was implemented in java netbeans 8.2 which run on the tagged corpus and the clustering were performed in weka 3.8 tool.

In this study the investigation of application of POS tagging on WSM of Afaan Oromo words was conducted. To tag the sentences that were used for training purpose we used improved brill tagger from [10] work. Brill tagger tag the words based on rules, or transformations, where the grammar is induced directly from the training corpus without human intervention or expert knowledge. The only additional component necessary is a small, manually and correctly annotated corpus (the training corpus) which serves as input to the tagger. The system is then able to derive lexical/morphological and contextual information from the training corpus and learns how to assume the most likely part of speech tag for a word. Once the training is completed, the tagger can be used to annotate new, unannotated corpora based on the tag set of the training corpus [10]. Since our study uses hybrid techniques, we choose clustering algorithms from Weka tools. The algorithms used for this research were best performer algorithms from previous work. They were simple k-means algorithms, which represent simple, hard and flat clustering methods and the Expectation Maximization algorithms also known as the EM which is probabilistic clustering algorithms to generate hierarchical clustering where clusters are described probabilistically.

1.7.4 Procedure

Procedure demonstrates the intended methodology of the major works and the flow of activities. In our work we have started by tagging the corpus using Brill tagger. To test POS tagging we used incremental and accuracy measurement approach .Then, we continued with clustering the extracted context terms using WSM model of Afaan Oromo words that was built in previous work. The extraction of the context terms was performed using machine learning and hybrid approach using unannotated corpus and annotated corpus. After that, the best performers clustering algorithms were used to cluster the senses. Based on the result, we did comparison of Machine learning and hybrid approach and results found when using annotated corpus and unannotated corpus.

1.8 Organization of the thesis

This thesis is organized into five chapters. The first chapter is introductory part of the thesis which includes statement of the problem, research question objective of the study, scope of the study and significance of the study. The second chapter is the literature review part. In this chapter the conceptual review of WSI and WSD, Approaches in WSD and review of related works in the area

of WSD for local languages are summarized. The third Chapter presents WSM system design and Architecture. Under this chapter the flow chart of brill tagger, the architecture of Afaan Oromo WSM system, Methods in the methodology and the WSM system of Afaan Oromo design are covered. The forth one discusses about Experimentation procedure, results and findings. And the fifth chapter draws conclusions and forwarded recommendations.

CHAPTER TWO

Literature Review

2.1 Basic Concepts of Word Sense Induction and Word Sense Disambiguation

Word sense induction (WSI) is the task of automatically discovering all senses of an ambiguous word in a corpus [2]. It is a modeling system which uses unsupervised approach. The inputs to WSI are instances of the ambiguous word with its surrounding context. And, the output is a grouping of these instances into clusters corresponding to the induced senses. Which means, senses represented by token clusters. Senses induced by WSI are more likely to represent the task and domain of interest. They can lead to improved performance for IR, MT and other NLP works [2].

Word sense is a definition or meaning of a word [1]. In natural language many words have more than one meanings, and to determine the meanings the context in which the word is found in is used. The automated process of recognizing word sense in context is known as Word Sense Disambiguation (WSD) .WSD is the process of selecting the appropriate meaning or sense for a given word in a document. It is also a task that automatically assigns a meaning, selected from a set of pre-defined word sense to an instance of polysemous word in particular context [6]. The WSD task necessarily involves two steps. The first step is the determination of all the different senses for every word relevant (at least) to the text or discourse under consideration. This is followed by the design of a means to assign each occurrence of a word to the appropriate sense [12].

Disambiguating word senses has the potential to improve many NLP tasks, such as machine translation, information retrieval, question-answering, text categorization and speech synthesis [13]. Although WSD is an important task it is challenging problem in the area of NLP, because of basically two reasons. First, dictionary-based word sense definitions are ambiguous. Even if trained linguists manually tag the word sense, the inter-agreement is not as high as would be expected. That is, different annotators may assign different senses to the same instance. Second, WSD involves much world knowledge or common sense, which is difficult to verbalize in dictionaries [14].

2.2 Approaches in Word Sense Disambiguation

Different approaches have been used through the evolution of WSD research. Many approaches have been proposed for assigning senses to words in context. Currently, there are three main methodological approaches in this area: knowledge-based, corpus-based and hybrid approach [12]
.

2.2.1 Knowledge-based Approaches

Under this approach disambiguation is carried out using information from an explicit lexicon or knowledge base. Since corpus based approaches require considerable amount of work to create a classifier for each word in a language, as a result researchers tend to work on few words. Knowledge-based approaches use an explicit lexon like, Machine Readable Dictionaries (MRD), thesauri, computational lexicons such as Word Net or (hand-crafted) knowledge bases as information source to resolve lexical ambiguities for many words [15].

Lesk [16] created knowledge bases which associate each sense in a dictionary with a signature composed of the list of words appearing in the definition of that sense. Disambiguation was accomplished by selecting the sense of the target word whose signature contained the greatest number of overlaps with the signatures of neighboring words in its context. Because of the fact that dictionaries are created for human use, not for computers, there are some inconsistencies [16]. 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 [17]. 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 [12]. And this category then determines the correct senses that are used. Similar to machine readable dictionaries, a thesaurus is a resource for humans, so there is not enough information about word relations.

Computational Lexicons are a large electronic database containing useful lexical relations in linguistic Psycholinguistic and computational [18]. Lexicon like Word Net is used for sense evaluation and for similarity measure in WSD. For example [19] 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.

2.2.2 Corpus-Based Approaches

Corpus-based approaches are those that build a classification model from examples. These methods involve two phases: learning and classification [20]. The learning phase consists of learning a sense classification model from the training examples. The classification process consists of the application of this model to new examples in order to assign the output senses. Most of the algorithms and techniques to build models from examples come from the Machine Learning area of AI, such as supervised and unsupervised approach [20] .

2.2.2.1 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 [18]. In supervised disambiguation, a disambiguated corpus is available for training [21]. Usually, the classifier (often called word expert) is concerned with a single word and performs a classification task in order to assign the appropriate sense to each instance of that word. The training set used to learn the classifier typically contains a set of examples in which a given target word is manually tagged with a sense from the sense inventory of a reference dictionary. In most cases, supervised approaches to WSD have obtained better results than unsupervised methods [18].

2.2.2.2 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 [18]. Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck [22], 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].

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 anothermeasure [23].

2.2.3 Hybrid Approaches

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 [24]. Yarowsky [25]used Bootstrapping approaches where initial data comes from an explicit knowledge source which is then improved with information derived from corpora. Bootstrapping approaches He defines a small number of seed definitions for each of the senses of a word (the seeds can also be derived from dictionary definitions or lexicons such WordNet). Then the seed definitions are used to classify the obvious cases in a corpus.

2.3 Review of related works

2.3.1 Local Researches on Word Sense Disambiguation

Tesfa K. [3] 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, pretest has been done by the researchers using sense examples in test set and comparing the result with manually tagged test set. The pre-test has been conducted iteratively to increase prototype's performance. The errors encountered during this experimentation have been corrected and the experiment has been done iteratively until the result is found to be satisfactory. Finally, the actual test was conducted using sense examples in test set. During this process nine fold were used for training the developed system whereas the remaining tenth fold was used for testing the system that was trained on the previous nine folds. The process was repeated ten times by taking other nine as training and tenth one as testing. After each training phase, the system was tested on average of 124 Afaan Oromo sentence. Each of the corresponding training set contains an average of 1116 sentences. The result on test data set was obtained by comparing the result returned by the system with the corresponding test set which was manually tagged. The second experiment 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 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.

Feyisa [10] 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).A total of 1500 Afaan Oromo sense examples were collected for selected seven ambiguous words namely sanyii, karaa, horii, sirna and qophii, ulfina, ifa. Different preprocessing activities were applied on the sense example sentences to make it ready for experimentation. 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%.

Yehuwalashet [1] built context based prototype which gives related meaning of the ambiguous word. He designed and tested a hybrid system by combining unsupervised approach with rule based approach. He showed hybrid method can improve the accuracy of the system and, it is better to use when there is scarcity of training data. He conducted experiment using clustering algorithms with different window size and concluded that the window size for extracting Semantic contexts is window 1 and 2 words to the right and left of the ambiguous word achieved best result. The system yielded accuracy of 76.05% for the unsupervised and 89.47% hybrid approach respectively.

Solomon [23], applied a corpus based WSD so as to acquire disambiguation information automatically. This study reported experiments on five selected Amharic ambiguous words. A total of 1045 English sense examples for the five ambiguous words were collected from British National Corpus (BNC). Solomon used five selected unsupervised algorithms such as Simple k means, EM and agglomerative single, average and complete link clustering algorithms. In this research total of four experiments has been conducted. The first experiment was to check the effect of stemming and stop word removal, the second one was to investigate the effect of different context sizes, the third was to see the effect of sense distribution and the last experiment was to compare the accuracy of selected algorithms. The researcher 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.

Duretti [7] worked on WSD system that identifies a sense of an Amharic ambiguous word by using information from tagged example sentences and Word-Net. The system identifies the sense by measuring similarity between the input sentence and tagged example sentences. Two similarity measures are explored: Cosine similarity and Jaccard Coefficient similarity measure. She has collected 100 example sentences for each sense of the selected Amharic ambiguous words. The Word-Net is composed of words with their synonyms and gloss definition. She has tested the performance of the system using 9 nouns, 3 verbs, 3 adjectives and 2 adverbs, a total 17 words which are selected randomly. The experiments were done for disambiguating one target word in a given text. The experimental step is designed in such a way that, first the performance of Cosine similarity and Jaccard coefficient are checked individually for WSD, next Lesk algorithm is tested

on the third experiment and then experiments were conducted to check the performance of the two similarity measures as combined with Lesk algorithm. The result of their work showed that Jaccard coefficient combined with Lesk algorithm come up with the highest result, which is 89.83% accuracy.

Biruk [7] did research on Application of part of speech tag on the performance of WSD: the case of Amharic. The objective of his work was to investigate the application of corpus with part of speech tagged information for word sense disambiguation of Amharic text. He used 1031 Amharic sentences and five ambiguous words. He did three experiments for supervised, semi supervised and unsupervised approaches. In the first experiment he determined bench mark experimental result. In the second experiment he built WSD prototype using POS tag information to already available corpus and the last experiment has involved integration of POS tag information to already available corpus beyond the second experiment. He used EM and simple K-means from clustering algorithm and Adboosts, ADtree, Bagging, SMO and NaiveBayes from classification machine learning algorithms on Weka 3.6.4 Package. For tagging purpose he used CRF POS tagger model. His findings stated that, incorporation of POS tag information on the corpus used for semi-supervised machine learning algorithm has been found to yield better performance score unlike supervised and unsupervised machine learning methods. The performance improvement while POS information has been added is: 4.2% for ADtree, 8.4% for AdaboostM1, 1.1% for Bagging, 2.5% for SMO and 12.6% for Naïve Bayes while seen in comparison with the baseline score. In addition, it has been seen in his experiment that effect of one seed word to be better unlike that of two or three seed words. Lastly, he concluded that an optimal window size of 6-6 or 7-7 has been found to be enough for WSD using semi-supervised machine learning method using a corpus involving POS tag for each word of the text used during the experiment.

CHAPTER THREE

Word Sense Modeling System Design and Architecture

3.1 Afaan Oromo's brill Tagger

During Transformation based approach the rules or grammar is brought directly from the training corpus. These rules are induced using two major modules of Transformation-based errordriven learning. They are the lexical and contextual rules. Both modules use transformation-based error-driven learning. By using transformation-based error-driven learning, the lexical module produces lexical tagging rules and contextual module produces contextual tagging rules.

As a result, to induce the rules from Afaan Oromo training corpus we used the above two modules that means lexical and contextual rules. When using these rules to tag an unannotated corpus, first Afaan Oromo untagged corpus is given to the preprocessing components. The pre-processing of the untagged corpus is split the corpus into sentences, tokenizing the sentences into words and removal of punctuation marks. After preprocessing is performed the processed corpus is given to Initial State Annotator. The initial state annotator applies to the unknown words (i.e. words not being in the lexicon), then applies the ordered lexi-cal rules to these words. The known words are then tagged with the most likely tag and finally the ordered contextual rules are applied to all words. The Afaan Oromo Brill's taggers is taken from Getachew's work [10].

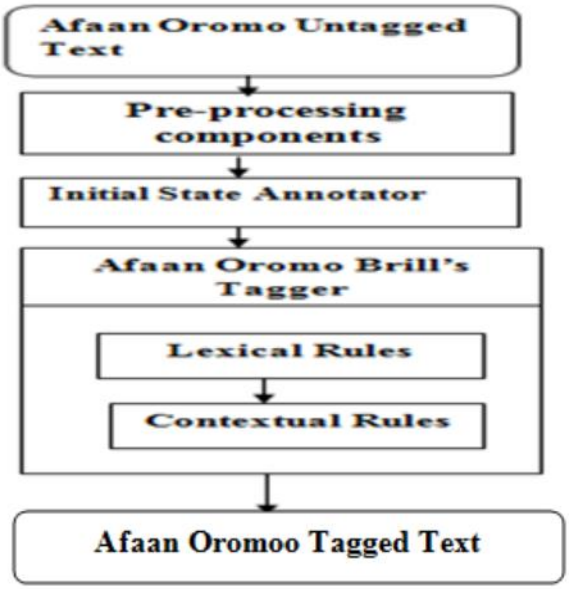


Figure 3.1 Flow chart diagram of Afaan Oromo Brill's tagger (getachew,2016)

3.2 Architecture of the Afaan Oromo WSM

In order to develop WSM model for Afaan Oromo the following steps are processed

- Text preprocessing which take inputs and corpus(unannotated and annotated), tokenize to remove stopwords and perform normalization.
- Extracting context terms providing clue about the senses of the ambiguous term using two techniques (window size and rules),
- 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 of the system with the underlying steps is presented in the following figure.

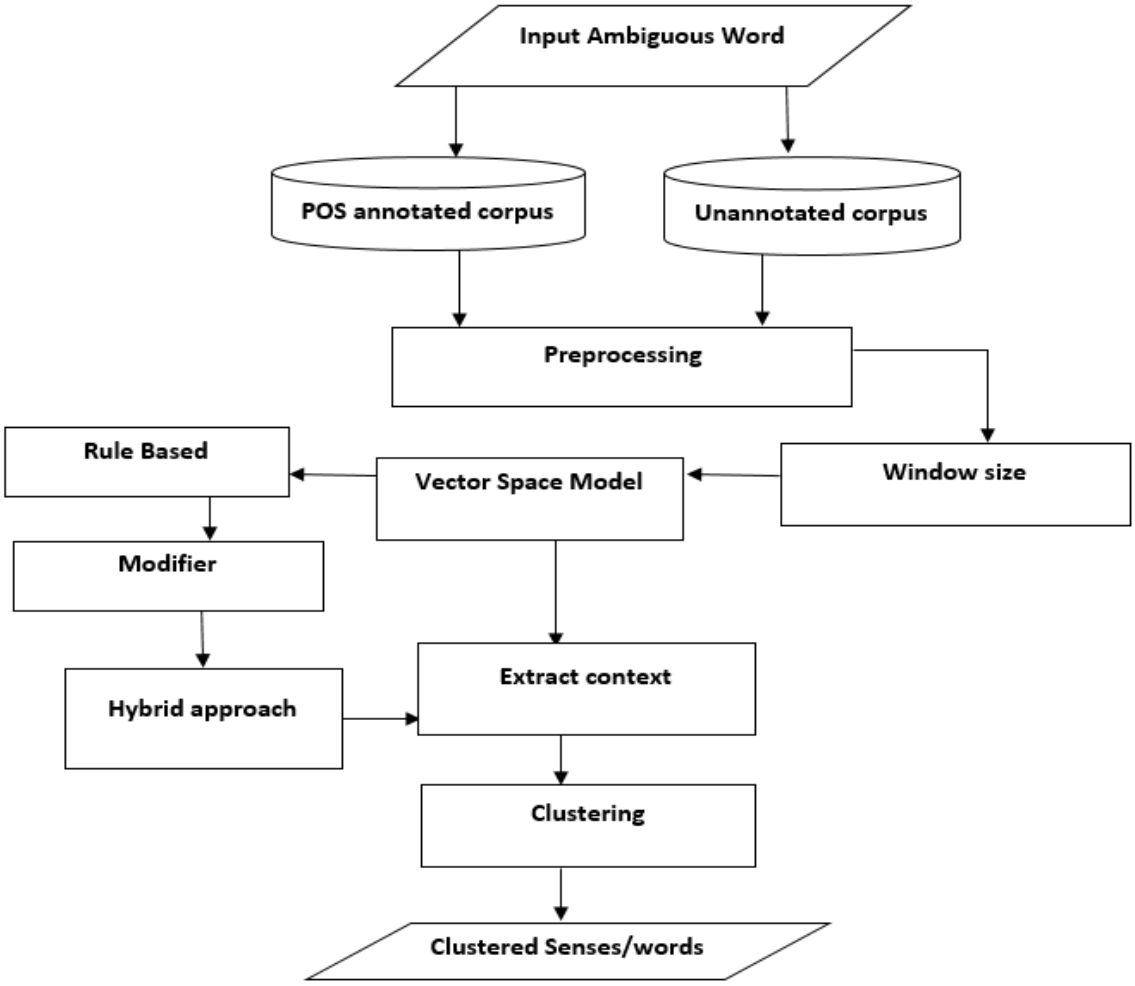


Figure 3.2 Afaan Oromo WSM Architecture

3.3 Methodology

The general methodology of the research is Quantitative Experimental and we followed the following methods in this methodology.

3.3.1 Data collection

We collect the corpus from [1]. He collected the corpus from newspapers, Bible, news (Oromia Radio and Television Organization (ORTO)), magazines, historical documents and bulletins.

3.3.2 Training and testing

The WSM system was trained and tested on unannotated corpus and annotated corpus, which constitutes ambiguous words. The work was tested by the most frequently 20 ambiguous words which were used in the previous work.

3.3.3 Implementation tools

To conduct POS tagging for Afaan Oromo words, an open source Natural Language ToolKit (NLTK) and Python programming language(python 3.6) were used. Both are suitable for processing different NLP tasks. NLTK is an open source tool that contains open source python modules, linguistic data and documentation for re-search and development in natural language processing [26]. Python is an easy to learn but powerful programming language especially for text processing in NLP applications. Using python the module was created and import to NLTK. Python has efficient high level data structures and a simple but effective approach to object-oriented programming [10].

To work with WSM model, the study was used Java NetBeans IDE 8.2 and weka 3.8 package. Java NetBeans IDE 8.2 and weka 3.8 package tool were selected due to the familiarity of the researcher to the tool and because of its accessibility, processing capability and language independent features [1]. And, for clustering purpose we used EM and k-means algorithms. The reason why we chose those algorithms was they performed well in the previous work.

The package has by default Euclidian distance which is used to measure the distance between clusters and Java NetBeans 8.2 programming language is used for preprocessing activities like

tokenization, stop word removal and normalization purpose to make sense example sentences ready for experimentation.

3.3.4 Corpus preparation

This study relied on the patterns learned from two corpus: unannotate corpus the one which was used in the previous work and annotated(part of speech tagged) corpus using unsupervised approach and hybrid (unsupervised with rule based)approach. The idea of this thesis was to investigate the effect of part of speech tagging on word sense modeling of afaan oromo words. for tagging purpose brill tagger is used.Since the study was based on the corpus and target words used in yehualashet's [1]work , we took 9518 words and performed POS tagging manually by three lingustis:

1. Eba Wecho (MA), Ambo University.
2. Fatuma jeldo (MA), Mada Walabu University.
3. Tujube Amansa (PHD candidate), Dambidollo University.

And, added it with and 1517 sentences (6812 words) have taken from Getachew [10].Then, we tagged 424397 words with birill tagger algorithim using 29845 words manually tagged corpus to train the algorthim. The tagged corpus which contains example sentenses has been prepared for sense of 20 selected Afaan oromo words, a total of 10 Nouns, 7 verbs, 1 adjective, 1 adverb and 1 word which is used as a Noun or as a Verb contextually.

Table 3.1 shows the word classes,possible senses and number of possible senses of the 20 selected Afaan Oromo words.

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

Table 3.1 the word classes,possible senses and number of possible senses of the 20 selected Afaan Oromo words.

3.3.4.1 Corpus preprocessing

In the preparation of POS tagged corpus we used manually tagged annotated corpus as an input to train the brill tagger model. And, preprocessing components were implemented on the corpus. Thos components were; sentence splitter , tokenizer and tagset analyizer.The sentence splitter module splits the document into sentences by using Afaa oromo's end punctuation marks like: ',' '?' '!'and '.'. The tokenizer module splits the string into words and punctuation marks. Then they were tagged in the form of 'word/tag' at the time of training phase. The tagset analyzer extracts the tagset from the out put of tokenizer module. This extracted tagset was used for brill tagger to tag a new text.

Then for both corpora tokenization, stopword removal, normalization had performed. In tokenization process a set of sentences were splited in to words using white space. After tokenization process took place, stop words were removed. Stop words are words are words that have no contribution to identify the correct sense of target words. In annotated corpus the POS tags of the words were written in uppercase. In addition,some characters of the same words might be represented in uppercase or lower case in the corpora aswell as in the user input. As a result, we had normalized them into lowercase.

3.4 The Afaan Oromo word Sense Modeling System

In this study two important works are needed to be performed for each corpus: the first one is determining all possible candidate sense words of the ambiguous words and the other one is to group these generated words, each group shows a specific meaning of the ambiguous word. To this end, two kinds of approaches towards the word sense Modeling were used. The first approach is unsupervised machine learning approach. The machine learning approach extracts the two important features (the various contexts of the ambiguous words and their clustering) automatically without supplying linguistic rule. The second approach combined machine learning algorithm and rule based.

Such algorithm is called hybrid approach, which is a mixture of rule based and corpus-based approaches applied in. In the hybrid approach the rule learned from a linguistic feature of Afaan Oromo with the semantic feature learned from corpus using machine learning approaches were combined. Additionally, the manually crafted linguistic rule to extract the various contexts assumed by the ambiguous word was used.Hybrid approach 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. the hybrid approach merges characteristics from both approaches.

3.4.1 Machine learning Approach

In this study two important features need to be extracted using both approaches.

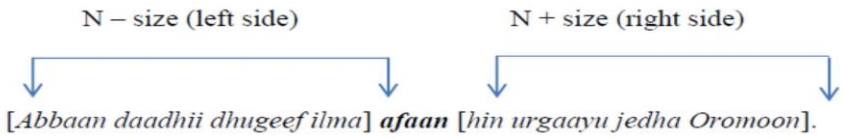
3.4.1.1 Extracting the Context Words

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 depend 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, [27] 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 [28]. 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 2, then there will be 2 words on the left side of the target word and 2 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 used three window sizes in this research. 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 the Contexts

For each context extracted the system constructs a vector space matrix from cooccurrences. 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 3.8 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 complete link, K-means and EM algorithms to merge the contexts based on the nature of the clustering algorithms using cosine similarity.

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, one, two, three 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 [29].

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. for 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 4 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 [30].

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

3.4.1.4 Clustering Algorithms

The clustering algorithms used in this study were partitional clustering include EM and Kmeans clustering. These clustering algorithms have their own unique nature of clustering. The partitional clustering algorithms have applied clustering as its own nature [31]. 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 [32].

3.4.1.5 Hybrid Approach

According to WSD heavily based 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.

3.4.1.6 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 [33].

The most common way of representing the language is a rule [34] . 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 [35]. 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 [35]. The words are described (modified) by the Noun or Verb preceding them.In an annotated corpus the modifiers can be identified as 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 [12].

3.4.1.7 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 [36].

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.

3.4.1.8 Evaluation Method

To our knowledge, there were no previous standard Afaan Oromo word sense Modeling dataset for evaluation. But, since we evaluate our system with previous work, the evaluations were undertaken on the basis of precision achieved in this work. Precision is defined as the percentage of correctly clustered words out the total of generated words [35] . As we explained before the model genarated words which shows possible meanings of the target words.Then those words are classified into a group of a given sense of ambigious word. After that, we checked manually if the words are classified under a given sense of word correctly.

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

CHAPTER FOUR

Experimental Results and findings

4.1 Data Requirements

In this work, three 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.

b) a corpus composed of 29845 words for POS tagging

c) a corpus composed of 424397 words for WSM.

4.2 Experimental procedures

The experiment started with prepocessing activity. After the pre-processing has been completed,the experiment would have been started on the following interface.

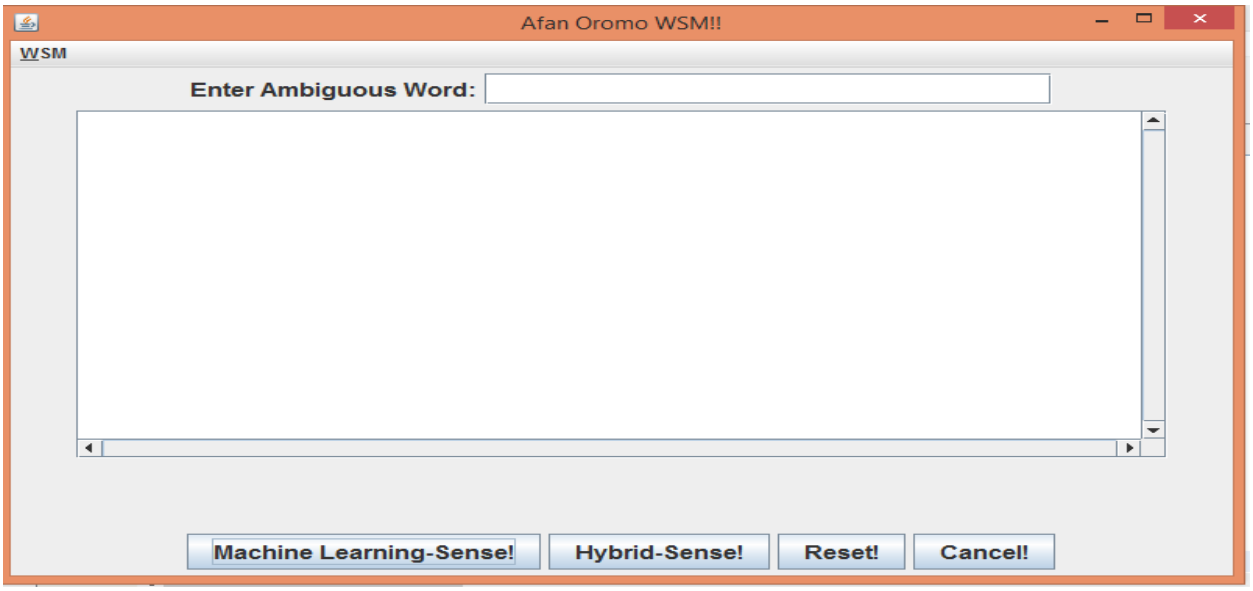


Figure 4.1 User Interface of Afaan Oromo WSM

The system works as follows:

We have to run the program from java NetBeans to display the above interface to work on. After that we assign window size number.Then, we enter the ambiguous word in to the given text box which is found next to the sentence ‘enter the ambigious word’on the interface. Subsequently, we choose the type of WSM techniques (Machine learning sense or Hybrid sense)that’s going to be applied by clicking on the buttons. At that time the system extract the context, by calculating cosine similarity and displaying on white space of the interface for ML approach. And, inaddition to cosine similarity value it adds the given rule to generate the instaances. These results are displayed in the format of CSV which are ready for clustering.

To extract the contexts from the corpus, it takes one ambiguous word at a time in the interface provided and produces the cosine similarity b/n the contexts.

As [1] stated window size for extracting semantic contexts for Afaan oromo words is window 2 words to the right and left of the ambiguous word achieved better result when unannotated corpus is used. As a result,we took window 1 to window 3 ($\pm$1 of the best window which is window 2) for our experiment. We did this experiment for using unannotated corpus too. The two clustering algorithms also have chose to cluster the extracted words because, they performed better result in the previous work [1].

4.3 Experiment on machine learning approach using unannotated corpus

In this section, we have presented the experimental procedure on EM, k-means algorithms and result on three windows size using unannotated corpus. 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 based on the provided window size. In this experiment, window sizes of $\pm$1 up to $\pm$3 words on both sides of ambiguous word have been used. The clustering algorithms were used to cluster contexts which are used to discover semantically related senses. These algorithms used the vector representations (cosine similarity measure) of extracted contexts of the ambiguous word using rules or the window of words as input. The basic idea of sense clustering is that related word senses must be allied and

grouped in the same cluster. Hence, it identifies groups of senses, which are assumed to represent different meanings.

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 distinguished, 12 words with 2 senses, 5 words with 3 senses, 1 word with 4 senses, and 2 words with 5 senses. Three different formations of clustering were run for each word by changing window size from 1-1 to 3-3 for each and every word. The results are presented as follows.

4.3.1 Experiment on the Machine Learning Approach using 1 upto 3 Context Window Sizes and EM clustering 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 2-2 it scored 70.35%.

Ambiguous words	Window Size		
	1-1	2-2	3-3
Horii	77	77	77
Ifa	60	60	60
Karaa	77	77	77
Qophii	60	60	60
Sanyii	77	77	77
sirna	77	77	60
ulfina	60	60	60
Afaan	70	70	70
Bahe	40	50	50
Boqote	85	85	85
Darbe	53	72	62
Diige	60	43	77
Dubbatate	95	95	74
Tume	70	70	60
Haare	70	75	85
Ija	50	60	50
Ji’a	83	83	83
Dhahe	87	70	60
Mirga	70	76	60
Waraabuu	70	70	60
Average	69.55	70.35	67.35

Table 4.1 Experiment result on ML approach using 1 upto 3 context window size,unannotated corpus and EM algorithm.

4.3.2 Experiment on the Machine Learning Approach using 1 upto 3 Context Window Sizes and k-means clustering 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 complete link clustering algorithm. It performed best accuracy, 66.1% with window size 1-1.

Ambiguous words	Window Size		
	1-1	2-2	3-3
Horii	80	80	80
Ifa	70	70	70
Karaa	80	80	80
Qophii	50	50	50
Sanyii	67	67	67
Sirna	67	67	67
Ulfina	67	67	67
Afaan	60	60	60
Bahe	50	50	50
Boqote	75	75	75
Darbe	82	82	82
Diige	60	60	43
Dubbatate	60	60	60
Tume	67	50	67
Haare	75	75	75
Ija	60	40	50
Ji’a	67	50	50
Dhahe	47	47	64
Mirga	61	64	70
Waraabuu	77	77	77
Average	66.1	63.55	65.2

Table 4.2 Experiment result on ML approach using 1 upto 3 context window size,unannotated corpus and K -means algorithm.

The above experiments was done on Machine Learning approach by using two algorithms EM and K-means using window size from one up to three to the left and to the right of the ambiguous words.

As can be seen from table, 4.1 and 4.2 EM algorithm performed better result than K-means algorithm. With window sizes 2-2 the algorthims have yielded accuracy result of 70.35% with EM using window size 2-2 and 65.2% with K-means using window 3-3.

4.3.3 Experiment on the Hybrid Approach using unannotated corpus

The second experiment in this work was on hybrid approach combining unsupervised machine learning and rule based approaches.The hybrid approach relies on both machine learning and rule based algorthms. The rule based algorithms relies on, hand-constructed rules that are acquired from the structure of language rather than automatically trained from data. This approach is recommended when there is a scarcity of data. The rule depend on domain knowledge to bridges the gap caused by data scarcity.

Since, unsupervised machine learning used window size only, the hybrid approach used rules to extract modifiers of the ambiguous word and consider them as contexts. These modifiers are therefore identified according to the developed rule planted. Similar to experiment performed on 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 experiments that have done on hybrid approach.

4.3.4 Experiment on the Hybrid Approach using 1 upto 3 Context Window Sizes and EM clustering algorithm

Most of the experiments we did here were resulted with 100% accuracy, which is false result. This means, weka couldn’t cluster the senses based on the number of cluster we gave to it.it puts all senses as one cluster. For example when we want to cluster the senses of the word “ji’a” into two clusters the final result look like this.

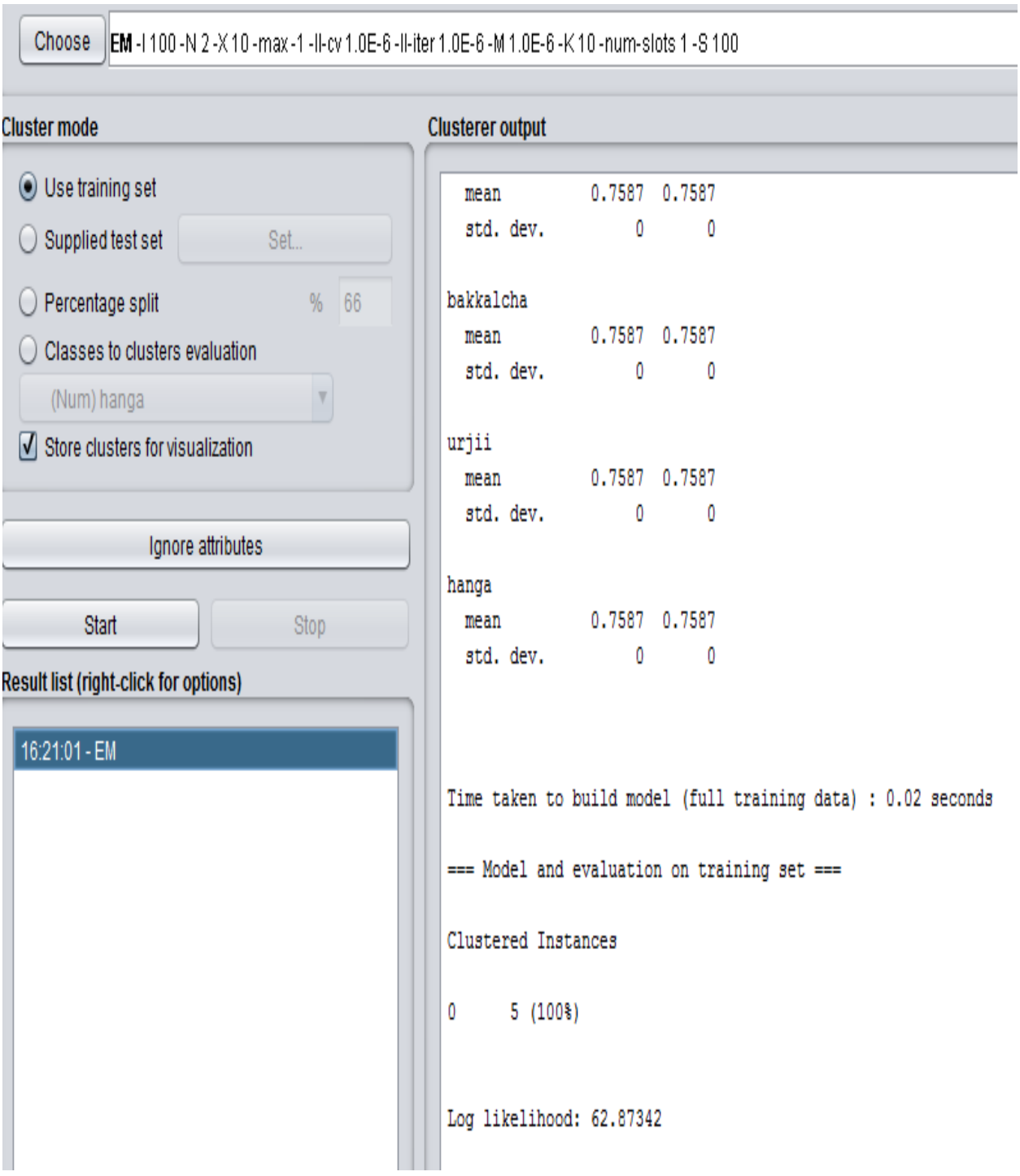


Fig 4.2 experiment result of the hybrid approach for the word "ji'a" using EM algorithm using unannotated corpus with window three.

4.3.5 Experiment on the Hybrid Approach using 1 upto 3 Context Window Sizes and k-means clustering algorithm

As shown on table 4.3 K-means using hybrid approach achieved an accuracy of 74.85% for window size one-one. It achieved better performance when compared with its values on machine learning which was 66.1%.

Ambiguous words	Window Size		
	1-1	2-2	3-3
Horii	75	75	75
Ifa	80	80	80
Karaa	80	80	80
Qophii	60	70	60
Sanyii	80	80	80
Sirna	75	75	70
Ulfina	75	75	75
Afaan	75	75	75
Bahe	50	50	50
Boqote	75	75	75
Darbe	86	86	86
Diige	80	80	80
Dubbatate	75	75	75
Tume	70	80	80
Haare	75	75	75
Ija	60	60	60
Ji'a	80	80	80
Dhahe	70	70	70
Mirga	86	86	86
Waraabuu	70	70	70
Average	73.85	74.85	74.1

Table 4.3 Experiment result on hybrid approach using 1 upto 3 context window size,unannotated corpus and k- means algorithm.

The above experiments was performed on hybrid approach by using two selected algorithms EM, K-means algorithim using window size from 1-1 upto 3-3 to the left and to the right of the ambiguous words.

In this experiment EM algorithm showed error result. And, the results observed from table 4.3 displayed that K-means algorithms showed better performance here (74.85%) when compared with using ML approach(66.1%)the same as the performance shown from Machine Learning approach.

4.4 Experiment using Machine learning approach on annotated corpus

In this section we added Part of Speech tagging annotated corpus to train the the system and performed experiment on both Unsupervised machine learning approach and Hybrid approach. The list of stopwords used in prerpocessing stage should also be POS tagged. In this research, we attempt to answer "Does the addition of Part of Speech tagged annotated corpus, improve the performance of a word sense Modeling system for Afaan Oromo ambiguous words?". As the experiment result shows the obtained result was more robust than the approaches which used unannotated corpus.

4.4.1 Experiment on the Machine Learning Approach using 1 upto 3 Context Window Sizes and EM clustering algorithm

As shown on table 4.4 EM algorithm has shown better result the same as on previous experiment results. It scored best accuracy 74.8% for window 2-2.

Ambiguous Words	Window Size		
	1-1	2-2	3-3
Horii	70	75	70
Ifa	70	75	75
Karaa	80	75	75
Qophii	70	70	70
Sanyii	75	80	75
Sirna	60	80	80
ulfina	60	75	60
Afaan	70	80	80
Bahe	70	70	65
Boqote	75	75	75
Darbe	75	75	75
Diige	70	78	78
Dubbatate	75	78	60
Tume	70	75	73
Haare	70	75	75
Ija	70	75	60
Ji'a	60	70	70
Dhahe	70	70	63
Mirga	70	75	75
Waraabuu	70	70	50
Average	70	74.8	70.2

Table 4.4 Experiment result on ML approach using 1 upto 3 context window size, annotated corpus and EM algorithm.

4.4.2 Experiment on the Machine Learning Approach using 1 upto 3 Context Window Sizes and k-means clustering algorithm

Table 4.5 presented that K-means executed best accuracy of 73.7% for window 2-2.

Ambiguous Words	Window Size		
	1-1	2-2	3-3
Horii	72	70	70
Ifa	80	75	75
Karaa	73	85	77
Qophii	70	80	60
Sanyii	77	70	70
Sirna	60	60	60
Ulfina	70	85	70
Afaan	60	67	70
Bahe	75	77	67
Boqote	50	75	50
Darbe	70	75	62
Diige	66	70	65
Dubbatate	75	75	50
Tume	70	75	60
Haare	75	75	75
Ija	60	70	70
Ji'a	60	75	40
Dhahe	70	70	70
Mirga	60	75	71
Waraabuu	70	70	70
Average	68.15	73.7	65.1

Table 4.5 Experiment result on ML approach using 1 upto 3 context window size,annotated corpus and k-means algorithm.

The experiments above, was done on Machine Learning approach by using two algorithms EM and K-means algorithm using window size from one up to three to the left and to the right of the ambiguous words and POS tagged annotated corpus.

The results presented from table, 4.4 and 4.5 proved that, the performance of EM algorithm was better than the performance of K-means algorithm. With window sizes1-1, 2-2 and 3-3 the algorthims have produced accuracy result of 70%, 74.8% and 70.2% with EM, and 68.15%, 73.7% and 65.1% with K-means.

4.4.3 Experiment on the hybirid approach usingannotated corpus

4.4.3.1 Experiment on the Hybrid Approach using 1 upto Context Window Sizes and EM clustering algorithm

As shown on table 4.7 EM algorithm has shown better accuracy result than k- means and complete link algorithms.

The experiment with EM algorithm produced false results (100%) yet again.

Example: To cluster the possible senses of the word “afaan” into two clusters the result seems like this.

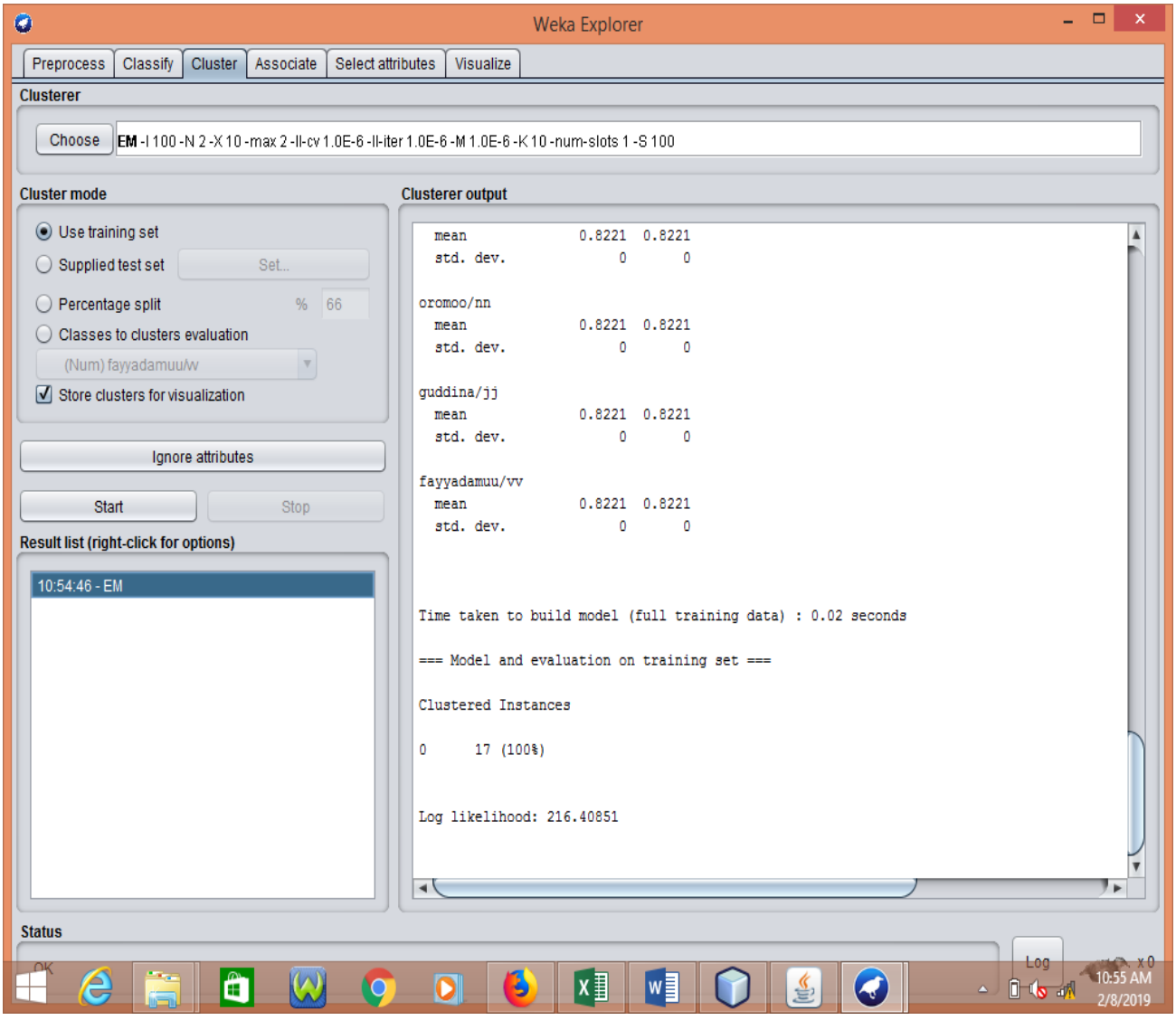


Fig 4.3 experiment result of the hybrid approach for the word “afaan” using EM algorithm using annotated corpus with window three.

4.4.3.2 Experiment on the Hybrid Approach using 1 upto 3 Context Window Sizes and k-means clustering algorithm

As displayed on table 4.6 K-means showed best accuracy of 79.1% for window 2-2. It scored better result when compared with its value scored from ML approach.

Ambiguous words	Window Size		
	1-1	2-2	3-3
Horii	75	80	75
Ifa	77	81	75
Karaa	83	85	83
Qophii	75	80	75
Sanyii	75	81	75
sirna	75	83	75
ulfina	70	82	70
Afaan	75	75	75
Bahe	75	77	73
Boqote	75	75	75
Darbe	75	86	70
Diige	75	82	75
Dubbatate	75	75	75
Tume	75	80	75
Haare	75	75	75
Ija	70	70	70
Ji'a	72	80	70
Dhahe	75	75	75
Mirga	74	80	75
Waraabuu	76	80	76
Average	74.85	79.1	74.35

Table 4.6 Experiment result on hybrid approach using 1 upto 3 context window size,annotated and k-means algorithm.

Under this section experiments was performed on hybrid approach by using two selected algorithms EM and K-means using window size from 1-1up to 3-3 to the left and to the right of the ambiguous words and POS tagged annotated corpus.

The experiment results demonstrated that EM algorithm showed false result and K-means performed 79.1% accuracy result which is better than the ML approach(73.7%).

4.5 Comparision of algorithms using unannotated corpus and annotated corpus with ML approach

The way we tested Machine Learning approach and Hybrid approach using untagged corpus was similar to the way we tested Part of Speech tagged annotated corpus. was similar to the way used in unsupervised machine learning. Though, performance hey have shown was different from each other. To evaluate the accuracy, we considered the precision value they have yielded for each algorithm they were tested with.

When we used unannotated data, for EM algorithm, window 2-2 presented better result which was 70.35% . Window 1-1 yielded better result which was 66.1% for K-means algorithm. When POS annotated corpus was used, window 2-2 again showed better accuracy value(74.85) for EM and window 2-2 yielded better result (73.7%) for K-means algorithm. As a result, the experiment results presented above indicated that, EM algorithm achieved better accuracy value of; 74.7 % when tested with ML approach using annotated data.

Machine learning approach							
Clustering algorithm		unannotated			Annotated		
EM	Window size	1-1	2-2	3-3	1-1	2-2	3-3
	Accuracy value	69.55	70.35	67.35	70	74.85	74.1
K-means	Accuracy value	66.1	63.55	65.2	68.15	73.7	65.1

Table 4.7 Comparision of the algorithms using unannotated corpus and annotated corpus with machine learning approach.

4.6 Comparision of algorithms using unannotated corpus and annotated corpus with hybrid approach

When we come to the hybrid approach, the experiment results shown from the test done on hybrid approach using unannotated corpus and POS tagged annotated corpus with K-means algorithm by applying context window size 1-1, 2-2 and 3-3, the hybrid approach with unannotated corpus achieves accuracy average of; 74.85% and 79.1% with annotated corpus. EM algorithm couldn't classify the words when we use hybrid approach .

Hybrid approach							
Clustering algorithm		unannotated			Annotated		
	Window size	1-1	2-2	3-3	1-1	2-2	3-3
K-means	Accuracy value	73.85	74.85	74.1	74.85	79.1	74.35

Table 4.8 Comparision of algorithms using unannotated corpus and annotated corpus with hybrid approach approach

4.7 Walk-through Using an Example

a) Machine learning approach

For instance assuming the ambiguos word "afaan", unsupervised machine learning WSM with unannotated corpus, assinging 2-2 window size and using EM clustering algorithm extracts the following result .

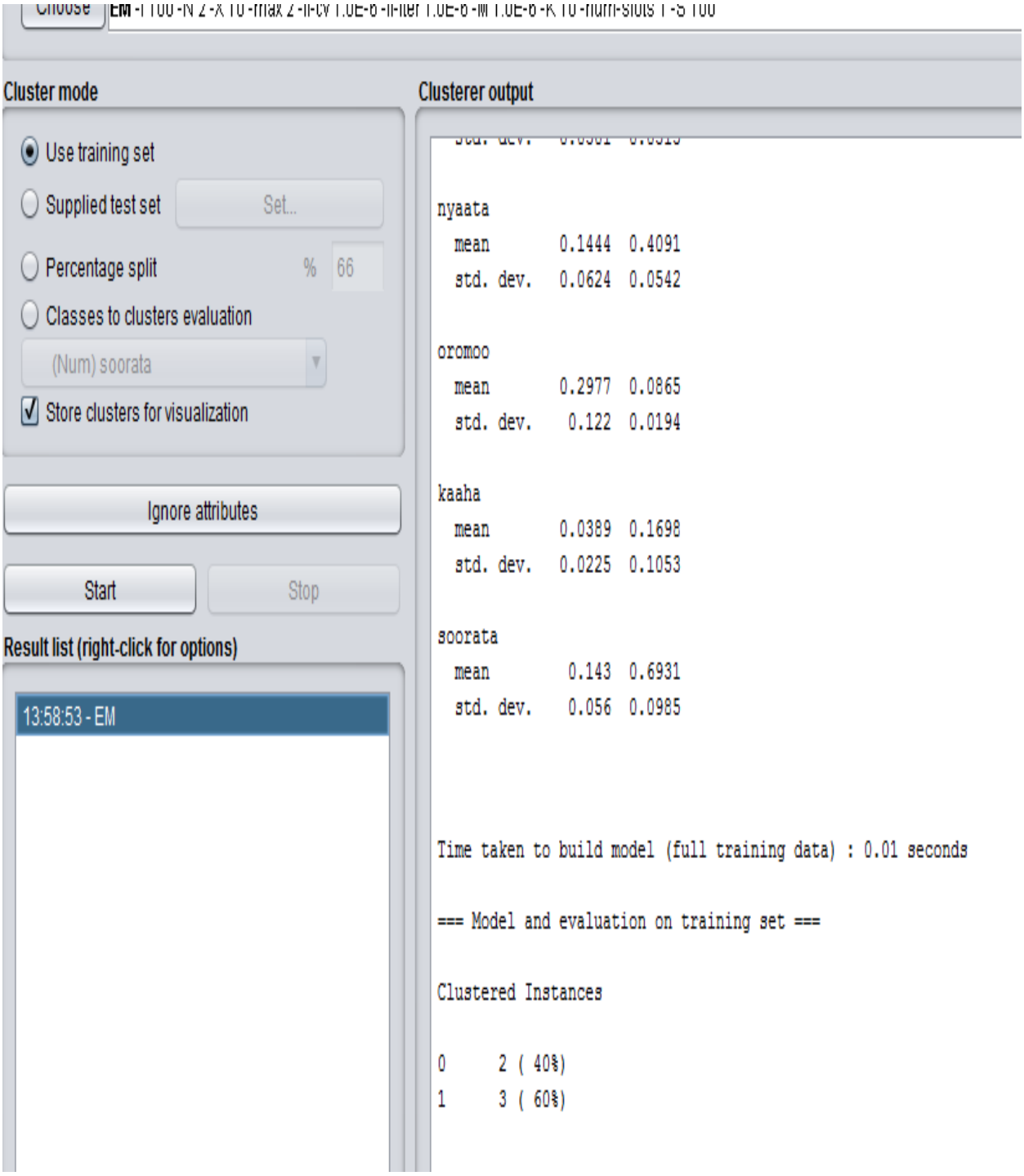


Figure 4.4 Example of WSM using Machine learning approach

b) Hybrid approach

For the same ambiguos word “afaan”, with annotated corpus, assinging 2-2 window size and using K-means algorithm the hybrid approach extracts the following result .

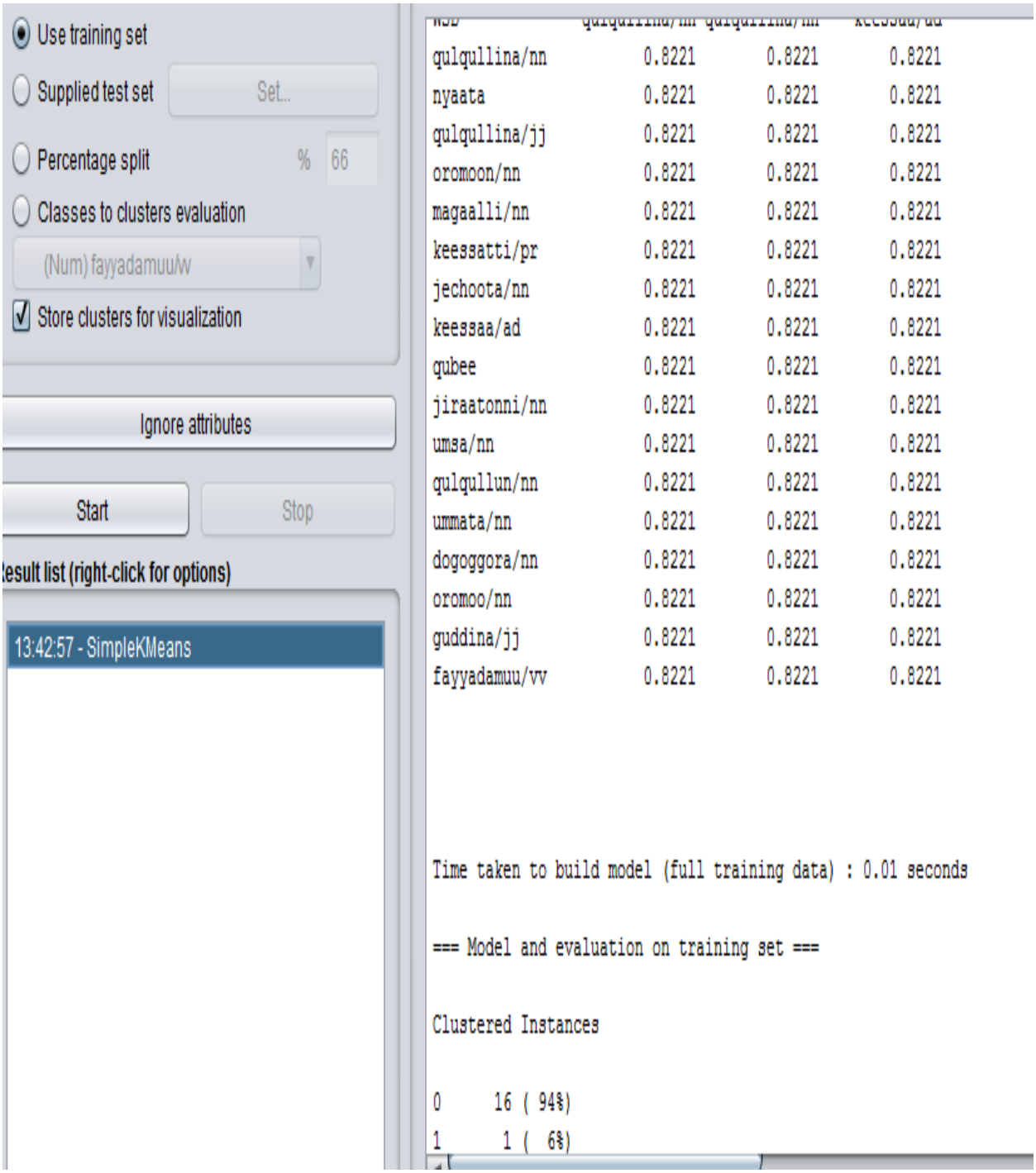


Figure 4.5 Example of WSM using hybrid approach

4.8 Findings and challenges

4.8.1 Findings

In this study unsupervised machine learning approach and hybrid approach for Afaan Oromo were presented to investigate the effect of Part of Speech tagging on Word Sense Modeling of Afaan Oromo words.We described the experiment we conducted to compare the performance achieved from the two approaches by using POS tagged annotated corpus and unannotated corpus.

We evaluated the Word Sense Modeling system with all the 20 ambiguous words. These words have two senses to five senses.

At the end of the experiments we achieved the following results:

- Machine learning approach using annotated corpus for EM and K-means algorithm achieved better result which was 74.85% with window size 2-2 and 73.7% with window size 3-3 when compared with Machine learning approach using unannotated corpus which was 70.35% and 65.2% respectively.
- Hybrid approach using annotated corpus achieved better result with k-means algorithm which was 79.1% when compared with hybrid approach using unannotated corpus which was 74.85% for all window size. But EM algorithm genarated false results for this approach.
- Hybrid approach which used POS tag annotated corpus with k-means algorithm produced greater accuracy value than ML approach that used POS tag annotated corpus. Because, as [1] stated, it used hand-constructed rules that are acquired from lingustis rather than automatically trained from data.

Generally experiments that were done on annotated corpus had relatively higher performance when compared with experiments that were done on unannotated corpus.This proved that, adding POS tagged annotated corpus on WSM system had improved tha performance of the system.

4.8.2 Challenges

The major challenge in conducting this research was getting volenter linguists on part of speech taagging work.In addition there is no annotated data for Afaan Oromo WSM system. Consequently, we found only 29845 tagged words to train and test the Afaan Oromo brill tagger. Another challenging thing was the main corpus which was used in the previous work [1] was not well organized and,it took as long time to reorganize and start part of speech tagging by linguists.

CHAPTER FIVE

Conclusion and Recommendation

5.1 Conclusion

The overall focus of this research was to investigate the effect of Part of Speech tagging on Afaan oromo Word Sense Modeling (WSM), which addresses the problem of deciding the correct sense of an ambiguous word based on its surrounding context's and the modifiers. We used the corpus from the previous work and 20 ambigiuos words which have 2-5 senses to test the Model. NLTK, Python programming language with brill tagger was used to tag the corpus. Manually tagged data was used to train the brill tagger algorithm. To run the WSM model we used java netbeans 8.2 and sense clustering were performed in weka 3.8 tool.

When we performed the experiments on the approaches of WSM system we used untagged corpus and POS tagged annotated corpus discretely.In addition to that, to find out the well performer context window size; 1-1 upto 3-3 window size were used. Preprocessing tasks; stopword removal, tokenisation and normalization were implemented on both corpora. Following that, The system had genarated the possible senses of the target words. These possible related senses should be clustered to generate the correct sense of a given word. Clustering algorithms; EM and K-means algorithms used for clustering purpose.

Then, System is evaluated based on the accuracy value which was generated from those algorithms. At the last, we collected accuracy value from each algorithms and performed comparision between Unsupervised Machine Lerning approach and Hybrid Approach when they used unannotated corpus and POS annotated corpus.

The experiment results presented that, Machine learning approach using annotated corpus with EM algorithm achieved better result which was 74.85% when compared with Machine learning approach using unannotated corpus which was 70.35%. And, Hybrid approach using annotated corpus with k-means achieved better result which was 79.1% when compared with hybrid approach using unannotated corpus which is 74.85%.

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

- ✓ Adding POS tagging to Afaan Oromo Word Sense Modeling will improve the performance of Afaan Oromo Word Sense Modeling.
- ✓ Hybrid approach which used POS tagged annotated corpus is the best approach for Afaan Oromo Word Sense Modeling.

5.2 Recommendations

.

- ➢ Other languages have standard sense annotated data for Word Sense Modeling researches. So, researchers should improve the POS tagged annotated data we used in this research to achieve good improvement on Afaan Oromo Word Sense Modeling system.

- ➢ Afaan oromo lacks knowledge resourses like WordNet, Lexion, machine readable dictionary. Since those resources have high role in WSM researches, reserchers shall work on developing them.

- ➢ For researchers who are interested to work on WSM for other local language we suggest to follow this approachs and investigate more.

- ➢ Using the Afaan Oromo WSM system developed in this research, the reseachers can proceede the work to developing other NLP systems such as: Word Sense Disambiguation, Information Retrieval and Machine Translation .

References

[1] Y. B. Tesema, "Hybrid Word Sense Disambiguation Approach for Afaan Oromo," June 2016.

[2] Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, Clement T. Yu, "A Sense-Topic Model for Word Sense Induction with Unsupervised Data Enrichment," vol. 3, p. 59–71, 2015.

[3] T. Kebede, "WORD SENSE DISAMBIGUATION FOR AFAAN OROMO LANGUAGE," November 2013.

[4] Sruthi Sankar K P, P C Reghu Raj, Jayan V, "Unsupervised Approach to Word Sense Disambiguation in Malayalam," 2015.

[5] T. Degeneh Bigiga, "The development of Oromo Writing System.," 2015.

[6] F. G. Shoga, "Unsupervised Corpus Based Approach for Word Sense Disambiguation to Afaan Oromo Words," Addis Ababa, 2015.

[7] B. Retta, "Application of Part of Speech tagged corpus to improve the performance of WSD:the case of Amharic," 2015.

[8] S. Mohammad, "Combining Lexical and Syntactic Features for SupervisedWord Sense Disambiguation," August 2003.

[9] T. Gaustad, "The Importance of High Quality Input for WSD:An Application-Oriented Comparison of Part-of-Speech Taggers".

[10] G. Emiru, "Developement of Part of Speech Tagger Using Hybrid approuch," october 2016.

[11] D. S. Bekeli, "A Generic Approach towards All Words Amharic Word Sense Disambiguation," 2017.

[12] Nany Ide and Jean Veronis, "Word Sense Disambiguation :The state of art,Computitional Lingustics," vol. 24, no. 1, March 1998.

[13] Andres Montoyo,Armando Suarez,German Rigau,Manuel Palomar, "Combining Kowledge and courpus Based Word Word Sense Disambiguation methods," *Journal of Artificial Intellegence Reseach,* vol. 23, no. 1, January 2005.

[14] Stchutze, Mainning Christopher and Hinrich, "Faundation of Statistical Natural Language Processing," 1990.

[15] J. H. a. M. Kamber, Data Mining :Concepts and Techniques, 2nd ed., University of Illions at urban-Champain, 2006.

[16] M. Lesk, "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.

[17] D. Tatar, "Word sense disambiguation by machine learning approach: a short survey.," vol. XLIX(2)., no. 2, 2004..

[18] R. NAVIGLI, "Word Sense Disambiguation: A Survey," *ACM Computing Surveys,* February 2009.

[19] A. F. Sementon, "Linguistic Approaches to Text Management: An Appraisal of Progress.," *Journal of Document & Text Management,* vol. 2, no. 2, 1995.

[20] G. E. Bakx, "Machine Learning Techniques For Word Sense Disambiguation.," 2006.

[21] Yogita Rani and Harish Rohil, "Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9"," vol. 2, no. 1, 2014.

[22] William A. Gale. Kennet W. Church and David Yarowiski, "A method for disambiguating word senses ina large corpus.," *Computers and the Humanities,* vol. 26, no. 5,6, December 1912.

[23] S. Mekonin, "Word Sense Disambiguation for Amharic Text:A Machine Learning," 2010.

[24] S. Assemu, "Unsupervised Machine Learning Approach For Word Sense Disabgiguation to Amharic Words," June,2011.

[25] D. Yarowsk, "unsupervised word sense disambiguation rivaling supervised methods," June,1995..

[26] Steven Bird, Ewan Klein, and Edward Loper, "Natural Language Processing with Python".

[27] Katrin Erk and Sebastian Padó , "A structured vector space model for word meaning in context.," *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language,* 2008.

[28] H. Schutze, "Dimensions of meaning. In Supercomputing '92:," *Proceedings of the1992 ACM/IEEE Conference on Supercomputing. IEEE Computer Society Press, Los Alamitos,,* 1992.

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

[30] Joseph Reisinger and Raymond J Mooney, "Multi-prototype vector-space models of word meaning," *In Proceedings of HLT-NAACL,* 2010.

[31] M. Hearst, " "Noun Homograph Disambiguation Using Local Context in Large Text Corpora," in Using Corpora," 1991.

[32] Nancy Ide and Jean Véronis, "Word Sense Disambiguation Algorithms and Applications," vol. 33, 2007.

[33] Dagan, Ido and Alon Itai, "Word Sense Disambiguation Using a Second Language Monolingual Corpus,," *Computational Linguistics ,* vol. 20, 1994.

[34] T. Guya, "CaasLuga Afaan Oromoo: Jildii-1,Gumii Qormaata Afaan Oromootiin," 2003.

[35] F. D., "Natural Language Engineering-Efficient Processing with HPSG:Methods, Systems, Evaluation.," 2015.

[36] D. Megersa, "An Automatic Sentence Parser For Oromo Language Using Supervised Learning Technique," 2002.

[37] Ravi Mante, Mahesh Kshirsagar, Dr. Prashant Chatur, "A Review Of Literature On Word Sense," *(IJCSIT) International Journal of Computer Science and Information Technologies,,* vol. 5, no. (2), pp. 1475-1477, 2014.

[38] G. E. Bakx, "Machine Learning Techniques for Word Sense Disambiguation," May 22, 2006.

[39] S. H. Yesuf, "AMHARIC WORD SENSE DISAMBIGUATION USING WORDNET," March 2015.

[40] T. D. Bijiga, "The Development of Oromo writing system," november 2015.

[41] R. Navigli, " Word Sense Disambiguation: A Survey. ACM Computing Surveys,," vol. 41., no. 2, Feburary 2009..

[42] S. Mohammad, "Combining Lexical and Syntactic Features for SupervisedWord Sense Disambiguation," August 2003.

APPENDIXES

Appendix I: Sample of Manually tagged corpus

Boora'u/VV malee/CC hin/NG taliilu/VV ./PN

Oduun/NN soora/NN (nyaata/NN) gurraati/VV ./PN

Waan/JJ biyyaa/NN bilbilli/NN iyya/VV ./PN

Quuqqaa/NN cinaacha/NN jalaa/PRI abbaatu/NN beeka/VV ./PN

Namni/NN mataa/NN isaatii/ PS hin/NG tolle/VV,/PN namaaf/NP hin/NG tolu/VV ./PN

Bakka/NN hin/NG rafnetti/VV hin/NG mugan/VV ./PN

Garaa/NN nagaa/NN qabu/VV sukkuumuun/VV,/PN dhukkuba/NN itti/PR fiduudha/VV ./PN

Ejersa/NN halaalaarra/NP qobboo/NN qe ee/NN wayya/JJ ./PN

Bara/NN bofti/NN nama/NN nyaate/VV lootuun/NN nama/NN kajeelti/VV ./PN

Bareedde/VV jedhanii/VV obboletti/NN hin/NG fuudhani/VV ./PN

Laaftuun/JJ dubaraa/NN obbolessa/NN irraa/PRI ulfoofti/VV ./PN

Malannee/VV bolla/NN lama/JN qotanne/VV jette/VV hantuunni/NN ./PN

Wallaalaan/NN bishaan/NN keessa/PRI dhaabatee/VV dheebota/VV ./PN

Yaadni/NN hamaan/NN nama/NN huqqisa/VV ./PN

Fuula/NN ilaali/VV na/PP elmadhu/VV jette/VV harkuun/NN ./PN

Kan/JJ leenci/NN naasise/VV hindaaqoo/NN baqata/VV ./PN

Hunda/JJ dubbatan/VV garaan/NN duwwaa/JJ hafa/VV ./PN

Eelan/VV eelee/NN dhaabbatan/VV ./PN

Gowwaan/NN bakka/NN rafe/VV hunda/JJ mana/NN se a/VV

Nyaataa/NN fi/CC lola/NN abbaatu/NN tolfata/VV ./PN

Ilmoon/NN huntuutaa/NN gumbii/NN uraa/VV haadha/NN jalatti/PRI barti/VV ./PN

Sossobbiin/VV madaa/NN hin/NG foyyessu/VV ./PN

Harree/NN fi/CC gadadoon/NN nama/NN irraa/PR hin/NG gortu/VV ./PN

Nama/NN jechi/NN hin/NG madeessine/VV waraanniyyuu/NC hin/NG madessu/VV ./PN

Naman/NN garaa/NN ga aatti/VV garaan/NN lafa/NN nama/NN ga a/VV ./PN

Kan/JJ dhiqantu/VV nama/NN dhiqa/VV jatte/VV waciitiin/NN ./PN

Namni/NN mana/NN tokko/JN ijaaru/VV citaa/NN wal/PRE hinsaamu/VV ./PN

Ariifataan/NN horii/NN gata/VV ./PN

Appendix II: Sample of unannotated corpus

Ilmaan koo biyya fagoodhaa, intaloonni koos andaara lafaatii haa dhufan! Warri maqaa kootiin waamaman hundinuu, warri ani kiiloo ulfina kabaja kootiif uummadhe, warri ani tolchee bifa itti godhe haa dhufan!' jedhe. Waaqayyo, 'Yaa saba nana, isin dhuga-baatuu koo ti, garbichi ani fo'adhes isinuma. . . . Saba kana ofii kootiif uummachuun koo, akka inni na galateeffatuuf.

Obboloonni tokko tokko kana gochuurraa duubatti jedhaniiru. Kana gochuu dhiisuuf mormiiwwan gara garaa kan kaasan yoo ta'eyyuu, rakkinnisaanii inni guddaan tajaajilli manaa gara manaa kiiloo ulfina kabaja keenya nu jalaa hir'isa jedhanii yaaduusaanii ture.

Kun ta'us ta'uu baatus, yommuu biyyi lafaa kun badu, maqaan Waaqayyoofi Mootummaansaa kiiloo ulfina kabaja sanaan dura argatee hin beekne yommuu argatu ija keenyaan ilaalla.

Furaan saanii adiin seenaa darbe muxannoo hawaasaa fi hariiroo akaakilee waliin inni qabu; diimaan yeroo fi haala amma keessa jirru, si'oomina, qabso fi wareegumma hawaasa irraa eegamu; gurraachi egeree hin beekamne kan abdiin dorobee hawaasi dharra fi kiiloo ulfina kabaja guddaan eegatu jechuu dha.

Gulantaa mallatoqaama ulfina kabaja bultii

Kiiloo ulfina kabaja argisiisuun, obboloota keenya kabajuu kan dabalatudha.

Keessumaa namoonni gaa?ela godhachuuf yaaduudhaan jaalalaan walitti dhihaatan kiiloo ulfina kabaja waliif qabaachuunsaanii barbaachisaadha.

Bohaartiiwwan biyya lafaa, mataan maatii tokko akka itti ga?isamuufi kiiloo ulfina kabaja akka hin arganne gochuudhaan namoonni akka gammadan godhu.

Akka warra kaaniif kiiloo ulfina kabaja akka kenninu argisiisuu kan dandeenyu akkamitti

Haati manaasaa nama Waaqayyoo taate, waan isa gargaartuufi kiiloo ulfina kabaja guddaa waan isaaf kennituuf dhugumaan jaallatamtuudha.

Hunda caalaammoo gaa'ellisaanii inni fakkeenya ta'e, Yihowaa isa galanni isaaf maluuf kiiloo ulfina kabaja argamsiisa.

Gaafa morkaa kiiloo ulfina kabaja guddaa dhiirarraa arganna jechuun ibsu.

Sirna gadaa keessatti dubartiin ulfaa akkasumas deessun baay'ee sodatamtuudha. Dubartii ulfaa hintuqan. Qoraan cimallee ishii biratti hin baqaqsan. Yoo itti dheekkamanis handaara wayaatiin dhaanan malee miidhaa hamaa irraan hin gahan. Walumaagalattii namni dubartii ulfaa tuqe sa'a qala.

Appendix III: Sample of Stopwords

uumuu

irra

keessa

kana

yoo

ofirratti

jiraate,

abdatu

milkiqabeettiin

dhitamu

irratti

akeekkatee

dabalata.

qofa

noolaa

osoo

abba

abbicha

gad

uggee

jiru

guruu

ka'u,

hundi

warri

tokkee

unkutaawa.

ta'u

godhee