JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Semantic Distinguishing To Identify The Context Of Words In Telugu Sentences

[1]K. Neha, [2]P. Shivani, [3]M. Indu, [4]S. Arjun
[5]Dr.M.V.Vijaya Saradhi
[1,2,3,4] B.Tech (IV-CSE), Department of Computer Science and Engineering
[5]Head of the Department (HOD), Department of Computer Science and Engineering,
ACE Engineering College, Ghatkesar, Hyderabad, Telangana, India

Abstract : *Natural Language Processing domain is mainly used in fields such as Information retrieval and Text mining. These do not have a built-in semantic disambiguate and thus, it has to be performed explicitly. This has been largely performed on sentences in the English language to identify the context of a word in that sentence. This project's main idea is to implement semantic distinguishing in sentence by giving examples for regional language Telugu. Since, Telugu vocabulary is huge, occurrences of ambiguity is common and this project is aimed at resolving them. To achieve this, a synset is used, which is an interface for the Telugu WordNet. The algorithm used in this project is the Lesk algorithm. It is based on the assumption that the correct meaning or sense of a word is dependent upon the words present in its neighborhood. More specifically, context of a word is determined by comparing the meaning of each sense of the word with the meanings of the neighboring words. This project aims to get the examples for each ambiguous word in the sentence using Lesk algorithm.*

IndexTerms **- Semantic distinguishing , Telugu, Lesk algorithm, WordNet, Python.**

I. INTRODUCTION

In Natural Language Processing, Word Sense Disambiguation (WSD) is the problem of determining which meaning of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people. WSD is a natural classification problem: Given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context into one or more of its sense classes. The features of the context (such as neighboring words) provide the evidence for classification.

A famous example is to determine the sense of pen in the following passage: Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

1. WordNet lists five senses for the word pen: pen — a writing implement with a point from which ink flows.
2. Pen — an enclosure for confining livestock.
3. Playpen, pen — a portable enclosure in which babies may be left to play.
4. Penitentiary, pen — a correctional institution for those convicted of major crimes.
5. Pen — female swan.

Semantic Distinguishing is the task that automatically identifies the exact sense of the ambiguous word based on the context in which the word is used and give examples in the same context to understand the sense of the sentence. The proposed system aims to perform semantic distinguishing for Telugu .The input is taken in English and translated to Telugu using googletrans in python. Then required pre-processing steps are performed on the sentence. Stop words are removed, Stemming is performed, Lesk algorithm is applied and various examples are given as output.

II. LITRATURE SURVEY

The research papers and their brief summaries are as follows:

1. J.Sreedhar, in their paper spoke about the various methods which can be used to perform Word Sense Disambiguation for Telugu. The paper began by first talking about the taxonomy of Word Sense Disambiguation in NLP and explained in brief about each of them. In their own words, Word Sense Disambiguation is defined as follows: "Word Sense Disambiguation (WSD) is the ability to computationally determine which sense of a word is activated by its use in a particular context." In the following sections of the paper, current state of the art methods used for Word Sense Disambiguation are briefly explained. Some of the methods discussed in the paper are: - Method proposed by Walker, in which a thesaurus is used. Each word is assigned a particular subject categories in the thesaurus. Each subject assigned to a particular word is assumed to be a particular sense and WSD is performed based on this assumption. - Method proposed by Quillian in mid 1960s in which he proposed the use of semantic network representation of a machine readable dictionary.

In this, a node represents the meaning and this node is used to connect words. Many such methods have been specified in this paper along with a brief description of each method. Lesk is one of the proposed method which uses a WordNet. Improvements in this project: Even though the paper discusses various methods for Word Sense Disambiguation, it does not provide any implementation details about it. Lesk has been chosen to perform WSD in this project, as it has not been researched extensively for Telugu, unlike other algorithms. Many methods were analysed in this paper and provided a concise explanation. Since no implementation details were provided, this project aims at implementing Lesk algorithm described in the paper.

2. Ritesh Panjwani , in their paper spoke about their python-based Application Programming Interface (API) for Indian WordNets. This can be used as a module by getting it from pypi.org, where it is present as an open source project. In their own words, the aim of their project: “With our work, we aim to provide an accessible, robust, easy-to-use API for Indian language WordNets. The paper starts off by talking related work that already exists, such as ‘The Java WordNet Library’, which is a Java API to access the WordNet. Their motivation was to create an API for IndoWordNet which was python based so that it could be easily integrated with projects.

3. Word Sense Disambiguation for Telugu Using Lesk Sudheendra Poluru, A.Brahmananda Reddy, Rahul Manne, Lokesh Bathula & Nikhilender B. Reddy .

III. EXISTING SYSTEM

At present almost all word sense disambiguation focus on English and English synsets. It cannot work for the regional languages. It would be a great problem to find the ambiguity for the regional languages. Theme or context finding is an important part in processes such as page ranking and very few systems exist that perform word sense disambiguation in Telugu.

The drawbacks of the current system include the following:

Mostly applicable to English language: Most of the work on Word Sense Disambiguation is being performed in the English language. All new algorithms are first applied to the English language and only then, they are adapted to other languages slowly. The stemming required to perform Word Sense Disambiguation is performed by many python packages or similar only for English. There are no open source stemmers for Indian regional languages, for example, as a python module. There is no interface to the IndoWordNet provided by the ‘Pyiwn’ package which can be used to view the different synsets of a word, its hyponyms, gloss, etc.

IV. PROPOSED SYSTEM

Our solutions aims at performing word sense disambiguation on Telugu sentences to help gauge the theme of the sentence. It gives example for each word in the searched sentence in Telugu with same sense (meaning in the sentence).

The proposed system aims to perform Semantic distinguishing for regional Telugu in the following ways:

1. The input is taken in English and translated to Telugu using googletrans package in the python.
2. Pre-processing steps are performed on the sentence like removing stop words, Stemming.
3. Pyiwn python package is used, which provides an API to access the IndoWordNet.
4. The relevant synsets are used and provided as input to the Lesk algorithm.
5. Lastly, it outputs example for each word in the searched sentence in Telugu with same sense (meaning in the sentence).

LESK ALGORITHM

The Lesk algorithm is a classical word sense disambiguation algorithm introduced by Michael E. Lesk. It is based on the assumption that the sense of a word is dependent upon the words present in its neighborhood. More specifically, sense or context of a word is determined by comparing the meaning of each sense of the word with the meanings of the neighboring words. This comparison is made by matching the words in meanings and the sense whose meaning matches the maximum number of times is taken as the resultant sense of that particular word in a given sentence.

Steps in Lesk Algorithm:

1. Consider a word for processing.
2. Next, take the remaining words and create a list of words, which consist of the gloss and examples of those words. Let us call it ‘matcher_list’.
3. Remove stop words from these lists to reduce irregular matches.
4. Now for every synset of the word to be processed, create a similar list of words containing the gloss and examples of that synset. Let us call this list ‘synset_list’.
5. Match the above two list to find overlap of words.
6. The synset_list, which has the maximum overlap with the matcher_list, is taken as the best synset and the gloss of the synset is taken as the best sense.

V. IMPLEMENTATION

In Implementation different steps involved are Reading Input, Translating Input, Removing stop words, Getting synsets, Applying Lesk, Displaying results (examples).

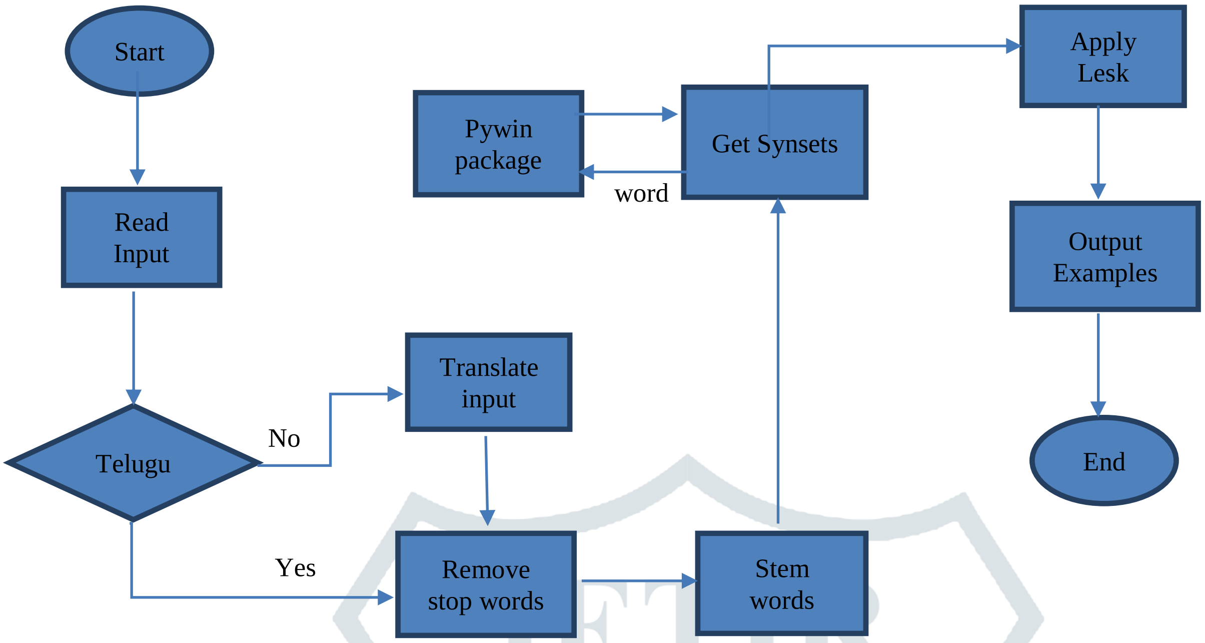
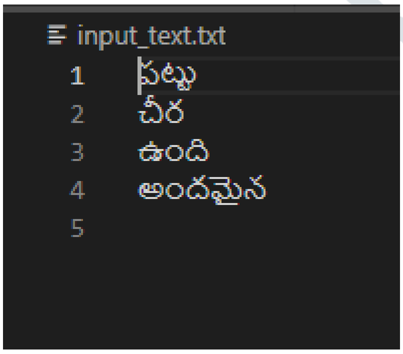


Fig 1: Implementation flow

To achieve this knowledge based algorithm is used which includes finding the overlap between the features of different senses of ambiguous word and the features of the word in its context. That is performed by taking words in the neighborhood of the target word and the sense, which has the maximum overlap, is selected as the context appropriate sense. Semantic distinguishing in Telugu is achieved after performing necessary pre-processing steps such as translating the words, stop words filtration, stemming, etc. In the below input sentences the word grip and silk have same translation 'pattu' in telugu but the meaning according to the context of sentence is different . So, In the output examples are given for each word in the context of the sentence.

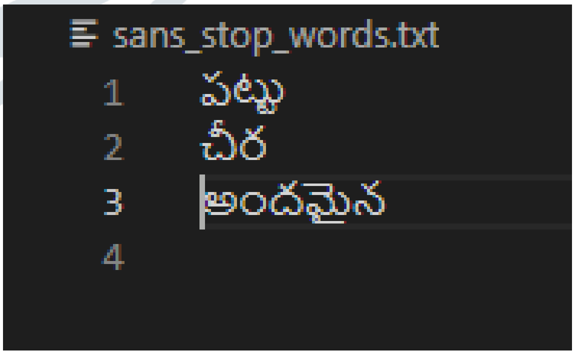
Example 1:

Input text : Silk saree is beautiful



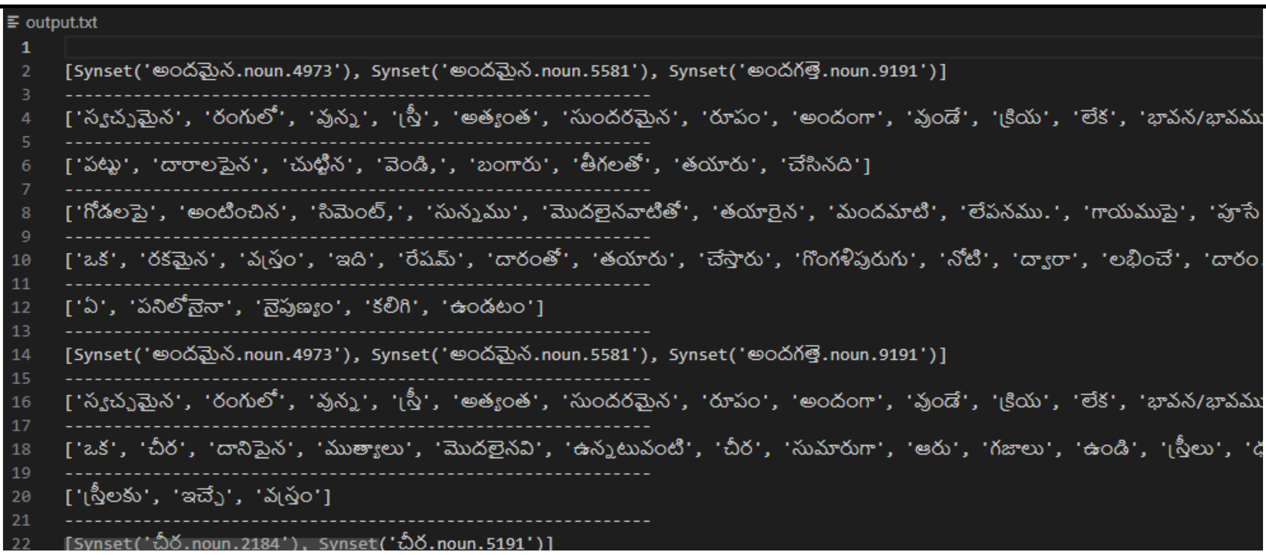
```
input_text.txt
పట్టు
చీర
ఉంది
అందమైన
```

Fig 2.1: Input in telugu



```
sans_stop_words.txt
పట్టు
చీర
అందమైన
```

Fig 2.2: Input after removing stop words

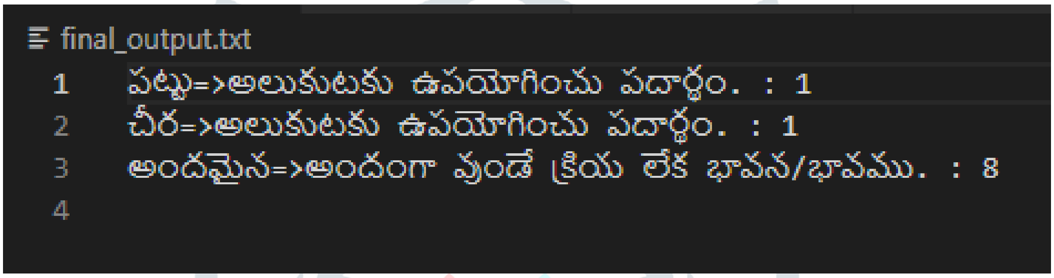


output.txt

```

[Synset('అందమైన.noun.4973'), Synset('అందమైన.noun.5581'), Synset('అందగత్తె.noun.9191')]
------------------------------------------------------------
['స్వచ్ఛమైన', 'రంగులో', 'వున్న', 'స్త్రీ', 'అత్యంత', 'సుందరమైన', 'రూపం', 'అందంగా', 'వుండే', 'క్రియ', 'లేక', 'భావన/భావము
------------------------------------------------------------
['పట్టు', 'దారాలపైన', 'చుట్టిన', 'వెండి,', 'బంగారు', 'తీగలతో', 'తయారు', 'చేసినది']
------------------------------------------------------------
['గోడలపై', 'అంటించిన', 'సిమెంట్,', 'సున్నము', 'మొదలైనవాటితో', 'తయారైన', 'మందమాటి', 'లేపనము.', 'గాయముపై', 'పూసే
------------------------------------------------------------
['ఒక', 'రకమైన', 'వస్త్రం', 'ఇది', 'రేషమ్', 'దారంతో', 'తయారు', 'చేస్తారు', 'గొంగళిపురుగు', 'నోటి', 'ద్వారా', 'లభించే', 'దారం.
------------------------------------------------------------
['ఏ', 'పనిలోనైనా', 'నైపుణ్యం', 'కలిగి', 'ఉండటం']
------------------------------------------------------------
[Synset('అందమైన.noun.4973'), Synset('అందమైన.noun.5581'), Synset('అందగత్తె.noun.9191')]
------------------------------------------------------------
['స్వచ్ఛమైన', 'రంగులో', 'వున్న', 'స్త్రీ', 'అత్యంత', 'సుందరమైన', 'రూపం', 'అందంగా', 'వుండే', 'క్రియ', 'లేక', 'భావన/భావము
------------------------------------------------------------
['ఒక', 'చీర', 'దానిపైన', 'ముత్యాలు', 'మొదలైనవి', 'ఉన్నటువంటి', 'చీర', 'సుమారుగా', 'ఆరు', 'గజాలు', 'ఉండి', 'స్త్రీలు', 'ధ
------------------------------------------------------------
['స్త్రీలకు', 'ఇచ్చే', 'వస్త్రం']
------------------------------------------------------------
[Synset('చీర.noun.2184'), Synset('చీర.noun.5191')]
```

Fig 2.3: Matching synset



final_output.txt

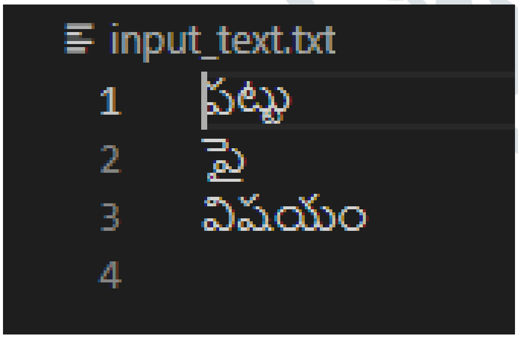
```
పట్టు=>అలుకుటకు ఉపయోగించు పదార్థం. : 1
చీర=>అలుకుటకు ఉపయోగించు పదార్థం. : 1
అందమైన=>అందంగా వుండే క్రియ లేక భావన/భావము. : 8

```

Fig 2.4: Final output examples

Example 2:

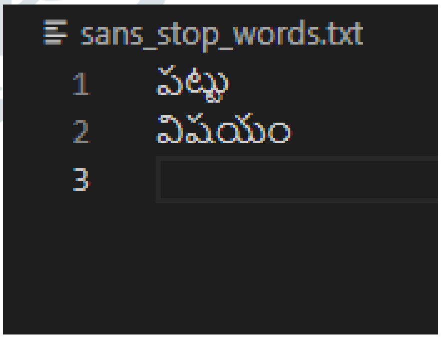
Input text : Grip on subject



input_text.txt

```
పట్టు
పై
విషయం

```



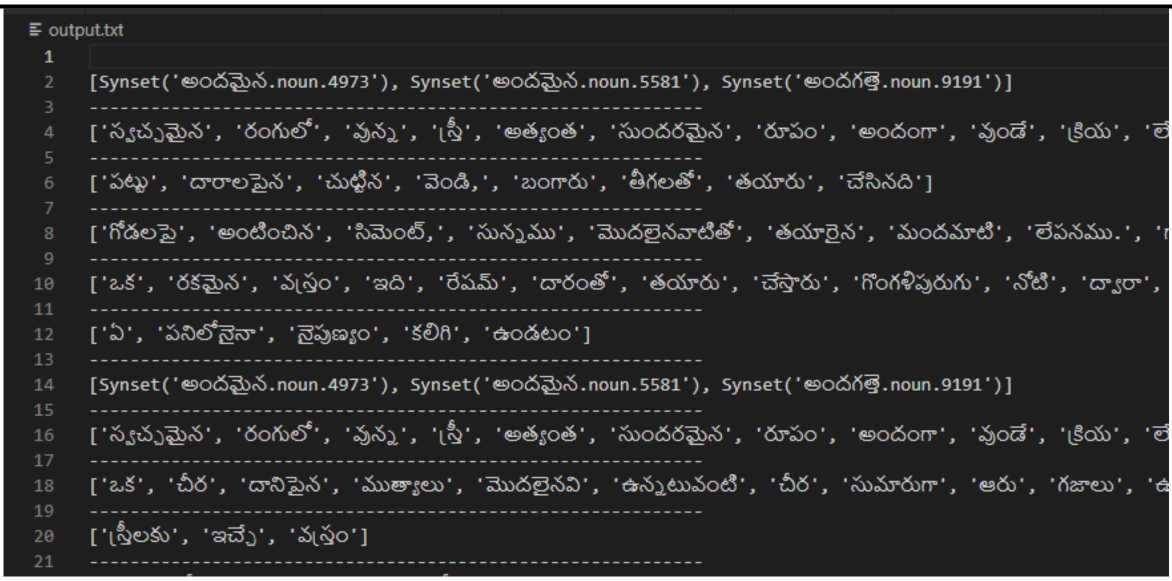
sans_stop_words.txt

```
పట్టు
విషయం

```

Fig 3.1: Input in telugu

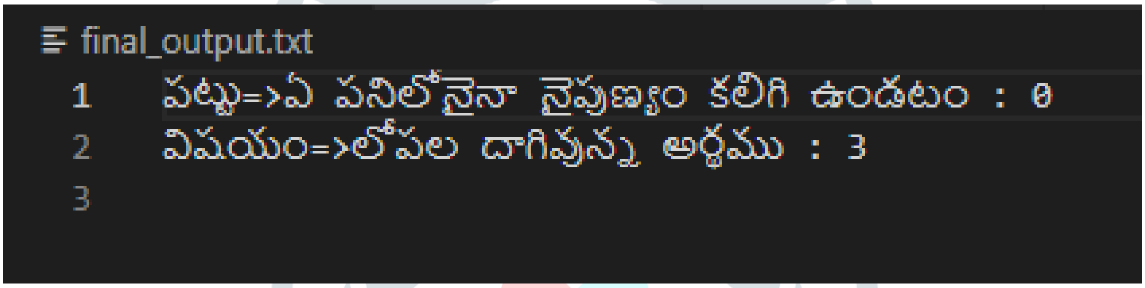
Fig 3.2: Input after removing stop words



```

[Synset('అందమైన.noun.4973'), Synset('అందమైన.noun.5581'), Synset('అందగత్తె.noun.9191')]
------------------------------------------------------------
['స్వచ్ఛమైన', 'రంగులో', 'వున్న', 'స్త్రీ', 'అత్యంత', 'సుందరమైన', 'రూపం', 'అందంగా', 'వుండే', 'క్రియ', 'లే
------------------------------------------------------------
['పట్టు', 'దారాలపైన', 'చుట్టిన', 'వెండి,', 'బంగారు', 'తీగలతో', 'తయారు', 'చేసినది']
------------------------------------------------------------
['గోడలపై', 'అంటించిన', 'సిమెంట్,', 'సున్నము', 'మొదలైనవాటితో', 'తయారైన', 'మందమాటి', 'లేపనము.', '
------------------------------------------------------------
['ఒక', 'రకమైన', 'వస్త్రం', 'ఇది', 'రేషమ్', 'దారంతో', 'తయారు', 'చేస్తారు', 'గొంగళిపురుగు', 'నోటి', 'ద్వారా',
------------------------------------------------------------
['ఏ', 'పనిలోనైనా', 'నైపుణ్యం', 'కలిగి', 'ఉండటం']
------------------------------------------------------------
[Synset('అందమైన.noun.4973'), Synset('అందమైన.noun.5581'), Synset('అందగత్తె.noun.9191')]
------------------------------------------------------------
['స్వచ్ఛమైన', 'రంగులో', 'వున్న', 'స్త్రీ', 'అత్యంత', 'సుందరమైన', 'రూపం', 'అందంగా', 'వుండే', 'క్రియ', 'లే
------------------------------------------------------------
['ఒక', 'చీర', 'దానిపైన', 'ముత్యాలు', 'మొదలైనవి', 'ఉన్నటువంటి', 'చీర', 'సుమారుగా', 'ఆరు', 'గజాలు', 'ఉ
------------------------------------------------------------
['స్త్రీలకు', 'ఇచ్చే', 'వస్త్రం']
------------------------------------------------------------
```

Fig: Matching synset



```
పట్టు=>ఏ పనిలోనైనా నైపుణ్యం కలిగి ఉండటం : 0
విషయం=>లోపల దాగివున్న అర్థము : 3

```

Fig : Final output examples

VI. CONCLUSION

This paper mainly focuses on what Semantic Disambiguation. Many approaches were made in solving them, which mainly focused on English language. So, this paper focuses on Semantic distinguishing for Telugu which is an Indian regional language spoken by the people of states of Telangana and Andhra Pradesh. Semantic distinguishing to identify the context of a word in Telugu sentences is achieved after performing necessary preprocessing steps such as translating the words, stop-words filtration, stemming, etc., and later implementing the Lesk algorithm on the processed words and outputs examples to understand the context of the sentences. This works for the input, which is given in single sentences form.

VII. ACKNOWLEDGMENT

We would like to thanks to our guide Dr.M.V.Vijaya Saradhi, Head of the Department of Computer Science and Engineering, ACE Engineering College and Mrs. Soppal Kavitha for their continuous support and guidance. Due to their guidance, we could complete our work successfully.

REFERENCES

[1] A. Brahmananda Reddy “Integrated Feature Selection Methods for Text Document Clustering”, Integrated Feature Selection Methods for Text Document Clustering, 2015.

[2] A. Brahmananda Reddy, A.Govardan “Ontology for an Education System and Ontology Based Clustering”, The Fifth International Conference on Fuzzy and Neuro Computing (FANCCO - 2015), IDRBT, Hyderabad, 2015.

[3] J.Sreedhar , Dr.S.Viswanadha Raju, Dr.A.Vinaya Babu “A Study of Critical Approaches in WSD for Telugu Language Nouns: Current State of the Art” International Journal of Scientific &Engineering Research, Volume 5, Issue 6, ISSN 2229-5518, June 2014.

[4] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[5] George A. Miller (1995). WordNet: A Lexical Database for English Communications of the ACM Vol. 38, No. 11: 39-41.

[6] Ritesh Panjwani, Diptesh Kanojia, Pushpak Bhattacharyya “pyiwn: A Python based API to access Indian Language WordNets” IIT Bombay, 2018.