International Journal of Advance and Innovative Research
Volume 9, Issue 4 October - December 2022

ISSN 2394 - 7780

A STUDY ON DIFFERENT APPROACHES OF WORD SENSE DISAMBIGUATION FOR DIFFERENT INDIAN AND FOREIGN LANGUAGES

Chingakham Ponykumar Singh and Dr. H. Mamata Devi
Department of Computer Science, Manipur University, Indo Myanmar Road, Canchipur, Imphal, Manipur – 795003

ABSTRACT

A disambiguating method known as Word Sense Disambiguation (WSD) uses the words around the ambiguous word to determine which sense is most pertinent in the given context. The varied methods and tactics utilised to solve WSD problems in certain widely spoken Indian and foreign languages were covered in this paper. Depending on the types of the materials available and how well they fit with the language's nature, different strategies or approaches are used. Same approaches with different techniques and algorithms are also employed depending on the nature and the type of the languages. Because of differences in data size and language characteristics, it has been found that a single technique does not produce the same accuracy when employed on various groups of languages. Comparative studies among the different approaches used in Indian and foreign languages are also further discussed. Different evaluation metrics like precison, recall, F-score etc are used across different language as per the accuracy calculation is concerned. The uncertainty that exists in many kinds of Indian and foreign languages needs to be pleasantly resolved by selecting an acceptable approach with an effective method.

Keywords: Word Sense Disambiguation, Ambiguity, Natural Language Processing, WordNet

INTRODUCTION

Speaking or writing, or expressing one's thoughts in any other way requires the use of an organised linguistic system. Any language may occasionally have words with several meanings, which can cause confusion and misinterpretation. There are three categories of lexical ambiguities: (i) Polysemy ambiguity: This refers to a word or phrase that has more than one meaning yet is connected to another. (ii) Homonymy ambiguity: This refers to the use of a term or phrase with numerous senses that are wholly unconnected to one another. (iii) Categorical ambiguity: This refers to a word or phrase with more than one meaning but diverse grammatical meanings[17]. Humans are capable of determining the appropriate meaning of a word in light of the context in which it is used. However, until some rule-based characteristics are incorporated into the computers' memory, they will not be able to handle such a situation[21]. How one word is used in a human or machine context, knowledge of the word is essential to determining its proper meaning. Word Sense Disambiguation (WSD) is a method for separating vague words into their most pertinent senses by using the words around the equivocal word[4, 25]. Finding the proper meaning of a term behind an argument is assisted by syntactic, positional, and contextual factors[2]. A specific word in a phrase can be disambiguated, or the whole word can be (All Word Sense Disambiguation). There are three main approaches of WSD:

- **Knowledge based approach:** Knowledge-based NLP systems will also define and execute the text and the conceptual domain using their various techniques to address NLP difficulties like confusion resolution[27], which are necessary for understanding text meaning. Any small amount of information can be viewed as a collection of cues, including the words used, their tone, their arrangement, etc. [3]. There are numerous knowledge-based methodologies that rely primarily on knowledge resources like collocations, WordNet, thesaurus, ontology, etc. [5]. Additionally, for the sake of disambiguation, grammar rules, hand-coded rules, explicit lexical information, etc. may be employed. One of the most well-known knowledge-based algorithms, the Lesk algorithm, operates by determining the number of times a word occurs more than once in a context[6].
- **Machine learning based approach:** The corpus proof is the primary source that this strategy uses. The model, which is a probabilistic/statistical model, is often trained using labelled or unlabelled corpora. The classifier's primary responsibility in this case is to learn the features that are needed to retrieve and assign the proper meaning of the term in the example phrase. Three different machine learning-based methodologies exist:

a) Supervised Techniques: With hand-labeled, sense-annotated data sets, this approach is effective[1]. The data are trained and then classifier is applied. Neural Networks, decision trees, decision lists, and Naive Bayes are examples of supervised WSD methods[22].

b) Unsupervised Techniques: The word senses are generated from word clusters, and the newly derived cluster is then searched for new instances of the word. Context clustering, co-occurrence graphs, word clustering, and other unsupervised WSD techniques are included[22].

c) Semi – supervised Techniques: In this method, a reduced data collection that only contains the essential details can be utilised to teach the system important traits. In some cases, this method performs better than unsupervised methods.

- **Hybrid approach:** The above two mentioned approaches are used together in this approach. Dictionary data which are machine readable are used to correlate the relations among the senses and the corpus.

WSD applications include speech recognition, machine translation, information extraction, lexicography, parsing, information retrieval, automatic text summarization, hypertext navigation, classification of documents, spelling correction, reference resolution etc[4, 11, 22,30].

The paper is organised as follows: Section 2 gives the insights of the literature review in WSD. Section 3 elaborates the details about the various WSD evaluation metrics for measuring the system and last section deals about the conclusion.

LITERATURE REVIEW

Numerous researchers in WSD have produced impressive works in a variety of languages.

Foreign Languages

Mohannad AlMousa et.al[18] examined an English WSD system using SCSMM thereby recording the maximum sentence context by preserving the internal word order. Further, a study on different aspects like rate of ambiguity, size of sentence, granularity were also discussed. Farag Ahmed and Andreas Nurnberger[13] used a methodology to deal with ambiguity of Arabic words based on statistical co-occurrence. The user query phrases, topic context, and word inflection forms were used to pick features. The cohesion of the Arabic words and a unique similarity score were utilised to translate the word into its accurate English meaning. Myung Yun Kang et al.[25] suggested a brand-new supervised model with sense space incorporated for disambiguating Korean language. The huge training set was used for a ten-fold cross validation after the design was trained with a baseline frequency of 5 (five). Boon Peng Yap et.al[10] developed a Neural Network based English WSD system which excellently well by using the resources available in WordNet but ignored the combination of gloss and context. Ali Saeed et.al [15] created a unique benchmark Urdu corpus by to carry out all word sense disambiguation. The created dataset will be extremely effective because the annotation of the uncertain word was allocated following the consolidation of the results from the three observers. The accuracy of the corpus was further improved using a voting-based approach. Manish Kumar et.al[24] developed a WSD system that disambiguate the inputted sentence using the surrounding words information using the WordNet data. Here, the raw inputted data is converted into useable data by adding the additional information like POS, subject, object etc. which allow the classifier to disambiguate the ambiguous word contained in the sentence. Nyein Thwet Thwet Aung et al. [16] employed a Myanmar WSD system by using a concurrent dataset of Myanmar and English to handle word ambiguity. The machine translation method between Myanmar and English was improved using the newly built module. Tang Shancheng et al. [14] created a WSD system which uses TextCNN, TextLSTM, and TextMultiTask instances. When compared to the other best practises currently in use, it was found that there was an improvement of 11.48%. Alok Ranjan et al. [21] utilising both supervised and unsupervised methods determine the appropriate meaning of words based on real-world circumstances. In order to determine the proper meaning of the unfamiliar word, it presented a mixed methodology that included the "Modified Lesk" and "Bag-of-Words" methodologies.

Table 1: Implication of various Approaches used in different Foreign languages

Authors & Year	Language Used	Approaches / Techniques	Merits	Demerits
Boon Peng Yap et.al(2020)	English	Machine Learning Approach (Supervised Technique)	Excellent WSD system.	Ignored context gloss combination
Ali Saeed, Rao et.al(2008)	Urdu	Knowledge Based Approach	Clear representation of data and methods with great level of presentation.	Used limited data resource with very low accuracy rate of 57.71%.
Nyein Thwet Thwet Aung	Myanmar	Machine Learning Approach	Achieved a great accuracy of 89% considering the	Apply bag – of – words features to noun and verb

et.al(2011)		(Supervised Technique)	limited resource available.	words only.
Tang Shancheng et.al(2018)	Chinese	Machine Learning Approach (Supervised Technique)	Deals with all ambiguity types and minimises feature extraction process.	Relatively small data size.
Myung Yun Kang et.al(2018)	Korean	Machine Learning Approach (Supervised Technique)	Word2Vec's CBOW architecture was utilised. Carry out sense embedding and evaluation of it.	Hybrid Space model cannot perform evenly on micro and macro precision.
Farag Ahmed et.al(2019)	Arabic	Machine Learning Approach (Supervised Technique)	Arabic's unique linguistic characteristics were taken into account. Choose the optimal features for the data training.	Many-sensed words were not taken into consideration, creating a significant gap.
Manish Kumar et.al(2020)	English	Knowledge Based Approach	Reduces comparison steps	Uses only 50 ambiguous words with 2000 example sentences.
Mohannad AlMousa et.al(2021)	English	Hybrid Approach	Outperformed existing system	Cannot handle sentences with less word and cannot point to the topic of the document.
Alok Ranjan et.al(2015)	English	Hybrid Approach	Slight improvement in performance from the existing system	Used too lengthy steps

Indian Languages

Arindam Roy et al.[28] created a straightforward yet efficient Nepali WSD algorithm to address the issue of ambiguity in the Nepali language that include overlap-based, conceptual distance-based and semantic-based techniques. This algorithm demonstrated how metaphysical distance-based and semantic-based techniques were utilised to dramatically improve the efficiency of the algorithm when compared to the coincide approach. Alok Ranjan Pal et al. [22] used the Naive Bayes probabilistic model as a starting point for sense classification when a supervised technique is used for word meaning disambiguation in the Bangla language. The Bengali text corpus was first normalised and the Naive Bayes criteria were then applied. P. Iswarya and V. Radha's[18] employed unsupervised learning method to deal with the very ambiguous word using clustering and POS. With this method, the best k-value in the k-cluster may be automatically selected and a sense collocation dictionary can be created. In this case, POS taggers were employed to disambiguate the word senses, and the clustering and sense collocation dictionaries were used to improve the overall efficiency of the system. Purabi Kalita et al.[7] used the Walker algorithm in conjunction with a customized version of the Assamese WordNet to disambiguate Assamese words. By including an element called FEATURE, which specifies the subject category or the word domain, Assamese WordNet was modified. To make the extraction process and Walker Algorithm execution easier, the WordNet data were encoded in XML format. According to the result obtained, the system's accuracy out shadowed the old method by 1.86%. Sruthi Sankar K P et.al[11] advocated an unsupervised learning algorithm to distinguish between different word senses using context matching. A dataset was compiled from numerous Malayalam web data and the training examples included the majority of co-occurring words as well as collocations. In order to clarify the confusing word, seed sets and sense clusters produced from the database were used. Richard Laishram et al. [12] made the first attempt of developing a Word Sense Disambiguation system in the Manipuri language based on a decision tree. Ttraditional positional along with context-based features were merged and using CART technique the classifier is trained. Himdweep Walia et al. [8] employed a supervised technique for word sense disambiguation in the Gurmukhi language. For determining the k-nearest neighbour in relation to the provided test vector, the distance between the test sets was calculated to create two lists. A unique cross validation called 5-fold was implemented to evaluate the operation of the system. Manisha Gupta et al.[23] present an answer by conducting a study on Hindi WSD system using Hindi WordNet to decipher the word senses. The definition of the ambiguous words, along with the ten words around them, were produced and enhanced for the intention of disambiguation using a modified

LESK technique. Anidhya Athaiya et.al[20] developed a Supervised Hindi WSD system using Genetic algorithm using Hindi WordNet data. The developed system is very flexible that there are rooms for expandability to Hybrid approach.

Table 2: Implication of various Approaches used in different Indian languages

Authors & Year	Language Used	Approaches / Techniques	Merits	Demerits
Manisha Gupta et.al(2013)	Hindi	Knowledge Based Approach	Valuing each and every word in the sentences by taking 10 as window size.	Work mainly on the dictionary's glosses length. Of noun, verbs and adjectives.
Gauri Dhopavkar et.al(2015)	Marathi	Knowledge Based Approach	Well designed rules and accuracy is satisfactory	Disambiguate only word level ambiguity.
Himdweep Walia et.al(2018)	Gurmukhi	Machine Learning Approach (Supervised Technique)	Satisfactory result	Small data size.
Alok Ranjan Pal et.al(2015)	Bengali	Machine Learning Approach (Supervised Technique)	Improved performance with the usage of lemmatization and bootstrapping.	Further better algorithm can be applied to increase the performance of the system
Richard Laishram et.al (2014)	Manipuri	Machine Learning Approach (Supervised Technique)	Employed the decision tree approach, which is most effective for agglutinative languages like Manipuri. The sense of the word was captured using conventional placement and context-based characteristics.	Small data size with non - unicode font.
P. Iswarya and V Radha(2016)	Tamil	Machine Learning Approach (Unsupervised Technique)	Collocation and clustering were employed to boost performance. Processing speed improved.	Reduce performance due to the lack of context word tagging and collocation.
Sruthi Sankar K P et.al(2017)	Malayalam	Machine Learning Approach (Unsupervised Technique)	Very simple but very effective	Small data size and all word senses may not have been covered by the generation of seed sets and sense clusters.
Purabi Kalita et.al(2015)	Assamese	Knowledge Based Approach	Improve accuracy rate.	Limited to noun and adjective phrases.
Anidhya Athaiya et.al(2018)	Hindi	Machine Learning Approach (Supervised Technique)	Improve accuracy and further expandability to Hybrid approach is there.	Limited to noun words only.
Arindam Roy et.al(2014)	Nepali	Knowledge Based Approach	Used mixed different concepts of Overlap, conceptual distance & semantic graph thereby giving good performance.	Performance can be improved and word coverage can be expanded to verb also.

Figure 1: Comparison of usage of different approaches based on the size of the data sources

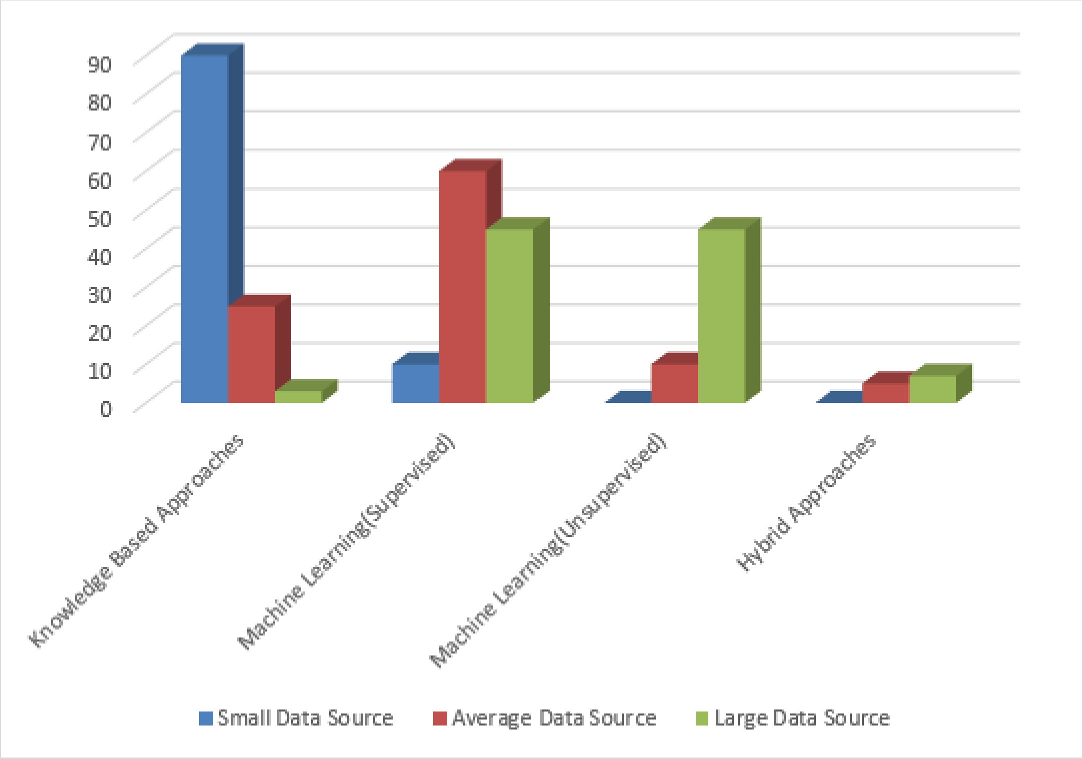

Figure 1 shows that different approaches were used depending on the availability of the resources in a particular language. Supervised approaches are the most widely used approach for an average data source language.

EVALUATION METRICS

Performance measurements were used to assess how successfully a computer taught itself through the process of machine learning, which entails training computers how to teach themselves how to solve problems. Depending on the problem's nature, the right approach should be taken. The accuracy of the outcomes a designed system generates is a good indicator of its efficacy. So, the majority of the systems that have been developed evaluate effectiveness in terms of accuracy. Accuracy establishes an assessment of reality that is unbiased and free of errors. Additionally, several proposed systems quantify their effectiveness using precision, recall, and F-Score[31, 7, 25]. F1-score assessment metrics, which perform best for classes with uneven distributions, have also been seen in some systems[16, 21].

CONCLUSION

Based on the amount of data availability and how well they matched the language's characteristics, various approaches were used. The majority of knowledge-based methods have been found to be resource constrained, while machine learning techniques have shown to be more accurate than the previous approach[18]. For Indian languages, which are morphologically rich, uncommon WSD systems were available[13, 28, 18]. For languages with small dataset, knowledge based approached is mainly used while machine learning approaches are mainly used for language with average or large data sets. It has been noted that the identical method does not produce the same accuracy when used with different types of languages. Therefore, it is necessary to select an acceptable method and strategy in order to conveniently resolve the discrepancy that exists in many types of Indian languages and foreign languages.

REFERENCES

[1] Kavi Mahesh, Sergei Nirenburg, “Knowledge-Based Systems for Natural Language Processing”, CRC Press Handbook of Computer Science and Engineering

[2] C.Leacock, G.Towell and E.Voorhees, “Corpus-Based Statistical Sense Resolution,*”ARPA Workshop on Human Language Technology, 1993”*

[3] Purabi Kalita, Anup Kumar Barman, “Word Sense Disambiguation: A Survey”, “*International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 4 Issue 5 May 2015 Page 11743-11743*”

ISSN 2394 - 7780

[4] Yarowsky, D. (1992). "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora", *Proceedings of the 14th International Conference on Computational Linguistics (COLING, Nantes, France), 454-460.*

[5] Grigori Sidorov and Alexander Gelbukh, "Word Sense Disambiguation in a Spanish Explanatory Dictionary", unpublished

[6] Lesk, M.,(1986) "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", *Proceedings of SIGDOC.*

[7] Purabi Kalita, Anup Kumar Barman, "Implementation Of Walker Algorithm In Word Sense Disambiguation for Assamese Language", "International Symposium Of Advanced Computing and Communication(ISACC), 2015"

[8] Aditi Salodkar, Mrunali Nagwanshi, Ms Bhavana Gopchandani, "Supervised Approach to Word Sense Disambiguation", *IOSRJEN ISSN€: 2250-3021, ISSN(p) 2278-8719 Vol. 09, Issue 5(May 2019), ||S(IX) || PP 80-85.*

[9] Tanveer Siddiqui, U. S Tiwary, "Natural Language Processing and Information Retrieval", "*Pages 156 – 173*"

[10] Boon Peng Yap, Andrew Koh, Eng Siong Chng(2020), Adaptive BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentence, *"Finding of the Association for Computational Linguistic MNLP 2020, 41-46"*

[11] Alok Ranjan Pal, Diganta Saha, Niladri Sekhar Dash, Antara Pal "Word Sense Disambiguation in Bangla Language Using Supervised Methodology with Necessary Modifications"," *J. Inst. Eng. India Ser. B (October 2018) 99(5):Pages 519 – 526*"

[12] Richard Laishram, Krishnendu Ghosh, Kishorjit Nongmeikapam and Sivaji Bandyopadhyay, "A decision tree based word sense disambiguation system in Manipuri language". "*Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014*"

[13] Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha, "A hybrid approach to word sense disambiguation combining supervised and unsupervised learning", "*International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, July 2013*"

[14] Tang Shancheng, Ma Fuyu, Chen Xiongxiong, Zhang Puyue, "Deep Chinese Word Sense Disambiguation Method Based on Sequence to Sequence", "*2018 International Conference on Sensor Networks and Signal Processing (SNSP)*"

[15] Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, Paul Rayson, "A Sense Annotated Corpus for All Words Urdu Word Sense Disambiguation", "*ACM Trans. Asian Low-Resource. Lang. Info. Process, Vol. 18, No. 4, Article 40. Publication date: May 2019*"

[16] Nyein Thwet Thwet Aung, Khin Mar Soe, Ni Lar Thein, "A Word Sense Disambiguation System Using Naïve Bayesian Algorithm For Myanmar Language", "*International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011 ISSN 2229-5518*"

[17] Alok Ranjan Pal, Diganta Saha, "Word sense disambiguation: a survey", "*International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, July 2015*"

[18] Mohannad AlMousaa,, Rachid Benlamria and Richard Khoury, A Novel Word Sense Disambiguation Approach Using WordNet Knowledge Graph, "*Preprint submitted to Elsevier"(2021)*

[19] Daniel Jurafsky, James H Martin, "Speech and Language Processing" "*Pages 22 – 24*"

[20] Anidhya Athaiya, Deepa Modi, Gunjan Pareek, "A Genetic Algorithm Based Approach for Hindi Word Sense Disambiguation","*Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018) IEEE Xplore Part Number: CFP18AWO-ART; ISBN:978-1-5386-4765-3*"

[21] Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha, "A hybrid approach to word sense disambiguation combining supervised and unsupervised learning", "*International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, July 2013*"

[22] Myung Yun Kang, Tae Hong Min, Jae Sung Lee, "Sense Space For Word Sense Disambiguation", *"2018 IEEE International Conference on Big Data and Smart Computing"*

[23] Alok Ranjan Pal, Diganta Saha, Niladri Sekhar Dash, Antara Pal "Word Sense Disambiguation in Bangla Language Using Supervised Methodology with Necessary Modifications"," *J. Inst. Eng. India Ser. B (October 2018) 99(5):Pages 519 – 526*"

[24] Manish Kumar, Prasenjit Mukherji, Manik Hendre, Manish Godse, Baisakhi Chakraborty(2020), "Adaptive Lesk Algorithm Based Word Sense Disambiguation using the Context Information", "*International Journal of Advance Computer Science and Applications, Vol. 11, No 3*"(2020)

[25] Himdweep Walia, Ajay Rana, Vineet Kansal "A Supervised Approach on Gurmukhi Word Sense Disambiguation using k-NN Method", "*978-1-5386-1719-9/18/$31.00 c 2018 IEEE*"

[26] Gauri Dhopavkar, Manali Kshirsagar, Latesh Malik, "Application of Rule Based Approach to Word Sense Disambiguation of Marathi Language Text", "*IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication systems(ICIIECS) 2015*"

[27] Mohammas Shibli, Md. Asif Bin Khaled, Mahady Hasan, Mohammad Ibrahim Khan, "Word Sense Dismabiguation of Bengali Words using FP – Growth Algorithm", "*International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February,2019*"

[28] Arindam Roy, Sunita Sarkar, Bipul Syam Purkayashtha, "Knowledge Based Approaches To Nepali Word Sense Disambiguation", "*International Journal of Natural Language Computing(IJNLC) Vol.3 No.3 June 2014*"

[29] Pooja Sharma and Nisheeth Joshi, "Knowledge-Based Method for Word Sense Disambiguation by Using Hindi WordNet", "*International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019*"

[30] P. Iswarya, V Radha, "Unsupervised Approach to Word Sense Disambiguation in Malayalam", "*International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)*"

[31] Sruthi Sankar K P, P C Reghu Raj, Jayan V(2017) "Word Sense Disambiguation for Tamil Language using part – of – speech and clustering technique". "*Journal of Engineering Science and Technology, Volume 12, No. 9 2017, Pages 2504 – 2512*"