ORIGINAL ARTICLE

Towards the Sense Disambiguation of Afan Oromo Words Using Hybrid Approach(Unsupervised Machine Learning and Rule Based)

Workineh Tesema[1], Debela Tesfaye[1] and Teferi Kibebew[1]

Abstract

This study was conducted to investigate Afan Oromo Word Sense Disambiguation which is a technique in the field of Natural Language Processing where the main task is to find the appropriate sense in which ambiguous word occurs in a particular context. A word may have multiple senses and the problem is to find out which particular sense is appropriate in a given context. Hence, this study presents a Word Sense Disambiguation strategy which combines an unsupervised approach that exploits sense in a corpus and manually crafted rule. The idea behind the approach is to overcome a bottleneck of training data. In this study, the context of a given word is captured using term co-occurrences within a defined window size of words. The similar contexts of a given senses of ambiguous word are clustered using hierarchical and partitional clustering. Each cluster representing a unique sense. Some ambiguous words have two senses to the five senses. The optimal window sizes for extracting semantic contexts is window 1 and 2 words to the right and left of the ambiguous word. The result argued that WSD yields an accuracy of 56.2% in Unsupervised Machine learning and 65.5% in Hybrid Approach. Based on this, the integration of deep linguistic knowledge with machine learning improves disambiguation accuracy. The achieved result was encouraging; despite it is less resource requirement. Yet; further experiments using different approaches that extend this work are needed for a better performance.

Keywords*: Afan Oromo, Ambiguous Word, Hybrid, Rule Based, Word Sense Disambiguation.*

---

[1]Jimma University, College of Natural Sciences, Department of Information Science, Jimma, Ethiopia

INTRODUCTION

In today's world, where World Wide Web technology is keeping on growing very fast, many users go to the web to search for information, for entertainment, to read documents and electronic books. Sometimes it is observed that the result of a search is not appropriate. The reason behind is, there is an ambiguous word in the query (Salton, 2015). Information Retrieval (IR) can potentially benefit from the correct senses of words provided by Word Sense Disambiguation *(*WSD). Ambiguity is a cause of poor performance in IR systems. The queries may contain ambiguous words (terms), which have multiple senses (Ide and Jean, 2010). Therefore, the objective of word sense disambiguation (WSD) is to identify the correct sense of a word in context. It is one of the most critical tasks in most natural language processing (NLP) applications, including information retrieval, information extraction, and machine translation.

Natural languages have ambiguous words which need to be disambiguated and thus the appropriate sense of an ambiguous word in a given context can be identified. Each word may have more than one sense that is why a single word sometimes can have many senses (Sreedhar *et al*., 2012). An ambiguous word can take several senses depending on the context in which it appears. The same form and pronunciation can take different senses in different contexts (David and Radu, 2002).

This work assigns the sense and accomplished by using major sources of information contained within the context (Adam, 2007). All disambiguation work for Afan Oromo was used only machine learning either supervised or unsupervised methods. Still a few researches on sense disambiguation for this language are focus only on corpus information which is less accurate and limited to disambiguate few ambiguity words. However, in this study, we have used the context information derived from the corpus and ruled based approaches. One of the good news for the Afan Oromo disambiguation is that the developed system was open to take any ambiguity words provided by users unlike other systems which are limited to a few words. It was integrated deep linguistic knowledge with algorithms markedly improves disambiguation accuracy. Also argue that to find senses in Afan Oromo, this work was significantly increased word sense disambiguation performance and applicable for different applications like machine translation. This sense disambiguation has been used for many NLP applications and pursued as a way to improve retrieval systems, and generally get better information access. The motivation behind this study was to allow the users to make ample use of the available technologies.

In Afan Oromo identifying the correct senses of the ambiguous words is easy for human being, basically, sometimes it is difficult. However, it is too tough for the machine to identify the correct sense of these words. Nowadays, as the development of technology is increasing rapidly, like Afan Oromo has also started to use the technology for different purposes (Nancy and Jean, 2007). Like other natural languages in Afan Oromo there exists same form of words, which has more than one different meaning. The challenge is to find out which particular sense is appropriate in a given context. Sense disambiguation is the problem of determining which sense of a word is active in a particular context (Agirre and Martinez, 2000). Additionally, the machines have no ability to decide such an ambiguous situation unless some practices have been planted into the machines' memory (Shaikh Samiulla, 2013). Here the bag of words is the context that considered as words in some window surrounding the ambiguous word in terms

of a distance it has from ambiguity. The surrounding word of the ambiguous word decides correct choice of word. Assume, the ambiguity word *afaan* has surrounded with the following contexts. So, these contexts give us the clue of information what the sense of this ambiguity word is. It seems that this ambiguity word identifies its senses with the help of its contexts as shown the in Figure 1 below:

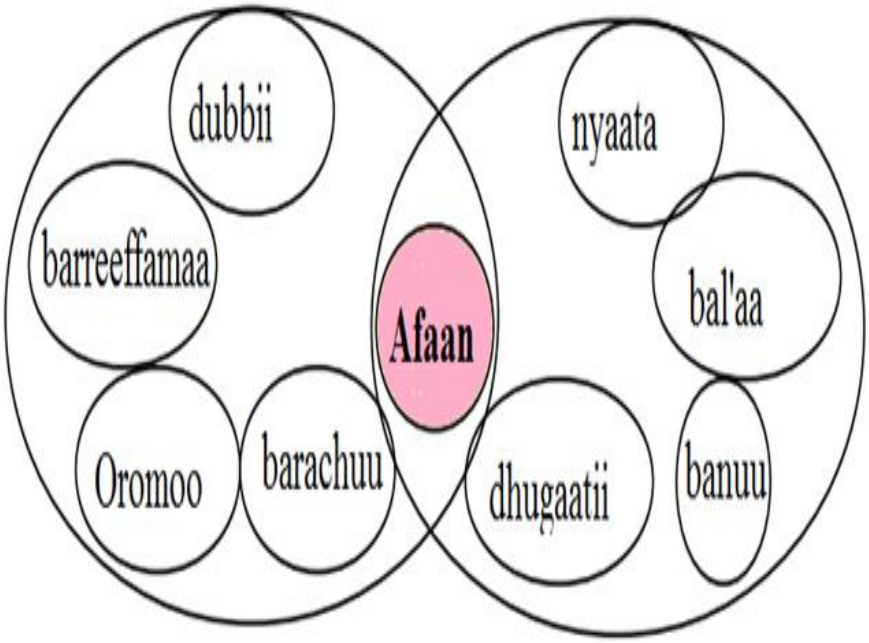


Figure 1. Contexts Identification

The work done by (Kebede, 2013) on Word Sense Disambiguation for Afaan Oromo language is based on an annotated corpus. However, the performance of the system was limited to only five ambiguities words (such as *sanyii*, *karaa*, *horii*, *sirna* and *qoqhii*). Hence, he manually annotates the corpus, the approach was exposed to a bottleneck of training data because of there is no standardized labeled dataset for this language. Additionally, his work was only relied on the machine learning approach which its performance was not much surprise unlike hybrid approaches.

However, in our case the system was open to disambiguate any ambiguity word given by users rather than limited to few ambiguity words. Since, the developed corpus was large in size it is suitable to capture the surrounding *N* contexts to left and right unlike the supervised approach. Additionally, different algorithms and developed hybrid (rules + machine learning) are used because it overcomes the

limitation of algorithms by rule based (linguistic knowledge). This makes integration of different algorithms and rules which was suitable for Afan Oromo which has lack of tools and resources. Therefore, unlike the study conducted by Tesfa, our study focused on hybrid approaches which is unsupervised machine learning and ruled based to overcome the limitation of machine learning.

The unsupervised machine learning for sense disambiguation was trained on unannotated corpus, rivals that the performance of supervised techniques which requires time-consuming hand annotations (Yarowsky, 2007). So, unsupervised Word Sense Disambiguation method is based on unlabeled corpora, and does not exploit any manually sense-tagged corpus to provide a sense choice for a word in context unlike supervised method (Roberto, 2009). It has the potential to overcome the knowledge acquisition bottleneck (Yarowsky, 2007), that is, the lack of large-scale resources manually annotated with word senses. This approach to WSD is based on the idea that the same sense of a word has similar neighboring words. It's able to induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters. It does not rely on labeled training text and, in their purest version; do not make use of any machine-readable resources like dictionaries, thesauri and ontology.

The information about the ambiguous word to be disambiguated, words that are syntactically related, and words that are topically related to the ambiguous word (Jurafsky and Martin, 2009). The limited availability of resources in the form of digital corpora and annotated, the rule based method is applied. All senses are discovered using a set of rules and knowledge base for later use in the disambiguation process. The hybrid shows an improvement in assigning correctly the corresponding disambiguation over the baseline method (Francisco *et al.,* 2005).

An unsupervised machine learning approach for Amharic using five selected algorithms was used; these are Simple k-means, EM (Expect Maximization) and agglomerative single, average and complete link clustering algorithms. The tested unsupervised machine learning method that deals with clustering of contexts for a given word that express the same sense. The work concluded that simple k means and EM clustering algorithms achieve higher accuracy on the task of WSD for selected ambiguous word, provided with balanced sense distribution in the corpus (Wassie, 2014).

The sense of the words is extracted 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. The 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. The mixed methodology having rule based approach and Bag-of-Words having enriched bags using learning methods. The approach establishes the superiority over individual rule based and Bag-of-Words approaches based on experimentation (Ranjan Pal *et al.,* 2013).

Rule based approach exploit the hand craft rule for WSD task. The rule based require extensive work of expert linguists and thus can result in near human accuracy. The (Tesfaye, 2010) Afan Oromo rule based Afan Oromo Grammar Checker, showed a promising result. The results show that rule

based is an approach used in the morphologically rich language like Afan Oromo. This rule based approach for languages, such as Afan Oromo, advanced tools has been lacking and are still in the early stages. In this work, (Stefan *et al.*, 2011) a model that represents word sense in context by vectors which are modified according to the words in the target's syntactic context (Adam, 2007).

OVERVIEW OF AFAN OROMO

Afan Oromo, also called Oromiffaa or Afaan Oromoo, is a member of the Cushitic branch of the Afro-Asiatic language family (Gragg and Gene, 2006). It is the third most widely spoken language in Africa, after Hausa and Arabic. Its original homeland is an area that includes much of what is today Ethiopia and some parts of other East African countries like northern Kenya, Somalia, Tanzania and Sudan (Abera, 2001). Currently, it is an official language of Oromia Regional State (which is the biggest region among the current Federal States in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which constitute about 50% according to the estimate of 2007 (Gragg and Gene, 2006). With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afan Oromo from 1991 (Guya, 2003). As this language has more than half of the country's population, there are no standard dataset and natural language tool. This study was one of the contributors for this under resourced language.

METHODS AND MATERIALS

This section describes the methods and tools employed in this study. The study relies on the patterns learned from the corpus (unsupervised approach) in combination with the manually crafted rules for clustering similar contexts of ambiguous word and extracting the contexts respectively. The motivation behind the use of hybrid is mainly aroused form the fundamental problem of corpus-based approach in relation to the sparseness of the training contexts. The idea of this research is therefore to combine both the rule based and unsupervised machine learning approaches into a hybrid approach. Such method of word sense disambiguation as also employed in this work, combine the advantages from machine learning and rule based, potentially yielding better results. Hence, the reason why the hybrid approach was used is taking the availability and reliability of linguistic knowledge on the top of the semantic techniques and training methods learned from a corpus for learning the role of the words in its context.

Hence, there is no annotated corpus for Afan Oromo the study was limited to use free developed corpus (Unsupervised Method, which is suitable when there is scarcity of training data). However, Unsupervised Method was enough than other methods when there is a small training dataset. Contrary, the hybrid method was more confortable than Unsupervised Method due to it combines a set of rules with machine learning. Therefore, the integration of both methods was overcome the problems of each other and improve the performance of the system

For implementation of the study, we have used Java; Net Beans 8.0.2 which runs on the prepared corpus and the clustering were performed in Weka 3.6.5 tool (compatible with Java package). This package is a general-purpose and open source programming language. Moreover, it is optimized for program portability and component integration. This makes more prefer for the study than other software packages.

Dataset

For this study, new corpus prepared for the Sense Disambiguation of Afan Oromo. However, in such a work, it is very difficult to obtain standardized dataset for under resourced language like Afan Oromo. The procedure for collecting and preprocessing corpus is described here below:

Corpus Preprocessing and Acquisitions

The corpus acquisition and preparation are set of techniques required for gathering and compiling data for training and testing algorithms. The lack of resource has led to use of unannotated raw corpus 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.

The mechanism to acquire resource is to use the data from various sources and hence it has not previously used in any research on Afan Oromo. The collected data are machine readable free text. It is collected from newspapers (Bariisaa, Kallacha Oromiyaa and Oromiyaa. Bariisaa is a weekly newspaper, whereas the rest two come out once in two weeks), bulletins, news (ORTO), government and official websites. Moreover, Oromia Radio and Television Organization (ORTO) found in Adama releases daily news through radio and television broadcast and on its official website (Debele, 2014). To reduce the data sparse, the data used from these sources since they are believed to represent texts addressing various issues of the language. Actually, the collected data were not directly used for the purpose.

SENSE DISAMBIGUATION ALGORITHMS

The developed disambiguation model for Afan Oromo involve followed three step process:

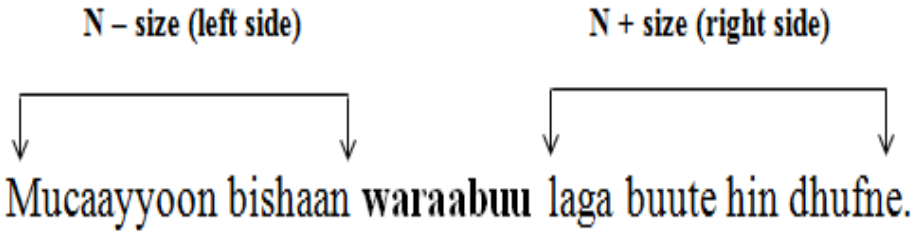
a. Text preprocessing which takes input and corpus, tokenize to remove stop words and perform normalization.

b. Extract context terms providing clues about the senses of the ambiguous term using two techniques (window size and rule based).

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 the degree of similarity computed using the vectors constructed from co-occurrence information.

As described in the above, one of the method used is unsupervised machine learning:

i. In the unsupervised machine learning, the surrounding contexts were extracted by sliding a window of *n* words. The words that occur in similar contexts tend to have similar senses. In order to extract the contexts from a set of sentences the role of the window is great. The senses and contexts can be captured in terms of the frequency co-occurrence neighborhood, i.e. words co-occurring.



Context is the only means to identify the sense of an ambiguous word. 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. In order to disambiguate a given word, a small and wider context should be considered in the performance of the system to rise overall.

Once the context words are extracted, the next step cluster similar contexts based on their inherent semantics, the number of the cluster representing the number of senses assumed by the ambiguous word. The underlying idea of the clustering of word contexts provides a useful way to discover semantically related senses.

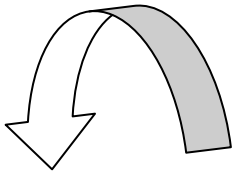
For each context extracted, vector space matrix constructed from co-occurrences. After the co-occurrence matrix, the cosine similarity was computed based on the angle between vectors of the contexts. These cosine similarity values were used to cluster similar contexts.

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

The context terms of the ambiguous words cluster using their similarity values produced. The clustering algorithms used in this study are hierarchical agglomerative clustering, which include single link, complete link, average link and EM and K-means clustering from partitional clustering.

ii. The other method used in this study was the hybrid machine learning approach. In this hybrid approach, the machine learns by the help of manually developed rule approach. The hybrid approach constitutes the unsupervised approach to cluster the contexts followed by hand crafted rule to extract the modifiers of the ambiguous word. The modifiers have a great role to decide on the word sense according to its role in the sentence. The modifiers can appear before the target word (the word, it modifies or describe).

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



For example, [*Mucayyoon **ija** akka boqqoolloo qabdi*].

Disambiguation is done by analyzing the linguistic features of the word and its preceding word. The rule-based section of this approach disambiguates word automatically using rules in order to complement the features learned from training data. This information is coded in

the form of rules. Based on this notion, the rule was developed was as follows:

- ⇨ If ambiguous word preceded by modifiers, then collect the modifiers to disambiguate.
- ⇨ If ambiguous word is a noun, the modifiers immediately following ambiguous Word.
- ⇨ If the ambiguous word was a verb, the modifiers immediately preceding ambiguous word.

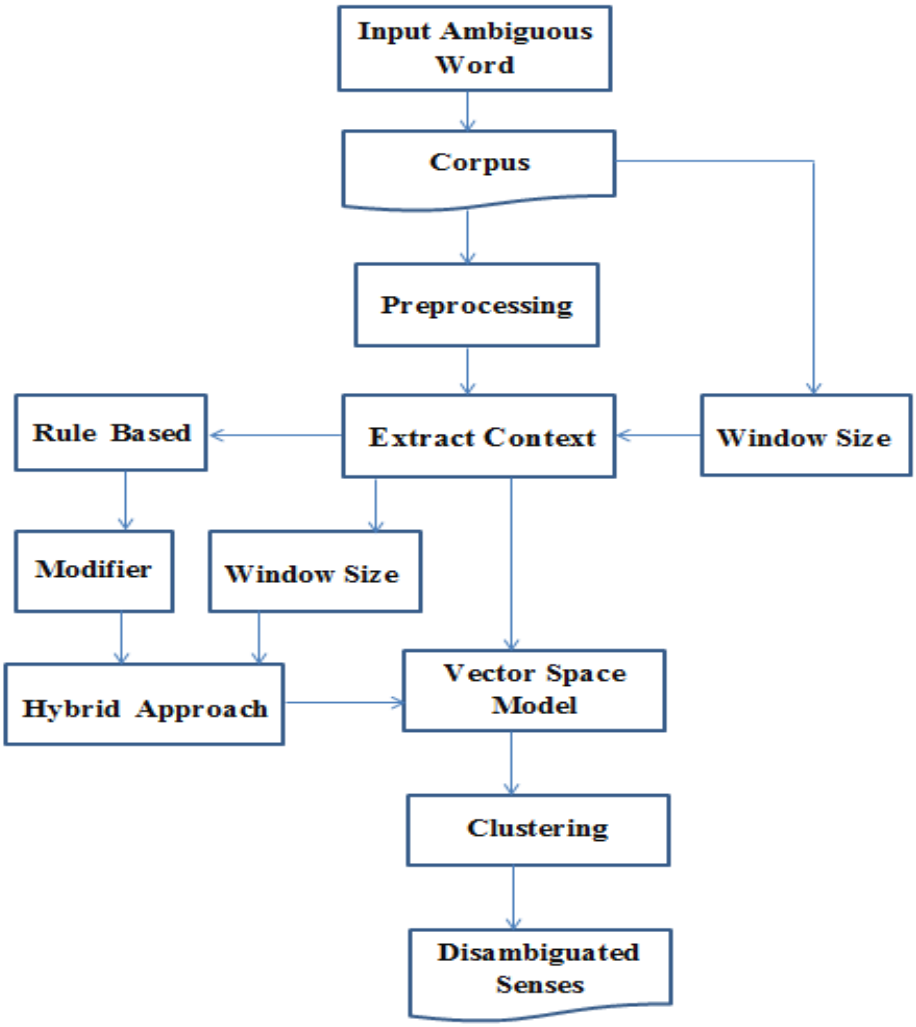


Figure 2. Architecture of the System

IMPLEMENTATIONS

In this work, two types of data were used: (a) the small list of ambiguous words to test the algorithms. Fifteen (15) highly frequent ambiguous words were selected from the language speakers using the questionnaire, (b) the big corpus containing thousands of sentences to extract contexts and its vectors to represent the group of contexts. This corpus is prepared for the purpose of this study as there is no standard corpus for Afan Oromo language. However, it is unlabeled data and never used in any researcher and it needs to be developed further.

The corpus, which is a set of sentences first tokenized into words. Since, Afan Oromo is Latin alphabet the sentences can split

using similar word boundary detection techniques like white space. After tokenization takes place, we have removed Afan Oromo stop words (non-content bearing words); hence it has no effect on the sense of the words. Finally, 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. 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. Then for clustering purpose, the Weka tool is used to cluster the cosine similarity planted to it. The following was the snapshot of the Word Sense Disambiguation system interface:

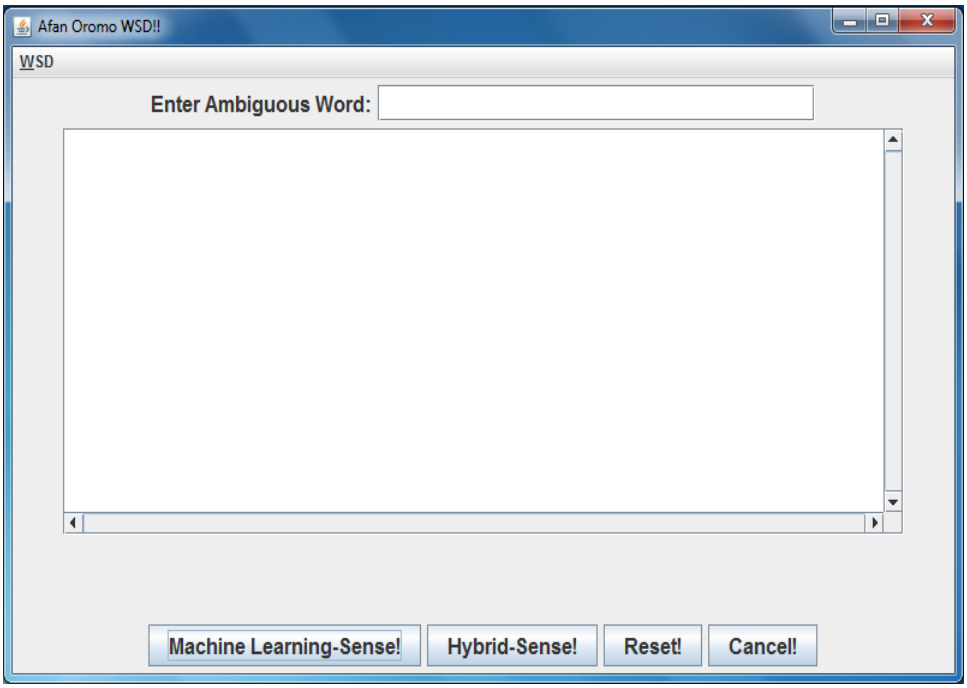


Figure 3. User Interface of the WSD

RESULTS

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 investigation, window size of up to ±10 words on both sides of ambiguous word has been tried for Afan Oromo. Starting with the first term in the test set, extract words appearing *N*-words left and right of the ambiguous words. In order to measure the performance of the window size for the disambiguation, we run the disambiguator using all the window sizes (starting from 1 to 10) and observe the differences in the performance of the disambiguator.

As evidenced by the experiment, differences in the window sizes yield different results. The result has proved that window of two-words to the left and right

of the ambiguous word achieved the best performance than other windows. This result converges with the result obtained by (Jurafsky and Martin, 2009). Table 1 below shows the performance of the disambiguate using different window sizes.

Table 1. Determining Optimal Window Size

Window Size (N)	Accuracy (%)
1	73.34
2	66.67
3	60.0
4	55.5
5	51.6
6	46.3
7	41.1
8	36.0
9	31.8
10	28.6

As can be seen from Table 1.1, a narrow window of context, one and two words to either side (73.34% and 66.67% respectively), was found to perform better than wider windows (28.6%). The accuracy is conducted by measuring the performance of WSD with varying the window size (the tested ambiguous word found in sentences). It is very likely that smaller window sizes have yielded significantly higher accuracy than other windows and different windows gave different results. From this experiment, we conclude that smaller window sizes usually lead to accuracy while bigger window sizes relatively low accuracy (Debele, 2014).

Table 2. Unsupervised and Hybrid WSD Results

Clustering Algorithms	Accuracy (%)	
	Unsupervised Machine Learning	Hybrid Approach
Single Link	54.4%	61%
Complete Link	54.4%	59.7%
Average Link	54.4%	61%
K-Means	56.9%	71.2%
EM	60.7%	74.6%

From the finding of the above table, 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. The modifiers contain

a lot of valuable clues for disambiguation (Gamta, 2005).

One thing that is clear from the experiment is that the senses are clustered where

cluster 0 is the pair of contexts which are *bilisa* and *qabsoo* clustered and make the sense of *got freedom*, cluster 1 is the pair of contexts *gaara* and *tabba* are grouped to make the sense of *highland* and cluster 2 is *daara* and *uccuu* are merged to make the sense of *cloth*, the other two clusters: cluster 3 and cluster 4 are incorrectly clustered and cannot make a sense. The figure 4 below Dendrogram shows the more description of these results:

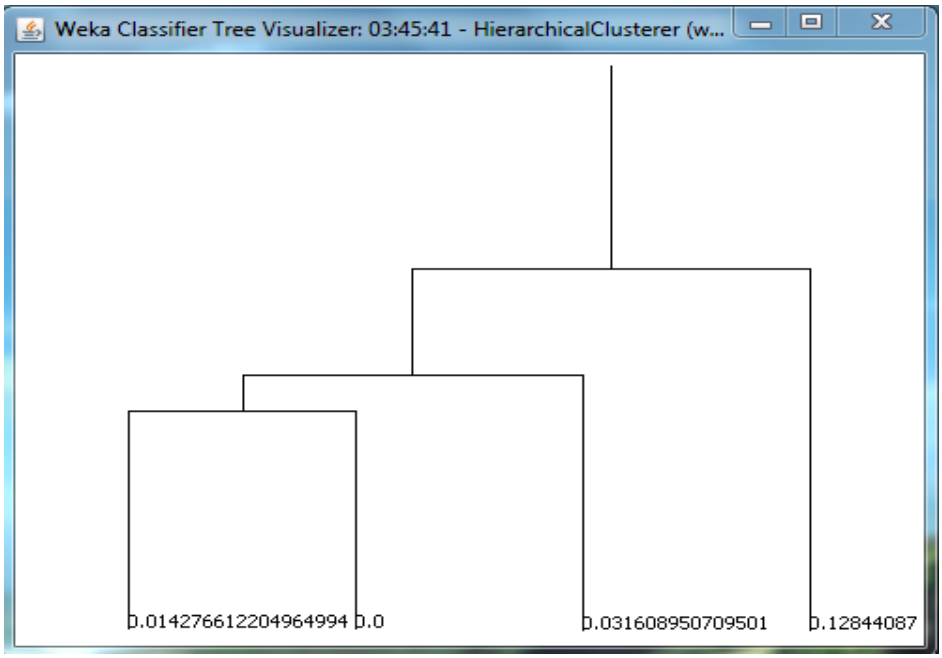


Figure 4. Dendrogram of the Senses Clustered for B*ahe*

Evaluation of Sense Disambiguation System

Hence, there were no previous standard Afan Oromo word sense disambiguation dataset for evaluation as presented in corpus preparation Section. For this reason, the system did not evaluate against the other systems. In this work, the evaluation was 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 (Flickinger, 2015).

$$\textbf{Precision}\ (\%) = \frac{\#\ \text{Correctly Disambiguated Words}}{\#\ \text{Disambiguated Words}}$$

$$\textbf{Recall}\ (\%) = \frac{\#\ \text{Correctly Disambiguated Words}}{\#\ \text{Total number of Ambiguous Words}}$$

In addition to Precision-Recall measures, the researchers have used extra evaluation, particularly in Hybrid method and Clustering as it clearly shown in the figure

2 system architecture. The reason why this evaluation mechanism needed is that on the rule based to identify whether the modifiers preceding ambiguous words are either Nouns or Verbs. And also to evaluate how much the produced clusters are comply with the clusters prepared by human experts as a benchmark. The researchers achieved, 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.

As the result shows in Table 2, the actual result of the study was 85.5%, which is encouraging with the under resourced language that of Afan Oromo language. The overall performance of the system on the machine learning and rule based yields that better achievement than before. An important point here is the study decide good clustering, since it is commonly acknowledged that there is best criterion of the final aim of the clustering.

DISCUSSION

The conducted experiment shows that, the semantic has come to the conclusion that the sense of words are closely connected to the statistics of word usage, which are working with window size and vector value derived from event frequencies; that is, we are dealing with Vector Space Model (cosine similarity) and clustering (Euclidean distance) (Kaplan, 2015). By using cosine similarity we include important semantic information in the purely statistical process of selecting the appropriate sense of a particular word. This benefits both unsupervised, hybrid approaches to WSD by increasing the chances of matching a particular context.

The result found that 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 values of the window size, which leads to the proper choice of sense of the target word. Based on this result, we conclude that for Afan Oromo window 2 was recommended unlike Amharic language (Solomon, 2011) which window size of 3 is recommended.

As shown in table 2, the result obtained by unsupervised machine learning and hybrid approach was 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 (Kaplan, 2015) that machine learning algorithms were a useful information source for disambiguation but that it's not as robust as a linguistic (modifiers in this case). The most likely reason for this is that our approach relies on automatically assigned immediately preceding words while machine learning are needs to left and right of unannotated data set. On the other hand, the machine

learning is noisy while the rule is more reliable and prove to be a most useful linguistic knowledge for WSD.

As the conducted experiment showed, each cluster has context group, where the sense of these context groups is hopefully different. The underlying assumption is that the senses found in similar contexts are similar senses. 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 as seen in table 2 (Fei Shao and Yanjiao Cao, 2005; Tesfaye, 2011).

From Table 2, 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 was different. Therefore, the linguistic knowledge (hybrid approach) the best approach to solve WSD than machine learning algorithms in Afan Oromo (Ravi Mante *et al.,* 2014; David, 2014) as shown in the experiment (Table 2). 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 (Gamta, 2005).

The WSD developed for Afan Oromo has its own strength and weakness sides. As the result showed that, the experiment attempts to disambiguate any ambiguous words, if it's running in corpus rather than limiting itself to treating a restricted ambiguous word. 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.

This system has the first work, that integrated different algorithm to find the appropriate sense of ambiguous words in Afan Oromo. However, there is some ambiguous word on which the performance of our approach is actually low. The system reported that the vector space model was affected by the data sparsity. The frequency of co-occurrence of most context words is zero due to the limited corpus size of the language. This result affects our cosine similarity values (as shown in figure 3 above). 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 (as shown in Table 2).

CONCLUSION

WSD has been based on the idea that the semantics of the context words belonging in the same sense of a word will have similar neighboring words. The context is hence a source of information and is the only means to identify the sense of an ambiguous word. The approach does not rely on labeled training text and does not make use of any expensive resources like dictionaries, thesauri, and WordNet (Adam, 2007).

For under resourced Ethiopian language like Afan Oromo the hybrid approach is recommended. Since there is no annotated corpus, hybrid approach plays a great role to disambiguate. The hybrid approach relies on hand-constructed rules that are

acquired from language specialists rather than automatically trained from data. In this study, we faced a significant challenge as Afan Oromo lacks Word Net, Sense Definition and annotated resources. Taking into account their contribution to WSD and other research concerned institutions should develop these resources.

LIST OF ABBREVIATIONS

IR: Information Retrieval
NLP: Natural Language Processing
WSD: Word Sense Disambiguation

ACKNOWLEDGMENT

I would like to acknowledge Jimma University for financial support in this work. Secondly, I would like to thank Oromia Radio and Television (ORTO) agency for allowance of the corpus/data collection from their studio.

REFERENCES

Abera, N. (2001) Long vowels in Afan Oromo: A generic approach, School of graduate studies, Addis Ababa University, Ethiopia.

Adam, Kilgarriff (2007) I don't believe in word senses: Computers and Humanities, London: Longman.

Agirre, E. and Martinez, D.(2000) Exploring automatic word sense disambiguation with decision lists and the web, Toronto, Ontario.

David Yarowsky (2014) Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French, USA: Stroudsburg.

David Yarowsky and Radu Florian (2002) Evaluating sense disambiguation across diverse parameter spaces. Natural Language Engineering.

Debele G. (2014) Afan Oromo News Text Summarizer, Master's thesis, Pohang University of Science and Technology, Pohang: Korea.

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

Flickinger, D. (2015) Natural Language Engineering-Efficient Processing with HPSG*:* Methods, Systems, Evaluation Coli website: from: http://www.coli.uni-sb.de/nlesi/ on 3/15/2015.

Francisco Oliveira, Fai Wong, Yiping Li, Jie Zheng (2005) Unsupervised Word Sense Disambiguation and Rules Extraction using non-aligned bilingual corpus, Tsinghua University: Beijing.

Gamta T. (2005) Seera Afaan Oromoo, Finfinnee, Boolee Press.

Gragg and Gene B.(2006) Oromo of Wollega: non-semetic languages of Ethiopia , East lansing, Michigan state University press.

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

Ide, Nancy and Jean Véronis (2010) Word

sense disambiguation: The state of the art, Computational Linguistics.

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

J. Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amjan Shaik, P. Pavan Kumar (2012) Word Sense Disambiguation: An Empirical Survey. Hyderabad: India.

Kaplan,A. (2015) An experimental study of ambiguity and context, Mechanical Translation.

Kebede T. (2013) Word Sense Disambiguation for Afan Oromo Language, Addis Ababa University, Addis Ababa, Ethiopia.

Nancy and Jean, (2007) Word Sense Disambiguation Algorithms and Applications, Ney work, Spring.

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

Ravi Mante, Mahesh Kshirsagar and Prashant Chatur (2014) A Review of Literature on Word Sense Disambiguation, Government college of engineering Amravati: Maharashtra.

Roberto Navigli (2009) Word Sense Disambiguation: A Survey, ACM Computing Surveys, USA: Washington Dc.

Salton G. (2015) The Measurement of Term Importance in Automatic Indexing: In Journal of the American Society for Information Science, vol.32.

Shaikh Samiulla Zakirhussain (2013) Unsupervised Word Sense Disambiguation, Indian Institute of Technology: Bombay.

Solomon, A.(2011) Unsupervised Machine Learning Approach For Word Sense Disambiguation To Amharic Words, Addis Ababa University. Addis Ababa: Ethiopia.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal (2011) Contextualizing semantic representations using syntactically enriched vector models: In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala: Sweden.

Tesfaye D. (2010) Designing a Stemmer for Afan Oromo Text: A hybrid approach, school of graduate studies, Addis Ababa University, Ethiopia.

Tesfaye D. (2011) A Rule-based Afaan Oromo Grammar Checker, Addis Ababa University, Addis Ababa , Ethiopia.

Wassie G. (2014) A Word Sense Disambiguation Model for Amharic Words using Semi-Supervised Learning Paradigm, School of graduate studies. Addis Ababa University: Ethiopia.

Yarowsky, D. (2007) Unsupervised word

sense disambiguation rivaling supervised methods, In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge: M.A.