International Journal of Research Studies in Science, Engineering and Technology
Volume 4, Issue 6, 2017, PP 10-15
ISSN : 2349-476X

Word Sense Disambiguation and Semantics for Afan Oromo Words using Vector Space Model

Workineh Tesema*[1], Debela Tesfaye[2]

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

[2] *Department of Information Technology, Jimma Institute of Technology, Jimma University, Ethiopia*

***Corresponding Author:** Workineh Tesema, Department of Information Science, College of Natural Sciences, Jimma University, Ethiopia*

Received Date: 05-07-2017 ***Accepted Date: 18-08-2017*** ***Published Date:02-09-2017***

ABSTRACT

This paper presents Afan Oromo semantics which is identifying the words semantically related. Semantic is one of the critical application in natural languages, hence it is a fundamental problem for many natural language technology applications. The aim of this work is to develop sense disambiguation which finds the sense of words based on surrounding contexts. Hence, this study used unsupervised approach that exploits sense in a corpus which is not labelled. The idea behind the approach is to overcome the problem of scarcity of training data. The context of a given word is captured using term co-occurrences within a defined window size of words. The similar contexts of target words are computed using vector space model and then clustered. From total clustering, each cluster representing a unique sense. Most of the target words have more than three senses. The result argued that the system yields an accuracy of 85% which was encouraging result. Therefore, for Afan Oromo semantic has come to the conclusion that the sense of words is closely connected to the statistics of word usage. Further study using different approaches that extend this work are needed for a better performance.

Keywords: *Semantic, Machine Learning, Sense Disambiguation, Afan Oromo, Target Word.*

INTRODUCTION

Semantics concerns the study of meaning as communicated through language and linguistics part that is concerned with meaning (Sebastian, 2002). The semantics is the study of the senses of a words or any text in human language. Semantics like syntax and phonology is an internalize enterprise concerned with linguistic expressions and the minds that generate them. We are concerned with narrower senses of semantics, such as the semantics based on the corpus. We present a survey of vector space model and their relation with the distributional hypothesis as an approach to representing some aspects of natural language semantics (Salton G, 2001).

Recent research in artificial intelligence has long aimed at endowing machines with the ability to understand natural language. One of the core issues of this challenge is how to represent language semantics in a way that can be manipulated by computers. Prior work on semantics representation was based on purely statistical techniques, lexicographic knowledge, or elaborates endeavors to manually encode large amounts of knowledge. The simplest approach to represent the word semantics is to treat the word as an unordered bag of words, where the words themselves become features of the textual object. The sheer ease of this approach makes it a reasonable candidate for many information retrieval tasks such as search and text categorization (Salton G, 2001).

The lexical meaning of words identified by using vector spaces and linear algebra (Stephen Clark, 2014). The meanings of words will be represented using vectors, as part of a high-dimensional “semantic space”. The fine-grained structure of this space is provided by considering the contexts in which words occur in large corpora of text. Words can easily be compared for similarity in the vector space, using any of the standard similarity or distance measures available from linear algebra, for example the cosine of the angle between two vectors. And also the hypothesis underlying distributional models of word meanings is so-called distributional hypothesis: the idea that

"Words that occur in similar contexts tend to have similar meanings" (Peter D. and Patrick Pantel, 2010). Now the basis vectors correspond to whole sequences of grammatical relations, relating the target word and context word. Which paths to choose is a parameter of the approach, with the idea that some paths will be more informative than others.

The words sense is in principle infinitely variable and context sensitive. It does not divide up easily into distinct sub-senses. Lexicographers frequently discover in corpus that overlapping word senses, and standard or conventional senses extended, modulated, and exploited in a bewildering variety of ways (Adam Kilgarriff, 2000). In lexical semantics, this phenomenon is often addressed in theories that model sense extension and semantic vagueness, but such theories are at a very early stage in explaining the complexities of word meaning (Lyons J, 2005).

Schutze (2008) states formulated a way to represent the vector space with words as dimensions. Arbitrary words can be chosen as axes, and the words in the corpus can be vectorized based on the counts of co-occurrences of these words with each of the axes. The occurrence of every word within a window size is counted. Each occurrence of ambiguous word in a corpus is represented as a context vector. The vectors are then clustered into groups, each identifying a sense of the ambiguous word. A historical approach of this kind is based on the idea of word space (Marco Baroni and Gemma Boleda, 2009) that is, a vector space whose dimensions are words.

The dimensionality of the vector is the number of words in the vocabulary (the number of distinct words occurring in the corpus). As all vectors under consideration, a cosine value of zero means that the word and corpus vector are orthogonal and have no match (i.e. the word does not exist in the corpus). It is the most common similarity measure in distributional semantics, and the most sensible one from a geometrical point of view. Its ranges from 1 for parallel vectors (perfectly correlated words) to 0 for orthogonal (perpendicular) words/vectors. It goes to -1 for parallel vectors pointing in opposite directions (perfectly inversely correlated words), as long as weighted co-occurrence matrix has negative values (Jurafsky, and Martin, 2009). The context is formally a text that surrounds a language unit (e.g. a word) and helps to determine its interpretation. It represents the occurrences of target words as word vectors. From these vectors, context vectors are formed and meaning similarity is found that is a function of cosine between the context vectors.

The rest of this paper will proceed as follows. Section 2 will discuss the different aspects of the proposed approach. Section 3 presents an experiment of the system.

Finally result and discussion and conclusion are presented in section 4 and 5 respectively.

PROPOSED METHOD

This section describes the methods employed in this study. In order to develop semantic model for Afan Oromo we followed three steps process which involve corpus preprocessing which tokenize and remove stop words and perform normalization. Extract context terms providing clue about the senses of the target word, and then clustering to group similar context terms of the given target words, the number of clusters representing the number of senses encoded by the target word. In order to cluster similar context terms we computed the degree of similarity using the vectors constructed from co-occurrence information. The more details of this method as follows:

Capturing Contexts

In this study, the surrounding contexts are captured using context window hence it is the only means to identify the sense of target word. The words that occur in similar contexts tend to have similar senses. 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.

Frequency and Co-Occurrence of Contexts

Once the context words are extracted, the next step cluster similar contexts based on their inherent semantics, which count frequency co-occurrence of contexts of senses assumed by the target word. We have proposed co-occurring words are the ones that appear together through the corpus in different sentences hence, co-occurring words are automatically extracted from a corpus. We have created co-occurrence matrix of contexts with contexts of the target word. Contexts-by-contexts co-occurrence matrices are instead typically populated by simple frequency counting: if word i co-occurs

x times with word j, we enter x in the cell which is cosine similarity frequency of i with i, i with j and j with j, j with i in the contexts-by-contexts co-occurrence matrix. The co-occurrences are normally counted within a context window spanning some usually small number of words. The contexts are sorted by their frequency which means that the most frequently co-occur words have used to decide the word senses. Finally, based on the real number of frequent occurrence, the contexts has similar senses are clustered.

IMPLEMENTATION AND EXPERIMENTS

As it discussed in section above, from the corpus the surrounding contexts are extracted by sliding window sizes. For example, when a target word *bahe* entered the system was captured the following contexts as in figure 1below. After the surrounding (left and right) contexts are identified, the total contexts around target word within determined window are counted which is the frequency of co-occurrence of the contexts as following figure 2:

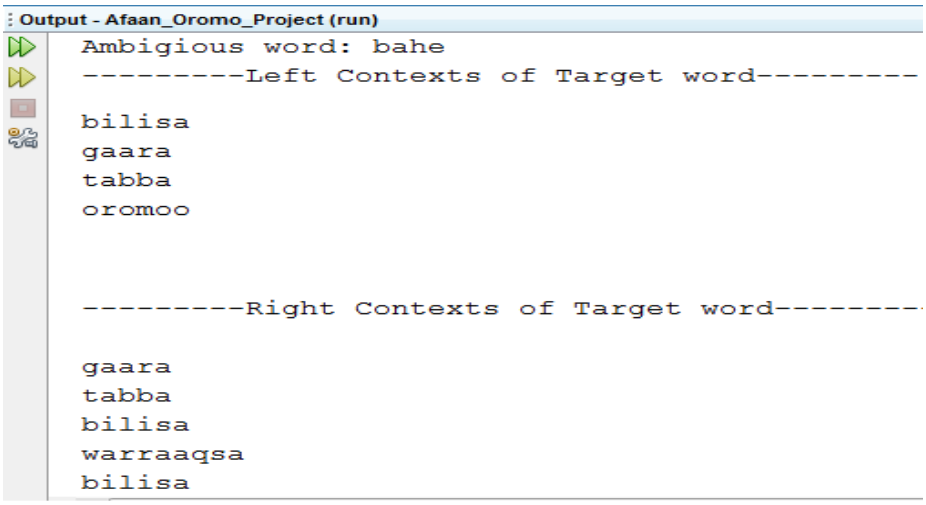

Figure1. *Identifying the Left and Right Side Contexts*

To identify the most co-occurrence of the contexts, both sides of target words are grasping together. As the following figure shows that, within the window one the frequency counted to compute the cosine similarity.

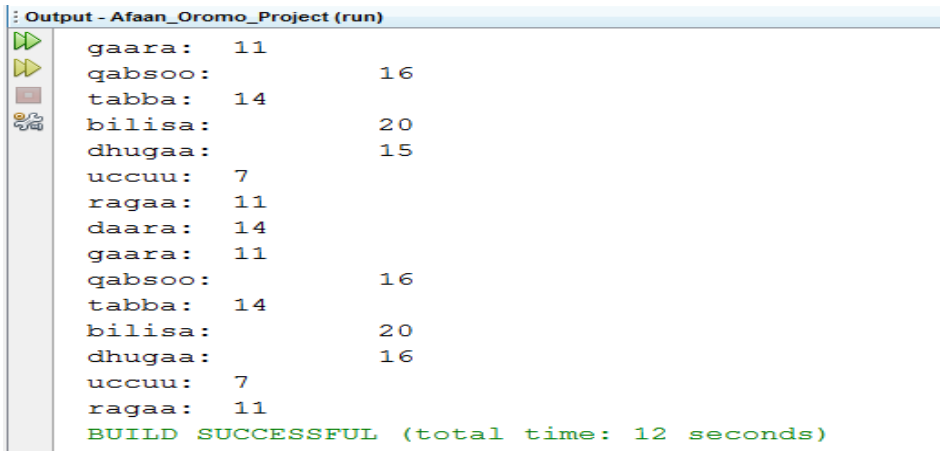

Figure2. *Frequency of Co-occurrences of the terms*

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.

$$\mathrm{Cos}(v,w) = \frac{v \cdot w}{|v| \cdot ||w|} = \frac{\sum_{i=1}^{m} v_i \cdot w_i}{\sqrt{\sum_{i=1}^{m} v_i^2 \sum_{i=1}^{m} w_i^2}}$$

The context terms of the target words cluster using their similarity values produced. The clustering algorithms used in this study are hierarchical and partitional clustering. Use

clustering to find words with similar context vectors. This can find words that are syntactically or semantically similar, depending on parameters (context words, window size).

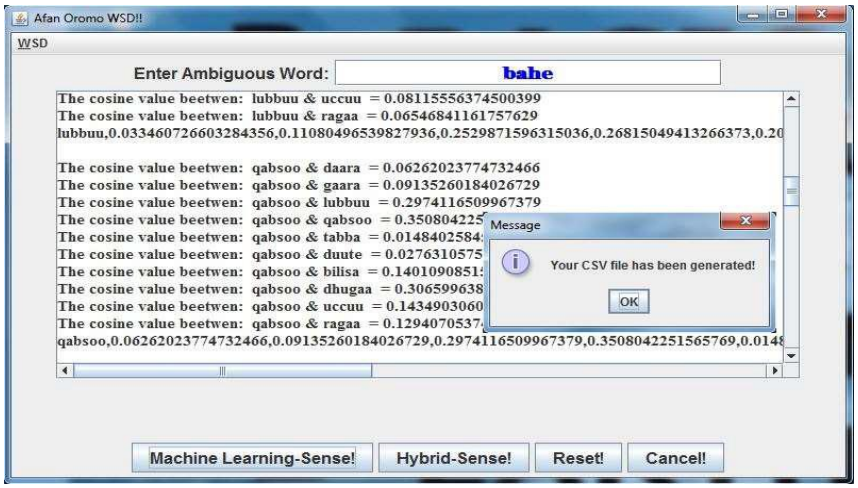

Figure3. *Word Sense Disambiguation System*

From the total contexts, the clustering algorithm provided five clusters for the target word *bahe*. Out of the total, it correctly clustered three senses. While the rest two pairs of cluster were incorrectly clustered with different senses. Based on the experiment shows that the target word *bahe* with given contexts has the following senses; the first cluster include *bilisa* and *qabsoo* at dissimilarity 1.15, the second cluster include *gaara* and *tabba* at dissimilarity 1.23 and the third cluster include *daara* and *uccuu* at dissimilarity 1.42, the two wrongly clustered contexts are *dhugaa* with *lubbuu*, and *ragaa* with *duute*. But it should cluster *dhugaa* with *ragaa*, *lubbuu* with *duute* to give the senses of witness and death/pass respectively as experts evaluated.

Table1. *Cosine Similarity Measure Representations*

	Bilisa	Qabsoo	Gaara	Tabba	Daara	Uccuu	Duute	Lubbuu	Ragaa	Dhugaa
Bilisa	1	**0.98**	0.09	0.05	0.01	0.15	0.04	0.17	0.07	0.09
Qabsoo	**0.99**	1	0.09	0.015	0.063	0.14	0.02	0.01	0.111	0.05
Gaara	0.07	0.009	1	**0.98**	0.062	0.027	0.004	0.20	0.11	0.010
Tabba	0.09	0.07	**0.97**	1	0.05	0.06	0.04	0.09	0.06	0.002
Daara	0.010	0.02	0.02	0.06	1	**0.98**	0.0	0.0	0.014	0.016
Uccuu	0.52	0.04	0.34	0.63	**0.98**	1	0.0	0.0	0.031	0.013
Duute	0.01	0.08	0.02	0.41	0.0	0.0	1	**0.91**	0.0	0.038
Lubbuu	0.08	0.247	0.11	0.20	0.033	0.08	**0.94**	1	0.066	0.07
Ragaa	0.03	0.14	0.287	0.49	0.083	0.009	0.0	0.12	1	**0.98**
Dhugaa	0.01	0.20	0.02	0.006	0.07	0.08	0.05	0.008	**0.95**	1

From this experiment, the semantic value using cosine is between 0 to 1. One and near to (0.98, 0.97, 0.91, 0.94, etc) means the two contexts are relatively have high similarity senses whereas zero means these contexts are dissimilarity. Corpus based similarity functions rely on the more often two words occur in similar contexts (and are used in similar way), the more semantically similar they are. Based on their frequency of the contexts in which they co-occur, a real number representing the similarity may be calculated.

RESULTS AND DISCUSSION

The conducted experiment shows that, the semantic has come to the conclusion that the meaning of words are closely connected to the statistics of word usage (Stefan Thater, *et.al.*, 2011), which are working with window size and vectors value derived from event frequencies; that is, we are dealing with cosine similarity and clustering. By using cosine similarity we include important semantic information in the purely statistical process of selecting the appropriate sense for a particular word. This

benefits the 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 value of the window size, which leads to the proper choice of sense for the target word. Based on this result, we conclude that for Afan Oromo window 2 was recommended unlike other languages. Figure 3 and Table 1 above shows, the context words for a target word *bahe* with it's computed the angle between vectors (Caropreso, *et.al.*, 2001).

As the experiment shows that, the average accuracy for test terms was 56.2% for the machine learning approach. Thus, in total there are 75 target words to be discriminated, 5 words with 2 senses, 30 words with 3 senses, 20 words with 5 senses, and 20 words with 4 senses. Five different configurations of clustering are run for each word, leading to a total of 375 experiments. For example, from test set if the target word *bahe* has entered the following senses are generated:

Table2. *Target Word of Bahe*

Target Word	No Senses	Nearest Neighbor Context (on Window Size =1)	Senses
Bahe	Bahe1	Bilisa, Qabsoo	Freedom
	Bahe2	Gaara, Tabba	Highland
	Bahe3	Daara, Uccuu	Cloth
	Bahe4	Dhugaa, Ragaa	Witness
	Bahe5	Lubbuu, Duute	Dead / Pass

As the evaluation of the system, indicates that an accuracy of 81% on the test. The below table 3 contains the evaluation performance of the WSD:

Table3. *Evaluation of Word Sense Disambiguation*

No	Accuracy	
	Precision	Recall
Correctly Disambiguated Senses	85%	80%

As the experiment shows that different windows bring different results which are gathered from left and right side of target word. To identify the best context window sizes, we tried until windows +10 on the developed model. The most optimal context window size for Afan Oromo was window +2 as the result shows. Therefore, the smallest window size is used to disambiguate the correct sense of the target word. This provides a clear motivation for further investigation in the area.

CONCLUSION & FUTURE WORK

The contribution of NLP in achieving a goal of Information Retrieval Systems has been clearly pointed out. Furthermore, it has been pointed out how NLP plays a significant role in enhancing the computer's capability to process texts. To that end, semantic is one component of NLP contributing a lot to the effort of solving the problem of Information Retrieval Systems in answering users' requests by introducing semantics of a query term and index terms. The window of +2 is the standard window applicable for disambiguation in Afan Oromo. The nearest words surrounding the target word give more information than words far from the target word and consequent surrounding words to the left and to the right provide any information for the purpose of disambiguation. This work helps further study on Afan Oromo semantics.

ACKNOWLEDGMENT

I would like to thanks to Jimma University for the financial support.

REFERENCES

[1] Adam Kilgarriff(2000) I don't believe in word senses: Computers and Humanities, vol.31(2).
[2] Caropreso, M. F., Matwin, S., & Sebastiani, F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization: Theory and Practice, Idea Group Publishing, Hershey,US.
[3] H. Schutze(2008) Automatic word sense discrimination. Computational Linguistics, vol.24(1).
[4] Jurafsky, D.,Martin, J.(2009) Speech and Language Processing (2nd Edition) Pearson Education.
[5] Marco Baroni and Gemma Boleda (2009) Distributional Semantics: Natural Language Processing.
[6] Löbner, Sebastian(2002). Understanding Semantics. Arnold: London: Blackwell.
[7] Lyons, John(2005). Linguistic Semantics: An Introduction. Cambridge, UK: Cambridge Press.
[8] Salton G(2001) The Measurement of Term Importance in Automatic Indexing: In Journal of the American Society for Information Science,vol.32.
[9] Salton, G. (2001). The SMART retrieval system: Experiments in automatic document processing. Prentice-Hall, Upper Saddle River, NJ.
[10] Stephen Clark (2014). Vector Space Models of Lexical Meaning, Handbook of Contemporary Semantics, 2nd ed, edited by Shalom Lappin and Chris Fox.
[11] 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.
[12] Turney, Peter D. and Patrick Pantel (2010), From frequency to meaning: Vector space models of semantics, Journal of Artificial Intelligence Research vol.37.

Citation: *W. Tesema and D. Tesfaye, "Word Sense Disambiguation and Semantics for Afan Oromo Words using Vector Space Model", International Journal of Research Studies in Science, Engineering and Technology, vol. 4, no. 6, pp. 10-15, 2017.*