International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-1, October 2020

Global Word Sense Disambiguation of Polysemous Words in Telugu Language

Suneetha Eluri, Vasu Kumar Pilli

Abstract*: Word Sense Disambiguation (WSD) is a significant issue in Natural Language Processing (NLP). WSD refers to the capacity of recognizing the correct sense of a word in a given context. It can improve numerous NLP applications such as machine translation, text summarization, information retrieval, or sentiment analysis. This paper proposes an approach named ShotgunWSD. Shotgun WSD is an unsupervised and knowledge-based algorithm for global word sense disambiguation. The algorithm is motivated by the Shotgun sequencing technique. Shotgun WSD is proposed to disambiguate the word senses of Telugu document with three functional phases. The Shotgun WSD achieves the better performance than other approaches of WSD in the disambiguating sense of ambiguous words in Telugu documents. The dataset is used in the Indo-WordNet.

Keywords: shotgun sequencing, Word sense disambiguation, Word embedding, Telugu.

I. INTRODUCTION

Natural language [1] [2] is full of ambiguity; numerous words can have various meanings in various contexts. Word Sense Disambiguation is the capacity of recognizing which sense of an ambiguous word is being used in a given context. A sense is a definition or meaning of a word. For example, in the Telugu sentence,

“మట్టి పాత్ర లో నీరు చల్లగా ఉంటుంది”/ the water in the clay pot is cold.

The word “పాత్ర” has multiple senses including:

Sense 1: పదార్థాలు నిల్వ ఉంచు వస్తువు/Dish or basin

Sense2: వేషం/the actions and activities assigned to a person.

In this example, the context of ambiguous word is “మట్టి” and “లో నీరు చల్లగా ఉంటుంది”. Using the context, the WSD system must decide which sense of the word “పాత్ర“.

Word sense Disambiguation [3] is a key issue in Natural Language Processing. WSD refers to the task of recognizing the sense of word in given context. It can be possibly improving numerous NLP applications, for example, machine translation, text summarization, information retrieval, or sentiment analysis. There are two common ways to deal with WSD are supervised and unsupervised machine learning methods. These two approaches may also be combined to form a third approach, semi-supervised. Among these, supervised methods give the nest disambiguation results, but the main disadvantage is that it needs a large number of labeled examples or data for the supervised learning stage. Large annotated corpora are difficult to obtain many researchers have turned their focus on developing unsupervised learning or knowledge-based WSD methods. In this paper implement the WSD concept using Telugu data [4] [5]. Telugu is one of the South Indian languages belongs to the Dravidian languages family recognized by the government of India. Telugu is the most communicated language in India after the Hindi language. As per the statistics, the Telugu is 15th most spoken language throughout the world. The Telugu language is more complex with high morphological features compared to other languages and Word sense disambiguation by word co-occurrence improves the recall of the information retrieval system. The use of Synset while applying sense count will improve the robustness of the system and Telugu data is collected from Indo-Wordnet

II. SURVEY OF LITERATURE

In the literature, many authors and researchers used various techniques for the past many years for Word Sense Disambiguation. This section will discuss the related studies for Word Sense Disambiguation.

Brief History of Word Sense Disambiguation: WSD [6] is one of the most testing positions in the field of Natural Language Processing. Exploration work in this area was begun during the last part of the 1940s.

In 1949, Zipf proposed his "Law of Meaning" hypothesis. This hypothesis expresses that there exists a force law connection between the more successive words and less incessant words. The more successive words have a greater number of faculties than the less regular words. The relationship has been affirmed later for the British National Corpus. In 1950, Kaplan verified that in a specific setting two words on either side of an equivocal word are identical to the entire sentence of the unique situation.

In 1957, Masterman proposed his hypothesis of finding the genuine feeling of a word utilizing the headings of the classes present in Roget's International Thesaurus.

In 1975 Wilks built up a model on "inclination semantics", where the sectional limitations and an edge based lexical semantics were utilized to locate the specific feeling of an equivocal word. Rieger and Small in 1979 developed the possibility of individual "word specialists".

During the 1980s there was a striking advancement in the field of WSD research as Large-scale lexical assets and corpora opened up during this time. Subsequently,

Revised Manuscript Received on October 25, 2020.
* Correspondence Author
Dr. Suneetha Eluri*, Assistant Professor, Department of CSE, JNTUK - UCEK, Kakinada, India. Email: suneethaeluri83@gmail.com
Vasu Kumar Pilli, MTech (IT), Department of CSE, JNTUK - UCEK, Kakinada, India. Email: vasukumarpilli@gmail.com

scientists began utilizing distinctive programmed information extraction methodology (Wilks et al.1990) corresponding with the handcrafting approaches.

In 1986, Lesk proposed his calculation dependent on covers between the shines (Dictionary meanings) of the words in a sentence. The most extreme number of covers speaks to the ideal feeling of the equivocal word.

In this methodology, the Oxford Advanced Learner's Dictionary of Current English (OALD) was utilized to get the word reference definitions. This methodology had demonstrated the path to the next Dictionary-based WSD works. During the 1990s, three significant advancements happened in the examination fields of NLP: online word reference WordNet opened up, the measurable strategies were presented in this area, and Senseval started. The creation of WordNet (Miller 1990) got a transformation this exploration field since it was both automatically available and progressively composed into word faculties called synsets. In 1991, Guthrie et al. utilized the subject codes to disambiguate the specific sense utilizing the Longman Dictionary of Contemporary English (LDOCE).

Today, WordNet is utilized as a significant online sense stock in WSD research. Measurable and AI techniques are likewise effectively utilized in the feeling of characterization issues. Today, strategies that are prepared on physically sense-labeled corpora (i.e., managed learning techniques) have become the standard way to deal with WSD. Corpus-based Word Sense Disambiguation was first actualized by Brown et al. in 1991. As the informational indexes, corpora, online Dictionaries change language to language everywhere on the world, there was no benchmark of execution estimation in this area at an early age. Senseval acquired a wide range of examination works this area under a solitary umbrella. The principal Senseval was proposed in 1997 by Resnik and Yarowsky. Presently, subsequent to facilitating the three Senseval assessment works out, everywhere on the world specialists can share and overhaul their perspectives in this examination field.

Author: R. Mihalcea, P. Tarau, and E. Figa [7].

This paper, Authors have proposed a solo chart based technique for WSD. The diagram portrayal is utilized to demonstrate conditions among word faculties in-text known as a semantic chart. Six unique proportions of word semantic closeness known as the Leacock &Chodorow, the Lesk, the Wu-Palmer, the Resnik, the Lin, and Jiang &Conrath are utilized to decide the reliance between word faculties spoke to as hubs in the diagram. Next, four diagram based centrality calculations the in degree, closeness, between's, and PageRank are utilized to dole out scores to vertices of the chart. At last, the hub that has the most noteworthy worth is allocated as the sense for the word. They accomplished an exactness of 61.22, 45.18, and 54.79 and review of 60.45, 40.53, and 54.14 for things, action words, and descriptors separately.

Author: G.Tsatsaronis, IraklisVarlamis, and Kjetil Norvag [8]. This paper, the Authors tentatively examined the exhibition of unaided chart based techniques including the development of a semantic diagram. They have chosen four diagram handling strategies in particular SAN, PageRank, HITS, and P-Rank for assessment. To acquire a similar assessment, a similar semantic portrayal is utilized for all strategies. The presentation is assessed dependent on two standards on Senseval. They are the exactness and the between arrangement rate in the sense choice level.

Author: L. Vial, B. Lecouteux, and D. Schwab [9].

In this paper, Authors build up another method of making sense vectors for any word reference, by utilizing a current word embeddings model and adding the vectors of the terms inside a sense's definition, weighted in capacity of their grammatical feature and their recurrence. These vectors are then utilized for finding the nearest faculties to some other sense, in this way making a semantic organization of related ideas, consequently created. This organization is consequently assessed against the current semantic organization found in WordNet, by contrasting its commitment with an information based technique for Word Sense Disambiguation. This strategy can be applied to whatever other language which needs such a semantic organization, as the formation of word vectors is solo, and the making of sense vectors just needs a customary word reference. The outcomes show that our produced semantic organization improves incredibly the WSD framework, nearly as much as the physically made one. Author: O. Dongsuk, S. Kwon, K. Kim, and Y. Ko [10].

Word sense disambiguation (WSD) is the assignment to decide the feeling of a questionable word as per its unique situation. Many existing WSD considers have been utilizing an outside information based solo methodology since it has less word set imperatives than regulated methodologies requiring preparing information. In this paper, the Authors propose another WSD technique to produce the setting of an uncertain word by utilizing similitude between a vague word and words in the information archive. Likewise, to use our WSD strategy, we further propose another word closeness estimation technique dependent on the semantic organization structure of BabelNet. Assess the proposed strategies on the SemEval-2013 and SemEval-2015 for the English WSD dataset. Exploratory outcomes exhibit that the proposed WSD strategy fundamentally improves the standard WSD technique. Besides, our WSD framework beats the cutting edge WSD frameworks in the Semeval-13 dataset. At long last, it has a better than the best in class unaided information based WSD framework in the normal exhibition of both datasets.

Author: Suneetha Eluri and L. Sumalatha [11].

In this paper, authors developed Rule Based Approach for Finding Lexical Morphemes in Telugu language. We have generated corpus of 25K compounds words from Telugu daily news paper and developed Sandhi splitting process (rule base forward model) an d Sandhi formation process(Rule base backward model) for all Sandhi rules including Telugu and Sanskrit Sandhi and also specified the name of Sandhi. We verified the system with pre-tagged corpus of size 1200 words and achieved performance about 90% in terms of accuracy. This can be used as a learning tool for Sandhi splitting process for Indian languages which shares syntax and semantics with Telugu language.

Author: Sumalatha Lingamgunta, Suneetha Eluri [12].

In this paper, authors proposed a POS tagger for Telugu language, a South Indian language is proposed. In this model, the lexemes are tagged with various POS tags by using pre-tagged corpus however a word may be tagged with multiple tags. This ambiguity in tag assignment is resolved with Stochastic Machine Learning Technique i.e. Hidden Markov Model (HMM)

Bigram tagger which uses probabilistic information built based on contextual information or word tag sequences to resolve the ambiguity. In this system developed a pre-tagged corpus of size 11000 words with standard communal tag sets for Telugu language and the same is used for testing and training the model.
This model tested with input text data consists of different number of POS tags at word level and achieved the average performance accuracy of 91.27% in resolving the ambiguity.

Author: Suneetha Eluri, Sumalatha Lingamgunta [13].
In this paper, author presents a hybrid statistical system for Named Entity Recognition in Telugu language in which named entities are identified by both dictionary-based approach and statistical Hidden Markov Model (HMM). The proposed method uses Lexicon-lookup dictionary and contexts based on semantic features for predicting named entity tags. Further HMM is used to resolve the named entity ambiguities in predicted named entity tags. The present work reports an average accuracy of 86.3% for finding the named entities.

Author: Meryem Hdni [14] This paper, the authors have proposed a novel methodology for Arabic WSD which includes the utilization of two outside assets Arabic WordNet (AWN) and English WordNet. Subsequent to preprocessing the given content, words are planned into ideas in the event that they are in AWN. Something else, the term-to-term Machine Translation System from Arabic to English is utilized to have the comparable word in English. At that point WordNet is utilized to plan the word into the idea. Their thought in choosing the most suitable idea is that it sets up a more semantic relationship with various ideas in the nearby setting. At that point the idea chose is made an interpretation of back to Arabic utilizing Machine Translation System from English to Arabic whenever required. They have utilized Wu and Palmer's closeness measures, Chi-Square insights for include determination, and nearby and worldwide weighting ideas. At long last, their WSD framework has accomplished an exactness of 73.2%.

Author: Neeraja Koppula, Dr. B. Padamaja Rani [15].
Artificial Intelligence (AI) is the area where the entire world is revolving around. In AI, NLP (Natural Language Processing) is a challenging area, where to develop Word Sense Disambiguation (WSD) systems, to develop WSD systems we can adopt three approaches, knowledge-based approach, supervise and unsupervised approach. WSD is the ability to computationally determine which sense of the word is initiated by its utilization in a specific context. In this, the proposed work is to build WSD systems for regional Telugu language, which is suffering from scarcity of data sets. Till no data sets are available in the Telugu language only training data is used for testing the WSD system. In this work knowledge-based methodology is used to build a WSD system for the Telugu language. Depending on the context words we have to identify the meaning of the disambiguate words.

III. PROPOSED SYSTEM

Word Sense Disambiguation (WSD) is a significant issue in Natural Language Processing (NLP). WSD refers to the capacity of recognizing the correct sense of a word in a given context. It can improve numerous NLP applications such as machine translation, text summarization, information retrieval, or sentiment analysis. Generally, WSD algorithms can be categorized into 2 methods that work at the local and the global levels. A local WSD method is designed to assign the corresponding sense for a target word in a given context window of only few words. A global WSD algorithm aims to select the appropriate sense for each ambiguous word in an entire text in an entire text document. But the complexity of time increases in the document.

We have proposes a third approach named Shotgun WSD. It overcomes the problem of WSD algorithms of two levels of methods. Shotgun WSD is an unsupervised and knowledge-based algorithm for global word sense disambiguation. The algorithm is motivated by the Shotgun sequencing technique, which uses the genome sequencing method [16]. Shotgun WSD is proposed to disambiguate the word senses of Telugu document with three functional phases.

The Architecture has shown in the figure.1 Gives a brief description of the system. The system is implemented by performing Global Word Sense Disambiguation of Polysemous Words in the Telugu Language.
The architecture consists of three modules:

- Data Preprocessing
- ShotgunWSD Algorithm
- Result

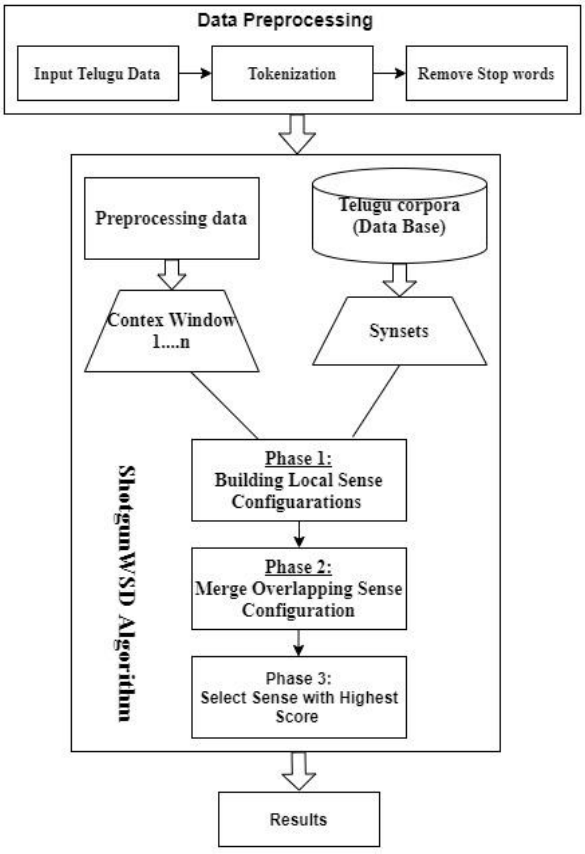

Figure.1: Architecture of the Proposed System

In the first module performs the Data Preprocessing, Telugu data i.e, nouns, verbs, adjectives, adverbs, and their words families similiarly senses of the words collected from the Indo-Wordnet and Develop a dataset that consists of ambiguous and unambiguous words and sentences.

From the user take an input sentence perform Data Preprocessing operations on the input data perform Tokenization and Remove the stop words. If the words having more than one sense then it is an ambiguous word i.e. target words.

In the second module apply the Shotgun WSD algorithm to disambiguate the word senses of Telugu document with three operational phases:

- **Phase1**: Building local sense configurations.
- **Phase2:** Consist of assembling shorter configuarations into longer configuarations by prefix-suffix matching.
- **Phase3:** Consist of ranking the configuarations obtaines by the relatedness score and choose highest score of configurations.

In the Third module obtain the Result and perform the performance analysis and visualization.

A.Shotgun sequencing technique: The ShotgunWSD algorithm is inspired by the Shotgun DNA sequencing technique. This technique takes a strand of DNA as input and outputs the most likely DNA sequence for that DNA strand.

First, copies of the input DNA strand are made, and many sample substrings of a fixed length are taken from these copies. Next each substring is sequenced. If any of these sequences are of low quality or difficult to read, they are typically removed before moving on. Ideally, this should not cause any gaps in the resulting DNA sequence because there are a large number of copies being sequenced. Once each small substring has been sequenced, the substring is pieced together by merging their overlaps the longer the better to produce the final DNA sequence. These matches are not perfect; the goal is to find the most likely sequence.

B. Word Embedding: Word embedding [17] is a technique used for mapping words to vectors of real numbers. Word embeddings can be generated using various methods like neural networks, co-occurrence matrix, probabilistic models, etc. Word Embedding is performed by using Word2Vec.

C. Word2Vec: word2vec [17] is an algorithm that takes an unlabeled source text as input to generate a word embedding for each word found in that source. Because the input data are not labeled, this is an unsupervised algorithm. Word2vec uses one of two models to generate embeddings: Continuous Bag-of-Words Model or Continuous Skip-gram Model. CBOW model predicts the current word given context words inside a particular window. The input layer contains the context words and the output layer contains the current word. The hidden layer contains the number of dimensions in which we want to represent the current word present at the output layer. Skip-gram predicts the surrounding context words within a specific window given the current word. The input layer contains the current word and the output layer contains the context words. The hidden layer contains the number of dimensions in which we want to represent the current word present at the input layer.

Cosine Similarity: Cosine similarity[177] measures the similarity between two vectors of an inner product space. It is estimated by the cosine of the angle between two vectors and decides if two vectors are pointing generally a similar way. It is regularly used to quantify report comparability in text examination.

Relatedness function: Relatedness scoring function for shotgun wsd, the semantic relatedness of two synsets is just given by the cosine comparability between their middle vectors:

$$\text{Relatedness}(A,B) = \frac{\sum_{i=1}^{m} a_i b_i}{\sqrt{\sum_{i=1}^{m} a^2{}_i}\ \sqrt{\sum_{i=1}^{m} b^2{}_i}}$$

Implementation of the Algorithm: The ShotgunWSD algorithm follows the concept as Shotgun DNA sequencing, but with a different goal. The goal of ShotgunWSD is to take a text document as input and to output a sense configuration for that document that matches the sense configuration a human would produce. Here, the text document to be disambiguated corresponds to the long DNA strand being sequenced and short context windows within this document correspond to the many short substrings taken from the DNA sequence. Consider the following example sentence: "మట్టి పాత్ర లో నీరు చల్లగా ఉంటుంది" / the water in the clay pot are cold.

Implementation of algorithm step by step process as follows:

Step 1:"మట్టి పాత్ర లో నీరు చల్లగా ఉంటుంది" / the water in the clay pot are cold. This sentence will serve as input to our entire document.

Step 2: The algorithm begins by selecting one window of up to n words at every possible location in the document, resulting in overlapping context windows covering the entire document. Selecting windows of up to 5 words at every ambiguous word from our example produces the following two context windows:

Context1:'[మట్టి పాత్ర లో నీరు ఉంటుంది]'/Clay contains water

Context2:'[పాత్ర లో నీరు ఉంటుంది]'/pot contains water

Step 3: Follow the brute-force approach is used to compute all possible sense configurations for each of these windows. Possible senses are chosen from the machine-readable dictionary Indo-Wordnet and replacing the senses in place of a target word in the both contexts.

"పాత్ర"is a Target word.

Senses of target word are:

Sense 1:'మట్టి తో తయారు చేసిన వస్తువు'/an object made of clay

Sense2:'కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'/the feeling of taking a person out of a story, novel, movie, etc.

Step 4: Replacing senses in place of a target word and following configurations are obtained:

Replacing senses in place of a target word in context1:

1.'మట్టి ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'] /An object made of clay (Target word sense)

2.'మట్టి ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది'/the feeling of taking a person out of a story, novel, movie, etc. (Target word sense)

Replacing senses in place of a target word in context2:

1. ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'] /an object made of clay

2. ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది /the feeling of taking a person out of a story, novel, movie, etc.

Step 5: Each of these possible sense configurations are assigned a score based on the semantic relatedness between the word senses within that configuration. This is computed using word embedding.

Semantic relatedness is calculated by using the Relatedness function by using Cosine Similarity.

Here, for simplicity or understanding purpose assign relatedness scores of 1 for related synsets and 0 for unrelated ones.

1 -> 'మట్టి ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'] / an object made of clay (1 for related synset)

0 -> 'మట్టి ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది'/ the feeling of taking a person out of a story, novel, movie, etc. (0 for unrelated synset)

1 -> ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'/ an object made of clay (1 for related synset)

0 -> ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది'/ the feeling of taking a person out of a story, novel, movie, etc. (0 for unrelated synset)

Step 6: Next merge the overlapping configurations. To do this, the algorithm checks if the suffix of one sense configuration matches the prefix of the next.

Below are possible sense configurations from two consecutive windows.

Suffix of one sense configuration matches the prefix of the next below shows the matching of two sentences of contexts:

1 -> 'మట్టి ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'

1-> ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'

Another Suffix of one sense configuration matches the prefix of the next below shows the matching of two sentences of contexts:

0 ->'మట్టి ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది'

0 -> ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది'

After merging (assembling):

2 → 'మట్టి ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది'

0 -> 'మట్టి ['కథ, నవల, సినిమాలో మొదలైన వాటిలో ఒక వ్యక్తిని తీసుకొని చేసేటటువంటి భావన'] లో నీరు ఉంటుంది'.

Step 7: Finally, select the sense with the highest similarity score.

2 → 'మట్టి ['మట్టి తో తయారు చేసిన వస్తువు'] లో నీరు ఉంటుంది' /an object made of clay

IV. RESULT AND DISCUSSION

ShotgunWSD was tested on the Telugu data. ShotgunWSD does not go through a training phase before testing. Its results are compared to the Most Common Sense (MCS) baseline. ShotgunWSD performs on the Telugu data producing its accuracy score of 80%.

The evaluation measures:

Accuracy: The accuracy score is calculated using actual and predicted values. The actual values and predicted values are represented by 1s and 0s. Here, '1' represents the related sense and '0' represents the unrelated senses.

Accuracy is the ratio between the number of correct predictions made and the overall number of predictions made.

$$\text{Accuracy} = \frac{\text{Number of Correct predictions}}{\text{Total number of predictions made}}$$

Table 1: Accuracy score of the proposed approach with Telugu dataset.

Algorithm	Accuracy
ShotgunWSD	80%

F1 score: The evaluation measure used is an F1 score, the weighted consonant mean of P and R, where P is precision, the quantity of true positives over the quantity of true positives and true negatives. R recalls, the quantity number of true positives over the quantity of true positives and false negatives.

$$F1 = \frac{2PR}{P+R}$$

Table 2: F1 score of the proposed approach with Telugu dataset.

	Precision	Recall	f1-score
0	0.75	1.00	0.86
1	1.00	0.50	0.67

Table 3: Comparison with other Algorithms

Algorithm	Accuracy Score(%)
Shotgun WSD	80
Genetic Algorithm	78.53
Extended Lesk	75

The figure.3 shows proposed system algorithm Shotgun WSD algorithm is compare with other algorithms. Compare ShotgunWSD with the two algorithms Genetic algorithm and extended algorithm. The accuracy score of ShotgunWSD is 80% compared with Genetic algorithm with accuracy score is 78.53% and then compares with Extended Lesk with accuracy score is 75%. The result shows better performance than Genetic and Extended Lesk algorithms Accuracy Scores.

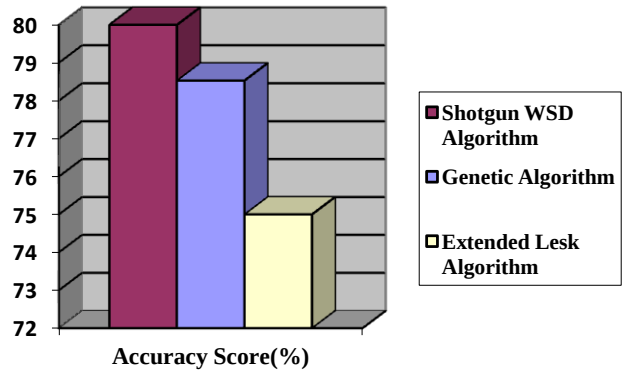

Figure.3: Comparison with other Algorithms

V. CONCLUSION

In this work, we have used ShotgunWSD approach on Telugu data and this approach is the combination of the unsupervised ML approach and knowledge-based approach. To disambiguate Telugu words, the proposed Methodology consist of estimating the semantic relation between the context of the utilization of the ambiguous word and its sense definitions to extract the senses of the ambiguous word from Telugu corpora. In this system shotgun sequencing technique, word embedding with word2vec and Cosine similarity methods are used to solve word sense disambiguation for the Telugu language and results are get the best rate of accuracy and precision. This whole work will be carried out on Telugu language and the Telugu data is collecting from the IndoWordNet and created own sense annotated corpora. Our approach is obtained an accuracy of 80%.

In the future work, we would like to apply our algorithm for other Regional Languages of India for which Wordnet is accessible.

REFERENCES

1. Anuja Bharate, Devendra Gadekar, "Survey Paper on Natural Language Processing". International Journal of Computer Engineering and Applications, Volume VIII, Issue III, Part I, December 14.
2. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa. Natural language processing (almost) from scratch. CoRR, abs/1103.0398, 2011.
3. Sense Disambiguation Techniques: A Survey Rekha Jain1, Sulochana Nathawat2, Dr. G. N. Purohit, Vol.1 Nov-Dec 2012.
4. Pushpak Bhattacharyya, IndoWordNet, Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.
5. Bhingardive, S., & Bhattacharyya, P. (2017). Word sense disambiguation using IndoWordNet. In The WordNet in Indian Languages (pp. 243-260). Springer, Singapore.
6. R.Navigli, Word Sense Disambiguation: a Survey, ACM Computing Surveys, Vol. 41, No.2, ACM Press, pp. 1-69 2009.
7. R. Mihalcea, P. Tarau, and E. Figa. PageRank on semantic networks with application to word sense disambiguation. In Proc. of COLING, 2004.
8. G.Tsatsaronis, IraklisVarlamis, and Kjetil Norvag, An experimental study on unsupervised graph-based word sense disambiguation, In Proc. of CICLING, 2010.
9. L. Vial, B. Lecouteux, and D. Schwab, ``Sense embeddings in knowledge-based word sense disambiguation," in Proc. IWCS, 2017
10. 10. O. Dongsuk, S. Kwon, K. Kim, and Y. Ko, ``Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph," in Proc. COLING, Aug. 2018, pp. 2704_2714.
11. Suneetha Eluri, Sumalatha Lingamgunta "Rule Based Approach for finding Lexical Morphemes in Telugu, an Indian Language" Published, Journal of Advanced Research Dynamic Control Systems, Volume 10, Issue 12, Page No: 419-420, August 2018. ISSN 1943-023X [Elsevier Scopus Indexed(Free) Impact Factor 0.11]
12. Suneetha Eluri, Sumalatha Lingamgunta "ARPIT: Ambiguity Resolver for POS Tagging of Telugu, an Indian Language" published in i-manager Journal on Computer Science, Volume 7, Issue 1, Page No: 25-35, ISSN Print: 2347-2227, March-May 2019 [Double Blind Peer Reviewed Free Journal with Impact Factor 0.750].
13. Suneetha Eluri, Sumalatha Lingamgunta "A Statistical Method for Named Entity Recognition in Telugu, an Indian Language" published in International Journal of Recent Technology and Engineering (IJRTE): ISSN: 2277-3878, Volume -8 Issue-2, Page No:4211-4216, July 2019. [Free journal with Scopus Indexing from 2018].
14. Meryeme Hdni et al, Word Sense Disambiguation for Arabic Text Categorization, IAJIT, Vol.13, 2016.
15. Neeraja Koppula, Dr. B. Padamaja Rani, Word Sense Disambiguation Using Knowledge based Approach in Regional Language, Vol. 10, 2018. Language Processing, NCRSTCST, Vol.4,2013.
16. A. Butnaru, R. T. Ionescu, and F. Hristea, ''ShotgunWSD: An unsupervised algorithm for global word sense disambiguation inspired by DNA sequencing,'' in Proc. EACL, Apr. 2017, pp. 916–926.
17. Orkphol, K.; Yang, W. Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet. Future Internet 2019, 11, 114.

AUTHORS PROFILE

Dr. Suneetha Eluri is working as Assistant Professor in the Department of Computer Science and Engineering at Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India. Her research interests are Natural Language Processing of regional languages with AI, Machine Learning and Deep Learning techniques. She is a Faculty champion of University Innovation Fellows programme at Stanford University. Currently she is working on NLP tasks and sentiment analysis of Telugu language. She has 16 years of academic experience. She has published a number of research papers in various reputed National and International Journals and Conferences. She has guided around 35 Post-graduates and 25 graduates of Computer Science and Engineering.

Vasu Kumar Pilli is a second year Post Graduate student at Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India, pursuing his M. Tech in Information Technology, Department of Computer Science and Engineering. He is currently working on his project on Natural Language Processing in Telugu Language. This is his first paper on Natural Language Processing Machine and its algorithms.