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

A Knowledge Based Word Sense Disambiguation in Telugu Language

Suneetha Eluri, Vishala Siddu

Abstract: ***Telugu ((తెలుగు) is one of the Dravidian languages which are morphologically rich. As within the other languages, it too consists of ambiguous words/phrases which have one-of-a-kind meanings in special contexts. Such words are referred as polysemous words i.e. words having a couple of experiences. A Knowledge based approach is proposed for disambiguating Telugu polysemous phrases using the computational linguistics tool, IndoWordNet. The task of WSD (Word sense disambiguation) requires finding out the similarity among the target phrase and the nearby phrase. In this approach, the similarity is calculated either by means of locating out the range of similar phrases (intersection) between the glosses (definition) of the target and nearby words or by way of finding out the exact occurrence of the nearby phrase's sense in the hierarchy (hypernyms/hyponyms) of the target phrase's senses. The above parameters are changed by using the intersection use of not simplest the glosses but also by using which include the related words. Additionally, it is a third parameter 'distance' which measures the distance among the target and nearby phrases. The proposed method makes use of greater parameters for calculating similarity. It scores the senses based on the general impact of parameters i.e. intersection, hierarchy and distance, after which chooses the sense with the best score. The correct meaning of Telugu polysemous phrase could be identified with this technique.***

Keywords: Natural Language Processing (NLP), Polysemous, IndoWordNet, Word Sense Disambiguation (WSD), Intersection, Hierarchy, Senses, Distance measure.

I. INTRODUCTION

Natural Language Processing (NLP) is directly related to human-computer interaction in the field of computer science, semantics and artificial intelligence in which computer has to understand and generate natural language. Building an NLP system is difficult because the human's natural language is not always characteristic and structured. It is carried in different processing steps namely sentence boundary detection, word segmentation, Parts Of Speech (POS) tagging [1], Named Entity Recognition [2], word sense disambiguation, etc.

In Natural Language Processing (NLP), many verbal languages will have many polysemous phrases. These polysemous phrases could have extraordinary meanings in exceptional contexts. The method of identifying the correct meaning of an ambiguous phrase is Word Sense disambiguation (WSD). Human beings can apprehend verbal communication of a language relying at the context where the polysemous word is used, however for machines that is a difficult hassle, as it involves an association of data structure. By thoroughly reading this data structure we can locate the suitable meaning of the polysemous word.

The precise meaning of the ambiguous word only depends on the surrounding phrases of the ambiguous phrase. WSD mainly relies on knowledge. It can be interpreted to work as follows: it takes input as a sentence, then NLP techniques [3] are applied which utilize at least one source of knowledge to identify the exact senses of the phrases regarding the context. For example, consider following two sentences:

1. సమాజం లో మార్పు రావాలంటే నువ్వు ముందుకు సాగు.
2. పత్తి సాగు లాభదాయకం.

The above two sentences carry a unique meaning and context for the word ‘సాగు’.

The primary sentence points at the sense that is associated with ‘PROCEED IN ADVANCE’ and the second sentence points at the sense that is related to ‘CULTIVATION’. These types of problems are word Sense Disambiguation (WSD). Like this when a word is mapped to more than one sense then this type of ambiguity arises. WSD is to define, analyze and identify the relationship between phrase, meaning and context. The answer for this WSD problem in its way enables to resolve a few other tough issues like common sense reasoning, encyclopedic Knowledge and finally natural language understanding. This WSD problem plays an essential role in machine translation, information retrieval, content evaluation, grammatical evaluation, thematic analysis, speech processing and text processing.

II. RELATED WORK

Knowledge based implementation of WSD (Word Sense disambiguation) using the concept of the intersection or overlapping phrases was primarily conceived by [4]. The Lesk algorithm chooses a sense whose gloss has the best variety of overlapping phrases with the surrounding text. Lesk algorithm claims 50-70% accuracy on short samples of textual content.

The similarity used in this paper to calculate the similarity between two phrases is a fairly improved version of the only used in adapted Lesk [5]. Rather than computing intersection with phrase families,

Revised Manuscript Received on October 26, 2020.

Suneetha Eluri*, Professor, CSE Department, JNTUK, Andhra Pradesh, India. Email: suneethaeluri83@gmail.com

Vishala Siddu, Academics, CSE department, JNTUK, Andhra Pradesh, India. Email: sidduvishala@gmail.com.

Adapted Lesk calculates intersection with the glosses of the sense and also the glosses of its associated words. Also adapted Lesk does not think about the stages of hierarchy i.e. levels of hypernyms of the phrase.

Therefore, we've stepped forward the method of computing overlapping words. The improvement over adapted Lesk is that in preference to just finding out the number of overlapping phrases in glosses, we've used the idea of the word family. Respective phrase families of target and near-by phrases include relevant phrases from the gloss and associated words like Synonyms, Hypernyms, Hyponyms, Homonyms, and many others. Finally, the algorithm supplied in this paper calculates the intersection of phrase families at each hierarchical level. At every level word family of a new hypernym is brought.

Adapted Lesk calculates a score for every candidate aggregate couple i.e. a couple of the target phrase and a nearby phrase. For each pair, a score is calculated. The candidate combination with the best score is the outcome, and the target phrase is assigned the sense given in that combination. The algorithm presented in this paper sums up the score of a particular sense for every nearby phrase, and the sense with the highest score is selected as the favored sense. Additionally in [5], the scores calculated are the squares of the no. of overlapping phrases. The accuracy claimed for the set of rules in [5] is 32%. But, it might be irrelevant to reach at a sense only on the basis of the parameter of overlapping phrases because the closeness of words can also be dependent on the level of hierarchy where an overlap has been encountered. Therefore the score in this proposed algorithm is calculated on the premise of intersection along the entire hierarchy of the target and close-by phrase's senses and the gap among the target and close-by words.

Another method to Knowledge based word sense disambiguation [6] plots the senses of nearby phrases within the hierarchy of senses of the difficult phrase. The hierarchy that is observed to be the densest is chosen as the suitable sense. despite the fact that [6] takes into account the hierarchical connection between two words, it does no longer draw from the Lesk algorithm and its variations for the concept of intersection among bags of words. For this dataset, the claimed precision is 47.6%. The language version presented in [7] a joint opportunity model for statistical system translation, which automatically learns phrase and phrase equivalents from bilingual corpora. A Tri-gram primarily based version proposed in [8] has given splendid effects for POS tagging in the Marathi language. In [11] the proposed work is to develop a WSD system using a machine learning technique and knowledge based approach for the Telugu language. The knowledge resource which is used to develop the WSD system is Lexical Knowledge Base (LKB). The performance of the WSD system is right while as compared with different unsupervised procedures.

The statistical method proposed [12] assigns an accurate sense to the focused polysemous word which has specific meanings in different contexts. The technique proposed in this paper which solves the well-known AI-complete problem is related to Natural language processing (NLP) that is called Word Sense Disambiguation (WSD). The polysemous word may also belong to any one parts-of-speech assigned via the POS taggers. However, there are a few phrases that belong to the same components of speech but their meaning differs primarily based on the context. Presently, the system disambiguates nouns and verbs. The system gives 100% coverage.

In [16] Lesk algorithm has been implemented for word sense disambiguation (WSD) of Punjabi text has been tested on approximately 500 Punjabi sentences and 100 Punjabi text corpora. However, the Authors had implemented Punjabi WordNet for the creation of bilingual dictionaries. In [16] Punjabi WordNet has been implemented for WSD. However, accuracy is less compared to [17] and [18].

III. PROPOSED METHODOLOGY

The goal of Knowledge based WSD is to find the Correct sense of a given ambiguous word that has the highest Score using the Lesk Algorithm. WSD makes use of dictionaries to achieve better identification of word senses. Knowledge resources can also vary depending on corpora of textual content, labeled or unlabeled with phrase senses or more arranged resources like machine-readable dictionaries, semantic systems, and so on. Knowledge of the words is the critical ingredient to differentiate the meaning. In this paper, we utilize the Lesk approach [4], which involves appearing for overlap among the words in given definitions with words from the text surrounding the word to be disambiguated. Our system consists of the following phases:

Step 1: Input Sentence is given
Step 2: Preprocess the sentence i.e., Tokenize and remove the stopwords from the given Input Sentence.
Step 3: Find the target polysemous word.
Step 4: Extract the Senses of target and near-by phrases from IndoWordNet.
Step 6: Apply the Lesk Algorithm
Step 7: Generate the Sense with the Highest Score.

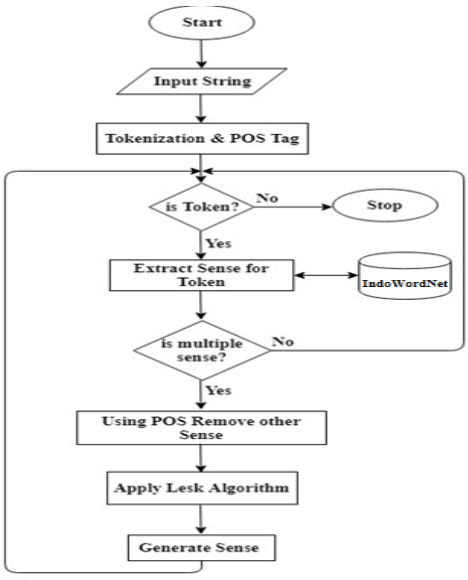

Fig 1. Flowchart diagram of WSD using Lesk algorithm

Retrieval Number: 100.1/ijeat.A19110109119
DOI:10.35940/ijeat.A1911.1010120
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A. IndoWordNet

Indo-WordNet [9] is a lexical database for the Indian language which combines phrases into sets of synonyms referred to as synsets. The idea of Indo-WordNet is stimulated with the English WordNet. English WordNet [10] is prepared by means of semantic relations. Since semantic relation is a relation among meanings, and meanings may be represented as synsets. Indo-WordNet consists of simplest open class phrases like nouns, verbs, adjectives and adverbs. It does contain close class words together with pronouns, prepositions, conjunctions. Indo-WordNet is prepared semantically as parts of speech. The main object in Indo-WordNet is synset, a set of synonyms. The Indo-WordNet is being used in this project. This was produced by researchers in the center of Indian Language Technology (CFILT), IIT, and Bombay. The famous English Wordnet [36] inspires its design.

- Synset: It is a collection of words with similar meanings or synonyms. For Example "□□□□, □□□□□, □□□□□□□□□□" describes a thing for playing with a ball(□□□□) The words are arranged in the synset according to the popularity of usage.
- Gloss: Gloss gives the definition of a word of that particular sense. For Example,
  "□□□□□□□□□ □□□□ □□□□□□□ □□□□□□□ ఒక□□□□□□□"
  (A spherical object used as a plaything)
  Is the definition meaning of the word "□□□□" (ball)
- Semantic Relation: It affirms the connection between the form and the meaning of a word. IndoWordNet Contains the Following Semantic relations: Hypernymy, Hyponymy, Holonymy, Meronymy, Antonymy, Troponymy, etc.
  For Example, Consider below Figure 2:

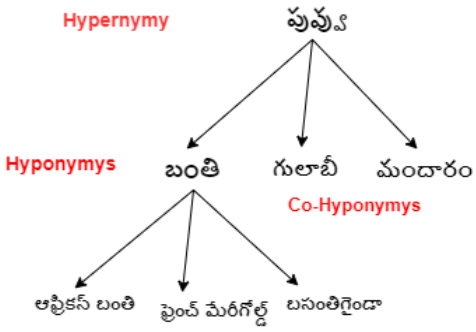

Fig 2: Semantic Relations

B. Lesk Algorithm

Lesk algorithm [2] is a traditional knowledge-based WSD Algorithm which removes the ambiguity of a word by choosing a sense whose definition overlaps the most with the words in its context? A knowledge-based approach utilizes outer lexical assets like IndoWordNet. The similarity system presented to calculate the similarity between phrases is a highly improved model of the only utilized in Adapted Lesk [5]. Instead of computing intersection with word families, Adapted Lesk computes intersection with the glosses of the sense and also the glosses of its associated words. Therefore, we have stepped forward the method of calculating overlapping phrases. The development over adapted Lesk is that in place of just finding out the range of overlapping words in glosses, we have used the concept of the phrase family. Respective phrase families of target and close-by words consist of applicable phrases from the gloss and associated phrases like Synonyms, Hypernyms, Hyponyms, Holonyms, Meronyms, Attributes, Topic members, Verb group, and many others. Subsequently, the algorithm provided in this paper calculates the intersection of phrase families at every hierarchical level. The idea of hierarchical level method that once the intersection is calculated among the phrase families of target and nearby words as shown in Fig 3, we keep on adding the hypernyms of each sense of target and near words as shown in Fig 4. At every level phrase family of a new hypernym is delivered.

C. Similarity Parameters

The Important task of word sense disambiguation is to compute the similarity/Closeness of two phrases. The following parameters are used for measuring the similarity.

- Intersection - For each sense, IndoWordNet offers a set of words that are similarly related to that sense. This set can encompass {Synonyms, Hypernyms, Hyponyms, Holonyms, Meronyms, Attributes, Verb Group and so forth} is referred to as the word family of that sense. Every word family additionally consists of words from the gloss, which have only one sense according to IndoWordNet. Therefore intersection is computed among the phrase families of target and nearby phrases.
- Hierarchical Relationship (Hypernyms and Hyponyms) - Hypernyms and Hyponyms refers super-subordinate relationship among two senses. It hyperlinks more general senses like {ఆట(ball) to increasingly specific ones like {□□□□}(hockey) and {□□□□}(kabaddi), such that {ఆట is a hypernym (parent) of both {□□□□} and {□□□□}, and that {□□□□} and {□□□□} are hyponyms (child) of {ఆట A hypernym is taken into consideration as a extra general phrase than its hyponym.
- Distance - This parameter refers to the distance among the target and the near-by phrase in the given textual content. It is primarily based on the fact that the farther the nearby phrase is from the target word in the input textual content; the weaker might be the relation between them. But, we've changed the method of measuring distance. Only include content material phrase i.e. phrases whose parts of speech are noun, verb, adjective, or adverb. Therefore, the inappropriate phrases like '□□' (in), '□□□□' (but), '□□' (with), and many others words. Do not unnecessarily impact the value of the parameter distance.

D. Formula for Calculating Score

$$\text{Score} = \frac{(\text{Intersection})^{1/K1}}{(\text{Level})^{K2} * (\text{Distance})^{1/K3}}$$

The values of K1, K2 and K3 had been experimentally decided by putting specific combos of K1, K2 and K3 values. We obtain a quality result for K1=K2=K3=3.

IV. ILLUSTRATION OF ALGORITHM

Let us illustrate the algorithm with the following Instance Sentence:

‘ □□□□ ఆటమన□□□□ □□□□□ ‘

(‘Hockey is our national sport’)

Assuming we need to disambiguate the phrase ‘ఆట we take the following steps:

Step 1: Filter the sentence by Tokenization and removing the stop words from the given sentence. After the preprocessing the sentence it looks like in the following manner

□□□□ ఆట □□□□□ □□□□□

noun noun adjective noun

Step 2: consistent with IndoWordNet the phrase ‘ఆట’ has the subsequent senses:

1. □□□□ □□□□□□□□ □□□□□ □□□□ □□□□□□□□□□ □□□□□□□□□□

2. □□□□□ □□□□□□ □□□□□□□□□□, □□□□□□□□□ □□□□□□□□□ □□□ □□□□ □□□.

3. □□□□□□ □□□□□□ □□□□□□□□□□ ఆట

4. □□□□□□ □□□□□□□□

5. □□□□□□ □□□□□□□□□□□□ □□□□□□□□□ □□□□□□ □□□□□□□□□ □□□□□□□□□□ □□□□□□ - □□□□□□ □□□□□□□□□□□□□□□□□□

Therefore the task is to find one sense out of these five senses as the sense of the phrase 'ఆట' in the given sentence.

Step 3: Now 'ఆటbeing the target phrase, we make pairs of nearby phrases i.e. □□□□, □□□□□ (National) and □□□□□ with 'ఆటin the subsequent manner:

{'ఆట□□□□'} {'ఆట□□□□□'} {'ఆట□□□□□'}

Step 4: let us take the primary pair: ఆటand □□□□. In this step, we merge all of the phrase families of senses of close-by phrase i.e. □□□□ in a single set. The phrase families of senses of target phrase i.e. ఆటare stored separately in five distinct sets.

We compute the intersection of every phrase family of the target word with the merged word family of the nearby word.

In Fig 3, the left side boxes represents the word families of senses S1, S2, S3, S4, and S5 of the target word 'ఆటThe right side box represents the combined set of senses of the nearby phrase ‘□□□□ ‘.

A phrase family additionally contains a variety of phrases; therefore, we’ve not mentioned all the phrases and have written just a few for representation purpose. The Intersection of each word family of the target phrase is computed with the whole set of word families of the close-by phrase. Fig 3 suggests the intersection with the double-ended arrows. **Step 5:** If a few intersections is found in Step 4, then we compute the score for each target experience. The intersection in Step 4 is at level l due to the fact the target phrase and nearby words families are being taken into consideration.

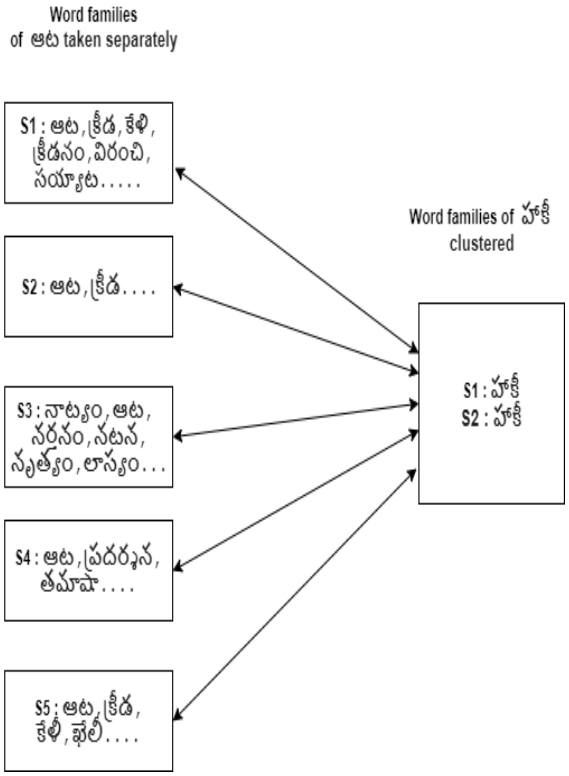

Fig 3: Intersection at Level 1

Step 6: At this step, we add the families of hypernyms or parent of the target phrase and near-by phrase's senses and we name it level 2.

In level 2 we include hypernyms of every word families of target and close by words. Consequently, Fig 3 is changed to Fig 4 after adding the parents or hypernyms of the phrase families. Parent of sense S1 is denoted by PS1; that of S2 is denoted via PS2 and so on. Hypernyms of word families of the near-by word are again merged in a single set much like within the previous step. PS1 can be a lever high than S1, so an arrow points from S1 to PS1 denoting that S1 is the kid and PS1 is the parent (hypernym). This is finished for all the phrase families as shown in the Fig 4 below. Intersection is computed at this stage as depicted by using the double-ended arrows in Fig 4.

Retrieval Number: 100.1/ijeat.A19110109119
DOI:10.35940/ijeat.A1911.1010120
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

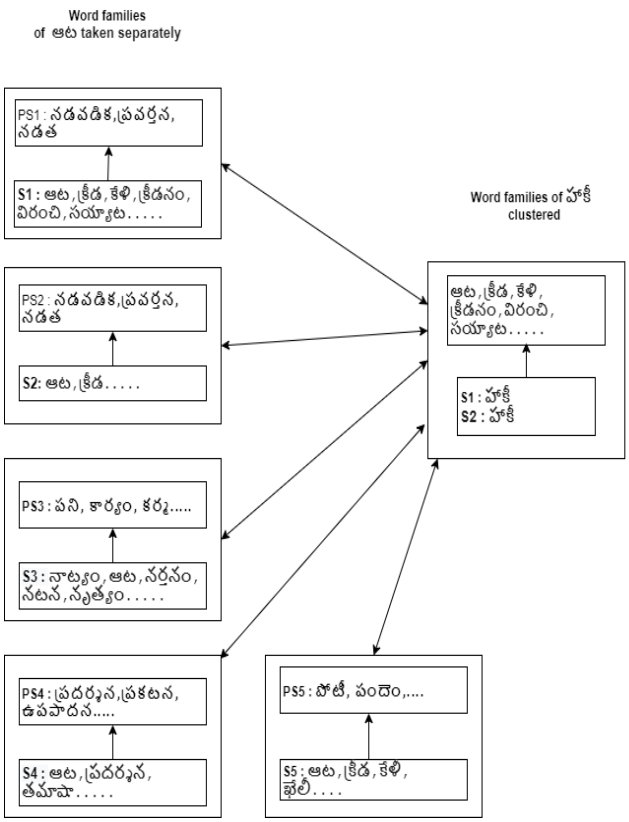

Fig 4: Intersection at Level 2

Step 7: The score is once more computed for this level or stage using the formula. These new scores are brought to those obtained on the preceding level for each sense. Likewise, we maintain incrementing the level by adding the successive units of phrase families of hypernyms of that respective sense. This process is continued until a hypernym exists for all the phrase families.

Step 8: According to the intersection computed at each stage, every sense may have a score correspond to a selected level. (If there is no intersection found then the score is 0)

Step 9: visit the next pair ('ఆట□□□□□') and repeat steps 4 to 8 for this pair and the subsequent pairs.

Step 10: When this procedure has been completed for all the pairs, the sense with the best score is chosen as the correct sense of the ambiguous words or target word.

Hence the overall impact of all the near phrases on the target phrase is computed using three parameters: intersection, level and distance.

V. RESULTS

To evaluate our system Telugu sentences with ambiguous words are collected manually from newspapers and some storybooks. From this, the accuracy of the system was calculated by dividing the correctly identified words by the total ambiguous words as shown in below Table 1.

We have considered 150 Telugu polysemous words. These belong to different part-of-speech like nouns and verbs Even though a word belongs to one part-of-speech it may have different senses (meanings) which are the WSD task.

$$Accuracy = \frac{Correctly\ Disambiguated\ Words}{Total\ Ambiguous\ Words} \quad (1)$$

$$Error = \frac{Incorrectly\ Disambiguated\ Words}{Total\ Ambiguous\ Words} \quad (2)$$

Table 1: Accuracy result for the proposed approach

Total Number of Ambiguous Words	150
Correctly Disambiguated Words	98
Incorrectly Disambiguated Words	52
Accuracy	65.4
Error	34.7

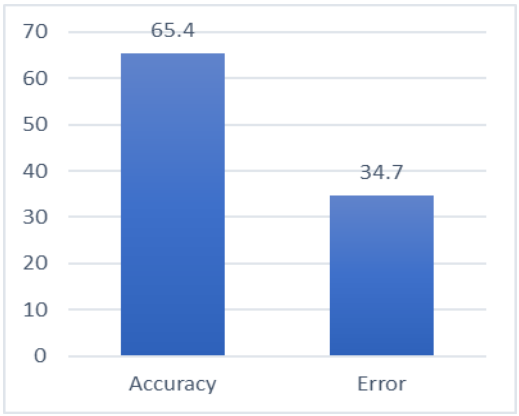

Fig 5: Accuracy in error measure

There are number of ambiguous words in Telugu but this algorithm is tested for 150 phrases which include nouns and verbs. For every phrase minimum of 2 senses and a maximum of 5 senses are considered. The metrics for the developed system are discussed below (Fig. 6).

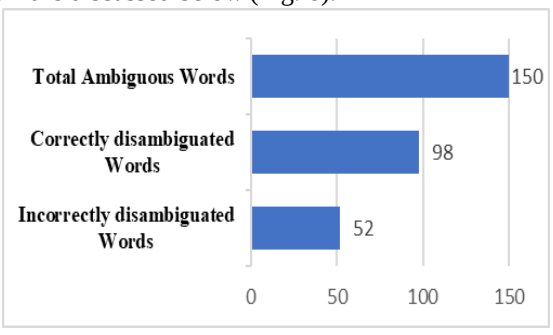

Fig 6: Statistics of disambiguated words

Figure 7 shows the Accuracy percentages of the proposed approach along with the Accuracy percentages of [14], [15], [16] and [17].

The approach suggested by [14] gets a higher accuracy than the algorithm introduced in this paper. However [14], Uses only nouns as target phrases. Although the proposed algorithm gets a lower accuracy value as compared to [14], it is tested on all content words unlike [14] and does not need any corpus in computing the correct sense of the objective word. Therefore the algorithm in this paper beats the results of the existing approaches when it is tested only for nouns and it is capable of disambiguating nouns, verbs, adjectives and adverbs and still achieves better results.

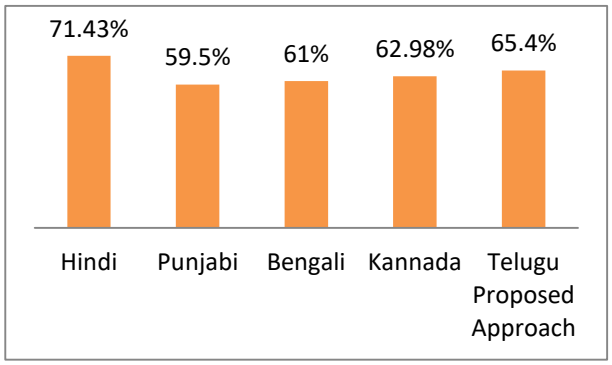

Fig 7: Comparison of results of proposed Indian language with other Indian languages

VI. CONCLUSION AND FUTURE SCOPE

The algorithm stated in this paper takes three parameters (intersection, hierarchy, distance) for calculating the closeness between targets and neighboring words. The similarity is enumerated by computing intersection among word families along the whole hierarchy of the target and nearby words. Also, the distance is merged with intersection and level to compute a score for all senses, representing to every target-nearby pair. The sense with highest score is chosen as the correct sense. The algorithm presented got a higher Accuracy when compared to various existing [13] knowledge-based methods on word sense disambiguation. The approach used in this paper is upgraded as compared to [13]

This concept may enable users to extract the correct meaning of the word from the given context but in the future, by merging more than one approach of machine learning it will give more accuracy and can be extended to enable users to get meaning more meaningfully and elaborately. Also, using in education and all other fields of a user, it can provide different types of templates and themes. Also, provide security to the users. As we mentioned that by merging the approaches will increase the complexity of the approach so complexity will be reduced as we apply the minimization to the system. Also, it implements the pdf documents.

REFERENCES

1. SuneethaEluri, SumalathaLingamgunta "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**].
2. SuneethaEluri, SumalathaLingamgunta "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**].
3. Ralph Grishman, "Natural Language Processing", Journal of the American Society for Information Science. 35(5): 291-296; 1984
4. Michael Lesk, "Automatic Sense Disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", Proceedings of SIGDOC'86. 1986
5. Satanjeev Bannerjee and Ted Pederson, "An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet," Third International Conference, CICLing 2002 Mexico City, 2002
6. Eneko Agirre and German Rigau, "Word Sense Disambiguation using conceptual density," Proceedings of the 16th conference on Computational linguistics - Volume 1, 1996
7. Daniel Marcu and William Wong, "A Phrase-base Joint Probability Model for Statistical Machine Translation", Proceedings of EMNLP, 2002.
8. Jyothi et al.," Parts of speech tagging of marathi text using trigram method", International Journal of Advanced Information Technology (IJAIT) Vol. 3, No.2, April2013.
9. Pushpak Bhattacharyya, "IndoWordnet" Department of Computer Science and Engineering Indian Institute of Technology Bombay http://www.cfilt.iitb.ac.in/indowordnet/index.jsp
10. Samhith.k, Arun Tilak.S, prof.G.Panda, " Word Sense Disambiguation using WordNet Lexical Categories " International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016.
11. Neeraja Koppula, B. Padmaja Rani and Koppula Srinivas Rao, "Graph-based word sense disambiguation in Telugu language" International Journal of Knowledge-based and Intelligent Engineering Systems 23 (2019) 55–60 55 DOI 10.3233/KES-190399 IOS Press
12. Palanati DurgaPrasad, K. V. N. Sunitha and B. Padmaja Rani, " Context-Based Word Sense Disambiguation in Telugu Using the Statistical Techniques "Springer Nature Singapore Pte Ltd. 2018 V. Bhateja et al. (eds.),
13. Pradeep Sachadev, Surabi Verma, Sandeep Kumar Singh, "An Improved Approach to the Word Sense Disambiguation" 978-1-4799-1812-6/14/$31.00 ©2014 IEEE
14. Pooja Sharma, Nisheeth Joshi "Knowledge-Based Method for Word Sense Disambiguation by Using Hindi WordNet" Vol. 9, No. 2, 2019, 3985-3989
15. Preeti Rana and Parteek Kumar, "Word Sense Disambiguation for Punjabi Language Using Overlap Based Approach" Springer International Publishing Switzerland 2015 El-Sayed M. El-Alfy et al. (eds.), Advances in Intelligent Informatics.
16. Alok Ranjan Pal and Diganta Saha, "Word Sense Disambiguation in Bengali language using unsupervised methodology with modifications" Indian Academy of Sciences
17. S.Parameswarappa and V. N. Narayana, "Target Word Sense Disambiguation System for Kannada language" © 2011 lET. Proc. of Int. Conf on Advances in Recent Technologies in Communication and Computing 2011.

AUTHORS PROFILE

Dr. Suneetha Eluri is working as Assistant Professor in the Department of Computer Science and Engineering at Jawaharlal Nehru Technological University Kakinada (JNTUK), Andhra Pradesh, India. Her research interests are Natural Language Processing (NLP) of regional languages with AI, Machine Learning and Deep Learning techniques. She is a Faculty champion of University Innovation Fellows programmer 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.

Vishala Siddu is a second year student at Jawaharlal Nehru Technological University Kakinada (JNTUK), Andhra Pradesh, India. She is pursuing her M. Tech in Computer Science and Engineering. She is currently working on her project on Natural language Processing (NLP) and Machine Learning. This is her first paper on NLP.