# Assamese Word Sense Disambiguation using Genetic Algorithm

**Arjun Gogoi**
Dibrugarh University
Dibrugarh-786004, India
agogoi440@gmail.com

**Nomi Baruah**
Dibrugarh University
Dibrugarh-786004, India
baruahnomi@gmail.com

**Shikhar Kr. Sarma**
Gauhati University
Guwahati-781014,India
sks001@gmail.com

## Abstract

Word sense disambiguation (WSD) is a problem to determine a word according to the context in which it occurs. There are plenty amount of works done in WSD for some languages such as English, but research work on Assamese WSD remains limited. It is a more exigent task because Assamese has an intrinsic complexity in its writing structure and ambiguity, such as syntactic, semantic, and anaphoric ambiguity levels. A novel unsupervised genetic word sense disambiguation algorithm is proposed in this paper. The algorithm first uses WordNet to extract all possible senses for a given ambiguous word, then a genetic algorithm is used taking Wu-Palmer's similarity measure as the fitness function and calculating the similarity measure for all extracted senses. The winner sense which will have the highest score declared as the winner sense.

## 1 Introduction

The specific task that resolves the lexical ambiguity is commonly referred to as Word Sense Disambiguation(WSD). It is a process to assign a meaning to a word based on the context in which it occurs. WSD is quite important for resolving some problems in such NLP applications as text clustering and classification, word similarity computation, and so on.

Suppose there is a word জিকা/Jika. The word জিকা/Jika is an ambiguous word which has multiple senses.
Sense 1: জিকা/Jika= To Win
Sense 2: লতাজতীয় গছত লগা তৰকাৰী হিচাপে খোৱা এবিধ দীঘল আৰু পাতল ফল /Lota jatio gosot loga torkari hisape khuwa abhid dighol aru patol fol = a kind of Vegetable
**A.** মায়ে আজি জিকাৰ তৰকাৰী ৰান্ধিছে/Maa'e aaji jikar torkari randhise/Mother is cooking jika curry today.
→ This জিকা/Jika gives a sense as a kind of vegetable
**B.**ৰাহুলে বাজী জিকিল/Rahul e baji jikil/Rahul won the toss
→ This জিকা/Jika gives a sense as to win.

Various WSD Methods uses external knowledge source, for example, Word Net/ Thesaurus as a basement to disambiguate the ambiguous words. There are also some hybrid methods, such as an Iterative Approach to WSD, Sense Learner which first used textual definitions of senses from a machine-readable dictionary and then used a corpus to identify and calculate the relatedness between two senses. Another kind of approach to disambiguate words using machine learning-based algorithms (Tatar, 2005 ) rather than taking it directly from an explicit knowledge source such as:

**i**. Supervised WSD: These types of approaches use machine learning techniques and sense annotated corpora to train the data.
**ii**. Unsupervised WSD: These types of methods are based on unlabeled corpora, and make groupings of similar words based on their similarity.

In this paper, we propose to do word sense disambiguation using genetic algorithms which is a novel unsupervised based algorithm. The basic idea is to maximize the overall similarity of disambiguated senses. We propose a novel similarity measure which combines domain information with the Wu-Palmer similarity measure (Zhang et al., 2008) to



*Proceedings of the 17th International Conference on Natural Language Processing*, pages 303–307
Patna, India, December 18 - 21, 2020. 

calculate the similarity between senses in Word-net.

The structure of this paper is as follows: Related works of the proposed approach, Methodology of the proposed approach, Data set and Experiments of the proposed approach, Some close observations, Conclusion of the proposed approach.

## 2 Related works

As no work on the unsupervised approach is done to date in the Assamese language, we had considered work in Indian languages done to date as Assamese is one of the Indian languages. Most of the Indian language structure is the subject-object-verb(sov). The proposed works in the unsupervised approach are as follows:

A work on Malayalam language WSD have proposed by (Sankar et al .,2016) . In this paper, an Unsupervised Approach has been proposed and the proposed system makes use of a corpus which is taken from various Malayalam web documents. Based on the similarity between the given input and the sense clusters, the most similar sense is selected as the winner sense. The proposed system gives an accuracy of 72 percent.

A work on Gujarati language WSD have proposed by (Vaishnav and Sajja,2019) . In this paper, a Genetic algorithm has been proposed which uses a Knowledge-based approach where Indo-Aryan WordNet for Gujarati is used as a lexical database for WSD .

A work on Hindi language WSD have proposed by (Tayal et al., 2015). In this paper, an Unsupervised Approach has been proposed and applied the Fuzzy C-Means Clustering algorithm to form clusters. They test the approach on the corpus created using Hindi news articles and Wikipedia and then compare the approach with other methods available in all the previous works done for the Hindi Language. The training data used consists of 3753 words in total and found an accuracy of approximately79.16 percentage .

A work on Hindi language WSD have proposed by (Sangwan and Singh, 2013). In this paper, a Genetic algorithm has been proposed and this is the very first use of a Genetic Algorithm for the Hindi language WSD. Here, Hindi Wordnet has been used as a lexical database for WSD. To date, they have found some results which are under process.

A work on Gujarati language WSD have proposed by (Mamulkar and Nandanwar, 2020) have proposed "Word Sense Disambiguation for the Marathi language". In this paper, they are proposing the Genetic Algorithm Approach for the disambiguation of ambiguous words. In their survey, it is found that a genetic algorithm gives better results than other approaches

A work on Tamil language WSD have proposed by (Priya et al .,2018). In this paper, the Hierarchical Fuzzy Clustering Algorithm is applied along with the WordNet dictionary. To identify the disambiguated words, sense identification is performed for the adjectives, and comparison is performed. On comparing with previous methods for WSD in the Tamil language, their work gives a better result with a percentage of 91.2 percentage .

A work on Bengali language WSD have proposed by (Pal and Saha, 2019). In this paper, Word Sense Disambiguation has been done using the unsupervised methodology. In this work, clustering has been implemented using the weka tool and the test sentences are extracted from the Bengali text corpus developed in the TDIL (Technology Development for Indian Language) project of the Govt. of India.

A work on Assamese language WSD have proposed by (Sarmah and Sarma,2016) using a supervised Machine Learning approach "Decision Tree-based". In this approach, a few polysemous words with different real occurrences in Assamese text with manual sense annotation were collected as the training and test dataset. They have been able to get an average F-measure of .611 when 10-fold cross-validation evaluation when performed on 10 Assamese ambiguous words.

A work on Assamese language WSD have proposed by (Borah et al., 2014). This is the first work to the best of our knowledge on developing an automatic WSD system for the Assamese language using Naive Bayes approach. They have conducted the experiment

using 25 highly polysemic words using the articles collected from the Assamese textbook of class X.

A work on Assamese language WSD have proposed by (Kalita and Barman, 2020) in which the Walker algorithm, a thesaurus based algorithm, is applied that makes use of the category of each word as defined in the thesaurus. But for a language like Assamese, WSD is far from being a topical approach. Moreover, due to the non-existing thesaurus with a tagged category particularly for the Assamese language.

## 3 Methodology

The initial idea of Genetic Word Sense Disambiguation(GWSD) comes from disambiguating a set of domain words extracted from a domain corpus. Because the terms can belong to the same domain and we can assume that they tend to be similar in semantics with each other. So a good disambiguation solution should be a solution that assigns one sense to each term and increases the overall similarity on the set of selected senses. Each sense of a word corresponds to a synset present in the WordNet and hence we use genetic algorithms to find an optimal set of senses, which increases the overall similarity as it is a global search heuristic. For the WSD task, in a given population each chromosome which is equal to the all possible sense for a given term in the WordNet will be tested using a fitness function, and the most fitted chromosome or sense will be declared as the correct sense.

We give a simple example of initialization. If we want to disambiguate a word, the population size is equaled to the collection of all the possible senses from the wordnet, and for each sense, we calculate the fitness function, and the fittest sense is declared as the winner sense i.e which sense scores the maximum similarity using the fitness function. A fitness function is an objective function that is used to summarize, as a single figure of merit, how close a given design solution is to achieving the set aims. Fitness functions are used in genetic programming and genetic algorithms to guide simulations towards optimal design solutions. In our proposed approach, we are using conceptual similarity between two senses using Wu-Palmer's similarity measure as the fitness function as described in section 3.1.

For example: ভেল/Bhul/Vegetable.

There a total of 2 sense for ভেল/Bhul/Vegetable so 2 will be the population size and then we calculate the fitness function as mentioned above.

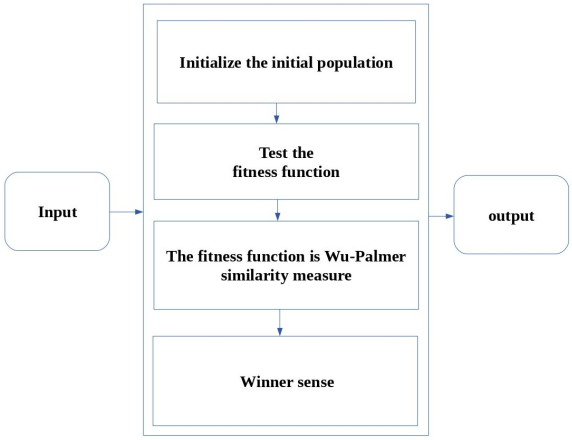


Figure 1: Diagram of the proposed approach

In Figure 1, after we take the input at first initialize the initial population i.e collect all the possible senses from the wordnet. For each chromosome/Sense in the population, test the fitness function(Wu-Palmer's similarity measure). The sense/ chromosome which gets the maximum similarity to assign declared as the winner sense i.e required output.

### 3.1 Conceptual similarity

In our approach, the conceptual similarity between two senses is measured using Wu-Palmer's similarity measure (Zhang et al. , 2008). In Wu-Palmer's similarity measure, Conceptual similarity between a pair of terms C1 and C2 in a hierarchy is defined by (Slimani, 2013) as:

$$SIM(C1, C2) = (2 \times N/(N1 + N2 + 2 * N)) \quad (1)$$

Where N1 and N2 are the distance that separates, respectively, the concept C1 and C2 from the specific common concept and

N is the distance which separates the closest common ancestor of C1 and C2 from the root node.

In our proposed approach, at first to disambiguate a word from an input sentence we draw a tree based on their similarity measure considering the Syn-sets(meaning) from the WordNet and calculate its Wu-Palmer's similarity measure.

For example. Wu-Palmer's similarity measure for the two senses জ্বৰ/Jor/Fever and বেমাৰ/
Bemar/Illness) is given by
SIM(জ্বৰ/Jor/Fever,বেমাৰ/Bemar/Illness)=1.2

## 4 Data set and Experiment

We evaluate the genetic algorithm developed and applied a sense annotated Assamese corpus as shown in the table below. The sense inventory used in this research has been derived from the Assamese WordNet. There are a total of 50000 words and among them, 15606 words are ambiguous (Baruah et al., 2020).

Table 1: Details of the corpus

| Details | Values |
|---|---|
| Total Number of Words | 50001 |
| Ambiguous Words(Nouns) | 15606 |
| Total Number of Unique Words | 12282 |
| Total number of Instances | 12282 |
| Total number of word senses | 50001 |
| Total number of instances per word | 4.07 |
| Total number of senses per word | 1.0 |

Using precision, recall metrics we have done the experiment for genetic algorithm as follows:

$$Precision = correct\ senses/Sentences\ taken \quad (2)$$

$$Recall = correct\ senses/Total\ Sentences\ taken \quad (3)$$

Table 2: Results

| | |
|---|---|
| **Precision** | 81.25 |
| **Recall** | 74.28 |

As no unsupervised works are done earlier, our work is compared with the available works related to supervised approaches in the Assamese language.Though we have been able to get a few papers only which have calculated their results in precision and recall metrics, we have been able to get quite better results while comparing.

## 5 Some close observations

**a**.Very short sentence: Having sentences too short in length, the proposed system could not retrieve sufficient data and it creates difficulty in the case of the similarity measure.
**b**.Spelling error: It is a very important factor as spelling errors in words can decrease the performance of the system.
**c**.Scarcity of information in Assamese WordNet: In this dictionary, synonymous sense definitions of the common Assamese ambiguous words are absent and it is a great difficulty in the proposed approach.
**d**.Same contextual words with different senses: various sentences that are not similar through their similar contextual words. For example:

I. মায়ে আজি ভোলৰ চব্জি ৰান্ধিছে/Maa'e aaji bhulor sobji randhise/Mother is cooking bhul curry today.
This ভোল/Bhul/Vegetable gives a sense as a kind of Vegetable
II.এইটো মুৰ ভুল/Aitu mur bhul/This is my fault.
This ভোল gives a sense as To do something wrong.

## 6 Conclusion

In this paper, the WSD system for the Assamese language using a genetic algorithm has been proposed and analyzed. Works on Assamese language using genetic algorithm

are almost none. Therefore we have an attempt to disambiguate the ambiguous words using a genetic algorithm. This proposed system has been tested on a manually sense-tagged corpus of 30 most ambiguous words. Wu-Palmer similarity measure method has also been applied as a fitness function to the algorithm and found that Precision is 81.25 percentage and Recall is 74.28 percentage.

In the future, the scalability of the proposed system can be improved by adding more ambiguous words to the Assamese language. This proposed system is one target word WSD and can be extended to an all-word WSD system and will show more good results once the Assamese will fully complete.

## 7 Refrences


Deborah Gail Tatar. 2005. Word Sense Disambiguation by Machine Learning Approach. *Fundamental Informaticae*, pages 433-442.

ChunHui Zhang, Yiming Zhou, and Trevor Martin. 2008. Genetic Word Sense Disambiguation Algorithm. *Second International Symposium on Intelligent Information Technology Application*, pages 123-127.

Sruthi Sankar, Reghu Raj, Jayan. 2016. Unsupervised Approach to Word Sense Disambiguation in Malayalam. In textitproceedings of the International Conference on Emerging Trends in Engineering, Science and Technology, pages 1507 – 1513.

Zankhana Vaishnav and Priti Sajja. 2019. KnowledgeBased Approach for Word Sense Disambiguation Using Genetic Algorithm for Gujarati. *Information and Communication Technology for Intelligent Systems*, pages 485-45.

Devendra Tayal, Leena Ahuja, Shreya Chhabra. 2015. Word Sense Disambiguation in Hindi Language Using Hyperspace Analogue to Language and Fuzzy C-Means Clustering. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 49–58.

Shabnam Sangwan and Paramjit Singh. 2013. Genetic Algorithm Based Hindi word sense disambiguation. *International Journal of Computer Science and Mobile Computing* , 2(5):139-144.

Kalyani Mamulkar and Lokesh Nandanwar. 2020. Marathi Word Sense Disambiguation Using Genetic Algorithm. In *IJACEN* , pages 16-18.

Mohana Priya, Krishna Priya, Geetha Harini, Ragavi. 2018. Handling WSD using Fuzzy Hierarchical Clustering of Sentence Level Text. *International Journal of Research in Advent Technology*, 6(11):3209 -3214.

Alok Ranjan Pal and Diganta Saha. 2019. Word Sense Disambiguation in Bengali language using unsupervised methodology with modifications. *Sådhanå*, pages 44:168.

Jumi Sarmah and Shikhar Kumar Sarma. 2016. Decision Tree based Supervised Word Sense Disambiguation for Assamese. *International Journal of Computer Applications*, 141(1):42-48.

Pranjal Protim Borah, Gitimoni Talukdar, Arup Baruah. 2014. Assamese Word Sense Disambiguation using Supervised Learning. *International Conference on Contemporary Computing and Informatics*, pages 946-950.

Purabi Kalita and Anup Kumar Barman. 2015. Implementation of Walker algorithm in Word Sense Disambiguation for Assamese language. In *ISACC*, pages 136-140.

Thabet Slimani. 2013. Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications*, 80(10):25-33.

Nomi Baruah, Arjun Gogoi, Shikhar kumar Sarma, Randeep Borah. 2020. Utilizing corpus statistics for Assamese word sense disambiguation. *CoCoNET'2020.*