International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Exploiting Artificial Immune System to Optimize Association Rules for Word Sense Disambiguation

Mohd Shahid Husain[1*]

Submitted: 04/08/2021 Accepted : 31/10/2021

Abstract: Requirement specification is the major activity in software development. Since requirements are gathered from the customers in natural languages they are prone to ambiguities. Ambiguous requirements give many interpretations of the same word or sentence. In order to reduce the problems faced due to requirement ambiguities, many techniques have been proposed in the past. Word Sense Disambiguation is a bottleneck in most of the Natural Language Processing (NLP) applications. The approaches to deal with WSD aims to provide the best possible meaning for the target word which is lexically ambiguous. To reduce the ambiguities and optimize the association mining rules for Word Sense Disambiguation (WSD), this paper proposes a new approach based on Artificial Immune System and Association Rule Mining. The approach shows significant results when tested on a collection of many Software Requirement Specifications (SRS) Documents. The average accuracy provided by the system is 89.2725%. outperforms state of the art methods.

Keywords: *Artificial Immune System, Association Rule Mining, Lexical Ambiguity, Word Sense Disambiguation*

1. Introduction

Conventionally Natural languages are widely used to represent ideas, like user requirements due to flexibility but time and again they introduce ambiguity in the document, as different readers may interpret a sentence differently and assumes that the interpretation they carry is the only and correct meaning of the sentence.

Ambiguity means the same statement or a word can be interpreted differently by different stakeholders. For example: John went to the bank. Here for the word “Bank”, the WordNet 2.1 has contained 18 interpretations (10 senses for noun & 8 senses for verb) like a “financial institution”, a “river bank” (bay) etc. depending on the sense of the bank used different persons can understand the above statement differently. Another example is: The temperature of the room should be normal. Now for different readers the term “normal” may have different interpretation like for one user the ‘normal’ speed of internet is 10 Mbps and for the other user it may be 15 Mbps.

Ambiguity Resolution deals with providing the most likely interpretation of a statement or word given in the context. Several studies have explored and have proposed methodologies to overcome the ambiguities; however, their accuracies are limited to 60%-70%. The aim of this work was to provide an efficient method for resolving ambiguities in a document. I have taken a set of Software Requirement Specifications (SRS) documents. I got encouraging results of the proposed approach.

1.1. Problem Definition

One of the critical aspects in success of any software project is how well the user’s requirement is incorporated in design and implementation. Most of the stake holders are non IT persons and they express their requirements in natural languages which are documented as SRS document. For all the phases of SDLC (software development life cycle) this SRS is treated as a reference. As the natural languages are inherently ambiguous there are lots of possibilities that these SRS documents may have some ambiguities i.e. different users can interpret things differently. A poorly understood requirement statement can lead to a software design and eventually a software product which may not be acceptable by the users. So the SRS document should be clear, precise, complete and unambiguous.

There are various forms of ambiguities that may be present in a document like Lexical ambiguity, syntactic ambiguity, semantic ambiguity, pragmatic ambiguity and vagueness. In this work I have proposed an approach to deal with one of the important types of ambiguities namely lexical ambiguity (Word sense ambiguity). I have implemented an ensemble approach to deal with the above mentioned ambiguity. In first phase I have implemented Association Rule Mining to generate a pool of rules for providing the most appropriate sense of a given word/phrase based on the context. In the 2nd phase, the rules produced in phase one are optimized by using the concept of cloning in Artificial Immune System. I have taken a corpus of Software Requirement Specification Documents to test the proposed approach for resolving ambiguities present in an SRS document and the results are quite encouraging.

The rest of this paper is organized as follows: Section 2 deals with the prominent work done by the researchers in the field. In Section 3, I will discuss the concept and implementation of the proposed approach. Section 4 provides discussion about the results of the proposed approach. Section 5 concludes the research and introduces the scope and extension for future work.

[1] *Department of Information Technology, CAS - Ibri, University of Technology and Applied Sciences, Oman*
ORCID ID: 0000-0003-4864-9485
** Corresponding Author Email: siddiquisahil@gmail.com*

2. Related Work

Ambiguity is a problem because, if the readers' understanding of the document differs from that of the writers, then they are likely not to be satisfied with each other. Many times, ambiguity is not noticed by anyone looking at the document. Very often, each person subconsciously disambiguates the document to the first interpretation he or she finds and he or she thinks that this first interpretation is the only interpretation.

2.1. Dealing with Word Senses

Natural language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. Machines need to process unstructured textual information and transform them into data structures which must be analyzed in order to determine the underlying meaning. The computational identification of meaning for words in context is called word sense disambiguation (WSD). In natural language processing (NLP), word sense disambiguation (WSD) is defined as the task of assigning the appropriate meaning (sense) to a given word in a text or discourse. Word sense ambiguity is a central problem for many established Natural Language applications like Machine Translation, Information Retrieval Systems etc. For this reason, many international research groups are working on WSD, using a wide range of approaches. However, to date, no large-scale, broad-coverage, precise system for word sense disambiguation has been built [1]. With current state-of-the-art accuracy in the range 60–70%, WSD is one of the most important open problems in NLP. Main approaches to deal with word sense disambiguation or Lexical ambiguity can be categorized as follows [2]:

- Supervised WSD methods
- Unsupervised WSD methods
- Knowledge based WSD methods

2.2. Supervised WSD techniques

Supervised techniques make use manually sense-annotated data sets. The classifier is trained by applying various machine learning algorithms and the manually tagged dataset for predicting the best possible sense of each instance of an ambiguous word/phrase.

Table 1: Comparison of supervised WSD methods

Method	Avg. Precision	Avg. Recall	Corpus	Avg. Baseline Accuracy
Decision List	96%	--	Tested on a set of 12 highly polysemous	63.9%
Naïve Bayes	64.13%	--	Senseval-3 All Word Task	60.9%
Instance Based	68.6%	--	WSJ6 containing 191 content words	63.7%
SVM	72.4%	72.4%	Senseval-3 Lexical sample task, used for disambiguation of 57 words	60.9%
GA	79%	63%	Arabic Corpus	72%
KNN	--	--	Bengali Corpus	71%
Apriori	--	--	--	83%

Ronald L Rivest has defined an ordered set of rules to assign most applicable sense of an instance (target word/phrase) in the given context based on a parameter called score. Naïve Bayes classifier is based on conditional probability. It calculates the probability of all the possible senses of the target word/phrase given the context. The sense having the maximum probability is selected as the intended sense of the target word/phrase [1]. Researchers have implemented some instance based algorithms for word sense disambiguation (providing the best possible sense of target word). In these approaches we have a set of examples called instances in memory which are used to train the classifier. The new examples are added in the memory with time. One of the most common instance based approach used for Word Sense Disambiguation (WSD) is K-Nearest Neighbor (KNN) algorithm. Another supervised approach for WSD is the use of Support Vector Machine (SVM) [3]. In this approach learning of classifier is done with the help of training datasets to create a hyper plane which separates the positive and negative samples. In recent years many researchers have implemented approaches based on Genetic algorithms, Neural Networks and Association Mining for the purpose of word sense disambiguation. The results of these approaches as claimed by the researchers are very promising. Table 1 shows the performance (in terms of Accuracy) of some of the commonly used supervised methods for Word Sense Disambiguation [2]. These results (Accuracy %) is claimed by the researchers in their work.

2.3. Unsupervised WSD technique

These techniques do not have the luxury of any tagged dataset. The basic concept behind these approaches for WSD is that any target word having a particular sense will have similar context words in all occurrences. Context clustering, word clustering and co-occurrence graphs are the main WSD techniques which don't exploit any machine readable dictionary or thesaurus. Context clustering is based on the idea of word space where words in the corpus are represented as vectors. Clusters are then created by grouping the context vectors of the target word [3]. Word clustering approaches are based on the idea that the words which are semantically similar convey the same sense. A well-known approach based on word clustering is proposed by Lin. V´eronis proposed an ad hoc approach called HyperLex. In this method a co-occurrence graph is created where nodes represent the words occurring in paragraphs of the text document. The nodes are connected if the words represented by the nodes occur in the same paragraph. Weights are then assigned to these edges relative to the co-occurrence frequency of the two connected words. Minimum Spanning tree is then constructed to disambiguate a target word.

Table 2: Comparison of Unsupervised WSD methods

Method	Avg. Precision	Avg. Recall	Corpus	Avg. Baseline Accuracy
Lin's Approach	68.5%	Not reported	Trained using WSJ corpus containing 25 million words. Tested on 7 SemCor files containing 2832 polysemous nouns	64.2
Hyperlex	97%	82%	Tagged on a set of 10 highly polysemous French words	73%

There is another graph based approach called, PageRank algorithm. Although it is mainly used for ranking web pages, some of the reported work for WSD makes use of this technique [3]. Table 2 shows the Accuracy of some of the commonly used unsupervised methods for Word Sense Disambiguation [2].

2.4. Knowledge based WSD techniques

These methods make use of various knowledge resources like thesaurus, electronic dictionaries etc. for disambiguation purpose. A very simple algorithm based on overlapping of sense definitions is proposed by Lesk. This approach called gloss overlap or Lesk algorithm disambiguate the target word based on the number of words common in the sense definition and the context of the target word. Walker's approach is also based on

overlapping of sense definition where each of the senses are categorized on the basis of subjects. In Selectional preferences approach for WSD some constraints (based on grammatical relationships) are applied on the senses. Those senses which do not satisfy the constraints are ruled out and the senses which satisfy the grammatical ruling gets the preference. Table 3 shows the Accuracy of some of the commonly used Knowledge based methods for Word Sense Disambiguation [2].

Table 3: Comparison of Knowledge based WSD methods

Method	Accuracy
Lesk algorithm	50%-60%
Walker's approach	50% when tested on 10 highly polysemous English words
WSD using conceptual density	54% on Brown corpus
WSD using Selectional Preferences	44% on Brown corpus
Adaptive Lesk algorithm	79%

The analysis of these approaches shows that:

• Supervised approaches provide better results compare to Unsupervised and knowledge based approaches. The drawback of supervised algorithms is that there is knowledge acquisition bottleneck because they need manually tagged training dataset.

• Unsupervised approaches overcome this problem of having a manually tagged dataset, but their performance is not up to the mark.

• The knowledge based methods are also having low performance but they cover wide area of applications because of the use of knowledge resources.

3. Proposed Methodology

To deal with Lexical ambiguity present in a document, I have used Artificial Immune System (AIS) based Association Rule Mining.

Associative classification uses association rule mining for finding association rules in a transactional database. These methods are very effective to classify an entity based on the association of it with other entities in the context. To get the possible senses of the target word I have used WordNet as it can be used easily with good results [4]. The concept of Association Rule Mining for Word Sense Disambiguation is used to get the association (most frequent occurring item set) between each sense of the key word and the neighboring word. The sense having the highest confidence value is selected as most probable meaning (interpretation) of the key word [5]. The concept of Association rule mining is used by some researchers to deal with lexical ambiguity. S Kumar & Niranjan has implemented Apriori algorithm based Fuzzy Association Rule Mining approach to generate rules for providing the appropriate sense of a target word [6]. In their work J G Cho & H Huh proposed a Tree based Association Rule Mining (TAR) method for WSD [7].

Artificial Immune System is one of the emerging trends in the field of Computational Intelligence. AIS is based on the natural phenomena of Biological Immune System of guarding the system by identifying the foreign elements. There are different models of AIS like Clonal Selection, Immune Network model and Negative Selection. These models are effectively implemented by the researchers in various applications for example pattern recognition, malicious node detection in networks, fault tolerance etc. The study of Artificial Immune based methods and there reasonably good performance have encouraged me to incorporate them with Association rule mining approaches, which can be very useful in generating effective rules.

The proposed method for WSD executed in two phases. In first phase it makes use of Apriori algorithm to generate the rules for providing the possible sense of a target word. In the second phase CLONALG algorithm is used to the pool of rules generated in phase-1. Figure 1 shows the schematic diagram of the proposed system.

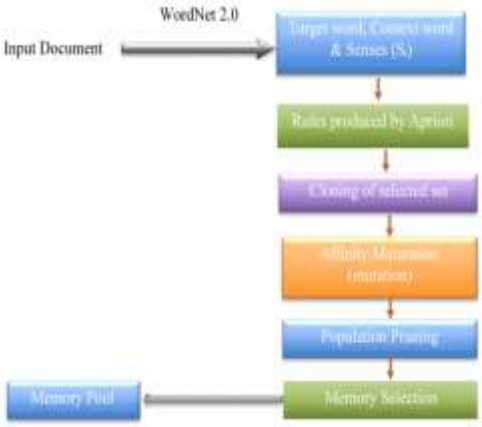

Fig 1. Block diagram of the proposed AIS based system

Steps involved in disambiguation of word senses

• Take a document as input.

• Apply Part of Speech (POS) Tagging.

• Identify the ambiguous words (Key words) with the help of WordNet.

• Get all the senses of the key word having the same Part of Speech, with the help of WordNet.

• Apply n-gram approach (to use context of the key word).

• Apply Association Mining Rule, to get association between the senses (S1, S2, ….) and the neighbouring context words (W1, W2, ….) in the statement.

• Apply CLONALG algorithm to get more frequent itemsets.

• At last the sense of the ambiguous word is determined by the rules committee voting. The sense which is decided by the most rules is selected as the sense of the ambiguous word in its context.

This section explains the working of the proposed system by taking a set of randomly taken 8 statements from a document in the corpus. In the preprocessing module with the help of WordNet 2.0, the target words (ambiguous words) were identified and all the possible senses of the target word was stored as Si. To get the context words, n-gram approach has been used, with n=2 to 5. In the next phase Apriori algorithm was used to generate the association rules for assigning the sense of the target word based on the association of it with the context words. These rules are stored in the memory pool. The final phase consists of applying CLONALG algorithm (an artificial immune based approach) on the set of rules stored in the memory pool. This phase generates more effective association rules by ignoring the rules having low confidence value below threshold and making clones of the rules having high confidence value. The more effective rules are then included in the memory pool. This process is repeated by a number of generations.

3.1. Data Preprocessing

Each statement in the input document is considered as a transaction record. These documents are referred to as datasets. A transaction in a dataset contains set of items (words) called itemset. For example, the following two statements randomly taken from a document in the corpus, are two transaction records

of the transactional dataset:

(i) Card will provide for user authentication and will store all data.

(ii) Student can purchase a textbook at the Co-operative store.

Target word (X): target word is the ambiguous words i.e. the word having multiple meaning/senses in the statement. For example, the word "store" in above statements.

Context words (Wi): these are the words apart from the target word in the statement which provide the context of the target word. For example, the words like authentication, data, card, purchase and textbook are used as context words.

Senses (Si): the senses are the different meanings of the ambiguous word (target word) system get from WordNet. For example, some of the senses of the word "store' (target word) we can get from WordNet are shop, stock, computer storage, a depository for goods.

Table 4: Example Set of Transactions (Sentences)

Transaction ID	List of item IDs	Sense
T1	w1 w2 w5	S3
T2	w2 w4	S1
T3	w2 w3	S3
T4	w1 w2 w4	S7
T5	W1 w3	S5
T6	w2 w3	S2
T7	w1 w3	S3
T8	w1 w2 w3 w5	S4

Table 4 represents the manually tagged senses of target word "X" for 8 sample sentences (Transactions) from a document. Each sample contains target word "X" and one or more context words represented by w1, w2, w3, w4, w5. The senses of the target word "X" are taken from WordNet 2.1 and represented by Si.

For representation, the presence of items (words Wi) in a statement is encoded as binary 1 and the absence of items (words Wi) in a statement is encoded as binary 0. Table 5 shows the binary representation of the occurrences of context words in example sentences.

Table 5: Data Representation in Binary

	W1	W2	W3	W4	W5
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1

3.2. Rule Generation

The proposed method makes use of association rule mining algorithm (Apriori) for finding frequent itemsets. The parameter support indicates the usefulness or generalization of a rule, i.e., how often the rule is observed. The selection process eliminates rules with support values below the support threshold.

Table 6: Frequent Itemsets with support threshold =22%

Items	w1	w2	w3	w4	w5	Sup. (%)	Conf. (%)
1	1	1	0	0	1	44	66.67
2	0	1	1	0	0	44	66.67
3	1	1	0	1	0	22	33.33
4	1	0	1	0	0	44	57.14
5	0	1	1	0	0	22	28.57
6	1	0	1	0	0	22	28.57
7	1	1	1	0	1	22	50
8	1	1	0	0	1	22	50

The support threshold value has been taken as 22%. The threshold value is decided on the experimental results. Increasing the threshold value for support, results in generation of very few rules. Table 6 shows the list of frequent itemsets based on support threshold and their corresponding confidence.

Table 7: Selection of the itemsets on the basis of Confidence

items	w1	w2	w3	w4	w5	Confidence (%)
1	1	1	0	0	1	66.67
2	0	1	1	0	0	66.67
3	0	1	1	0	0	57.14
4	1	1	1	0	1	50
5	1	1	0	0	1	50

The parameter confidence indicates the certainty of a rule. Ideally the support threshold value should be greater than or equal to 50%. Table 7 shows the rules obtained after applying Confidence threshold.

3.3. Cloning of selected rule set

Now to optimize the pool of rules generated in phase-1 using Apriori algorithm, the proposed method implements the COLONALG algorithm. The cloning process is carried out such that the clonal rate of a rule is directly proportional to the confidence value (i.e., affinity) of the rule and the average value of clonal rate of every rule is equal to clonalRate given by the user. In particular, the clonal rate of a rule is calculated as follows. Let us denote clonal rate of a rule R as cRate(R) and R1, R2, R3, ……, Rn are rules selected at a certain generation. As the clonal rate of a rule is directly proportional to the confidence value, the clonal rate of Ri will be equal to its confidence value multiply by a constant A.

$$cRate(R_i) = A \times conf(R_i) \quad (1)$$

As the average value of clonal rate of every rule is equal to clonalRate, I have

$$clone\ Rate = \frac{1}{n} \times \sum_{i=1}^{n} cRate(R_i) \quad (2)$$

or $$clone\ Rate = \frac{1}{n} \times A \times \sum_{i=1}^{n} conf(R_i) \quad (3)$$

Thus, $$A = \frac{n \times cloneRate}{\sum_{i=1}^{n} conf(R_i)} \quad (4)$$

After finding the frequent itemsets on the basis of support and confidence threshold, cloning of the rules have been performed on the basis of clonal rate.

Table 8 shows the number of clones which has to be produced for each itemsets in the sample. Thus the total no of clones produced is 8. Table 9 represents the clones generated for each Itemset.

3.4. Mutation

In optimization methods based on natural phenomena, we have the concept of mutation. Mutation is the process of performing some random changes in the result entities which most of the time results in improved output. This approach also implements mutation operation by doing some random changes in the cell (rules) expecting increase in its affinity.

Table 8: Numbers of Clones Generated

No of frequent items	*w1*	*w2*	*w3*	*w4*	*w5*	*No of clones*
1	1	1	0	0	0	3
2	1	0	1	0	0	2
3	0	1	1	0	0	1
4	1	1	1	0	0	1
5	1	1	0	0	1	1

Table 9: Clones generated for the selected itemsets

1	*2*	*3*	*4*	*5*
1 1 0 0 0	1 0 1 0 0	0 1 1 0 0	1 1 1 0 0	1 1 0 0 1
1 1 0 0 0	1 0 1 0 0			
1 1 0 0 0				

To provide more chances for mutation to a low affinity cell so that it can improve its affinity compared to other high affinity cells, here, the mutation rate has been taken inversely proportional to the affinity of the cell. In the proposed system, the mutation rate is equal to single bit i.e. "one word" for every rule. Table 10 shows the mutated clones by single bit mutation.

Table 10. Clones after Mutation

1	*2*	*3*	*4*	*5*
1 1 0 0 0	1 0 1 0 1	0 1 1 0 1	1 0 1 0 0	1 1 1 0 1
1 1 0 0 0	1 1 1 0 0			
1 1 0 0 0				

Now again the support and confidence values of the mutated clones has been calculated. Table 11 shows the mutated clones with their support and confidence.

After calculating support and confidence pruning is done to remove the rules from the pool of rules having lower affinity (confidence measure). Table 12 shows the pruned itemset on the basis of support and confidence threshold.

Table 11. Mutated Clones with their support and confidence value

Mutated Clones	*Support (%)*	*Confidence (%)*
11000	44	66.67
10101	11	25
11100	22	50
01101	11	25
10100	44	66.67
11101	22	50

Table 12. Pruned rules with their support and confidence value

Mutated Clones	*Support (%)*	*Confidence (%)*
10101	11	25
01101	11	25

Table 13 shows the best rules (based on affinity/Confidence) produced after generation one. The same process was continued for next generations. The system repeats each process and update the memory pool with new rules after each generation. After the specified no of generation, the set of rules in the memory will be the final set of rules.

Table 13. Best rules in memory pool after one generation

Rules	*w1*	*w2*	*w3*	*w4*	*w5*	*Confidence (%)*
1	1	1	0	0	0	66.67
2	1	0	1	0	0	66.67
3	1	1	1	0	0	57.14
4	1	1	1	0	1	50

3.5. Evaluation Measures

The proposed system works as a multi-class classifier to predict the best possible sense of the target word. The most common and effective parameters used to evaluate the performance of a classifier are TPR (True positive rate), FPR (False positive rate), TNR (True negative rate) and FNR (False negative rate).

TPR is the proportion of positive cases that were correctly classified.

The FPR is the proportion of negative cases that were classified incorrectly as positive.

The TNR is defined as the proportion of negatives cases that were classified correctly and

The FNR is the proportion of positive cases that were classified incorrectly as negative.

To evaluate the performance of the proposed system, one of the pragmatic performance measure, Accuracy is used. The Accuracy of the system can be defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

4. Experimental Results

This work provides a novel approach using Artificial Immune System based associative classification method for disambiguation, which uses Apriori algorithm for association rule mining and CLONALG algorithm for rules optimization. The proposed method was tested on a corpus of SRS documents. The dataset contains the SRS documents of 8 midsize projects.

The proposed technique was tested by varying clonal factors 0.1 to 0.9 & at different generations 10. 20, 30, 40, 50 and comparative study on the basis of accuracy is presented.

In first test scenario, the method was tested with 3-fold cross-validation for different generations. The result indicates maximum accuracy for 50 generations as shown in Table 14. The clonal factor is fixed at 0.4. Figure 2 shows the graphical representation of the classification accuracy (Identifying most suitable sense of the target word) of COLONALG on the dataset with 3-fold cross validation for varying generations and on fixed clonal factor. This figure shows that generation 50 has maximum classification accuracy.

Table 14: Accuracy (%) of CLONALG Vs No of Generations.

No. of Generation	Accuracy (%) by CLONALG
10	83.2584
20	82.1384
30	87.1913
40	84.3820
50	92.6966

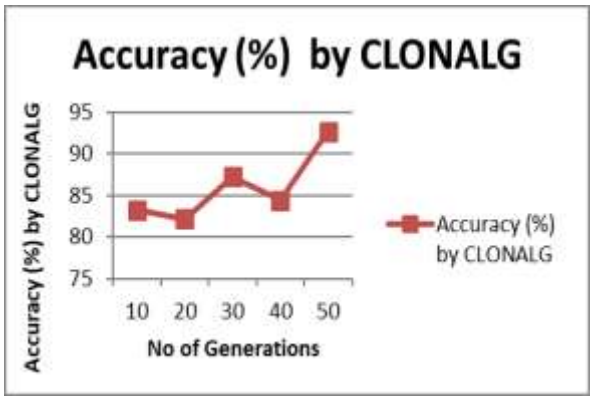

Fig 2. Accuracy of COLONALG Vs No of Generations

In second scenario, the number of generations have been fixed at 50 and the approach was tested by varying the clonal factor. The results of the proposed system with varying clonal factor with fixed number of generations is shown in Table 15. Figure 3 shows the graphical representation of the classification accuracy

on varying clonal factor at the maximum achieved accuracy on the generation 50 for the dataset.

Table 15: Accuracy (%) of CLONALG Vs CLONAL Factor

Clonal Factor	Accuracy (%) by CLONALG
0.1	82.1348
0.2	82.1348
0.3	85.5056
0.4	92.6966
0.5	85.9551
0.6	86.5164
0.7	86.5178
0.8	84.8315
0.9	85.9551

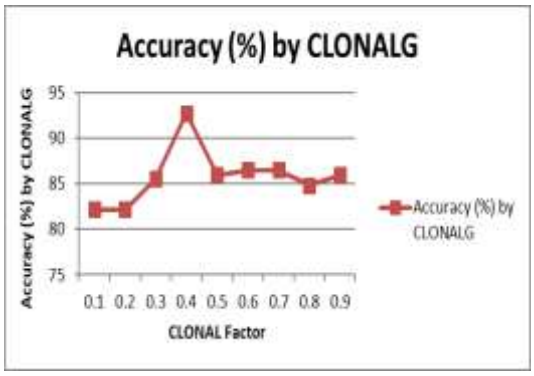

Fig 3. Accuracy of COLONALG Vs Clonal factor

Table 16: Accuracy (%) of CLONALG on different datasets

Data Set	Accuracy (%) by CLONALG
D1	92.13
D2	82.47
D3	95.51
D4	87.70
D5	89.96
D6	86.52
D7	88.52
D8	91.37

Finally, the proposed system is tested on different datasets. Table 16 shows the results of the proposed system when tested on 8 different datasets comprises of software requirement specifications documents for different software projects.

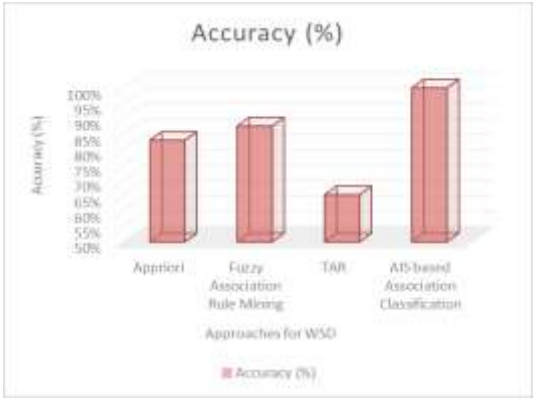

Fig 4. Comparison of Proposed approach with others

To analyze the performance of the proposed system, the method was compared with some existing approaches for WSD. S Kumar & Niranjan have applied Apriori Algorithm & Fuzzy Association rule mining for deducing the sense of ambiguous words in a dataset. There result shows maximum accuracy of 83% & 87.6% respectively [6]. Jung Gil Cho and Won Whoi Huh, have discussed a Tree based Association Rule Mining (TAR) approach dealing with lexical ambiguity present in a dataset. They achieved 65.3% accuracy for their approach [7]. Figure 4 shows the graphical representation of comparative result.

5. Conclusions

Ambiguity is one of the major issue in all the applications that involves NLP (Natural Language Processing). Although a lot of work has been done in this direction by the researchers, still this problem remains a bottleneck. In order to overcome this problem of resolving lexical ambiguities present in a document and improve the efficiency of the existing methods, this work proposed a new approach based on AIS and Association Rule Mining. The proposed method not only indicate the presence of a lexical ambiguous term in the document but also provide the best possible sense of the target word based on the context. The idea behind this proposal was based on:

- That Association Rule Mining approach is quite effective in dealing with WSD according to the results claimed by different researchers.
- CLONALG algorithm can optimize the set of rules generated by Association Rule Mining by making clones of rules having high affinity value i.e. confidence value.

The proposed approach has been tested in different scenarios by varying to important features i.e. number of generations and clonal factor. By going through different runs it is found that the approach best work for clonal factor value of 0.4 and also it gives the best performance if method goes through iteration up to 50 generations.

The proposed approach gives encouraging results compared to some state of the art methods. The approach was applied to a corpus of freely available documents. For comparative analysis purpose, some of the existing techniques were also tested on the same corpus. The result indicates that the proposed approach shows improved accuracy by generating effective association rules. It gives the average accuracy of 89.2725%, which is much better than the state of the art techniques like Fuzzy association rule mining (accuracy of 87.6%), Apriori algorithm (accuracy of 83%), Tree based Association Rule mining (accuracy of 65.3%). The reason behind this improvement is that the association rules were optimized by generating clones of best rules available in a generation and pruning out the rules which are not effective. After iterating the process by a set number of generations, the system gets a pool of results which is considerably better than the set of rules produced by the Apriori algorithm.

In this work simple association rule mining is used to get the association rules. In future, we can apply fuzzy set based association rule mining algorithm for finding association rules and check the performance of the system. With respect to the optimization capabilities of Artificial Immune System, in this work Clonal Selection Algorithm, CLONALG, have been used. In future one can apply other immune system based algorithms like Artificial Immune Network, Negative selection based algorithms for classification. Other nature inspired optimization techniques like BAT search, Dragonfly optimization etc. can also be tested for effective word sense disambiguation.

6. References

[1] M. R. B. Mohd. Shahid Husain, "Advances in Ambiguity less NL SRS: A review," in *IEEE International Conference on Engineering and Technology (ICETECH)*, 2015.

[2] M. S. HUSAIN, "An Approach Towards Ambiguity Resolution in Software Requirement Specifications", Doctoral dissertation,

INTEGRAL UNIVERSITY, LUCKNOW, INDIA, 2017.

[3] M. R. B. M S Husain, "Word Sense Ambiguity: A Survey," *International Journal of Computer and Information Technology,* pp. 1161-1168, 2013.

[4] H. M. S. Yadav Preeti, "Study of Hindi Word Sense Disambiguation Based on Hindi WorldNet," *International Journal for Research in Applied Science & Engineering Technology*, vol. 2, no. 5, pp. 390-395, 2014.

[5] M. S. &. K. M. A. Husain, "Word Sense Disambiguation in Software Requirement Specifications Using WordNet and Association Mining Rule," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, 2016.

[6] S. N. S Kumar, "Design and Implementation of Association Rules Based System for Evaluating WSD," *International Journal of Science and Research 3(6),* pp. 2791-2796, 2014.

[7] W. W. H. J G Cho, "Word Sense Disambiguation Using Tree-based Association Rule," *Advanced Science and Technology Letters, Vol.65 (Database 2014),* pp. 41-44, 2014.

[8] S. &. H. M. S. Sinha, "Proposal for avoiding ambiguity in requirement engineering using artificial intelligence," in *Proceedings of the ACEIT* , 2016.

[9] M. S. Husain, "Nature Inspired Approach for Intrusion Detection Systems," *Design and Analysis of Security Protocol for Communication*, pp. 171-182, 2020.

[10] J. Daudi, "An Overview of Application of Artificial Immune System in Swarm Robotic Systems," *Advances in Robotics & Automation*, 2015.

[11] A. S. D. G. B. D. X. Wang, "Application of Clonal Selection Algorithm in Construction Site," in *International Conference on Sustainable Design, Engineering and Construction*, 2016.

[12] W. M. R. M. P. G. C. Silvaa, " fault detection and isolation: A brief review of immune response-based approaches and a case study," *Applied Soft Computing*, 2017.

[13] "WordNet," [Online]. Available: https://wordnet.princeton.edu.

[14] M. B. Minai, "Word Sense Disambiguation using Evalutionary approach," *Informatica*, vol. 38, pp. 155-169, 2014.

[15] M. S. Husain, "Critical Concepts and Techniques for Information Retrieval System.," in *Natural Language Processing in Artificial Intelligence*, Apple Academic Press, 2020, pp. 29-51.

[16] M. H. Minai A.F., "Metaheuristics Paradigms for Renewable Energy Systems: Advances in Optimization Algorithms," Metaheuristic and Evolutionary Computation: Algorithms and Applications. Studies in Computational Intelligence, vol. 916, 2021.

[17] M. H. I. A. O. N. M. H. MH Adnan, "Modified ISR hyper-heuristic for tuning automatic genetic clustering chromosome size," IOP Conference Series: Materials Science and Engineering, vol. 932, 2020.

[18] M. A. M. S. H. Mohammad Suaib, "Digital Forensics and Data Mining," in Critical Concepts, Standards, and Techniques in Cyber Forensics, IGI Global, 2020, pp. 240-247.

[19] M. S. Husain, "Social media Analytics to predict depression level in the users," in Early Detection of Neurological Disorders Using Machine Learning Systems, IGI Global, 2019, pp. 199-215.

[20] M. A. H. A. F. M. A. N. K. M. A. J. D. K. &. A. I. M. Naseem, "Assessment of Meta-Heuristic and Classical Methods for GMPPT of PV System," Transactions on Electrical and Electronic Materials, pp. 217-234, 2021.