[1]Madhuri P. Karnik

[2]Dr. D. V. Kodavade

Homonym Detection Using WordNet and Modified Lesk Approach for English Language

Abstract: - Word sense disambiguation (WSD) is a basic and persistent problem that has existed since its inception in the natural language processing (NLP) area. The process of determining the accurate meaning of a word within a specific context is referred to as word sense disambiguation, commonly known as WSD. In NLP, a single word can have two or more meanings, with each meaning being distinguished by its context. This is known as word polysemy. Its applications span a wide range of fields, such as question answering systems, machine translation, information retrieval (IR) etc. Ontology and NLP are still struggling with ambiguity. Homonyms, which are ubiquitous in most languages, are words that have the same spelling but a different meaning. This method's fundamental premise is to select the appropriate sense by comparing a word's context in a sentence to contexts generated from WordNet. The primary goal of this study is to employ WordNet and the Lesk algorithm for WSD. After the algorithm was put into practice and tested on a collection of sentences that included ambiguous words, the synset was able to determine the proper interpretation for most of the sentences. The Lesk algorithm relies on finding the highest number of shared words (maximum overlap) between a word's context, prepositions and the definitions for its different meanings(glosses). This approach helps in identifying the most accurate interpretation for a given word within a specific context. According to experimental findings, the suggested strategy considerably boosts performance while identifying homonyms.

Keywords: WSD, Homonym, WordNet, Context, Lesk, Natural Language Processing.

I. Introduction

The amount of information generated has increased tremendously in recent years. Finding relevant information among vast volumes of data is therefore quite difficult. An essential component of human-machine communication is NLP. WSD is a critical problem in NLP, focuses on determining the sense of a word used in a sentence in order to obtain the precise and accurate meaning of the sentence. WSD is a crucial field in natural language processing, and one of its applications is determining whether a word used in a sentence is ambiguous. The English language contains a vast number of terms with diverse meanings and interpretations. In natural language, a term can have multiple meanings. This is a problem in NLP that WSD can help with. NLP still has this unsolved issue. It is a component of NLP communication. Uncertainty is a typical occurrence in human language. By writing or reading the other words in the context, humans are able to determine the correct meaning of a word. WSD is approached using two different methods: the Machine-Learning Based approach and the Knowledge-Based approach.

A. Machine-Learning Based approach

WSD is a job that is learned into systems using a machine learning approach. This method trains the system to identify the appropriate sense. The technique involves providing the system with the ambiguous word and its surrounding content as input. Supervised, unsupervised, and semi-supervised procedures are the three categories of machine learning-based methodologies.

- Supervised approach: This method uses sensible annotated corpora as a training set. Using a tagged dataset of word senses, a model is trained using supervised approaches for WSD. The target word's sense in a fresh text is then clarified using the model. The training data set for the classifier includes examples pertaining to the target word. Among the techniques that are commonly used are SVM, naive bayes, decision trees and neural networks.
- Unsupervised approach: The unsupervised technique for offering potential meanings for a word in context relies solely on raw annotated corpora, without utilizing any sense-tagged corpus. These techniques include cooccurrence graphs, word clustering and context clustering. The fundamental idea is that senses can be

[1]Ph.D Scholar, Shivaji university, Kolhapur, Maharashtra, India

madhuri.chavan@viit.ac.in

[2]Professor, D.K.T.E. Society's Textile & Engineering, Ichalkaranji, Maharashtra, India

dvkodavade@gmail.com

inferred from the text by grouping word occurrences based on a metric of context similarity. This is because comparable senses occur in similar circumstances.

- Semi-supervised method: This method combines supervised and unsupervised machine learning techniques. A technique known as semi-supervised learning combines a large number of unlabeled instances with a limited number of sense-labeled examples.

B. Knowledge Based approach

knowledge-based strategy relying on external lexical resources such as thesaurus, corpus, and other machine-readable dictionaries. Its foundation is the notion that words are related to one another when they are employed in a text and that this relationship is evident when looking up definitions and meanings. To distinguish between two or more words, the dictionary senses with the highest word overlap in their definitions are utilized. The traditional approach based on knowledge-based WSD is the Lesk algorithm.

Consider the examples of the word “park” for which different meanings exist.

1) We went for a walk in the **park.**
2) He found a place to ***park*** the car.

The term “park” appears and indicates that its meaning is distinct. For instance, in the first phrase, it means “an open area in a town or a place filled with greenery” and in second, it means “to leave the vehicle that you are driving somewhere for a period of time”. Therefore, the researchers' consideration of this matter is vital and significant for the accurate translation of the statement as well as many other applications like text summarization, Question answering system etc. These words with several meanings are referred to as ambiguous words, and WSD is the process of determining an ambiguous word's precise meaning in a given situation. WSD is the process of automatically assigning a polysemous word in a particular context its proper meaning. Nonetheless, a number of technical issues, such as homophones, can seriously impair the viability and usefulness of systematic reviews. As a result, finding homophones is one of the most crucial aspects of text mining and has been thoroughly researched across a number of fields. The topic of automatic WSD for the English language has been the focus of many studies. Here, we're concentrating on the English language, and within it, there are a great deal of unclear terms whose meanings are revealed by their context and sentence structure. This paper is organized into several sections, beginning with a summary of previous studies related to Word sense disambiguation (WSD). Following this, the paper discusses a system that has been proposed and an algorithm that has been modified for use in WSD, analyses the adapted algorithm's performance and concludes with suggestions for further research.

II. Related Work

A substantial portion of the many research on homophone word identification have been done in a particular context. The Lesk algorithm, first presented by Michael E. Lesk , in the paper [1] for WSD. The underlying premise of the Lesk algorithm is that words inside a specific textual “neighbourhood” will typically have a common topic. An adaption of the Lesk Algorithm for WSD is presented in the study [2]. This expands these comparisons to include the glosses of words that are connected to the words in the text being disambiguated, whereas the original method relied on identifying overlaps in the glosses of nearby terms. A comparison of two supervised approaches to the problem of ambiguity has been performed in the study [3]. They have compared the new Lesk algorithm and Support Vector Machine (SVM) and looked at how it affects the Hindi language. 10 Hindi words were used in the algorithm comparison.

An algorithm for performing word disambiguation in a given context utilizing Lesk via WordNet has been described in the work [4]. In the paper [5], authors have presented an efficient WSD model. Their method makes use of a LSTM network that is shared by all words and is bidirectional. As a result, the model may scale well with vocabulary size and share statistical strength. Word order is efficiently utilized by the model, which is trained from start to finish. They have assessed their method using the same hyperparameter settings on two standard datasets, and then fine-tune it on a third set of held-out data. Moreover, the system was created with the intention of generalizing to complete vocabulary WSD by sharing the majority of the word properties. The optimization of the computational complexity linked to the Lesk-based algorithm, a well-liked and successful knowledge-based algorithm, has been studied by the authors in the study [6]. Their research shows that good performance can be obtained while significantly reducing complexity.

The Lesk technique, which uses the polysemy word of the verb in the Hindi sentence, is the foundation of the study work [7]. To get the best overall performance, both stop word removal and stemming deletion are done. The verb words that have the highest value are allocated the correct sense. The research presented in [8] proposes the incorporation of automatic detection of homophones and homographs as a new feature for humour recognition systems. The approach integrates ambiguity-based features with style-properties from earlier research on humor recognition in brief text. Two possible practical homograph identification techniques are compared utilizing crowdsourced annotations as the ground truth. The authors of the paper [9] have described and assessed a classifier that uses a straightforward multilayer perceptron structure to identify homonymous and synonymous author profiles. By extracting a gold-data collection of profiles from previous years' active manual curation, their classifier was made possible. In the study [10], authors have employed artificial neural networks in conjunction with automated content analysis to efficiently and precisely sort through enormous collections of scholarly articles and assign them to various subjects. They have looked into the usage of the term "reintroduction" in academic writing, for instance.

The authors [11] have proposed a system that consists of three units - WSD classifier, pre-processing, and input query. The input query receives an unstructured query from the user while preprocessing unit transforms this query into structured form which is then transformed to the WSD classifier. WSD classifier employed context information from the query and a lexical database to uniquely identify the sense of polysemous words. WordNet was used as the knowledge source in their work.

The paper [12] compares various WSD approaches in supervised, unsupervised, and knowledge-based algorithms to provide an overview of WSD approaches in popular AI-NLP techniques. Moreover, through the comparison of accuracy and the identification of strengths and weakness in diverse surveyed systems, this aims to offer a gap analysis within surveyed systems. In the paper [13], the fundamental concept is selecting the appropriate sense by comparing the word's context in a sentence to contexts generated from WordNet. The Marathi WordNet and Lesk technique was employed by the authors of this research to disambiguate Marathi words. At the moment, their system only handles nouns. The problem of homonym identification was examined in the project [14], where an experiment was conducted to verify the validity of a hypothesis underlying homonyms, which states that contextual information in the form of word embeddings is adequate for homonym identification. A unique solution to the issue of differentiating between homonymy and polysemy has been put forth by authors [15]. Their techniques for proving semantic relatedness are based on formal theories of senses, synonymy, and translation and they make use of sense translation data from a multi-wordnet.

Sequential Contextual Similarity Matrix Multiplication (SCSMM), a unique knowledge-based WSD technique, is presented in the study [16]. The proposed algorithm uses the sentence's local context, past knowledge of the term's usage, and the document's global context—represented, respectively, by the terms' semantic similarity, term frequency heuristics, and document context—to simulate the disambiguation process of the human brain. A hybrid method for determining the word sense based on the collocation score has been provided in the work [17]. The suggested approach blends knowledge-based and corpus-based methodologies. The Senseval and SemEval datasets were used for the experimental evaluation.

In the paper [18], the authors propose to construct unified sense representation using Babel synsets and transfer annotations from rich source languages using alignment and machine translation tools in order to build feasible knowledge and supervised based systems for multilingual WSD. The work [19] provides a detailed description of the steps involved in identifying homonymy between groups of grammatically related Uzbek words using a naive Bayes classifier.

A novel annotation layer for the Princeton WordNet has been introduced by the authors [20]. This layer divides senses into lemmas and allows for the distinction between polysemy and homonymy. Their techniques functioned by establishing a connection between WordNet and the Oxford English Dictionary, which has the necessary data. They have linked definitions according to how close they are in an embedding space generated by a Transformer model in order to carry out this alignment.

III. METHODOLOGY

Following Fig. 1 depicts the proposed system architecture.

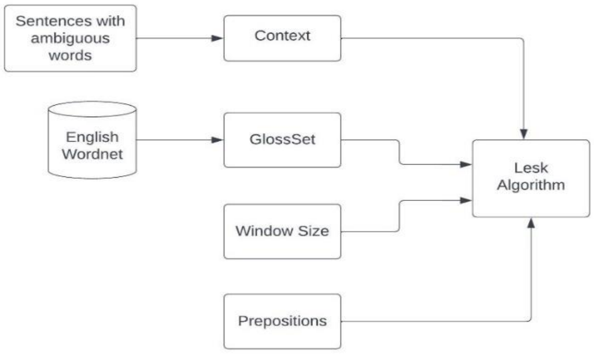


Figure.1. Block diagram of the proposed system

The proposed method leverages WordNet to disambiguate word definitions using the Lesk algorithm. Ambiguous words in sentences are inputs for the system. The suggested system's gloss set is a collection of semantic relations for ambiguous words gathered from WordNet, and the context set is a set of words from the surrounding window that contain ambiguous terms. The algorithm also takes English prepositions into account.

*A. **WordNet***

The English language's lexical database, or dictionary, is called WordNet. It is the most widely utilized resource for knowledge-based approaches that aim to clarify the meanings of words that have many meanings. For numerous applications involving NLP, Wordnet has shown to be a valuable lexical resource. Cognitive synonyms, also known as synsets, are collections of nouns, verbs, adjectives and adverbs that collectively convey a particular idea [21]. Synset is a unique type of straightforward interface that is included in NLTK for WordNet word searches. WordNet is used to help resolve ambiguities about the meaning of polysemy terms. Related words from synset, gloss, and other hypernym levels are gathered from the WordNet database and analyzed to identify overlaps utilizing WSD [4].

The suggested method identifies the appropriate sense definition in a given context for every ambiguous word in the synset. In this case, our focus is on homonym WSD within sentences. “Any word which shares identical spelling or pronunciation with another word” is the definition of a homonym. Words that have the same sound but differ in spelling or meaning are called homophones while homographs have the same spelling but different pronunciations or meanings.

Following are some examples:

1.Homographs (minute/minute, present/present)

- minute 1: small   or
  minute 2: measurement of time
- present 1: gift   or
  present 2: to bring forth

2. Homophones

- roll and role
- steal and steel

*B. **Lesk Algorithm***

Michael E. Lesk developed the Lesk algorithm in 1986[1], which is a traditional method for WSD. It is predicated on the notion that words that occur together in a text have a relationship of some kind, and that by looking up the definitions of the words of interest and the phrases that surround them, one may determine this relationship and the context in which the words fit. To put it simply, Lesk's method counts the number of times a word of interest's dictionary definitions overlap with all the definitions of the words that surround it, or what is called a "context window". The meaning of the word with the highest number of overlaps is then inferred. Word and context word matching is the only basis for overlap-based methods such as Lesk and Extended Lesk. Satanjeev Banerjee, Ted Pedersen [2] recommended this strategy. The dictionary-based approach to sense disambiguation that has received the greatest research attention is the Lesk algorithm.

C. Modified Lesk Algorithm

The process of detecting homonym is given in the algorithm - *Modified Lesk Algorithm.* Let sd be the sense definition and se be the sense examples extracted from wordnet. Stop words are removed.

s = individual sense of ambiguous word within a synset

wW = dictionary where weight of each sense is stored

u = each word of input sentence

sd = sense definition

se = sense examples

c = window size

d=distance of word from ambiguous word

w = weight of that particular word in the sentence for a particular sense

x = variable used to store the summation of weight for a particular word

max-weight=0

best sense = correct sense

sentence=total weight of input sentence.

Modified Lesk Algorithm

```
initialize wW
for each sense s in synset do:
extract sense’s definition sd, and sense’s examples se
sentenceW=0
     for each word u in direction € {left, right} of the input
     sentence do
          initialize x = 0
     calculate distance d from ambiguous word
   if d < = c then
             calculate weight w based on distance
             w=(1/(d+1))
        if u is a preposition on the left side of the ambiguous
        word then
                 w = w*2 (double the weight)
                 x += w
        if u is in sense’s definition sd then
        x += w
    if u is in sense’s example se then
        w = w*0.5 (half the weight)
        x += w
        sentenceW += x
  end
    append sentenceW to weights list wW
    if sentenceW > max-weight then
        max-weight = sentenceW
        best-sense = sense
end
return (best-sense)
```

In this approach, every word in the input sentence at a distance d from an ambiguous word is denoted by u. Every word u is being compared to terms from sense's example and definition. The input sentence's weight is determined and contrasted with the weights of the other senses.

IV. Results

A set of sentences with ambiguous terms were used to test and implement the suggested modified Lesk method. We examined various window sizes and assessed the proposed algorithm in the experiment. Table 1 shows F1-score for various Window-sizes and Table 2 shows the results of implementation of proposed modified algorithm. Fig. 2 displays the accuracy and F1-score for the given dataset.

Table 1 Results for various Window-sizes

Window-Size (WS)	Accuracy	Precision	Recall	F1-Score
WS-1	52.0	43.3	52.0	45.9
WS-2	56.0	49.3	56.0	51.3
WS-3	68.0	61.3	68.0	63.3
No WS	76.0	70.0	76.0	72.0

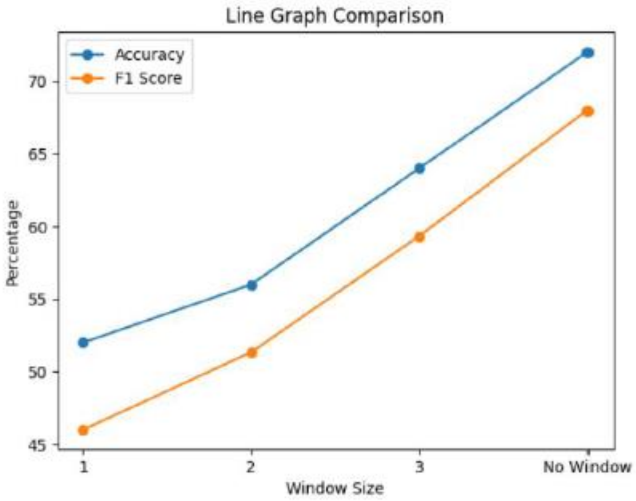


Figure 2. Accuracy and F1-score

V. Conclusion

The open problem of word sense disambiguation (WSD) deals with determining which sense of a polysemous word is valid. The accuracy of the systems doing this task needs to be increased in order to close the gap between humans and computers and to create better interfaces. In this research, we have implemented the modified Lesk algorithm with WordNet for English language homonym word sense detection. In our sample evaluation set, the synset has deduced the correct meaning for many of the sentences and in very few instances, it was unable to produce the desired outcomes. When recognising homophones, the suggested approach performs better when prepositions are included. Currently, our technology handles homophones and homographs in the sentences. In the future, we can expand the dataset, take into account a greater number of sentences including ambiguous words, and identify the proper sense of homophones.

References

[1] Lesk, Michael. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." In *Proceedings of the 5th annual international conference on Systems documentation*, pp. 24-26. 1986.

[2] Banerjee, Satanjeev, and Ted Pedersen. "An adapted Lesk algorithm for word sense disambiguation using WordNet." In *International conference on intelligent text processing and computational linguistics*, pp. 136-145. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

[3] J Garg, Mr Sandy, and Er Anand Kumar Mittal. "A Comparative Study of Svm and New Lesk Algorithm for Word Sense Disambiguation in Hindi Language.", International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 2, Issue 5, May 2015, PP 24-28.

[4] B. Surekha. Dr. K. Vijaya kumar and S. Siva skandha, “WORD SENSE DISAMBIGUATION USING LESK”, International Journal of Latest Trends in Engineering and Technology IJLTET Special Issue- ICRACSC-2016 , pp.063-066.

[5] Kågebäck, Mikael, and Hans Salomonsson. "Word sense disambiguation using a bidirectional lstm." *arXiv preprint arXiv:1606.03568* (2016).

[6] Ayetiran, Eniafe Festus, and Kehinde Agbele. "An optimized Lesk-based algorithm for word sense disambiguation." *Open Computer Science* 8, no. 1 (2016): 165-172.

[7] Gautam, Chandra Bhal Singh, and Dilip Kumar Sharma. "Hindi word sense disambiguation using Lesk approach on bigram and trigram words." In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, pp. 1-5. 2016.

[8] van den Beukel, Sven, and Lora Aroyo. "Homonym detection for humor recognition in short text." In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 286-291. 2018.

[9] Ackermann, Marcel R., and Florian Reitz. "Homonym detection in curated bibliographies: learning from dblp's experience." In Digital Libraries for Open Knowledge: 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10–13, 2018, Proceedings 22, pp. 59-65. Springer International Publishing, 2018.

[10] Roll, Uri, Ricardo A. Correia, and Oded Berger-Tal. "Using machine learning to disentangle homonyms in large text corpora." *Conservation Biology* 32, no. 3 (2018): 716-724.

[11] Kumar, Manish, Prasenjit Mukherjee, Manik Hendre, Manish Godse, and Baisakhi Chakraborty. "Adapted lesk algorithm based word sense disambiguation using the context information." *International Journal of Advanced Computer Science and Applications* 11, no. 3 (2020): 254-260.

[12] Bhattacharjee, Krishnanjan, S. ShivaKarthik, Swati Mehta, Ajai Kumar, Snehal Phatangare, Kirti Pawar, Sneha Ukarande, Disha Wankhede, and Devika Verma. "Survey and gap analysis of word sense disambiguation approaches on unstructured texts." In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 323-327. IEEE, 2020.

[13] Kharate, Namrata G., and Varsha H. Patil. "Word sense disambiguation for Marathi language using WordNet and the lesk approach." In *Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020*, pp. 45-54. Springer Singapore, 2021.

[14] Saha, Rohan. "CMPUT 600 Project Report: Homonym Identification using BERT." arXiv:2101.02398v1 [cs.CL] 7 Jan 2021

[15] Habibi, Amir Ahmad, Bradley Hauer, and Grzegorz Kondrak. "Homonymy and polysemy detection with multilingual information." In Proceedings of the 11th Global Wordnet Conference, pp. 26-35. 2021.

[16] AlMousa, Mohannad, Rachid Benlamri, and Richard Khoury. "A novel word sense disambiguation approach using WordNet knowledge graph." *Computer Speech & Language* 74 (2022): 101337.

[17] Rahman, Nazreena, and Bhogeswar Borah. "An unsupervised method for word sense disambiguation." *Journal of King Saud University-Computer and Information Sciences* 34, no. 9 (2022): 6643-6651.

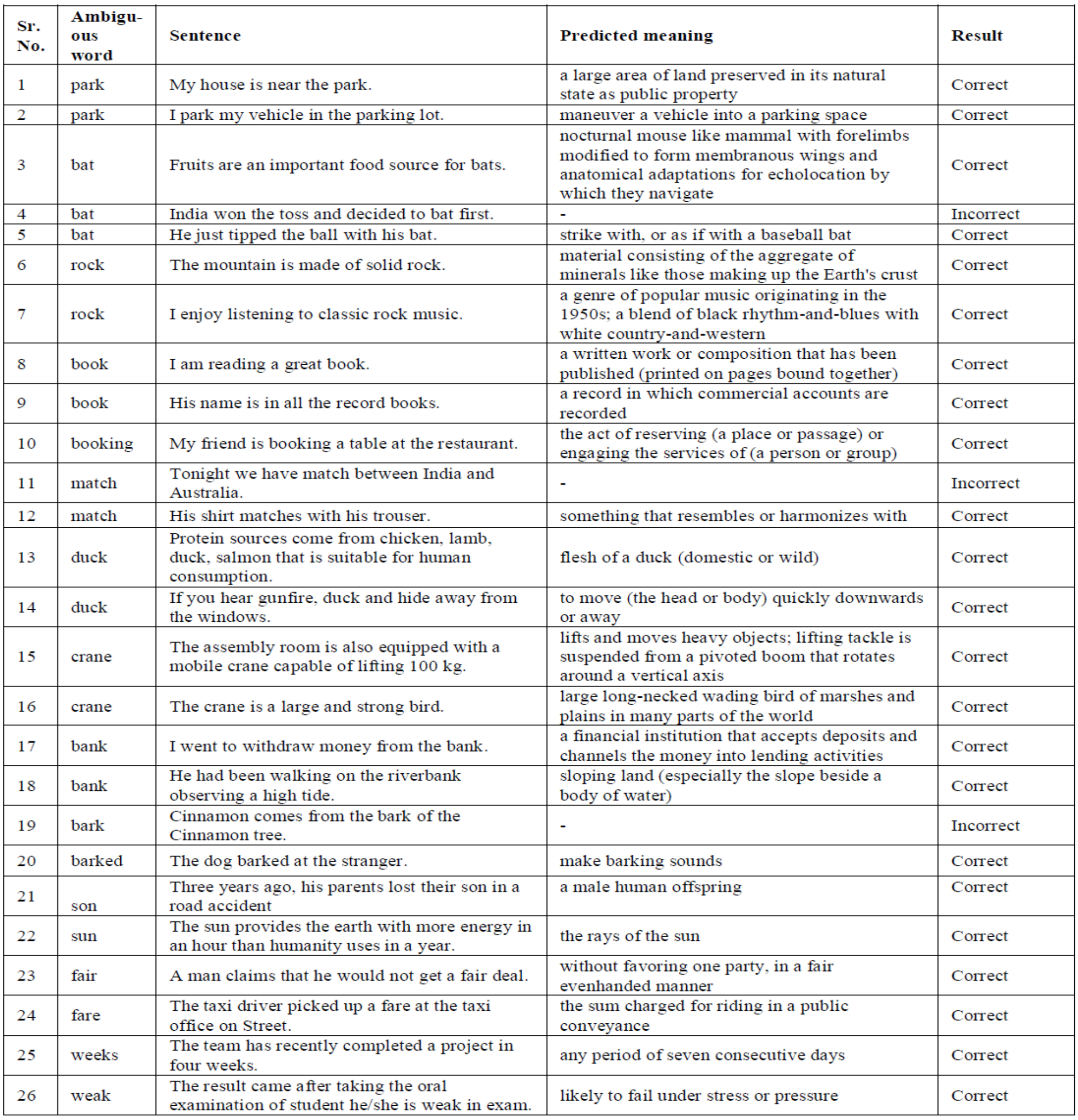
[18] Su, Ying, Hongming Zhang, Yangqiu Song, and Tong Zhang. "Multilingual word sense disambiguation with unified sense representation." *arXiv preprint arXiv:2210.07447* (2022).

[19] ILHOMOVNA, ELOV BOTIR BOLTAEVICH1&AKHMEDOVA HOLISKHON. "HOMONYMY DETECTION USING A NAÏVE BAYES CLASSIFIER.", Journal of Computer Science Engineering and Information Technology Research (JCSEITR) ISSN(P): 2250-2416; ISSN(E): Applied Vol. 13, Issue 1, Jun 2023, 5–20.

[20] Hall Maudslay, Rowan, and Simone Teufel. "Homonymy Information for English WordNet." *arXiv e-prints* (2022): arXiv-2212.

[21] Wordnet- https://wordnet.princeton.edu/

TABLE I. RESULTS OF IMPLEMENTATION OF PROPOSED ALGORITHM



Sr. No.	Ambiguous word	Sentence	Predicted meaning	Result
1	park	My house is near the park.	a large area of land preserved in its natural state as public property	Correct
2	park	I park my vehicle in the parking lot.	maneuver a vehicle into a parking space	Correct
3	bat	Fruits are an important food source for bats.	nocturnal mouse like mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate	Correct
4	bat	India won the toss and decided to bat first.	-	Incorrect
5	bat	He just tipped the ball with his bat.	strike with, or as if with a baseball bat	Correct
6	rock	The mountain is made of solid rock.	material consisting of the aggregate of minerals like those making up the Earth's crust	Correct
7	rock	I enjoy listening to classic rock music.	a genre of popular music originating in the 1950s; a blend of black rhythm-and-blues with white country-and-western	Correct
8	book	I am reading a great book.	a written work or composition that has been published (printed on pages bound together)	Correct
9	book	His name is in all the record books.	a record in which commercial accounts are recorded	Correct
10	booking	My friend is booking a table at the restaurant.	the act of reserving (a place or passage) or engaging the services of (a person or group)	Correct
11	match	Tonight we have match between India and Australia.	-	Incorrect
12	match	His shirt matches with his trouser.	something that resembles or harmonizes with	Correct
13	duck	Protein sources come from chicken, lamb, duck, salmon that is suitable for human consumption.	flesh of a duck (domestic or wild)	Correct
14	duck	If you hear gunfire, duck and hide away from the windows.	to move (the head or body) quickly downwards or away	Correct
15	crane	The assembly room is also equipped with a mobile crane capable of lifting 100 kg.	lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis	Correct
16	crane	The crane is a large and strong bird.	large long-necked wading bird of marshes and plains in many parts of the world	Correct
17	bank	I went to withdraw money from the bank.	a financial institution that accepts deposits and channels the money into lending activities	Correct
18	bank	He had been walking on the riverbank observing a high tide.	sloping land (especially the slope beside a body of water)	Correct
19	bark	Cinnamon comes from the bark of the Cinnamon tree.	-	Incorrect
20	barked	The dog barked at the stranger.	make barking sounds	Correct
21	son	Three years ago, his parents lost their son in a road accident	a male human offspring	Correct
22	sun	The sun provides the earth with more energy in an hour than humanity uses in a year.	the rays of the sun	Correct
23	fair	A man claims that he would not get a fair deal.	without favoring one party, in a fair evenhanded manner	Correct
24	fare	The taxi driver picked up a fare at the taxi office on Street.	the sum charged for riding in a public conveyance	Correct
25	weeks	The team has recently completed a project in four weeks.	any period of seven consecutive days	Correct
26	weak	The result came after taking the oral examination of student he/she is weak in exam.	likely to fail under stress or pressure	Correct