International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 1 132 - 139

Named Entity Recognizer for Telugu language using Hybrid approach

Dr. M. Humera Khanam
Department of Computer Science and Engineering
Sri Venkateswara University College of Engineering
Tirupati, India
email: humera_svec@yahoo.co.in

Miss. P. Sindhu Sree
Department of Computer Science and Engineering
Sri Venkateswara University College of Engineering
Tirupati, India
email: sindhusree79@gmail.com

Abstract:- The main goal of Named Entity Recognition (NER) is to classify all Named Entities (NE) in a document into predefined classes like Person name, Location name, Organization name and Miscellaneous. This paper outlines Named Entity Recognizer using hybrid approach i.e., combination of Rule based approach and one of the Machine learning technique i.e, Conditional Random Field (CRF). In Rule based approach we have prepared Gazetteer lists for names of persons, locations and organizations; some suffix and prefix features and dictionary consisting 350266 words to recognize the category of named entities. If ambiguity is rised while we are using Rule based approach, we use Machine learning technique i.e., CRF in order to improve the accuracy.

Keywords:- *Named Entities (NEs), Natural Language Processing (NLP), Named Entity Recognition (NER), Conditional Random Field (CRF), Human Computer Interaction (HCI).*

*****

1. INTRODUCTION

NLP is a component of Artificial Intelligence (AI). NLP is the ability of a computer program to understand human speech as it is spoken. The ultimate goal of NLP is to provide Human Computer Interaction (HCI) i.e., a human can communicate with a computer in his native language even without having any knowledge about computer languages like C, C++, Java, Oracle, etc.,[14].

In English language, lot of work has been done in NER, where capitalization is a major clue for developing rules[16]. Developing Named Entity Recognizer is a very difficult process in Indian languages such as Telugu, Hindi, Tamil, Bengali, Kanada, Urdu etc., due to lack of capitalization concept and sufficient gazetteers and annotated corpora are unavailable compared to English. We recognize that NEs are usually nouns, so we maintain a dictionary for Noun Identification.

1.1. HISTORY

The history of NLP started in the year 1950, although work can be found from earlier periods. In 1950, Alan Turing published his famous article "Computing Machinery and Intelligence", he proposed now popularly called the Turing test as a criterion of intelligence. Some successful NLP systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum around 1964 to 1966. In 1970's many programmers written 'conceptual ontologies', which structured real-world information into computer-understandable data[15]. Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Many of the notable early successes occurred in the field of machine translation, due especially to work at IBM Research, where successively more complicated statistical models were developed. Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms[2].

1.2. APPLICATIONS of NER

NER is a crucial tool in almost all of the NLP applications,

- Information Retrieval (IR),
- Information Extraction (IE),
- Question Answering (QA),
- Machine Translation (MT),
- Text Summarization (TS) etc.,

NER is a two phase problem:- one phase is Identification of Proper Noun and another is classification of the Proper Noun into a set of predefined classes.

1.3. Approaches of NER

There are three approaches for developing NER,

[1] **Rule based approach:** In Rule based approach we maintain gazetteer lists and need to maintain rules which was developed by linguistic experts [17].

Disadvantage: It is very difficult to maintain gazetteer lists and to develop rules since Telugu is a resource poor language.

[2] **Machine learning approach:** We use so many Machine learning techniques where no need to maintain any gazetteer

lists and rules. But, here we have to maintain trained data, based on trained data it will give required output to test data.

- Hidden Markov Models (HMMs)
- Maximum Entropy Markov Models (MEMMs)
- Conditional Random Field (CRF)
- Support Vector Machine (SVM)
- Decision Tree (DT)

Disadvantage: Maintaining trained data is not so easy and we can't resolve ambiguity completely.

[3] Hybrid Approach: Hybrid approach is a combination of Rule based approach and Machine learning techniques. Since it is not so easy to resolve ambiguity either by using Rule based approach or any one of the Machine learning technique. So, we use Hybrid approach in order improve accuracy to 100%.

1.4 Motivation

Named Entity Recognition (NER) is an essential technology for a number of propelled majority of the data management applications, such as search engines, question answering systems, text mining and business intelligence. All of these applications can profit from exact NER such as search engines and question answering systems can be improved by permitting searches and inquires for particular persons, organizations or locations. In text mining, exact NER will permit the development of databases with information wrenching out regarding specific entities. Multilingual NER applications are cross-language information retrieval, and business intelligence applications, where information about a specific person or organization has to be wrenched from textual sources in different languages.

1.5. Challenges

Very minimum work has been completed in NLP, especially in Indian Languages still 1950's. NER is a subtask of many applications of NLP; so, it plays a major role in NLP. It is very difficult to develop NER in Indian Languages, especially in Telugu, due to unavailability of resources, dictionaries, linguistics rules, gazetteer lists. So, it becomes a great challenge to maintain resources for developing Named Entity Recognizer. Ambiguity is also major challenge of NER.

2. PROBLEMS IN INDIAN LANGUAGES

In English, due to capitalization concept and availability of web resources it is somewhat easy to develop NLP applications [15]. But, in Indian languages, problems faces are

- No capitalization
- Brahmi script
- Non-availability of large gazetteer lists
- Lack of standardization and spelling
- Indian names are ambiguous
- Lack of labeled data
- Scarcity of resources and tools
- Free word order language
- Lack of dictionaries

3. PROBLEMS IN TELUGU LANGUAGE

Telugu language faces many problems since it is an inflectional language and problems faces are,

3.1. Variations of NEs:

Indian names are also more diverse i.e., there is lot of variations for a given named entities **Example:**

[1] వై.యస్.ఆర్, వై.యస్.రాజశేఖ రెడ్డి, వై యస్ ఆర్,

వై .యస్, రాజశేఖర్ రెడ్డి, ...

[2] టి.ఆర్.యస్, తెరాసా, టి ఆర్ యస్,టి.ఆర్.యస్ పార్టీ,

తెలంగాణా రాష్ట్ర సమితి, ...

3.2. Ambiguity with NE type: Words in Indian language have more than one meaning [11].

Person name Vs Organization name:

[1] డా.రెడ్డి (Person name)

Vs

డా.రెడ్డి ల్యాబ్స్ (Organization name)

Person name Vs Location name:

[1] రంగారెడ్డి (Person name)

Vs

రంగారెడ్డి జిల్లా (Location name)

[2] తిరుపతి (Person name)

Vs

తిరుపతి (Location name)

Location name Vs Organization name:

[1] విశాకపట్నం (Location name)

Vs

విశాకపట్నం థర్మల్ పవర్ స్టేషన్ (Organization name)

 ISSN: 2321-8169
132 - 139

3.3. Ambiguity with common noun:

[1] బంగారు (Person name)

Vs

బంగారు (Common noun)

[2] రాజు (Person name)

Vs

రాజు (Common noun)

4. MODULES:

- User interface
- Noun Identification
- Named Entity List
- NE Identification
  - Using Rule based approach
  - Using Conditional Random Field

4.1. User interface

User interface is visible to the users who can request their queries and can see the result for their queries.
For example,

Input: యస్వియూనివర్సిటి తిరుపతిలో ఉంది

Output: యస్వియూనివర్సిటి (Organization name)

తిరుపతి (Location name)

4.2. Noun Identification

Noun identification is very easy in English language because of the number of dictionaries, inflection lists, suffix lists and other resources that are available in the Web. In Indian languages like Telugu, these resources are unavailable. We have to develop gazetteer lists and language-dependent rules for the specific language.
Nouns are identified by using two approaches[3],

(a) Using Dictionary
(b) Using suffix list for common nouns

4.2.1. Using Dictionary

This approach requires the loading of the computer system with a dictionary of Telugu language. This dictionary contains 350266 words with root forms. We can also update dictionary with new words. It is helpful for identification of closed class words, such as, adjectives, adverbs, verbs, conjunctions and root forms for nouns. Nouns may also appear with various inflected forms. We have to identify the nouns by using various suffix features and rules.

4.2.2. Using suffix list for common nouns

The rules for Noun identification are framed after studying various articles, news papers, periodicals and Telugu grammar books. Some of the noun suffix lists are used for the identification of nouns[19]. For Example:

Words in Telugu	Transliterated form	Meaning
చేత	cheta	by
గుండా	gunda	through
కి	ki	to
కు	ku	to
లచేత	lachetha	with
లచే	lache	with
లద్వారా	ladwara	through
లకి	laki	for
లకు	laku	for
లాగుండా	lagunda	through
లలోపలకు	laloapalaku	into
లలోపలనుండి	lalopalanundi	from
లలోపల	laloapala	inside
నుండి	nundi	from
ని/లని/ను/లను	ni/lani/nu/lanu	is
తో	tho	with
లతో	tho	with

Table1: Suffix list for nouns

4.3. Named Entity List

Three NE Lists (Named Entity list) are maintained,

(a) Person_name list
(b) Location list
(c) Organization list

4.4. NER Identification

4.4.1. Using Rule based Approach

Suffix features

Every language uses some particular patterns which may act as end words in proper names and list of this type of words is called suffix list [4].
Examples:

Person suffixes:

శర్మ, రాజు, నాయుడు, చౌదరి, మూర్తి, రెడ్డి, రావు, శాస్త్రి, బాబు, గారు,.....

Example: [1] రవిశాస్త్రి

[2] రామారావు

Location suffixes:

వాడ, పట్నం, జిల్లా, రాష్ట్రం, పురం, పల్లి, మండలం, వీది, దేశం,.....

Example: [1] చిత్తూరుజిల్లా

[2] భారతదేశం

Organization suffixes:

యూనివర్సిటి, సంస్థ, అకాడమీ, పార్టి, బ్యాంకు, లిమిటెడ్, కాలేజీ, గ్రంధాలయం, కార్పొరేషన్, కళాశాల, లైబ్రరీ, రేసోర్ట్, హోటల్, రెస్టారెంట్లు, కంపెనీ,.....

Example: [1] మహిళాయూనివర్సిటీ

[2] సాంస్క్రిట్విద్యాపీట్

Context features

Every language uses some particular patterns which may acts as clue words and the list of this type of words is called as context features [18].

Example:

Surname context:

మిశ్రా, సోమగుట్లు, అరమటి, పూతలపట్లు, కానుగ, త్రివేణి, పూరి, కదిరి, పుట్ట, రాశి, చింతకాయల, దొంగల, బొడ్డు,.....

Example: [1] పూతలపట్లు శ్రీరాములురెడ్డి

[2] అరమటి ఈశ్వరమ్మ

Abbreviation form of person:

యస్.వి.ఆర్.ఆర్, వై.యస్.ఆర్, పి.యస్.ఆర్, యన్.టి.ఆర్, ఎ.యన్.ఆర్, పి,యస్.యస్,.....

Middle name context:

చంద్ర, రమణ, గోపాల్, కృష్ణా, లక్ష్మి, పార్వతి,.....

Example: [1] రామ చంద్ర శాస్త్రి

[2] రామ్ గోపాల్ వర్మ

Relational context:

మిత్రుడు, బాబాయ్, అమ్మ, నాన్న, పిన్ని, మిత్రురాలు, పెద్ద్ద న్న, పెద్దమ్మ, అత్త, మామ,.....

Location name:

గ్రామం, పట్నం, జిల్లా, నగర్, కాలనీ, మందిరం, మందిర్, మాల్, సూపర్ మార్కెట్, మార్కెట్, సూపర్ మాల్,ఆఫీస్,.....

Example: [1] బిర్లా మందిర్

[2] బాలాజి నగర్

Organization name:

శాఖ, పీటము, సంస్థ, కేతన్, స్కూల్, బడి, కాలేజీ, ఇన్స్టిట్యూట్, కళాశాల,.....

Example: [1] విద్యానికేతన్

[2] తెరెసా సంస్థ

Prefix features

Every language uses some particular patterns which may act as starting words in proper names and the list of this type of words is called prefix list.

Person name:

శ్రీ, అధ్యక్షుడు, మాస్టర్, మిస్సెస్స్, మిస్, డాక్టర్, శ్రీమతి, సర్, మేడమ్, డా., గౌరవనీయులైన, అధ్యక్షురాలు, పద్మశ్రీ.

Example: [1] మిస్ అనిత

[2] మాస్టర్ విశ్వనాధ్ ఆనంద్

Morphological features

Indian languages are rich in morphology. Words are inflected in various form depending on its number, tense, person, case, etc., Root word identification is very difficult in Indian languages especially Telugu [9].

Example:

(Common noun)

[1] ఆవులతో → ఆవు + లు + తో

(With cows) root word number case marker

[2] రామారావుతో → రామారావు + తో

(with ramarao) root word case marker

[3] నెల్లూరునుండి → నెల్లూరు + నుండి

(From Nellore) root word case marker

4.4.2. Using CRF

Many CRF algorithms like CRF Suite, CRF++, FlexCRFs, MALLET, HCRF library, Wapiti, etc., are available. We train the algorithm with some trained data and it also maintain word handler for unknown words and disambiguation rules [10].

CRF++ (used in NER, IE, Text chunking) is a simple, customizable and implementation of CRFs for segmenting / labeling sequential data. The first version of CRF++ is CRF++ 0.1 released on 28-05-2005 and the latest version is CRF++ 0.58 released on 13-02-2013 [7].

4.4.3. CRF learn command

% crf learn template_file train_file model_file

Where template_file and train_file are the files you need to prepare in advance. “crf learn” generates the trained model file in model_file [13][7][12].

 ISSN: 2321-8169
132 - 139

4.4.4. CRF test command

% crf test –m model_file test_file

where model_file is the file crf learn creates. In the testing, you don't need to specify the template file, because model file has the same information for the template. Test_file is the test data you want to assign sequential tags. This file has to be written in the same format as training file[7][12]

5. ARCHITECTURE:

The architecture of the project shows entire flow of our NER using Hybrid approach[16][17].

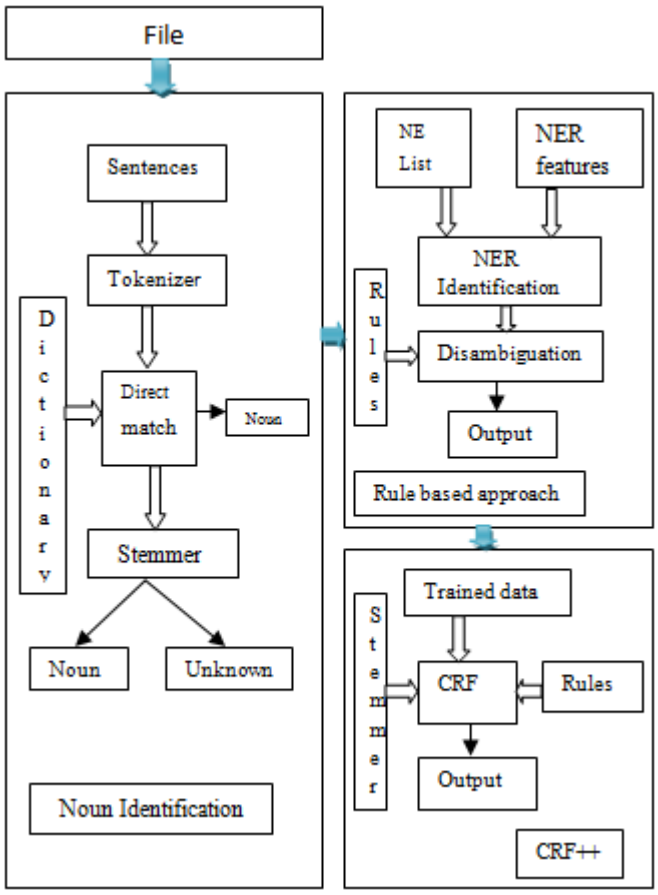

Fig1: Named Entity Recognizer Architecture

6. INPUT AND OUTPUT

Input:

నా పేరు సింధు. నేను యస్వియునివేర్సిటిలో చదువుతున్నాను. యస్వియునివేర్సిటి తిరుపతిలో ఉంది. నేను మా గైడ్ హుమెర మేడమ్ కింద ప్రాజెక్ట్ చెస్తున్నాను.మా హెచ్వోడి పేరు వెంకటసుబ్బారెడ్డి సర్. మా కళాశాల 8:౩౦కి మొదలవుతుంది.

Step1: divide file into sentences

నా పేరు సింధు.

నేను యస్వియునివేర్సిటిలో చదువుతున్నాను.

యస్వియునివేర్సిటి తిరుపతిలో ఉంది

నేను మా గైడ్ హుమెర మేడమ్ కింద ప్రోజెక్ట్ చెస్తున్నాను.

మా హెచ్వోడి పేరు వెంకటసుబ్బారెడ్డి సర్.

మా కళాశాల 8:౩౦కి మొదలవుతుంది.

Step 2: Tokenization

నా | పేరు | సింధు | .

నేను |యస్వియునివేర్సిటిలో | చదువుతున్నాను| .

యస్వియునివేర్సిటి | తిరుపతిలో | ఉంది | .

నేను | మా | గైడ్ | హుమెర | మేడమ్ | కింద |

ప్రోజెక్ట్ | చెస్తున్నాను | .

మా | హెచ్వోడి | పేరు | వెంకటసుబ్బారెడ్డి | సర్ | .

మా | కళాశాల | 8:౩౦కి | మొదలవుతుంది | .

 ISSN: 2321-8169
132 - 139

Step 3: If tokens directly match with dictionary, assign as noun

సింధు (noun)

తిరుపతి (noun)

హుమెర (noun)

వెంకటసుబ్బారెడ్డి (noun)

Step 4: Otherwise, do stemming.

యస్వియునివేర్సిటిలో[svuniversitylo] - లో[lo] → యస్వియునివేర్సిటి [svuniversity]

(noun)

తిరుపతిలో[tiruapatilo] – లో [lo] → తిరుపతి[Tirupati]

(noun)

8:౩౦కి[8:30ki] – కి[ki] →8:30 (noun)

Step 5: If the noun match with NER list then assign its tag, otherwise use NER features and Disambiguation rules

సింధు (Person name)

తిరుపతి (Location name Vs Person name)

యస్వియునివేర్సిటి (Organization name)

హుమెర (Person name)

వెంకటసుబ్బారెడ్డి (Person name)

8:30 (time)

Step 6: Still have ambiguity and unknown words, and then go for CRF

తిరుపతి (Location name)

Output:

సింధు (Person name)

తిరుపతి (Location name)

యస్వియునివేర్సిటి (Organization name)

హుమెర (Person name)

వెంకటసుబ్బారెడ్డి (Person name)

8:30 (time)

7. ALGORITHMS

7.1. Algorithm for Noun Identification

Step1: Read input text file and split into sentences
Step2: Read each sentence and split into tokens
Step3: Read each token
Step4: For each (token) Loop
check with Telugu dictionary
Step5: If direct match with dictionary then
assign noun
Step6: else if no match with dictionary then
check with suffix list for nouns
Step7: if suffixes are found and root is found in
Telugu dictionary then
assign noun
Step8: else if suffix matches and root is not found
then token may be noun
Step9: else if token ending with consonant then
the word may be loan word
assign noun
Step10: else
assign the category “unknown”
End loop

7.2. Algorithm for NER Identification

Step1: Read list of nouns identified by above
Algorithm 7.1
Step2: Check gazetteer lists for NER features
Step3: For each (noun) Loop
If suffix features found then
assign NER tag
Step4: else if prefix features found then
assign NER tag
Step5: else if context features found then
assign NER tag
Step6: else if found in NER list then
assign NER tag
Step7: else
assign “Miscellaneous word”
Step8: If ambiguity is found then
Call disambiguation rules
Remove ambiguity
Step9: Else if still ambiguity and unknown words
found then
Call CRF
Step9.1: Generate a template_file and train_file for
Each token
Step9.2: Train file by using command:
% crf learn template_file train_file model_file

Step9.3: Test file by using command:
% crf test –m model_file test_file
End Loop

 ISSN: 2321-8169
132 - 139

8. EXPERIMENTAL RESULTS

Fig2: GUI for Named Entity Recognizer

Fig3: Dictionary and Gazetteer lists are loaded

Fig4: Input file

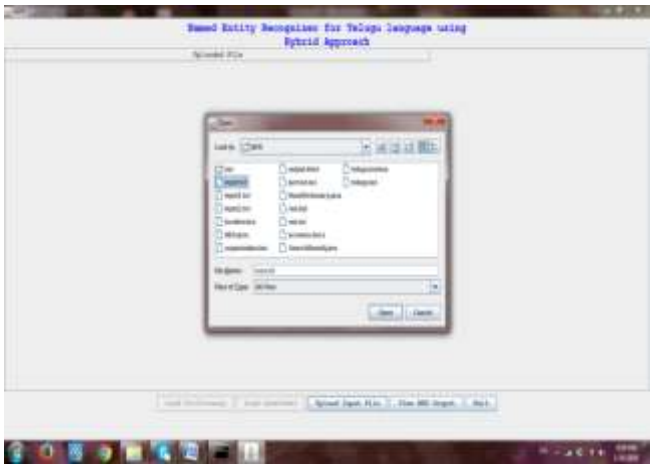

Fig5: Uploading input file into NER system

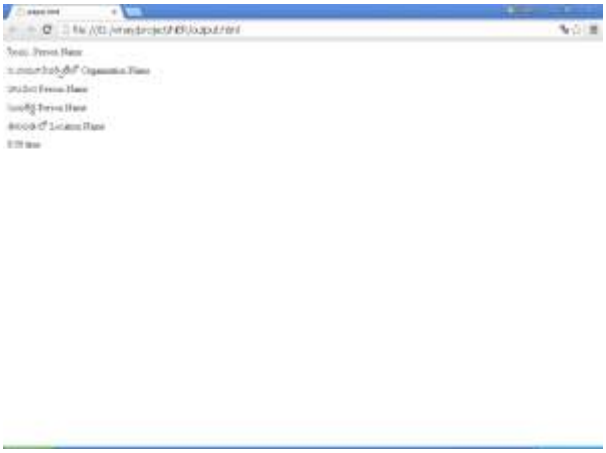

Fig6: Output by Named Entity Recognizer

9. CONCLUSION AND FUTURE WORK

This Named Entity Recognizer uses Hybrid approach i.e., combination of Rule based approach and one of the Machine learning technique, Conditional Random Field and it categorize universally accepted entities i.e., Person name, Location name and Organization name and give 91% -94% accuracy. Future work is to improve accuracy to 100% by using high standard dictionaries, gazetteer lists and the most accurate algorithms.

10. REFERENCES

[1] Krishnamurti, B., A grammar of modern Telugu. 1985, Delhi; New York: Oxford University Press.

[2] R.Grishman. 1995. “The NYU system for MUC-6 or Where’s the Syntax” in the proceedings of Sixth Message Understanding Conference (MUC-6), pages 167-195, Fairfax, Virginia.

[3] McDonald D. 1996. Internal and external evidence in the identification and semantic categorization of proper

names. In: B.Boguraev and J. Pustejovsky (eds), Corpus Processing for Lexical Acquisition, pp. 21-39.

[4] Brill E., 1992. A Simple Rule-Based Part of Speech Tagger. In Proceedings of the 3rd Conference on Applied NLP,152-155.

[5] D. Appelt, J.Hobbs, J. Bear, D.Israel, M. Kameyama, A.Kehler, D. Martin, K.Meyers and M. Tyson SRI: International FAS TUS system: MUC-6 test result and analysis, 1993.

[6] Satoshi Sekine, Newyork University: Named Entity: History and Future.

[7] CRF++: http:// crfpp.sourceforge.net / Yet Another.

[8] Nadeau, D., Turney, P. and Matwin, S. Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In Proceedings of Canadian Conference on Artificial Intelligence, 2006.

[9] Ekbal, A., Naskar, S., Bandyopadhyay, S.: Named Entity Recognition and Transliteration in Bengali. Named Entities: Recognition, Classification and Use, Special Issue of Lingvisticae Investigationes Journal 30 (2007) 95–114.

[10] Praneeth M Shishtla, Karthik Gali, Prasad Pingali and Vasudeva Varma. 2008. "Experiments in Telugu NER: A conditional Random Field Approach" in the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 105-110, Hyderabad, India.

[11] http://www.te.wikipedia.org/wiki.

[12] **https://taku910.github.io/crfpp/**

[13] https://courses.cs.washington.edu/courses/cse454/09sp/crf.html

[14] **http://nlp.lsi.upc.edu/freeling/doc/userman/html/node66.html.**

[15] P.Srikanth and K. N. Murthy, Named Entity Recognition for Telugu. In Proceedings of the IJCLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, Jan 2008, pp. 41–50.

[16] Darvinder Kaur, Vishal Gupta: A Survey of Named Entity Recognition in English and other Indian languages, IJCSI International Journal of Computer Science Issues, vol 7, Issue 6, November 2010; ISSN(online): 1694 – 0814.

[17] B. Sasidhar: A Survey on Named Entity Recognition in Indian languages with particular reference to Telugu: IJCSI International Journal of Computer Science Issues, vol 8, Issue 2, Mar 2011; ISSN(online): 1694 – 0814.

[18] B. Sasidhar, P.M. Yohan, Dr. A. Vinaya Babu, Dr. A. Govindhan: Named Entity Recognition in Telugu language using Language Dependent Features and Rule based approach: International Journal of Computer Applications (1975-8887), vol 22-No 8, May 2011.

[19] **http://shodhganga.inflibnet.ac.in/bitstream/10603/8258/18/18_synopsis.pdf".**