Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach

B. Sasidhar
Associate Professor
Dept. of MCA,
CM Engineering College,
Secunderabad

P. M. Yohan
Associate Professor
Dept. of MCA,
Wesley P.G. College,
Secunderabad

Dr. A. Vinaya Babu
Director, Admissions,
JNTU,
Hyderabad

Dr. A. Govardhan
Principal
JNTUH College of Engg
Nachupally (Kondagattu),
Karimanagar Dt., A.P.

ABSTRACT

The objective of Named Entity Recognition (NER) is to categorize all named entities in a document into predefined classes like person, organization, location, brand names and others. Named Entity Recognition is a difficult process in Indian languages like Telugu, Hindi, and Bengali, Urdu etc., where sufficient gazetteers and annotated corpora are not available compared to English language? A rule based systems is very difficult to implement because of lack of grammatical and linguistic analysis to make rules in Indian languages like "Telugu". In this paper we describe the identification of Named Entities using various features, gazetteer lists using language dependent features and rule based approaches for Telugu language. Here we described two phase representation of Named Entity Recognition. The first phase describes the noun identification using Telugu dictionaries, noun morphological stemmer and noun suffixes. The second phase identifies the Named Entities using transliterated gazetteer lists related to different Named Entity tags, various Named Entity suffix features, context features and morphological features.

Keywords: *NER, morphological stemmer, NER features.*

1. INTRODUCTION

NER is a subtask of information extraction, where we locate and classify proper names in text into predefined categories. NER has many applications in NLP Viz. Data classification, more accurate internet search engines, automatic indexing of documents, automatic question answering, cross language information access, and machine translation system etc,.[1], [4]. NER involves in identification of named entities such as Person name, Organization name, Locations, Designations, Measures, Time, Abbreviations and Brand. Construction of a Named Entity Recognition (NER) system becomes challenging if proper resources are not available. Gazetteer lists are often used for the development of NER systems [5], [16]. In many resource-poor languages like Telugu gazetteer lists of proper size are not available, but sometimes relevant lists are available in English. Proper transliteration makes the English lists useful in the NER tasks for such languages.

In Indian languages Telugu is a most popular language in southern part of India. Telugu language occupied 15th position in the world, 2nd position in India and 3rd most spoken language. Telugu language belongs to Dravidian family. Telugu is a highly inflectional and agglutinative language [14]. Telugu is primarily suffixing language, in which several suffixes added to the right. Telugu is a verb final language (in general) and word free order language. The major difficulties of NER in Telugu language is no capitalization, agglutinative nature i.e. each word in Telugu is inflected for a very large number of word forms, ambiguity i.e. Some words ambiguous with normal text with each other like Person name vs. Organization name, Person name vs. Place, Persona name vs. Common nouns, Appearance in various forms like prajaa raajyaM paarTii (praja rajyam party), pi.aar.pi (P.R.P) pra.raa.paa, pi aar pi, pra raa paa, piaarpi, praraapaa. In this paper, we have explored different features applicable for the Telugu NER task [2]. We have incorporated some gazetteer lists and suffix list in the system to increase the performance of the system. These lists are collected from the Telugu wikipedia and other websites. To make these English lists useful in the Telugu NER task, we have proposed a two-phase transliteration methodology. A considerable amount of improvement is observed after using the gazetteer lists in the system.

2. RULE BASED APPROACH

A rule based systems needs more grammatical and linguistic analysis to make rules. We observed that Rule Based approaches may give good result with sufficient gazetteers lists, language dependent features and rules for purely particular language. Named entities are open class words, every day new words added to languages and gazetteers list is infinite to store all words is not possible, hence gazetteers are needed to divide into finite tests like suffix, prefix, context words etc. All rule based approaches are language dependent. We cannot implement language independent NER system for Indian languages because language independent rules vary from one language to another [3].

2.1 Gazetteer Preparation

2.1.1. Gazetteers

Gazetteers preparation is an important role for identification of nouns. In resource rich languages like English, gazetteers lists are openly available in the web [13]. But resource less languages like Indian languages do not have such openly available gazetteers lists. We have to collect the lists from various web resources and these lists cannot be implemented directly in Indian languages like Telugu.

In order to measure the impact of using external resources in the NER task we have used a NER gazette which consists of three different gazetteers Person, Location, Organization, all are built manually using web resources. Person Gazetteer contains list of 30000 complete names, person beginnings, endings, and contexts of people found in Telugu Wikipedia, BSNL telephone directories and other websites. Location Gazetteer consists of 27,000 names of villages, mandals, cities, and districts in Andhra Pradesh found in the Telugu and collected from wikipedia and other websites. Organizations Gazetteer consists of a list of nearly 2000 names of companies, cricket teams, political party names and other organizations in India collected from business articles, and various web resources. We are containing nearly 40000 of words manually collected from CP Brown [6], JPL Gwynns and other online Telugu dictionaries with root forms and modified according to our purpose to eliminate closed class words [11].

2.1.2. Transliteration

Transliteration is an important and useful technique to process different languages and different approaches. The transliteration is based on the phonemes and spelling. The language of English is non phonetic in nature. This means that what we read from a printed text need not necessarily be the same as the printed matter. e.g., PSYCOLOGY keeping P silent. We find numerous such examples in the language of English. In case of phonetic languages like Telugu, Sanskrit, etc., what is written is exactly read out during reading activity, it is because there is no silent syllable in the written text in these languages. In fact most of the Indian languages are phonetic. To avoid the ambiguities arising from out of non phonetic languages like English, we deploy a technique called Transliteration. In the process of transliteration the meaning of text being translated from one language to another language is properly contained and is the same conveyed [7]. In effect in case of translation only the language changes but not the translated text. In case transliteration the transformation of words from one language to another language is done in serial fashion. In practice transliteration is easier than translation.

Transliteration is more relevant and desired feature while dealing with non-phonetic languages like English. The transliteration from English to Telugu is very difficult to build. This is necessitated because languages like Telugu depend more on long vowels, short vowels, stressed words and unstressed words. This is not the case in phonetic languages like Telugu because there is no need for these features and also the pronunciation is unique. It is also because phonetic languages have larger sets of base syllables or alphabet. Transliteration technique is very helpful for the preparation of gazetteers lists in Indian languages. We have developed a transliteration based gazetteers collected from various resources.

In our work we have collected employee English names and students databases from various organizations and colleges and we have converted into simple semi automatic transliteration scheme.

3. PROPOSED WORK

3.1. Noun Identification

It is useful to recognize nouns and eliminate non-nouns. The Telugu morphological analyzer developed here has been used to obtain the categories. Structure of Telugu nouns is root stem along with number marker along with case markers. A closed class word list including function words has been collected from existing dictionaries and closed class words are removed. Words with less than three characters are unlikely to be nouns and so eliminated. Last word of a sentence is usually a verb (Telugu is verb final language, in ever sentence final word may be a verb) and is also eliminated. Digits are eliminated. Telugu words normally end with a vowel and consonant ending words (Tiicar (teacher), sTeeshan (station), meeneejar (manager), byaak (bank) etc.) are usually nouns. Existing dictionaries are also checked for the category. Using these features, a naive Bayes classifier is built using the available tool WEKA.

3.2. NER Identification

Features of Telugu Language can be exploited for development of a good Named Entity Recognizer. Some features considered are as:

3.2.1. Suffix features

Every language uses some specific patterns which may act as ending words in proper names and the list of this type of words is called as suffix list [12]. Examples like s`arma, raaju, naayuDu, caudari, muurti etc., are person names' suffixes and vaaDa, paTnaM, puraM, palli, jilla etc., are location name suffixes and yuunivars`iTee, saMstha, aKaaDami, paarTii., etc are organization name suffixes. Sometimes these suffixes may act as ending words also. These are the few suffix clue words for identification of names of person, location and organization.

E.g.

- (person)

  Naagaraaju
  chaMdrabaabunaayuDu

- (location)

  vijayavaaDa
  maciliipaTnaM

- (organization)
  prajaaraajyaMpaarTii
  aaMdhrabyaaMk

3.2.2. Context Features

Every language uses some specific patterns which may act as clue words and the list of this type of words is called as Context Lists [15]. Such a list is collected after analyzing Telugu text. The words like s`ree, adhyakshuDu, misTar, DaakTar etc.,for identification of person names, similarly for identification of places graamaM, paTTaNaM, jillaa etc., and s`aaKha, peeThamu, saMsta ,etc., for identification of organizations. With identification of optimal window size of tokens we can yield good results. In our experiment we have taken window size of four.

E.g. (Person)

s`ree naagees`vararavu gaaru
DaakTar varaprasaad

(location)

anaMtapuraM jilla
puliveMdula graamaM

(organization)

aaMdraprades` rooDDu ravaaNaa saMsta
s`raamikavidyaa peeThamu

3.2.3 Morphological features

Indian languages are morphologically rich. Words are inflected in various forms depending on its number, tense, person, case, etc. Identification of root word is very difficult in Indian languages like Telugu [8], [10].

E.g. **Common noun**

aavulatoo	aavu	+	lu	+	too
(with cows)	root word.	+	number	+	case marker

Person

iMdiraadeevitoo	iMdiraadeevi	+	too
(with raamaaraavutoo)	root	+	case marker

Place

haidaraabaadunuMdi	haidaraabaadu	+	nuMdi
(from haidaraabaadu)	root	+	case marker

3.3. Algorithm for Noun identification

```
 Read file and divide into sentences.
Read sentences and divide into tokens
Read each token
For each (word) Loop
    Match with Telugu dictionary
    If direct match with dictionary then
        assign category
    else if no match with dictionary then
        apply noun Morphological suffixes
        if suffixes are found and root is found in
        Telugu dictionary then
            assign category
        else if suffix matches and root is not found
        then
            token may be noun
        else if token ending with consonant then
            the word may be loan word
            assign noun
        else
            assign the category "unknown"
End loop
```

3.4. Algorithm for NER identification

```
Read list of nouns identified by noun identification
Check gazetteers lists for NER features
For each (noun) Loop
    if suffix features found then
        Assign NER tag
    else if context features found then
        Assign NER tag
    else if found in NER dictionary then
        Assign NER tag
    else
        Assign " Miscellaneous word "
    If ambiguity is found then
    call disambiguation technique
        remove disambiguation
End loop
```

3.5. Functional diagram for NER identification.

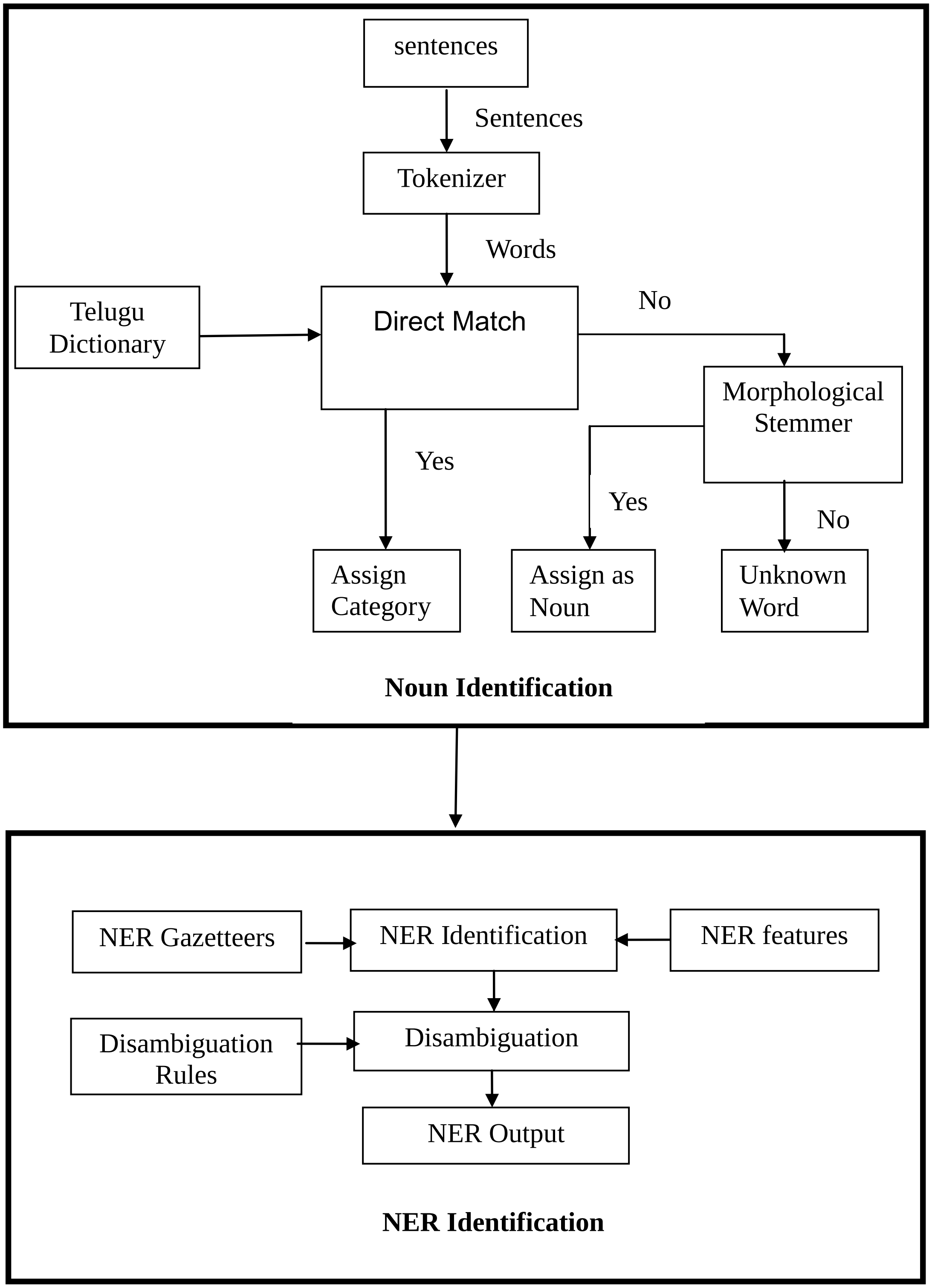

Fig 1: Functional diagram for NER identification.

4. TESTS AND RESULTS

We have collected different datasets on various domains collected manually from various web resources, EENAADU, VAARTHA, ANDHRA PRABHA NEWS PAPERS, TELUGU WIKIPEDIA and others [9].

In the First phase we have conducted various tests for noun identification and got very good performance by using dictionary gazetteers lists, morphological suffix mapping techniques and other features. A good number of nouns are identified in the first phase. These nouns may be common nouns or Named Entities or loan words. The identified nouns are are given as input to the second phase.

In the second phase we are checking each noun with gazetteers lists which contains beginnings, endings, contexts and suffixes of various tags. According to the category NE tags are assigned ambiguity is also resolved by using gazetteers lists and features. After conducting NER identification it is observed that good performance is achieved by the system.

Table1. Noun identification

	No. of . Words tested	**No.of nouns identified by the system**	**No.of nouns exactly identified by the system**	**Noun identification %**
Test 1	36850	17269	16181	93.7
Test 2	21626	12834	11165	87
Test 3	27878	16087	14639	91.1
Test 4	33923	15786	15201	96.3

Table2. NER identification

	No. of . Words tested	**No.of NERs identified by the system**	**No.of NERs exactly identified by the system**	**NER identification %**
Test 1	17269	16382	15624	95.37
Test 2	12834	11560	10591	91.61
Test 3	16087	15132	14281	94.37
Test 4	15786	14484	13634	94.13

5. CONCLUSION

Not much work has been done earlier in NER for Telugu. We discussed the various approaches available for NER including their positive and negative aspects. We also found that English NER features like capitalization cannot be used directly for Telugu Language. The noun identification technique to identify NERs by using morphological stemmer, suffix features, context features and gazetteers for identification of proper nouns giving good results. This method is to identify proper nouns correctly and to resolve the ambiguity using good language independent rules or any statistical approaches give very good performance. The Individual approaches also do not give good results for Telugu.

Resource less languages like Telugu can achieve maximum accuracy by using machine learning approaches like HMM, MEMM, CRF and SVM. More training data gives more accuracy. To create large training data manually very difficult process for resource fewer languages like Telugu. Our aim is to minimize manual effort and to create more training data using this noun identification approach. The test conducted on the selected data sets yield good results.

6. REFERENCES

[1] A. Borthwick.,"A Maximum Entropy Approach to Named Entity Recognition, Ph.D theis, New Yark University.

[2] 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.

[3] Asif Ekbal et. al. "Language Independent Named Entity Recognition in Indian Languages". IJCNLP, 2008

[4] Babych, B., Hartley, A.: Improving Machine Translation Quality with Automatic Named Entity Recognition. In: Proceedings of EAMT/EACL 2003 Workshop on MT and other Language Technology Tools. (2003) 1–8

[5] Bikel D. M., Miller S, Schwartz R and Weischedel R. 1997. Nymble: A high performance learning name-finder. In: Proceedings of the Fifth Conference on Applied Natural LanguageProcessing,pp.194

[6] Brown, C.P., The Grammar of the Telugu Language. 1991, New Delhi: Laurier Books Ltd.

[7] Daniel M. Bikel, R. Schwartz, Ralph M. Weischedel, "An Algorithm that Learns What's in Name", Machine Learning (Special Issue on NLP), 1999, pp. 1-20.

[8] Ganapathiraju, M., et al. OM: "One Tool for Many (Indian) Languages". in ICUDL: International Conference on Universal Digital Library. 2005. Hang Zhou

[9] Telugu Language website http://www.te.wikipedia.org/wiki/

[10] Morphological Analyzers – IIIT Hyderabad – http://www.iiit.net/ltrc/morph/morph_analyser.html

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

[12] 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.

[13] Nadeau, David; Turney, P.; Matwin, S. 2006. "Unsupervised Named Entity Recognition; Generating Gazetteers and Resolving Ambiguity" in the proceedings of Canadian Conference on Artificial Intelligence.

[14] 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.

[15] 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

[16] Wakao T., Gaizauskas R. and Wilks Y. 1996. Evaluation of an algorithm for the recognition and classification of proper names. In: Proceedings of COLING-96