Word Sense Disambiguation (WSD) using Neural Networks

Charushila Bhadane, Naresh Thoutam, Mansi Mahajan, Vrushali Suryawanshi

[1] Student, Department of Computer Engineering, SITRC, Maharashtra, India
[2] Assistant Professor, Department of Computer Engineering, SITRC, Maharashtra, India
[3] Student, Department of Computer Engineering, SITRC, Maharashtra, India
[4] Student, Name of the Department, Institute Name, State, Country

ABSTRACT

Word Sense Disambiguation (WSD) is the task of removing ambiguity in different senses of words. It is a core research field in computational linguistics dealing with the automatic assignment of senses to words occurring in a given context [11]. Humans are inherently good at WSD and distinguish senses used in words through spoken language. Computers on the other hand have difficulties identifying correct senses of words. Various advancements have been made in the task of disambiguation using mainly four approaches: Knowledge-based, Supervised, SemiSupervised, and Unsupervised. Better understanding of the human language will help computer's performance in various applications such as search engine optimization, information retrieval, information extraction, software assistants, and voice command interpretation. The objective of this work is to present a supervised neural network machine learning model using various algorithms dedicated to the task of maximizing accuracy of sense detection. The input layer of the neural network will consist of nodes having binary values depending on the presence or absence of frequently occurring context words related to the ambiguous words. The output layer will consist of nodes equal to the number of senses the ambiguous word has. Training and testing of the model will be done using lexical resources such as SemCor or OMSTI. Accuracy will be calculated based on All- Word tasks from SemEval International Workshops

Keyword: - *Machine Learning, Neural Network, Classification Algorithm*

1. INTRODUCTION

Word-Sense Disambiguation (WSD) is a branch of Natural Language Processing (NLP) which specifies some open problems concerned with identifying the correct sense of a word used in a respective sentence. Many words used in the English language have various different senses or meanings. WSD is concerned with the problem of selecting the correct meaning. The solution to this problem impacts improving relevance of search engines. The human mind is very proficient at word-sense disambiguation. Simple context is all that is needed for humans to understand the correct sense or meaning of a word. Human languages have developed due to the intellectual ability of neural networks in human brain. In computerscience it has been a long-term challenge to develop the ability in computers to perform language processing on the scale that humans do. For example, consider a word bass in English which has two meanings: any of various North American lean-fleshed freshwater fishes and the other: denoting the member of a family of instruments that is the lowest in pitch.

1.1 Introduction to Machine Learning

Machine Learning (ML) is defined as programming computers to optimize a performance criterion using example data or past experience. In our system, the performance criterion is the accuracy of the model on testing data. Supervised ML consists of 2 main parts: Training and Testing. Training comprises of feeding labelled data into the model to gain experience. Testingcomprises of predicting outputs by trained model based on experience.

1.2 Machine Learning in WSD

As stated before, the human brain is masterful at distinguishing between various senses of a word based on their context. The best way to reproduce this capability within machines is to make the computer think like humans do, allow it to learn from experience and make predictions based on this experience. The way this is implemented is MachineLearning, specifically using a Neural Network.

1.3 Artificial Neural Networks

A neural network is a network or circuit of neurons composed of artificial neurons or nodes. Artificial neural networks (ANN) are composite layers of compute units that process data individually in order to simulate the working of a human brain. Similar to human brains, ANNs learn with experience and show improvements in tasks when data available is increased. ANN consisting of one or two hidden layers are called shallow neural networks and those with more hidden layers are called deep neural networks.

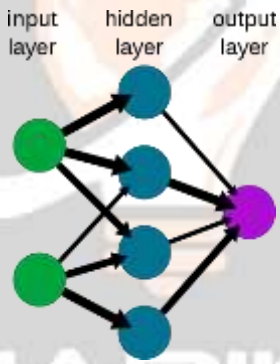

Fig. 1. Simple Neural Network

2. BACKGROUND

A large number of methods of Word Sense Disambiguation (WSD) have been studied and researched in thepast. These methods mostly include four different approaches to WSD: Knowledge based, Supervised, Semi-Supervised, and Unsupervised.

2.1 Knowledge-based Approach

Knowledge based algorithms use various lexical resources such as Machine-Readable Dictionaries (MRDs), WordNet to identify the correct sense of words. These Algorithms are easy to implement and were the first to be developed while trying to solve the problem of WSD. A knowledge-based system only needs access to commercial dictionary resources to start process of disambiguation.
Drawback of these algorithms is that theirperformance is limited on the speed of searching andretrieval of

these resources. As the size of theresources increase, so does the latency and hence performance decreases. [1]

2.2 Supervised Approach

Supervised methods are called so because they require human assistance. Large amount of labelled data is required to make supervised models perform as expected. The larger the data set available, the greater is the prediction accuracy of these system.

A learning set is prepared for the system to predict the actual meaning of an ambiguous word using a few sentences, having a specific meaning for that particular word. A system finds the actual sense of an ambiguous word for a particular context based on that definedlearning set [1].

Supervised approach always gives superior performance than any other methods. However, these supervised methods require large data sets, and are therefore limited in their capabilities. Such data requires manual tagging of senses which is an expensive and time-consuming task.

2.3 Semi-supervised Approach

Many word sense disambiguation algorithms use semi-supervised learning which provides a sort of compromise between supervised and unsupervised approaches. They allow both labelled and unlabelled data and are therefore useful when there is a lack of training data. The bootstrapping method starts from a small amount of seed data for each word: either a small number of sure fire decision rules (e.g., 'play' in the context of 'bass' almost always states the musical instrument) or manually tagged training corpus.

Using any of the supervised methods, small amount of tagged or labelled data is used to train an initial classifier. This classifier is then fed unlaballed data in order to extract a larger labelled dataset in which only the perfect classifications are included. Such processes are usually iterative each iteration training being done on a successively larger dataset. The obtained data set becomes larger and larger until we stop the process after a certain number of iterations have been reached or the maximum size of dataset is reached.

2.4 Unsupervised Approach

Unsupervised learning methods are the most difficultto implement for WSD researchers. Using Unsupervised ap- proaches, we basically mean to say that word senses can be deduced using other similar sentences. Using Clustering algorithms, such sentences with a certain degree of similarity can be grouped together with each cluster specifying one sense of a word. This process is called Word Sense Induction.

As expected, the performance of such algorithms has been shown to be less than the other methods of WSD due to lack of training data but it is hoped that in the future, the unsupervised techniques can successfully overcome the problem of scarcity of expensive manually tagged data in order to be the most efficient sense prediction approach.

3. LITREATURE SURVEY

1) 1986 Michael Lesk [5]: The paper written by MichaelLesk in 1986 has been proved to be revolutionary work in Word Sense Disambiguation (WSD). In this paper, he pre- sented his famous Lesk algorithm which has been the pivotal algorithm for knowledge-based approach WSD. The Lesk algorithm uses various machine-readable dictionaries (MRDs) to find correct senses of words. The algorithm searches for overlaps in various senses or signatures of a word. Senses having maximum overlap are chosen as the correct senses of the word. Lesk has concluded that the algorithm produces an accuracy of about 50-70% depending on the MRD used.

2) 2015 Udaya Raj Dhungana, Subarna Shakya, Kabita Baral and Bharat Sharma [4]: In this paper, they used the knowledge-based approach. They have used adapted Lesk algorithm to disambiguate the polysemy word in Nepali language. They grouped each sense of a polysemy word based on the verb, noun, adverb and adjective with which the sense of the polysemy word can be used in a sentence. The experiment is performed on 348 words (including the different senses of 59 polysemy words and context words) with the test data containing 201 Nepali sentences shows the accuracy of their system to be 88.05%.

3) 2016 Ignacio Iacobacci, Mohammad Taher Pilehvar, Roberto Navigli [3]: The main focus of this paper is on word embedding. i.e., is to collect the semantic information from the collection of the datasets. It is an example of knowledge- based approach. Word embedding is usually a collection of names for a set of language modelling and advanced learning techniques in the natural language processing.
In this the results are evaluated by using two methods:

1. Lexical Sample WSD Experiments.
2. All words WSD Experiments.

The main interests were on the training parameters of embed- ding and WSD features which were impacting on the WSD performance. The maximum accuracy observed during this experiment was 69.9%.

4) 2017 Pratibha Rani, Vikram Pudi, Dipti Misra Sharma [2]: In this paper, the authors have presented a generic Word Sense Disambiguation (WSD) method using semi-supervised approach. They explain that current WSD systems use ex- tensive domain resources and require advanced linguistic knowledge. Therefore, to improve these factors, they propose a system that extracts context-based list from a small amount of seed data containing sense tagged and untagged training data. Their experiments in Hindi and Marathi language domains show that the system gives good performance without lan- guage specific information with exception of sense IDs present in the training set, with approximately 60-70% precision.

5) Semi-Supervised WSD with Neural Models [10]: In this paper, the research team have used Supervised and Semi- supervised approaches to Word Sense Disambiguation (WSD) using Neural Nets and label propagation method respectively. They compare and contrast the 2 different ways of obtaining word embeddings including using Word2Vec model and a re- current neural network model using Long Short-Term Memory (LSTM) architecture.
This LSTM Recurrent Neural Network shows extremely high accuracy results when tested in SemEval International Workshop Tasks with training done on both SemCor [7] and OMSTI [8]

They also implement Semi-supervised WSD using label propagation method. This method works by representing la- belled and unlabelled examples as vertices in a connected graph. The label information is then circulated from any vertex to nearby vertices through weighted edges iteratively, finally inferring the labels of unlabelled examples after the propaga- tion process converges.

The accuracy results of this work are consistently higher than any previous research models. We plan to use the Long Short-Term Memory (LTSM) Architecture as the team have used to build Word Embeddings. But instead of testing data based on these word embeddings, we plan to use them to create feature vectors for our neural network. We hypothesize that using LTSM input vectors along with suitable cost reduction functions for our neural network will produce an increase in accuracy. Using principal component analysis (PCA), feature reduction can also be done to help visualize the data for better understanding.

4. DATASETS

List of approximately 1000 highly ambiguous words will be created from various dictionary sources such as dillfrog [9]. The words in this list have a minimum of
10 different senses and a maximum of 73. Each ambiguous word will have its own neural net and word embeddings.

Supervised Machine Learning requires several thousand datasets for effectively training and testing models. The more data that is made available to machine learning models, the more accuracy it can achieve. Neural Networks are no different. For the purpose of WSD using neural models, we will require massive amounts of labelled or tagged data which will help us to train and test our model and help it achieve high accuracy. For such purposes, we will be using 2 different labelled data sources: OMSTI and SemCor.
1) OMSTI: OMSTI (One Million Sense Tagged Instances) [8] and SemCor (Semantic Cortex) [7]. OMSTI is a dataset of a million examples of sense tagged sentences that was created for the purposes of supervised WSD.
2) SemCor: SemCor on the other hand has around a few hundred thousand instances but these datasets are manually tagged and hence more accurate.
3) Wordnet: All the resources mentioned above are anno- tated using WordNet [6] 3.0. It is a large lexical database of English words. WordNet is open source and is easily imported and executed in plenty of languages. Its primary use is in automatic text analysis and artificial intelligence applications.

A. Loading the Dataset

The datasets will originally be an xml file which will be loaded in the python environment with the help of Pandas library. The pandas will be used to label the dataset and split the dataset into Training set, Validation set and Test set. We intend to split the dataset in the ratio of 6:2:2

5. IMPLEMENTATION DETAILS

Various solutions to word sense ambiguity have been put forward. Most of these systems use one of the four approaches mentioned earlier. Out of these approaches, supervised ap- proach to WSD has been proven to produce maximum ac- curacy. Therefore, in our proposed model we will be making use of supervised approach as well. As mentioned earlier, Artificial Neural Networks mimic the functioning of a real human brain. Given the context words in which an ambiguous word occurs, the neural net should be able to successfully predict the correct sense of the ambiguous word. For creating a accurate neural netclassifier we also require large amounts of labelled data as such a model falls under the supervised approach to WSD. We will be using SemCor [7] and OMSTI [8] labelled data sets for this purpose. Once trained, we can judge the accuracy of the Neural Net byusing the test data set. The classifier should also be able to return the correct sense of an ambiguous word based on input given by user.

A. Word Embeddings

The input feature vector of an ambiguous word for the neural network will be created using its word embeddings. The vectors we use to represent words are called neural word embeddings. Word Embeddings are created using the words similar and most commonly used context words. They can be created using various methods such as Word2Vec, Recurrent RNN models such as LSTM, etc. Word Embeddings measure cosine similarity,i.e. no similarity is expressedas a 0, while total similarity is expressed as 1.
Example:

Figure 2 shows word embeddings for the word 'Sweden'. Since Norway and other Scandinavian countries are closely related to Sweden, their cosine values are closest to 1.

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

Fig. 2. Word Embeddings for the word 'Sweden'

B. Feature Vectors

The input feature vectors to the Neural Network will be created using both the input of dataset and the word embeddings of the ambiguous word. If embedding words are present in the input context then they will be represented in the feature vector with value '1', else with a '0'. After every word in input context is checked for its presence in word embeddings, the resultant feature vector will be passedto the Neural Net input layer. Therefore, number of nodes in the input layer of Neural Net will be equal to the number of word embeddings we use.

Example:
Assume that the word embeddings for the word 'crown' are: [jewels king teeth drill dentist]

and that inputis the sentence: “The dentist did a really good job putting the crown on my teeth”
Then the Input Feature vector will be:[0 0 1 0 1]

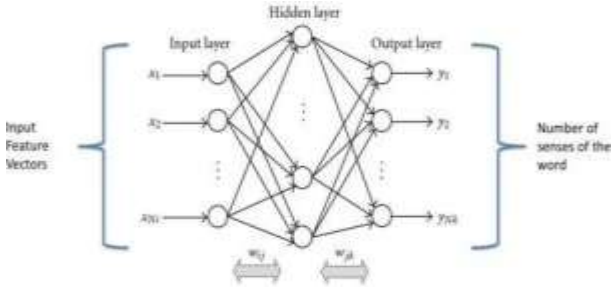

Fig. 3. Neural Network for an Ambiguous Word

C. Feed Forward Neural Networks

The following figure shows an example of a forward prop- agation step in a Feed Forward Neural Network. In vectorised implementations, input sets are represented in columns of a matrix which are multiplied by a weight matrix theta that represents each layer of a neural network. The use of matrices reduces time and increases efficiency for calculations as it does not require loops in the program structure to multiply each element individually.

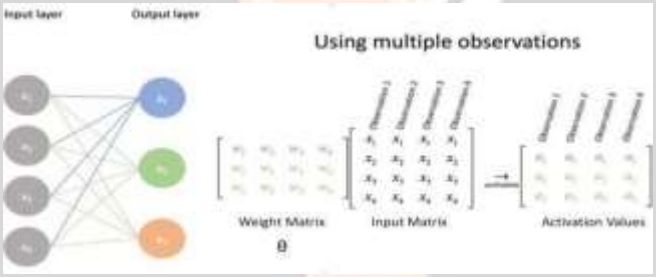

Fig. 4. One step of forward propagation

D. Cost Calculations of Neural Network

Calculating costs is the one definitive way of understanding that our Neural Network is working correctly. After every iteration the cost of the neural network is calculated using the cost function given below. Displaying the value of cost every few hundred iterations can help us accurately gauge whether our neural net is actually learning or not.

$$h_\Theta(x) \in \mathbb{R}^K \quad (h_\Theta(x))_i = i^{th} \text{ output}$$

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right]$$
$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2$$

Where:

- m = number of features
- K = number of output layer nodes
- L = number of layers in network
- θ represents weight matrix

Fig. 5. Cost Function for Classification Algorithms

E. Output of Neural Network

The output layer of neural network contains nodes equal to the number of different senses for the ambiguous word, according to the WordNet [6] dictionary. The node for which the highest numerical value is calculated among the different output nodes will represent the predicted sense. If the third node of output layer has the highest value, then it means that the system has predicted the third sense, all sensesbeing annotated by WordNet [6].

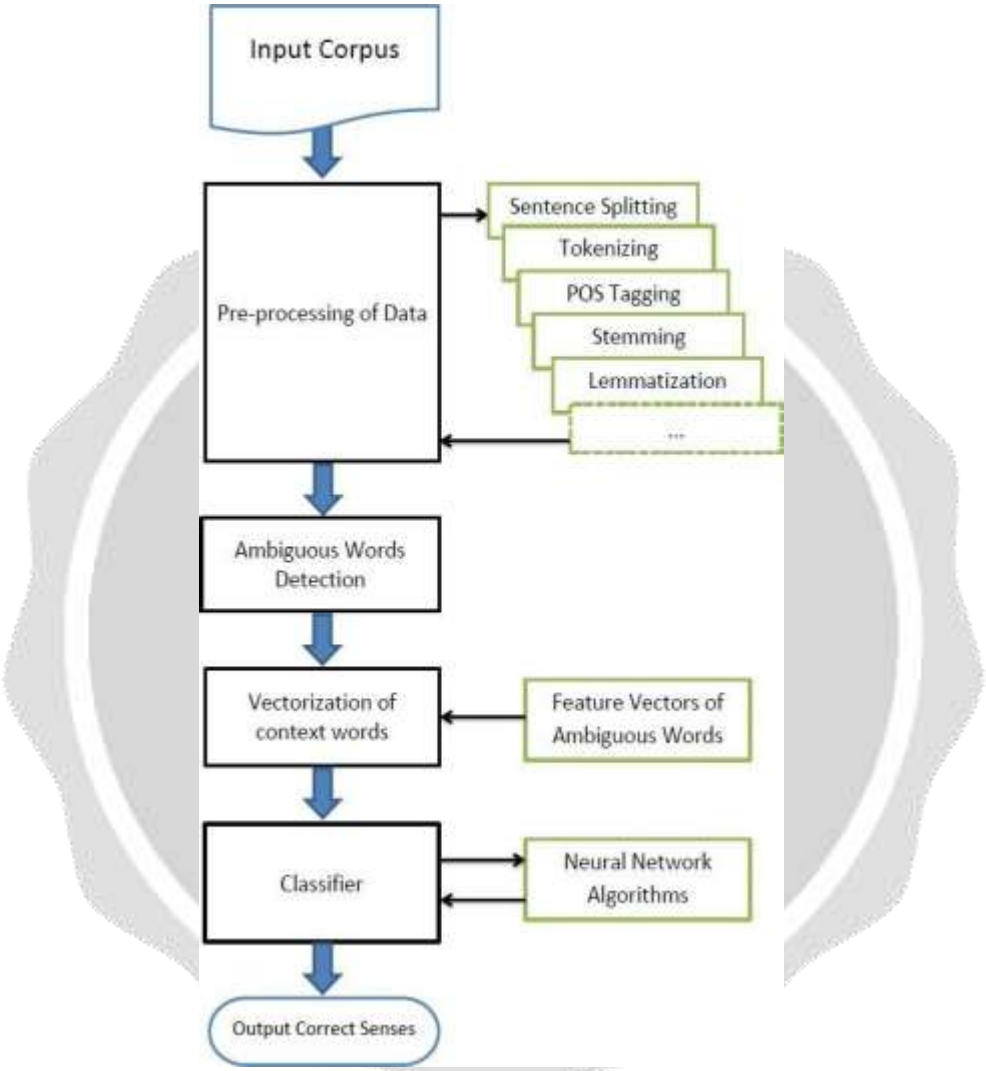

Fig. 6. System Architecture

6. CONCLUSIONS

This work proposed a Word Sense Disambiguation (WSD) Model using Neural Network Algorithms that aim to maxi- mize accuracy for the given natural language processing task. Building on previous research work, our system hopes to further improve sense prediction accuracy and help in human-computer interfacing for future applications.

7. REFERENCES

[1]. Kokane, Chandrakant D., and Sachin D. Babar. "Supervised Word Sense Disambiguation with Recurrent Neural Network Model."

[2]. Kokane, Chandrakant D., Sachin D. Babar, and Parikshit N. Mahalle. "An Adaptive Algorithm for Lexical Ambiguity in Word Sense Disambiguation." Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020. Vol. 169. Springer Nature, 2021.

[3]. Ignacio Iacobacci, Mohammad Taher Pilehvar, Roberto Navigli" Embedding for Word Sense Disambiguation: An Evaluation Study" 2016 Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 897-907

[4]. Udaya Raj Dhungana, Subarna Shakya, Kabita Baral and Bharat Sharma" Word Sense Disambiguation using WSD Specific WordNet of Polysemy Words" 2015 Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing

[5]. Michael Lesk// Automatic Sense Disambiguation using Machine Read- able Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone// Bell communications Research Morristown, NJ 07960 George A. Miller

[6]. WordNet: A Lexical Database for English.

[7]. Communications of the ACM Vol. 38, No. 11: 39-41 Bell communications Research Morristown, NJ 07960 https://wordnet.princeton.edu/ George A. Miller

[8]. SemCor: Semantically Annotated English Corpus Princeton University Kaveh Taghipour and Hwee Tou Ng

[9]. One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction 2015 Conference on Computational Language Learninghttps://muse.dillfrog.com/ Online Dictionary

[10]. Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, Eric Al- tendorf 2016 Semi-Supervised Word Sense Disambiguation with Neural Models Google, Mountain View CA, USA

[11]. Mohammad Taher Pilehvar, Roberto Navigli. A large- scale pseudo word- based evaluation framework for state-of-the-art word sense disambiguation 2014