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

# Word Sense Disambiguation: A Supervised Semantic Similarity based Complex Network Approach

**Chandrakant Kokane[1], Sachin Babar[2], Parikshit Mahalle[3] and Shivprasad Patil[4]**

**Submitted:** 06/06/2022 **Accepted:** 10/09/2022

***Abstract*** **-**Lexical ambiguity in machine translation and information retrieval is the challenge. Lexical ambiguity is caused by polysemous words where the word has multiple meanings. In Natural Language Processing before processing human commands the disambiguation of ambiguous commands should be done. The existing disambiguation methodologies disambiguate ambiguous sentences with available context information. The main identified problem is what if an ambiguous sentence doesn't have enough information for disambiguation. The proposed model elaborates an adaptive sentence semantic similarity based complex network approach for identification of ambiguity and resolving it using semantic information. The discussed model represents the sentences of ambiguous documents as a vertex. The weighted complex network is constructed with respect to semantic similarities. The complex network is further processed for the ambiguous sentences having lack of context information. The main goal of this model is to provide an adaptive solution to lexical ambiguity of the paragraph or large document.

***Keywords:*** *Lexical ambiguity, Semantic Similarity, Complex network, Sense disambiguation*.

## 1. Introduction

Word Sense Disambiguation(WSD) is the critical problem in Natural Language Processing(NLP). The WSD is mainly introduced by the polysemous word also known as ambiguous word. The ambiguous word has the same spelling and pronunciation but different sense values. The ambiguous words are disambiguated with available context information. A lot of research is done with word embedding and machine learning based classifiers for WSD. The important challenge is disambiguating ambiguous words without context information in large documents. A complex network is constructed by the S-3 model. The sense annotated data is further used for disambiguation. While processing large documents it is very important to store the information of the current sentence because the meaning of 28% sentences are dependent on the immediate adjacent sentences [1]. So while disambiguating sentences in the large document the information of adjacent sentences is required. This feature is provided by the Recurrent Neural Network(RNN) as it consists of Large Small Term Memory(LSTM) [2]. The proposed methodology is representing the ambiguous document in graphical representation, where individual sentence's are represented with vertex and based on semantic similarity the weight value is calculated. The weight value represents the closest meaning sentence to the ambiguous sentence so that the closest semantic similar sentence is considered for the disambiguation. The proposed adaptive model to resolve WSD mainly considers the context information of ambiguous words to discover its correct sense. If the ambiguous word doesn't have context information, then neighbourhood words (vertex) of the ambiguous sentence is considered for the disambiguation. The supervised RNN classifier is implemented with intelligent training and testing. The manual sense tagged data is used for the training. The proposed model uses adaptive word embedding technique for representing text into word vectors. The RNN model with LSTM used as a classifier. The goal of the research is disambiguating the ambiguous commands with correct commands to provide better communication between man and machine. The proposed Sentence Semantic Similarity S-3 based complex network approach for the disambiguation is invented for the ambiguous words or sentences having lack of context information.

[1]*Savitribai Phule Pune University & SKN College of Engineering, Pune*
[2]*Savitribai Phule Pune University & Sinhgad Institute of Technology, Lonavala*
[3]*Vishwakarma Institute of Information Technology, Pune*
[4]*Savitribai Phule Pune University & NBN Sinhgad School of Engineering, Pune*

The article is classified into four sections as, section-2 elaborates the available literature for lexical ambiguity with word embeddings and machine learning based classifiers. Section-3 elaborates the proposed methodology with semantic similarity based complex network and supervised machine learning based classifier. Section-4 describes the comparative result analysis of WSD for single sentences and large documents. Section-5 elaborates the conclusion and future work in the field of WSD.

## 2. Related Work

The WSD problem is solved with two methodologies, knowledge based and Machine learning based. The knowledge based approach is also known as the traditional approach in which standard lexical resource WordNet is used for the disambiguation. It works in a linear manner for disambiguation. The second discussed approach is the machine learning approach.

The unsupervised models discussed by authors R. Navigli and M. Lapata are compared and contrasted with various parameters of edge connectivity that determine the comparative importance of vertices in the graph. Graph theory is full of such steps and tests have been performed in connection with studying the structure of networked environments and as part of the analysis of social networks. [3]

The experiment attempts to determine if some of these features are mainly relevant to graph-based WSD, and examine the role of the selected lexical resource and its impact on WSD. Such comparative research is novel; previous work is limited to only one lexical resource and a possible scale specially designed for WSD or accepted in network analysis [4]. Author's contributions are threefold: a graph-based WSD framework, a realistic comparison of a wide range of graph connection steps using standard test data sets, and WordNet'sense impact index and graph structure to WSD. Authors have proposed WSD for regional language with shallow parser [5]. A database labelled Part of Speed (PoS) has been created. PoS marking is done with a shallow parser. PoS taggers play a vital role in determining the exact meaning of ambiguous words in a sentence. A shallow parser identifies the meaning of a polysemous word and the same information is stored in the generated database. The POS of each value of the ambiguous word is stored. The problem here is multiple statements with more than one result value. Authors solved the problem of lexical ambiguity and semantic ambiguity using a graph-based centrality model [6]. An ambiguous sentence graph is built by giving a sequence of words with corresponding valid meanings. When disambiguating, the algorithm tries to identify ambiguous words in the graph vertices. If the selected vertex has no context information to resolve the ambiguity, the next closest vertex is considered.

The methodology discussed is divided into preprocessing, PoS marking and normalization, and ambiguous word detection and disambiguation. We use the Stanford parser for PoS tagging. The result is a set of semantic values that can be used to assign values from WordNet and BabelNet lexical resources. The WSD discusses a sentence with available contextual information by the author [7]. An ambiguous command is being processed. An ambiguous sentence word was found. The available context information of the semantic word is stored in a locally generated database. The WSD procedure initiates by determining the similarity between a set of ambiguous words for each meaning and a sentence containing the ambiguous word. The cosine method is used for word embeddings because it takes the closest possible value of the ambiguous word. Considering the same moment, the number of meanings of the polymorphic words decreases. The problem lies in the function. Algorithm ShotgunWSD 2.0 proposed by the authors [8] of ShotgunWSD does WSD at the document level with three categories:

The first is to generate a list of possible configurations for every window using the WSD brute force algorithm in the short context of the window selected in the document. The second is to combine the local discovery configuration into a longer one by combining the prefix and the appropriate suffix. In the third step, the meaning of each ambiguous word is selected based on a mass voting system that measures the resulting construct by length and considers only the higher constructs in which the correct meaning appears.

Existing methods discussed are for one-sentence WSDs. One hot encoding, Word2Vec, Glove, BERT models are discussed for embedding in Word. Given the shortcomings of one hot encoding because it creates a large vector for ambiguous words. Word2vec, where two words with the same meaning are used to identify synonyms. BERT generates multiple word vectors for the same word. All methods can be applied to a single sentence containing contextual information. The proposed adaptive word embedding is used for WSD of large documents.

## 3. Methodology

### I. Graph Construction

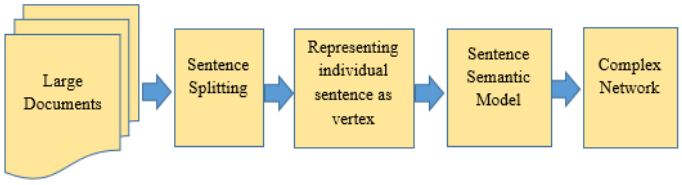


**Figure 1**- Functional Blocks of Methodology

Figure-1 shows the functional blocks of the proposed model. The ambiguous document is input to the model. The sentence splitting is performed on the input document and every sentence is separated. The dot separated comment is prepared from the input document. The local generated dot separated file is further taken for processing. Every sentence of the file is now represented as a vertex of Graph. Randomly the vertex with id is created and the same information is stored into the database with node and id. Once vertices are created the next task is to draw the edge. The edges are drawn with respect to the S-3 model. The semantic value of every vertex with every other vertex is calculated and based on semantic similarity the edges are drawn. Here some vertex may be isolated vertex because not a single sentence of the document is semantically similar with it. The output of this stage will be the complete annotated graph with weights and weights are nothing but the semantic similarities of sentences.

## II. Sentence (Vertex) Processing

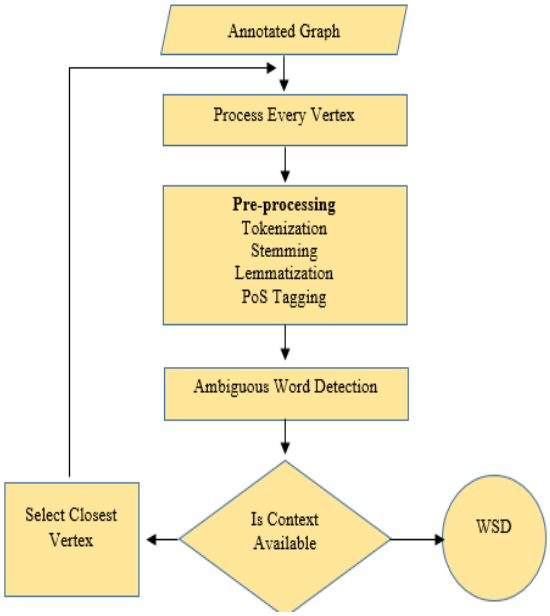


**Figure 2**- System Flow diagram

The system flow diagram shown in fig-2, describes the actual processing of the vertex. Here every vertex of the graph is considered for the processing. The sentence data associated with vertex is taken from the local generated database. The sentence is now pre-processed with tokenization stemming lemmatization and Part of Speech Tagging. The ambiguous/pronouns word from the sentence is identified. If the ambiguous word is having enough context information for disambiguation, then disambiguation will be done. If ambiguous words don't have enough information for disambiguation, then the next closest semantic similar node who is having less weight to the vertex is taken into consideration for disambiguation.

## III. System Architecture

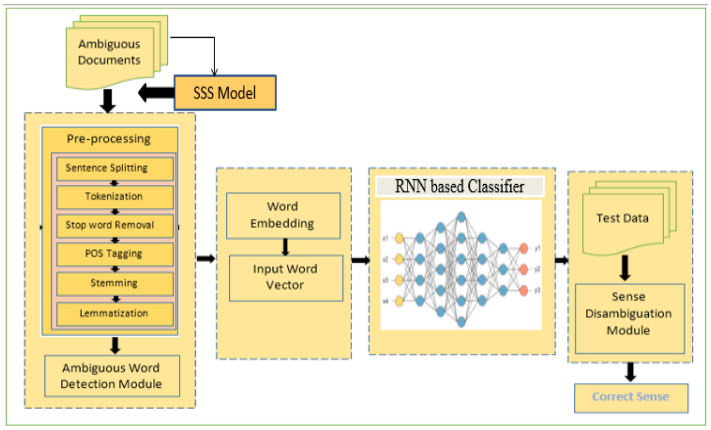


**Figure 3**- System architecture

Figure-3 shows the proposed WSD system architecture. If the ambiguous word has context information, select and configure the word vector. To represent an ambiguous word in word vector, the word vector with 897 features is constructed. The word vectors are provided as input to the RNN classifier. The output of the classifier will be the sense value. The sense values are further matched against freely available WordNet lexical resources to extract the exact meaning of ambiguous words.

## IV. Sentence Semantic Similarity(S-3) Model

The Proposed Semantic Similarity model calculates the semantic similarity of two sentences by taking the depths of the two-synsets in the freely available lexical resource WordNet.

SSS(V1, V2) = 2*depth_LCS / ( minimum {depth_lcs in depath_LCS}(depth_V1- depth_lcs)) + minimum{depth_lcs in depth_LCS}(depth_V2 - depth_lcs) ),

where depth_LCS(V1, V2) = arg_max{longext_comman_subsequence in LCS(V1, V2)}(depth_lcs).

SSS( cyclone#n#2 , hurricane#n#1 ) = 0.9565217391304348

“T1 = D_Trees( cyclone#n#2 ) =[1] *ROOT*#n#1 < entity#n#1 < physical_entity#n#1 < process#n#6 < phenomenon#n#1 < natural_phenomenon#n#1 < physical_phenomenon#n#1 < atmospheric_phenomenon#n#1 < storm#n#1 < windstorm#n#1 < cyclone#n#2”

“T2 = D_Trees( hurricane#n#1 ) =[1] *ROOT*#n#1 < entity#n#1 < physical_entity#n#1 < process#n#6 < phenomenon#n#1 < natural_phenomenon#n#1 < physical_phenomenon#n#1 < atmospheric_phenomenon#n#1 < storm#n#1 < windstorm#n#1 < cyclone#n#2 < hurricane#n#1”

Depth_LCS = depth( cyclone#n#2 ) = 11

Depth_1 = minimum(depth( {D_tree in T1 | tree contains LCS } )) = 11

Depth_2 = minimum(depth( {D_tree in T2 | tree contains LCS } )) = 12

Sementic_Score = 2 * Depth_LCS / ( Depth_1 + Depth_2 ) = 2 * 11 / (11 + 12) = 0.9565217391304348

## V. Complex Network

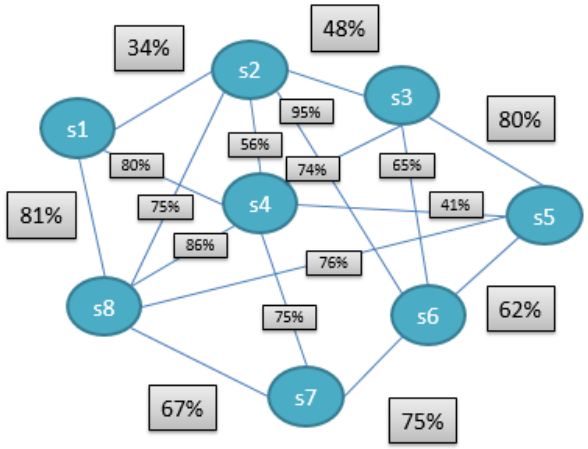


**Figure 4-** Complex Network Approach

Figure-4, shows the actual output of the S-3 model. For large documents or paragraphs, the complex network is constructed by considering individual sentences. The semantic similarity of every sentence with every other available sentence is calculated and based on the edge is drawn. The fully annotated graph is constructed with the weights. Weights describes the actual semantic distance of two sentences. If the vertex s1 does not have enough information for disambiguation so immediate vertex s-8 will be taken into the consideration for disambiguation sentence s1.

## 4. Results

The outcome of the proposed S-3 model is displayed in figure-5 as an output screen. The first output screen shows the semantic similarity for sentences or words. Output screen-2 shows the semantic similarity of all words of sentence-1 to all other words of sentence-2. Based on the semantic similarity the complex network is constructed. Figure-4 is the output with the ambiguous sentence that has been given as input, first ambiguous word detected in ambiguous sentence and finally available meanings of the ambiguous word with respect to context is shown

| | | |
|---|---|---|
| 1. | Input mode | ○ Word ◉ Sentence |
| 2. | Sentence 1 | Eventually, a huge cyclone hit the entrance of my house. |
| 3. | Sentence 2 | Finally, a massive hurricane attacked my home. |
| 4. | Submit | Calculate Semantic Similarity |

Output Screen:1 S-3 Model

The S-3 model initially starts with the sentence and finally goes for the paragraph or document.

| | Eventually /RB | , /, | a /DT | huge /JJ | cyclone /NN | hit /VBD | the /DT | entrance /NN | of /IN | my /PRP$ | house /NN | . /. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Finally/RB | - | - | - | - | - | - | - | - | - | - | - | - |
| ,/, | - | - | - | - | - | - | - | - | - | - | - | - |
| a/DT | - | - | - | - | - | - | - | - | - | - | - | - |
| massive/JJ | - | - | - | - | - | - | - | - | - | - | - | - |
| hurricane/NN | - | - | - | - | 0.9565 | - | - | 0.2857 | - | - | 0.3158 | - |
| attacked/VBD | - | - | - | - | - | 0.8571 | - | - | - | - | - | - |
| my/PRP$ | - | - | - | - | - | - | - | - | - | - | - | - |
| home/NN | - | - | - | - | 0.3529 | - | - | 0.6667 | - | - | 1.0000 | - |
| ./. | - | - | - | - | - | - | - | - | - | - | - | - |

Output Screen:2 S-3 Model

For the sentence-1 and sentence-2 shown in output screen-2, the semantic similarity values are calculated for all the words of sentence-1 and sentence-2.

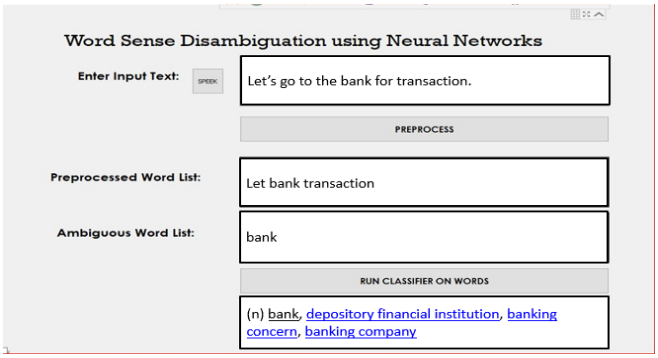


**Figure 5-** Result screen

Figure-5, shows the output screen where the sentence of paragraph or document is selected for the disambiguation. The command Lets go to the bank for translation is taken for processing because of ambiguity and ambiguity is introduced by the bank. According to WordNet bank is an ambiguous word having a sense value of 20. After pre-processing includes tokenization, stemming, lemmatization and PoS tagging ambiguous words are identified. Based on available context information of the polysemous word 'bank' the correct meaning as 'transnational unit' is extracted from the lexical resource WordNet.

## 5. Conclusion and future work

The proposed an adaptive sentence semantic similarity model to address the problem of lexical ambiguity in WSD. The result demonstrates that the disambiguated results for the paragraphs generated by the S-3 model can perform better for large documents with respect to the available context representations of polysemous words and semantic data of sentences. The proposed model calculates the dept_LCS and the semantic similarity of individual words of the sentence combinly helps to generate more accurate semantic keys. The S-3 model gives more accurate results for word documents with size 349 KB, but we still need enough data for the model to perform. Finally, further modifications to the model can provide the best performance, and we will explore

supervised learning using WSD for large documents in the future.

## References


[1]. B. S. Rintyarna and R. Sarno, "Adapted weighted graph for Word Sense Disambiguation," 2016 4th International Conference on Information and Communication Technology (ICoICT), 2016, pp. 1-5, DOI: 10.1109/ICoICT.2016.7571884.

[2]. Yadav, P. ., S. . Kumar, and D. K. J. . Saini. “A Novel Method of Butterfly Optimization Algorithm for Load Balancing in Cloud Computing”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 110-5, doi:10.17762/ijritcc.v10i8.5683.

[3]. C. D. Kokane, S. D. Babar and P. N. Mahalle, "Word Sense Disambiguation for Large Documents Using Neural Network Model," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5, DOI: 10.1109/ICCCNT51525.2021.9580101.

[4]. R. Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 4, pp. 678-692, April 2010, DOI: 10.1109/TPAMI.2009.36.

[5]. Ghazaly, N. M. . (2022). Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 01–04. https://doi.org/10.17762/ijfrcsce.v8i1.2063

[6]. Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labelling. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, USA, 411–418. https://DOI.org/10.3115/1220575.1220627

[7]. R. Rao and J. S. Kallimani, "Analysis of polysemy words in Kannada sentences based on parts of speech," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 500-504, DOI: 10.1109/ICACCI.2016.7732095.

[8]. N. A. Libre. (2021). A Discussion Platform for Enhancing Students Interaction in the Online Education. Journal of Online Engineering Education, 12(2), 07–12. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/49

[9]. F. Zait and N. Zarour, "Addressing Lexical and Semantic Ambiguity in Natural Language Requirements," 2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT), 2018, pp. 1-7, DOI: 10.1109/ISIICT.2018.8613726.

[10]. Rahman, Mohammad & Khan, Saeed & Hasan, K. M.. (2019). Word Sense Disambiguation by Context Detection. 1-6. 10.1109/EICT48899.2019.9068810.

[11]. A. M. Butnaru and R. T. Ionescu, "ShotgunWSD 2.0: An Improved Algorithm for Global Word Sense Disambiguation," in IEEE Access, vol. 7, pp. 120961-120975, 2019, DOI: 10.1109/ACCESS.2019.2938058.

[12]. Q. Nguyen, A. Vo, J. Shin and C. Ock, "Effect of Word Sense Disambiguation on Neural Machine Translation: A Case Study in Korean," in IEEE Access, vol. 6, pp. 38512-38523, 2018, DOI: 10.1109/ACCESS.2018.2851281.

[13]. Z. Li, F. Yang and Y. Luo, "Context Embedding Based on Bi-LSTM in Semi-Supervised Biomedical Word Sense Disambiguation," in IEEE Access, vol. 7, pp. 72928-72935, 2019, DOI: 10.1109/ACCESS.2019.2912584.

[14]. Kose, O., & Oktay, T. (2022). Hexarotor Yaw Flight Control with SPSA, PID Algorithm and Morphing. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 216–221. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/1879

[15]. H. Calvo, A. P. Rocha-Ramírez, M. A. Moreno-Armendáriz and C. A. Duchanoy, "Toward Universal Word Sense Disambiguation Using Deep Neural Networks," in IEEE Access, vol. 7, pp. 60264-60275, 2019, DOI: 10.1109/ACCESS.2019.2914921.