Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.9, No.6, 2018

Unsupervised Machine Learning Approach for Tigrigna Word Sense Disambiguation

Meresa Mebrahtu Reda
College of Computing and Informatics, Assosa University, PO box 18, Assosa, Ethiopia

Abstract
All human languages have words that can mean different things in different contexts. Word sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings (polysemy). We use unsupervised machine learning techniques to address the problem of automatically deciding the correct sense of an ambiguous word Tigrigna texts based on its surrounding context. And we report experiments on four selected Tigrigna ambiguous words due to lack of sufficient training data; these are መደብ read as "medeb" has three different meaning (Program, Traditional bed and Grouping), ሓለፈ read as "halefe"; has four dissimilar meanings (Pass, Promote, Boss and Pass away), ሃደመ read as "hademe"; has two different meaning (Running and Building house) and, ክበረ read as "kebere"; has two different meaning (Respecting and Expensive).Finally we tested five clustering algorithms (simple k means, hierarchical agglomerative: Single, Average and complete link and Expectation Maximization algorithms) in the existing implementation of Weka 3.8.1 package. "Use training set" evaluation mode was selected to learn the selected algorithms in the preprocessed dataset. We have evaluated the algorithms for the four ambiguous words and achieved the best accuracy within the range of 67 to 83.3 for EM which is encouraging result.
Keywords: Attribute- Relation File Format, Cross Validation, Consonant Vowel, Machine Readable Dictionary, Natural Language Processing, System for Ethiopic Representation in ASCII, Word Sense Disambiguation

1.1. Introduction

There is a need for people all over the World to be able to use their own language when using computers or accessing information on the Internet. This requires the existence of a variety of applications including local language spell-checkers, word processors, machine translation systems, word sense disambiguation, search engines, etc. [1].

Word Sense Disambiguation or discourse where this meaning is distinguishable from other senses potentially attributable to that word [28]. WSD is a natural classification problem: given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context into two or more of its sense classes.

WSD is an awkward problem [18]; most problems arise from the fact that the concept of a meaning is vague. Usually, there are no clear boundaries between one sense and the other. Typically, the problem of defining meaning is begun with using dictionaries, which is sense inventory in a context of WSD, i.e., from the algorithmic point of view sense inventories are used to specify all the meanings that a given word has. Now, the goal of WSD can be stated as choosing correct sense from sense inventory in a given context of a word [19].

As discussed in [2] corpus based approaches, information is gained from training on some corpus. A corpus provides a set of samples that enables the systems to develop some numerical models. In supervised WSD the training data is sense-tagged where as in unsupervised WSD the training data is a raw corpora which are not semantically disambiguated.

Current WSD systems are based on supervised learning methods which is still limited in that it does not work well for all words in a language. One of the main reasons is the lack of sufficient labelled training data that require expertise. Even though one can always label more examples to achieve better performance on a particular data set but the expense can be uncomforted [21]. A major problem with supervised approaches is the need for a large sense-tagged training set.

Unsupervised learning is the greatest challenge for WSD researchers. Unsupervised WSD approaches are composed of word sense induction or discrimination techniques aimed at discovering senses automatically based on unlabeled corpora and then applying them for WSD [5]. Unsupervised methods correspond to clustering tasks rather than sense tagging tasks.

Although different methods have been tested to find the correct sense of the polysemy words, accuracy at satisfactory level has not been obtained yet [15]. Among the different methods used for WSD; this research was focus its study on exploring unsupervised machine learning approach to WSD for Tigrigna words. Test the results in order to improve a bit further natural language understanding for Tigrigna word disambiguation.

1.2 Problem Formulation

Word sense disambiguation is a significant problem at the lexical level of natural language processing. The

philosophy is to determine the meaning of a word in a particular usage, by using sense similarity and syntactic context with corpus evidence as well as semantic relations from word Net [17]. As stated in [4], resolving the ambiguity of words is a central problem for large scale language understanding applications and their associate tasks. Humans are so skilled at resolving potential ambiguities that they do not realize they are doing it. There is considerable focus on how people resolve ambiguities; however it is still not known how exactly humans do lexical disambiguation [15]. Therefore, it is a difficult task to teach a computer to do the same thing. If there are more than one ambiguous words in a sentence, the number of potential interpretations of the sentence increases dramatically [2].

Correctly disambiguating words is a difficult problem [31]. When restricted to available on-line dictionaries like Word Net, it is sometimes impossible even for human beings to pick the right sense for words. Expecting a machine to resolve such ambiguities is not reasonable. But, a good online dictionary with example uses of words in each of their possible senses can allow a machine to disambiguate words accurately. Such dictionaries are not yet available. Incorrect disambiguation not only excludes correct synonyms from the query but it also introduces incorrect information to it reducing retrieval performance [27].

Ambiguities have been an issue in researches conducted in Tigrigna language. As discussed earlier; there are many uses for word sense disambiguation. The most common are application of WSD in machine translation, Information retrieval, speech processing, text processing, grammatical analysis, content and thematic analysis. The absence of automatic WSD would make it the development of such NLP and IR applications difficult [17]. A variety of WSD methods have been proposed over the last decade; however, such methods are still immature or undeveloped [25]. In response to this situation, the major concern of this research was to explore unsupervised machine learning approach for Tigrigna WSD, examine the outcomes in order to improve a bit further NLU.

1.3 Literature Review

As stated in [55], WSD serves as an intermediate step for computer science applications. Therefore, it has been a central problem since the earliest days of computational studies of natural language. Word Sense Disambiguation [51] is a technique to find the exact sense of an ambiguous word in a particular context.

According to Kolte [56], WSD is one of the most challenging jobs in the research field of NLP. Research work in this domain was started during the late 1940s. In 1949, Zipf proposed his “Law of Meaning” theory. This theory states that there exists a power-law relationship between the more frequent words and the less frequent words. The more frequent words have more senses than the less frequent words. In 1990s, three major developments occurred in the research fields of NLP: online dictionary Word Net [55] became available, the statistical methodologies were introduced in this domain, and Sense Val began.

Alternatively, techniques have been proposed for discovering senses of words automatically from unannotated text. This task of unsupervised word sense induction (WSI) can be conceptualized as a clustering problem [14].

Many approaches have been proposed for assigning senses to words in context, while early attempts only served as models for toy systems. Word Sense Disambiguation Approaches are classified into Knowledge based approach and corpus-based approach [16]. Knowledge-based methods utilize lexical and semantic knowledge bases such as machine readable dictionaries (MRDs), thesauri, computational lexicons. Corpus-based approaches provide an alternative strategy to overcome the lexical acquisition bottleneck observed in knowledge-based approaches by giving information necessary for WSD directly from textual data [57]. Corpus based approaches can be categorized into three sub classes based on the form of machine learning used for training [2]: Supervised Word Sense Disambiguation, unsupervised Word Sense Disambiguation and Bootstrapping Approach to WSD.

1.4 Tools and Techniques

As most NLP systems, a preliminary preprocessing of the input text is needed. Texts (sentences) preprocessing is a primary step to load the instances of data set into machine learning tool (WEKA) to develop WSD model for the study. The preprocessing task comprises tokenization, stop word removal, stemming and normalization.

In building the WSD model, the researcher was use five unsupervised algorithms that are found in the existing implementation Weka 3.8.1 package. But the researcher trying to choose algorithms representing a few different approaches to the problem of clustering [16]. The researcher was start with simple k-means algorithms, which represent simple, hard and flat clustering methods. The researcher was use agglomerative single, average and complete link algorithms for representative family of hierarchical clustering algorithms. Last but not least, we were test also the Expectation Maximization algorithms also known as the EM which is probabilistic clustering algorithms.

1.5 System Architecture

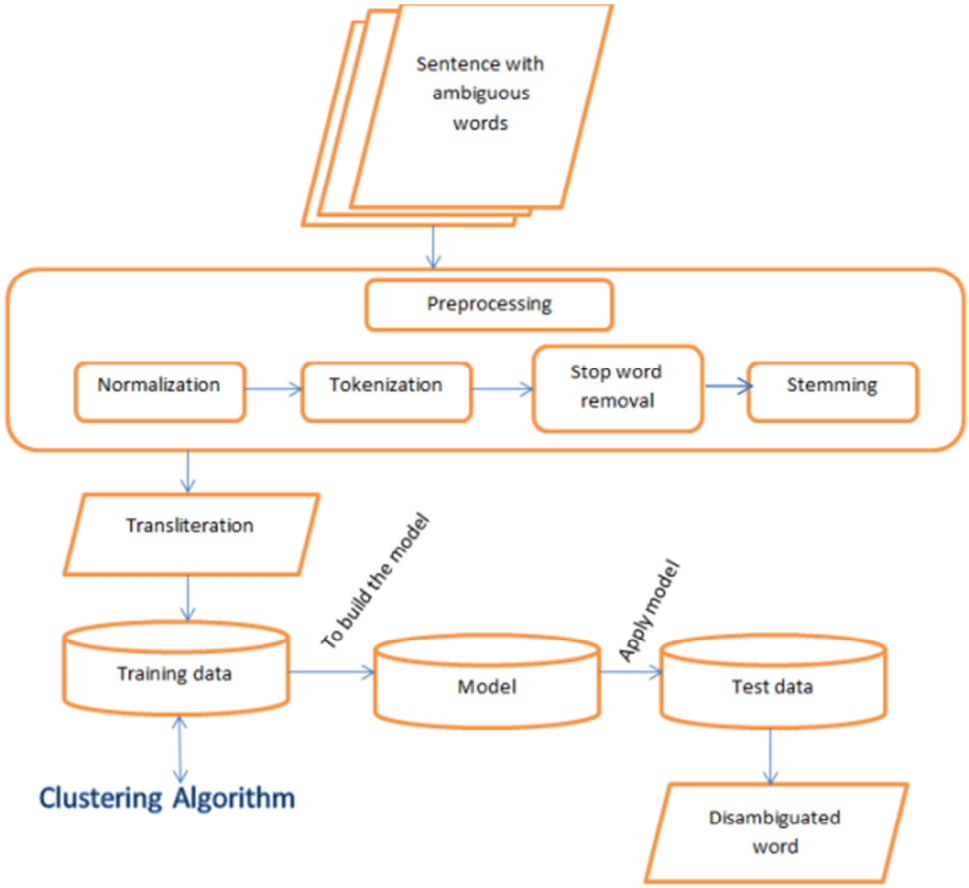

Figure 1. Unsupervised Word Sense Disambiguation System Architecture for Tigrigna.

1.6 Experimentation and Discussion

Unsupervised word sense disambiguation was selected to use a set of unlabeled data and automatically find sense distinctions for this study. Usually those methods involve some form of clustering. For WSD, learning the unsupervised machine learning procedures not required providing explicit sense labels, where each data set example is described by a feature vector within each target word and in their sense label. The feature vector comprises of attribute-value pairs, where the attributes are those contextual clues important for clustering. For these studies four ambiguous words namely መደብ read as “medeb”, ሓለፈ read as “halefe”, ሃደመ read as “hademe”, and ከበረ read as “kebere” are trained for each ambiguous word with their corresponding data sets.

There is a need to split the data into training and test sets for evaluation because of unsupervised nature of clustering algorithms [16], [17]. These features are extracted from text in the following process. First a text window surrounding ambiguous word of ±10 segments (word) is constructed. Then the occurrence of a target word is noted in a feature vector for every dimension corresponds to different word.

1.6.1 Experimentation Procedure

Step 1: Pre-processing - This involved reading the data into the program, cleaning the data, removing stop-words and stemming (see appendix 11).

Step 2: Generating an Arff file - This stage involved generating an Arff files for use with Weka. The features that were encoded into the Arff file were specified by the user (see appendix 13).

Step 3: Using Weka for Clustering - This stage involved loading the Arff file into Weka, Using a variety of clustering algorithms , and showing the bar graph of different classes.

Step 4: Evaluation of Clusters - This section involved running the evaluation program to gain the accuracy of the clusters. Such as:-

Check to what extent stemming and stop word removal of Tigrigna words in the corpus will affect the accuracy of unsupervised Tigrigna WSD.

Explore the outcome of dissimilar context sizes on disambiguation accuracy for Tigrigna ambiguous word. Here, different training data sets was organized for each ambiguous words, where the contextual information was gained from 1-left and 1-right to 10-left and 10-right following adjacent words are ready for each ambiguous

word.

Stemming has been found to give a significant upgrading on performance of WSD for morphologically rich languages. This investigation is performed to test whether this applies to unsupervised WSD for Tigrigna. The result of this experiment after stemming is presented as follows in figure 2:

Figure 2 Stemming input (241) and output screen (159)

Finally the experiment was showed effectively stemming on the data sets of containing ambiguous words. In successive using the stemmed data set the experiment was showed significantly improved the accuracy of the result before stemming.

1.7 Conclusion and Recommendation

1.7.1 Conclusions

The overall focus of this research is Word Sense Disambiguation which addresses the problem of automatically deciding the correct sense of an ambiguous word based on its surrounding context's. WSD is essential tool for NLP and IR applications .WSD is considered to be one of the most challenging of all NLP research areas due to its reliance on a varied range of linguistic and statistical knowledge.

The problem of WSD is addressed for Tigrigna which is one of less studied language. Though Tigrigna has many ambiguous words due to knowledge acquisition bottleneck, four ambiguous words are selected and clustering for each ambiguous word has been built.

In this study, unsupervised machine learning approach using five selected algorithms were used; these are Simple k means, EM and agglomerative single, average and complete link clustering algorithms. This method avoids the problem of knowledge acquisition bottleneck, that is, lack of large-scale resources manually annotated with word senses. This approach to WSD has been based on the idea that the same sense of a word will have 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, do not make use of any machine–readable resources like dictionaries, thesauri, ontology, etc. [16]. Based on selected algorithms, experiments on Weka 3.8.1 package, we conclude that simple k means, EM and CL clustering algorithms were achieved higher accuracy on the task of WSD for selected ambiguous word in corpus.

1.7.2 Recommendation

In this study we have only experimented with unsupervised machine learning approach but there are other approaches which performed well for WSD in other language. Therefore; the following recommendations are identified for further work in order to enhance WSD to Tigrigna texts, and the result useful in development of other NLP applications in Tigrigna:

Researches in WSD for other languages use linguistic resources like Thesaurus, Lexicon like word Net, machine readable dictionaries and machine translation software. In this study, we faced a significant challenge as Tigrigna lacks those resources. Taking into account their contribution to WSD and other researches concerned institutions should develop these resources.

For other language a standard sense annotated data are available for WSD research and also for testing a WSD systems. We don't have such data for Tigrigna language which makes the study to be limited for four ambiguous words. So, there need to be an initiative to prepare the data for WSD research.

Future research directions for WSD in Tigrigna include:
Extending this experimentation using Supervised and unsupervised WSD for other ambiguous words in addition to those covered in the research

This study experiment only five clustering algorithms that are implemented in Weka 3.8.1 package. But other algorithms like Clustering by Committee (CBC), Growing Hierarchical Self-Organizing Map (GHSOM) and Graph-based algorithms has been tested as they are used and found to yield impressive result for other language[16], [25].

In addition to corpus based approach, there are also knowledge based and hybrid approach (combination of knowledge base and corpus based approach) which are used for WSD for other language and found a good result [6], [34]. These approaches need to be investigated for Tigrigna as well.

Researchers can be study on how the word compounding and affixes affect Tigrigna words and their stems for WSD.

Develop an efficient full-fledged Dialect based Tigrigna language stemmer by including all dialect based irregular and exceptional words. And apply the stemmer in WSD from Tigrigna texts.

In this study some Tigrigna short forms are considered. But to enhance the accuracy of WSD including of all short words in Tigrigna language, and handles Tigrigna short forms that contain more than one slash or period should be studied.

Due to the reason that there are Tigrigna words that are dialect irregular words, short forms, the Tigrigna stemmer must be researched in order to bring these irregular words into their stem.

By incorporating necessary elements, the stemmer can also be used as a component for developing other computational tools like morphological analyzer, parser, machine translation, word frequency counting and other natural language applications.

Tigrigna language is a morphologically reach language and needs more morphological knowledge of the language. Therefore, researchers can enhance the stemmer by creating team with Tigrigna experts formally for full effective stemming.

1.8 Reference

Björn G. and, Lars A., Experiences with Developing Language Processing Tools and Corpora for Amharic, Kista, Sweden

Getahun W., A Word Sense Disambiguation Model for Amharic Words using Semi-Supervised Learning Paradigm, A Peer-reviewed Official International Journal of Wollega University, vol. 3, 147-155 November 2014, Ethiopia

Gerard E., Machine Learning Techniques for Word Sense Disambiguation, Barcelona, May 22, 2006

S.K.Jayanthi and S. Prema, Word Sense Disambiguation in Web Content Mining Using Brill's Tagger Technique, International Journal of Computer and Electrical Engineering, Vol. 3, No. 3, June 2011

Mohammad N., A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages, Univ. Grenoble Alpes, (n.d)

Omer O. and, Yoshiki M., Stemming Tigrinya Words for Information Retrieval, Nagaoka University of Technology, Nagaoka, Japan

David J., An Analysis and Comparison of Predominant Word SenseDisambiguation Algorithms, Faculty of Computing, Health and Science Edith Cowan University, 24th June 2011

Arindam R. et'al, Knowledge Based Approaches to Nepali Word Sense Disambiguation, Department of Computer Science, Assam University, Silchar, International Journal on Natural Language Computing(IJNLC) Vol. 3, No.3, June 2014

Philip R. and, David Y., A Perspective on Word Sense Disambiguation Methods and Their Evaluation, Dept. of Linguistics/UMIACS University of Maryland, Dept. of Computer Science/CLSP Johns Hopkins University

Ping C. and, David B., A Fully Unsupervised Word Sense Disambiguation Method UsingDependency Knowledge, Dept. of Computer and Math. SciencesUniversity of Houston-Downtown, The 2009 Annual Conference of the North AmericanChapter of the ACL, pages 28–36,Boulder, Colorado, June 2009

A. Eneko and, G. Rigau, Word sense using conceptual density, In Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, 1996 ndon, vol. A247, pp. 529–551, April 1996.

S.Weiss, Learning to Disambiguate Information Storage and Retrieval, 9:p:33-41.1973

Kelly F. and J.S.Philip, computer recognition of English Word Sense North Holland, Amsterdam 1975

Michael D., A Survey of Techniques for Unsupervised Word Sense Induction, Language Technologies Institute Carnegie Mellon University, December 4, 2009

Alok R., and Diganta S., Word Sense Disambiguation: A Survey, International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, July 2015

Solomon A., Unsupervised Machine Learning Approach For Word Sense Disambiguation To Amharic Words,

Adiss Ababa university, June, 2011
Solomon M., Word Sense Disambiguation For Amharic Text: A Machine Learning Approach, Adiss Ababa university, June, 2010
Clara et'al, A spreading-activation theory of semantic processing, Psychological Review, 1975. Vol 86(6).
Anderson, J. R., Language, Memory, and Thought.1976: Hillsdale, NJ.
Roberto, N., Word Sense Disambiguation: A Survey. ACM Computing Surveys, 2009. Vol 41(2).
Yarowsky, D. unsupervised word sense disambiguation rivaling supervised methods, in Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. 1995. Cambridge, M.A.
Hearst and Marti A., Noun homograph disambiguation using local context in large corpora, in presented at Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research,19 99: Oxford, United Kingdom.
Xinglong, W. and John, C. Word Sense Disambiguation Using Automatically Translated Sense Examples, In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). 2005. Association for Computational Linguistics.
Schutze, Automatic word sense discrimination, journal computational linguistics – special issue on word sense disambiguation, volume 24 issue 1, USA , march 1998
Hiroyuki K. and Yasutsugu M. "Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora", Central Research Laboratory, Hitachi, Ltd, Tokyo, Japan
Hagerie W. "Ensemble Classifiers Applied to Amharic Word Sense Disambiguation", Addis Ababa University, June 2013.
Bors K. et'al "Word Sense Disambiguation for Information Retrieval" Artificial Intelligence Laboratory Massachusetts Institute of Technology, Cambridge, Massachusetts 022139
Krister "Word sense discovery and disambigua tion", University of Helsink, 2005
Tsegay G., Towards performance improvement for Tigrigna language stemmer, university of Gondar, Ethiopia, 2016
Atelach A. and L.Asker, An Amharic Stemmer: Reducing Words to their Citation Forms",M.S Thesis, Department of Computer and Systems Sciences, Stockholm University, Sweden, 2010
Abhishek F. et'al, An Approach for Word Sense Disambiguation using modified Naïve Bayes Classifier, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 4, Nagpur , India, April 2014
Hammond, J., A Chronicle of the Revolution in Tigray region of Ethiopia, 1999, Red Sea Press, Eritrea.
ትግራይ, ማ.ባ., መፅናዕትታት ቀዳማይ ሲምፖዚየም ቋንቋ ትግርኛ 1990, መቀለ: ብርሃን እና ሰላም ማተሚያ ቤት(Birhan ena selam press).
Leslau, W., Documents Tigrigna, 1998, Paris: Libraire CKlincksieck.
Bender, M.L., Language in Ethiopia 1976, London Oxford University Press.
John M, Tigrigna Grammar. 1996, Lawrencevine, New Jersey: Red See Press.
Gebrehiwot A, a two-step approach for tigrigna text categorization, in Department of Information Science 2011, Addis Ababa.
Leslau, W., Documents Tigrigna, 1998, Paris: Libraire CKlincksieck.
Hailay Beyene, Design and Development of Tigrigna Search Engine in Department of Computer Science 2013, Adis Ababa
Agustina Situmorang and Tima Mariany Arifin, Derivational and Inflectional Morphemes in Pak-Pak Language
Martin H. and Andrea D. Sims, Understanding Morphology. 2nd edition ed. 2002, London Oxford University Press.
Getahun A, the Analysis of Ambiguity in Amharic. JES, 2001. Vol XXXIV (2).
Brown, Peter F., Pietra, Stephen A. Della, Pietra, Vincent J. Della, and L., Robert. Word-sense disambiguation using statistical methods. In Proceedings of the29th Annual Meeting of the ACL. 1991.
Yonnas F., Development of stemming algorithm for tigrigna text, in Department of Information Science, Addis Ababa, Ethiopia, 2011.
Wahiba Ben A, A New Stemmer to Improve Information Retrieval. International Journal of Network Security & Its Applications (IJNSA), 2013. Vol-5.
Yacob, D., System for Ethiopic Representation in ASCII (SERA). 1996 [cited 2016 Accessed on 3 may]; Available from:http://www.abyssiniacybergateway.net/fidel/
Zhao Y. and Karypis G., Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning, 2004. Vol 55(3): p. 311-331.
Forster R., "Document clustering in large German corpora using natural language processing", Ph.D. dissertation, 2006, University of Zurich.
Manning C. et'al, Introduction to Information Retrieval, 2008: Cambridge University Press
Witten and Frank, Data Mining: Practical Machine Learning Tools and Techniques. Second: ed2005: Morgan

Kaufmann publications.
Marine, C., Dekai, W.U.,(2005), Word Sense Disambiguation vs. Statistical Machine Translation, Proceedings of the 43rd Annual Meeting of the ACL ,Ann Arbor, June 2005
Miller G. A., WordNet: An On-line Lexical Database, Communications of the ACM, Vol.38 No. 11, 1995
Andres M. et'al, Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods, Journal of Artificial Intelligence Research 23 (2005) 299-330, Dept. of Software and Computing Systems, University of Alicante, Spain, published 03/05
Michael H., A Genetic Algorithm Using Semantic Relations for Word Sense Disambiguation, University of Colorado, 2011
Abhishek F. and, Dr.ManojB C., An Approach for Word Sense Disambiguation using modified Naïve Bayes Classifier, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 4, Nagpur , India, April 2014
Kolte, S. and, G.Bhirud, Word Sense Disambiguation Using WordNet Domains, First International Conference on Digital Object Identifier, pp. 1187-1191, 2008.
Jain, Data clustering: a review. ACM Computing Surveys, 1999. Vol 31(3): p. 264–323.
Han, J. and Kamber M., Data Mining – oncepts and Techniques, 2001: Morgan K.
Kaufmann, L. and Rousseeuw, P. J., Clustering by means of medoids, in In Dodge, Y. (Ed.) Statistical Data Analysis based on the L 1 Norm.1987: Elsevier/North Holland, Amsterdam. p. 405–416.
Dempster, A., laird, N., and Rubin, D., Maximum likelihood from incomplete data via the EM algorithm. . Roy. Statist. Soc. , 1977. Vol 39: p. 1-38.
Guha S. et'al, A robust clustering algorithm for categorical attributes. In Proceedings of ICDE, pp. 512–521. 1999. Sydney, Australia.
Anjali M. and Babu A., Ambiguities in Natural Language Processing, Vol.2, Special Issue 5, Kannur University, Kerala, India, October 2014.