Automatic Named Entity Identification and Classification using Heuristic Based Approach for Telugu

P. M. Yohan[1], B. Sasidhar[2], Sk. Althaf Hussain Basha[3], A. Govardhan[4]

[1]Associate Professor, Dept. of MCA, Wesley P.G. College, Secunderabad, Andhra Pradesh, India.

[2]Professor, Dept. of CSE, Mahaveer Institute of Science and Technology, Hyderabad, Andhra Pradesh, India.

[3]Professor, Dept. of School of Computing, GRIET, Hyderabad, Andhra Pradesh, India.

[4]Professor of CSE, SIT, JNTUH, Kukatpally, Hyderabad, Andhra Pradesh, India.

Abstract

Named Entity Recognition (NER) and Classification becomes more and more important in many natural language processing applications. It helps machine to recognize named entities in text and assign them with the appropriate categories. NER for Telugu is a challenging task since Telugu is very rich in morphology. Recent systems rely on machine learning approaches, but their performance is highly dependent on size and quality of training data. In this paper we proposed a rule based Named Entity Recognition and Classification system for Telugu language. In this paper we describe the identification and classification of Named Entities using word level features, work lookup features and contextual features. Further classification of identified Named Entities and ambiguity resolution is done through contextual rules and syntax information. The System is tested on different data sets of News paper and Teluguwiki corpus.

Keywords: *Heuristics, Named Entity, Gazetteers, Morphology.*

1. Introduction

The objective of Named Entity Reorganization (NER) is a classification problem. NER is to classify all tokens in a text document into predefined classes such as person, organization, location, miscellaneous. NER is a precursor to many natural language processing tasks. The creation of a subtask for NER in Message Understanding Conference [21], reflects the importance of NER in Information Extraction (IE). NER also finds application in question answering systems [2], [10], Information retrieval applications [28], [23], [21 and machine translation [6]. NER is an essential subtask in organizing and retrieving biomedical information [24]. NER can be treated as a two step process.

[1] Identification of proper nouns.
[2] Classification of these identified proper nouns.

Identification is concerned with marking the presence of a word/phrase as NE in the given sentences and classification is for denoting role of the identified NE.
NER systems have been developed for resource-rich languages like English with very high accuracies. But construction of an NER system for a resource-poor language like Telugu is very challenging due to unavailability of proper resources.
Telugu is a very popular language in the southern part of India and occupies the15th position in the world and 2nd position in India in terms of its popularity [29]. The language belongs to Dravidian family and is known to be highly inflectional and an agglutinative language. Every word in Telugu is inflected for a large number of word forms. Telugu is primarily referred to as suffixing language, where in several suffixes are appended to the right (Of any respective word). The language is also known to be a verb final language (in general) and a word free order language as well.

A few of the Various Named Entity classes identified in NER are

- Person Name
- Organization Name
- Location Name
- Designation
- Abbreviation
- Brand
- Title person
- Title object
- Number
- Measure
- Term
- Date and Time

2. Approaches on NER

There are diverse approaches used in NER system and to name a few, they are Rule based / Handcrafted Approach/Linguistic, Machine Learning / Automated / Statistical approach, and Hybrid Model.

2.1. Linguistic Approach

The NER system uses language based rules manually written by linguists and other heuristic to classify words. It needs rich and communicative rules and thereby gives produces effective results. The NER system requires an advanced knowledge of grammar and other language related rules and thus calls for a systematic knowledge and advanced skill arena related to the Language under consideration and they further need to come up with good rules and heuristic. There are several rule-based NER systems, containing mainly lexicalized grammar, gazetteer lists, and list of trigger words, which are capable of providing F-value of 88-92 for English [14], [19], [31]. The main disadvantages of these rule-based techniques are: they require huge experience and grammatical knowledge on the particular language or domain;

2.2 Machine Learning Based Approach / Stochastic Approach

In the recent years, when large number of data becomes electronically available, several methods of named entity recognition are proposed. Supervised learning methods require large amount of training data that has been annotated with correct class. Unsupervised learning methods assume that the correct classification of named entity training examples is not known and they classify the examples according to a similar metric.

2.2.1 Supervised learning

The current dominant technique for addressing the NERC problem is supervised learning. Supervised learning techniques include Hidden Markov Models (HMM) [5], Decision Trees [26], Maximum Entropy Models (ME) [7], Support Vector Machines (SVM) [3], and Conditional Random Fields (CRF) [18]. These are all variants of the SL approach that typically consist of a system that reads a large annotated corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features. A baseline SL method that is often proposed consists of tagging words of a test corpus when they are annotated as entities in the training corpus. The performance of the baseline system depends on the vocabulary transfer, which is the proportion of words, without repetitions, appearing in both training and testing corpus. D. Palmer and Day [22] calculated the vocabulary transfer on the MUC-6 training data. They report a transfer of 21%, with as much as 42% of location names being repeated but only 17% of organizations and 13% of person names. Vocabulary transfer is a good indicator of the recall (number of entities identified over the total number of entities) of the baseline system but is a pessimistic measure since some entities are frequently repeated in documents. A. Mikheev et al. [20] precisely calculated the recall of the baseline system on the MUC-7 corpus. They report a recall of 76% for locations, 49% for organizations and 26% for persons with precision ranging from 70% to 90%. Whitelaw and Patrick [32] report consistent results on MUC-7 for the aggregated enamex class. For the three enamex types together, the precision of recognition is 76% and the recall is 48%.

2.2.2 Unsupervised Learning

The typical approach in unsupervised learning is clustering. For example, one can try to gather named entities from clustered groups based on the similarity of context. There are other unsupervised methods too. Basically, the techniques rely on lexical resources (e.g. Word-Net), on lexical patterns and on statistics computed on a large un-annotated corpus. Here are some examples.
E. Alfonseca and Manandhar [1] study the problem of labeling an input word with an appropriate NE type. NE types are taken from Word-Net (e.g., location>country, animate>person, animate>animal, etc.). The approach is to assign a topic signature to each Word-Net synset by merely listing words that frequently co-occur with it in a large corpus. Then, given an input word in a given document, the word context (words appearing in a fixed-size window around the input word) is compared to type signatures and classified under the most similar one. In R. Evans [12], the method for identification of hyponyms/hypernyms described in the work of M. Hearst [15] is applied in order to identify potential hypernyms of sequences of capitalized words appearing in a document. For instance, when X is a capitalized sequence, the query “such as X”, is searched on the web and, in the retrieved documents, the noun that immediately precede the query can be chosen as the hypernym of X. Similarly, in P. Cimiano and Völker [9], Hearst patterns are used but this time, the feature consists of counting the number of occurrences of passages like: “city such as”, “organization such as”, etc. Y. Shinyama and Sekine [27] used an observation that named entities often appear synchronously in several news articles, whereas common nouns do not. They found a strong correlation between being a named entity and appearing punctually (in time) and simultaneously in multiple news sources. This technique allows identifying rare named entities in an unsupervised manner and can be useful in

combination with other NERC methods. In O. Etzioni et al. [11], Pointwise Mutual Information and Information Retrieval (PMI-IR) is used as a feature to assess that a named entity can be classified under a given type. PMI-IR, developed by P. Turney [30], measures the dependence between two expressions using web queries. A high PMI-IR means that expressions tend to co-occur. O. Etzioni et al. [17] create features for each candidate entity (e.g., London) and a large number of automatically generated discriminator phrases like "is a city", "nation of", etc.

2.3 Hybrid Model Approach

In this approach Rule Based approaches and Machine Learning approaches are mixed for more accuracy to identify NERs. Here several combinations are used.

- HMM approach and Rule Based approach.
- CRF approach and Rule Based approach.
- MEMM approach and Rule Based approach.
- SVM approach and Rule Based approach.

3. Proposed Method for Identification of Named Entity

Named entity recognition in Telugu is a difficult and challenging task due to highly inflectional and agglutinative nature of the language, scarcity of resources like gazetteers and labeled data. It lacks capitalization feature which plays a major role in identification of named entities in English.

3.1. Issues and challenges in Telugu NERC

- ***No Capitalization***

In Indian languages in particularly Telugu character concept is not there. There are called "akshara", and only one type.

- ***Agglutination***

Agglutinative and inflection nature of Telugu language. Each word is inflected for a very large number of word forms.

Example: nouns can take two types of numbers (singular and plural). After number it can take case and like suffixes. After case suffixes it can take vocatives and clitics, We can consider all suffixes in permutations and combinations, each noun root can containing above 10,000 word forms.

- ***Named Entity Ambiguities:***

In Indian languages each name uses particular purpose. Sometimes it gives different meaning or synonyms of particular word. For example Sun in Telugu different names are used like "Ravi", "Dhinakar", "Bhanudu"," Suryudu" etc.

a. Person name Vs Organization name:

"tata" as a person name as well as an organization, that creates ambiguity between person name and organization name.

b. Person name Vs Place name:

Ex: prakaashaM (Prakasham) Vs prakaashaM jillaa (Prakasham District)

raMgaareDDi (Rangareddy) Vs raMgaareDDI jillaa (Rangareddy District)
tirupati (Tirupati) Vs tirupati paTTaNaM (Tirupati town)

"Prakasham" and "Rangareddy" "Tirupati" both as a person name as well as location or place name, that create ambiguity between person name and place name.

c. Place Vs Organization:

Ex: aaMdra visvavidhyaalayaM (Andhra university) a place name or an organization.

vaijaag (Vizag) Vs vaijaag ukku karmaagaaraM (Vizag steel plant)

"andhra" and "Vizag" both as a place name as well as organization names, that create ambiguity between place name and organization name.

d. Person name Vs Common nouns:

Common noun sometimes occurs as a person name such as "geeta" which means line, thus creating ambiguities between common noun and proper noun.

- ***Lack of Standardization and spelling variations***

In Telugu, 20-40% daily life used words are loan words, means words barrowing form other languages like English, Hindi etc. Those words no proper writing conventions,
Example: English word bank in Telugu (byaaMku, baaMK, bhyaaMk). Same place names in English R. C. Puram in Telugu aarsiipuramu, aarsipuraM, aar.si.puraM. Some organization name in English T. D. P. in Telugu Ti.Di.Pi., TiDiPi, TiiDiiPii.

- ***Non availability of large gazetteer***

Like English named entity gazetteers are not available. Main problem is people are not used proper encoding techniques, large amount of data available in font encoding system. Other problem is machine readable text in English. Suppose name written in English "Koti" in Telugu "kooti", "koti'', "kooTi", "koTi" [17, 25] etc.. In English 26 characters are there, but in Telugu 56 characters are there [4, 8]. Mapping between English and Telugu is one-to-many problem. If you want prepare named entity gazetteers from English text we need proper transliteration program.

- **Lack of labeled data.**
- **Free word-order nature.**

3.2 NERC Architecture

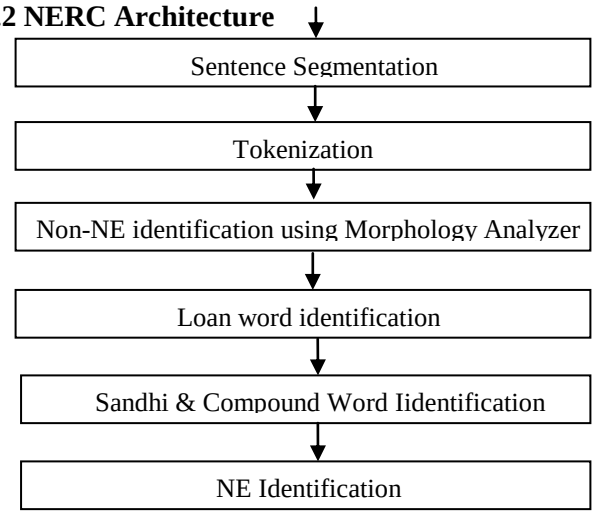

Fig 3.1: Functional diagram for NER

3.3 NE Identification

This NERC module has been developed to identify and classify named entities in given text. In this paper, we made an attempt to identify multi-word nested named entities. We choose to identify named entity at multi-token level because identification at token level may not help us in recognizing all named entities since many person names, location names and organization names are multi word named entities. These words taken together will form a named entity but individual words may or may not be named entities.

Example1: "aaMdhra visvavidhyaalaM" Andhra University
The two words taken together will form an organization name, but when looked at token-wise, first word appear as place name and second word appear as normal Telugu words. (Words which can be found in dictionary or common nouns)

Example 2:"baala tripura suMdari" bala tripura sundari
The above sentence first word "baala" "genitive form of boy or girl" and second word "Tripura" place name, third word "sundari" person name. Only when taken together they form a person name

Named entity features in Telugu:

1. Named Entities does not take plural forms unless it is preceded by another named entity of the same type.
2. Named Entities take case markers. But some case markers are never occurring.
3. Named Entities are not preceded by quantifiers like some, few, one, two etc.
4. Named Entities are hardly preceded by demonstrative adjectives like this and that.
5. Noun and noun, noun and verb compounding is natural. But Named Entity and verb compounds do exist.
6. Person entities hardly take locative case markers.
7. Organization entities do not include verbs or adverbs.

The above features helped us in recognition and classification of NEs and to eliminate non-NEs identified as NEs.

3.4 Gazetteers and Suffix list for Named Entity Identification

Non-named Entity Identification

This non-Named entity identification includes identifying common Telugu words which can be found in dictionary, Telugu words which undergone morphological inflections, numbers and alphanumeric words. This process goes in three stages. Different components constitute these filters. Each filter analyses the untagged word ignoring the tagged words.

- ***Dictionary***

Assumption 1: Words available in dictionary are not NEs or common Telugu words. Dictionary file contains headword and its part of speech. All the words of input text file that are matched with dictionary words are identified

and their respective tags are assigned. Dictionary file contains over 46,000 words. Dictionary words which are inflected with case markers are also recognized.

- ***Common Noun and Verb Suffixes***

List of common noun suffixes which do not generate NEs are prepared. If any word ends with any of the suffixes from list it is identified as common noun. These suffixes help in identifying common nouns. This list contains over 100 words.
List of verb suffixes were compiled from dictionary. If a word matches with this list or ends with any of words from dictionary it is identified as verb. These words help in identifying verb compounds.

- ***Symbols***

Words which are symbols i.e. special characters or combination of special characters are identified.
Ex: +, ->

- ***Morphological analyzer***

Morphology Analyzer analyzes[13] word internal structure. It is finding what type of suffixes is attached in the word form. It is useful for identification of stem and category of highly inflectional and derivational languages.

Assumption 2: Words that are analysed by morph are not NEs. Morphological analyser [16] developed at University of Hyderabad over the last many years has been used to obtain the root word and the POS category for the given word. Morphological analyzer helps us to recognize inflected forms which will not be available in dictionary.
Ex: For the word "puurtikaadu" its root word is identified as "puurti" and its tag is assigned as N words from morphological analyzer contain

- Named entities
- Loanwords.

These are the words borrowed from other languages and incorporated into Telugu language. Telugu borrowed many word from English language. Many regular use and domain specific words have been borrowed from English. Some of these words constitute named entities, such as kaMpenii (company), paarTii (party). And there are other loan words which are used by individuals like common Telugu words e.g. Teliphoon (telephone). And there are other domain specific words such as vairas (virus), eMDooplaasmik (endoplasmic).

Examples:

- kaaleeji => college
- aaksijan => oxygen
- kaMpenii => company
- DairekTar => director

- ***English Words***

Telugu-wiki documents contain English words as a part of text or to support the words which are not understandable and this supportive English text is often written in brackets.
Examples: cassettes, subjects

- prati (fascimile)
- vistiirNaM (area)
- kaMThasvaraM (Tone)

Assumption 3: Since they are not formed using Telugu aksharas, these words are not NEs.

External saMdhi or compound words

Words that contain external saMdhi are not handled by morphological analyser.

Ex: vaaraMdarikii, atanipeeru, viraamamerugaka
Nouns, verbs, noun compounds and verb compounds.

Few other nouns and compounds were also not analysed by morphological analyser.
Ex:satkaaraM,baalyamitruDu,sudhacellelu, bhiimaarjunulu

Stage-3:
In stage 3, we check for English word and orthographical information of word. If the word contains numbers then it not a Named entity.

- ***Numbers or alphanumeric word***

Numbers and alphanumeric characters cannot be a part of named entity. This information helps in identifying named entities at contextual level. All the words that contain numbers are identified and tagged. Words that contain numbers are checked for year pattern are identified and tagged as year. This information helps in identifying name entities. And all other words are tagged as numbers and alphanumeric words.

Our Named entity recognition and classification (NERC) system is built to identify person, location, organization entities. So any word that contains numbers cannot be part of Named entity.

- **NERC at Word Level:**

Untagged words from earlier non NE filers are now checked for named entities. Here, we tried to identify

person, location and organization names using NE gazetteer and NE suffix gazetteer and word level features such as punctuation marks, case markers to identify and classify named entities. We tried to handle Named entity variations that occur with the last akshara of word.

Ex: siitaaraaM, siitaaraam
maheesh. MaheeSh

We tried to identify abbreviations which may be or part of a person, location, organization entity.

Ex: Ti.Di.pi., en.Ti.raamaaraavu, es.aar.nagar. Person entities are identified by the postpositions which are particular to person case. Words ending in consonant are tagged as loanwords, which will be identified as NEs or as unknown words based on contextual information in the next steps.

SaMdhi or compound Filter:

SaMdhi is the fusion of sounds across word boundaries and the alternation of sounds due to neighboring sounds or due to grammatical function of adjacent words. SaMdhi can be both internal and external. Internal saMdhi features the alternations of sounds within words at morpheme boundaries. External saMdhi refers to changes found at word boundaries.

External saMdhi analyser has been developed as a part of spelling error detection module is used. It identifies saMdhi in a word and checks if a word contains external saMdhi based on saMdhi rules defined and checks if the saMdhi splits are valid or not by checking the resulting two words with dictionary and correct words analysed by morph. If the word split is valid then it is assigned a saMdhi tag and treated as common noun.

Examples: atanokkaDee (he is only)
"atanekkaDee" = "atanu+ okkaDee"

Assumption 4: The words that undergo external saMdhi are assumed to be common nouns.

External saMdhi analyser is placed after NERC word level module is because named entities also contain saMdhi. For example, consider "raamanna", this word can be split into two words "raamu" and "anna".

Proper Noun Identification using Morphological analyzer and Filters:

The assumptions (1) & (2) made in stage-1 and stage-2 are not necessarily true. Since sometimes common nouns and adjectives do act as proper nouns depending on context. And we have already tagged these words as common nouns. In this module we tried to identify the words that act both as common Telugu word and a proper noun. If a word is identified as NE then its previous 3 words and next 3 words that are marked as nouns or adjectives by Telugu dictionary or morphological analyser are checked for NE list and word features. If they act as NE corresponding tag is assigned. Common and proper noun disambiguation is done in the next module.

- **NERC at sentence level:**

In this phase, our main focus is on identifying person, location and organization entities and to identify multi-word named entities. We identified named entities using contextual information. NE patterns are identified from the corpus to recognize named entities. Rules are constructed based on these patterns. All the tagged and untagged words are considered here and rules make use of part of speech and NE tag information of previous and next words and word position. NEs recognized using list and word features are checked for their correctness using disambiguation and non-NE identification rules. (Here, non-NE identification means recognizing NE that is actually a Non-NE). Named entity features that were observed during earlier stages of this project helped in building disambiguation and non-NE identification rules. These rules make use of previous and next word POS information. The identified NEs are stored in another lexicon along with their category for future use. This will help if the next sentence contains same word then tag will assigned to it.
Disambiguation rules resolve the issue of NE ambiguity discussed in Telugu NERC issue. Non-NE identification rules can handle common and proper noun disambiguity to some extent.

Some common issue (6) is addressed here.

The above process is repeated for several iterations to extract patterns to identify more named entities. And we checked documents of different nature. This way we have developed a named entity database which includes 26,188 are person entities, 26,468 are location entities and 1600 are organization entities. NE suffix gazetteer consists of 3,369 person entity suffixes, 211 location entity suffixes, 409 organization entity suffixes and 8 person designation suffixes together constitutes 3,948 named entity suffixes. Tags used in NE gazetteer and suffix gazetteer and no. of NEs corresponding to each type are listed below.

Tags used in NE gazetteer:

TAG Description	Size
N-PER Person name	21,296
N-PER-SUF Person suffix	1034
N-PER-FN Person first name	3806
N-PER-Ti Person context	74
N-PER-DE Person designation	577
N-LOC-F Location name	26290
N-LOC-SUF Location suffix	211
N-LOC-PRF Location prefix	2
N-ORG-F Organization name	1172
N-ORG-SUF Organization suffix	412
N-ORG-PRF Organization prefix	18
N-OTH Miscellaneous names (Ex: iMDiyan)	202
N-DIR Directions	22

Table3.1: List of tag names used in named entity gazetteer

Tags used in NE suffix gazetteer:

TAG Description		Size
N-PER-S	Person suffix	*3362*
N-LOC-S	Location suffix	221
N-ORG-S	Organization suffix	409
N-PER_DES	Designation suffix	7

Table 3.2: List of tag names used in named entity suffix gazetteer

4. Tests and Results

We use two standard measures, Precision and Recall. Here precision (P) measures the number of correct NEs in the answer file (Machine tagged data) over the total number of NEs in the answer file and recall (R) measures the number of correct NEs in the answer file over the total number of NEs in the key file (gold standard). F-measure (F1) is the harmonic mean of precision and recall.

The current NERC system handles multi-tokens entities. Though we may not be able to identify accurate named entity boundaries, we are able to identify part of the named entity. Partial matches are also considered as correct in our analysis here. Nested entities are not fully focused here, it is just an attempt. We built nested entities for only person names. Our main focus here is to identify all people, location and organization entities present in the text.

The notation used is as follows:

NE – Named Entity,
PER – Person Entity,
LOC – Location Entity,
ORG – Organization Entity

#Files	NE	PER	LOC	ORG	#Words	%NEs in files
1	526	301	79	146	3689	14.25
2	483	149	198	136	9586	6.67
3	587	370	181	36	7978	13.48
4	502	318	146	38	14699	9.92

Table 4.1: NERC word level performance

Test File	# of NEs Identified by the System				# of NEs Correctly Identified by the System			
	NE	PER	LOC	ORG	NE	PER	LOC	ORG
F1	48	38	8	2	42	32	8	2
F2	235	71	119	45	218	57	115	38
F3	440	272	69	98	412	258	68	81
F4	518	329	161	25	481	304	148	19
F5	129	97	23	9	106	77	20	5
F6	305	173	100	30	243	133	90	13
F7	32	7	14	11	26	6	13	7

Table 4.2: NERC sentence level performance

5. Conclusions

In this paper we discussed various approaches available for NER and we proposed a platform independent rule based named entity recognizer and classifier system for Telugu. The system is built using newspaper corpus and Teluguwiki corpus. Different types of experiments are conducted and results are explained. Since Telugu language suffers from lack of labeled data, this system can be used to produce training data which can be used for machine learning techniques or hybrid techniques to produce much more reliable annotated data. We can say this system can work across all different domains with satisfactory results. We an attempt is made to identify multi-word named entity which helps in identifying the multi word organization names which is not possible with token level named entity recognition. This system may be executed on entire sentences database to get a named entity tagged data with precision in the range of 79%-94%.

References

[1] Silviu Cucerzan and David Yarowsky 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence.

[2] Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, Vishnu Vyas 2009. Web-Scale Distributional Similarity and Entity Set Expansion.

[3] Nancy Chinchor 1997. MUC-7 Named Entity Task definition. Technical report, 1997.

[4] Antonio Toral, Elisa Noguera, Fernando Llopis, and Rafael Mu.noz. 2005. Improving Question Answering Using Named Entity Recognition. In Proceedings of the 10th International Conference of Applications of Natural Language to Information Systems, pages 181-191, 2005.

[5] Diego Molla and Menno van Zaanen and Daniel Smith 2006. Named Entity Recognition for Question Answering. Australasian Language Technology Workshop 2006.

[6] Bogdan Babych & Anthony Hartley. 2003. Improving Machine Translation Quality with Automatic Named Entity Recognition. 7th EAMT Workshop, 13th April 2003, Budapest, Hungary; pp. 1-8.

[7] Richard Tzong-Han Tsai 2006. A hybrid approach to biomedical named entity recognition and semantic role labeling. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: doctoral consortium. pages 243-246

[8] Telugu language website http:// www. te.wikipedia.org/wiki

[9] Grishman R. 1995. Where's the syntax? The New York UniversityMUC-6 System. In: Proceedings of the Sixth Message Understanding Conference.

[10] 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.

[11] 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*

[12] Bikel, Daniel M.; Miller, S.; Schwartz, R.; Weischedel, R. 1997. Nymble: a High-Performance Learning Name-finder. In Proc. Conference on Applied Natural Language Processing.

[13] Sekine, Satoshi. 1998. Nyu: Description of the Japanese NE System Used For Met-2. In Proc. Message Understanding Conference.

[14] Borthwick, Andrew; Sterling, J.; Agichtein, E.; Grishman, R. 1998. NYU: Description of the MENE Named Entity System as used in MUC-7. In Proc. Seventh Message Understanding Conference.

[15] Asahara, Masayuki; Matsumoto, Y. 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis. In Proc. Human Language Technology conference – North American chapter of the Association for Computational Linguistics.

[16] McCallum, Andrew; Li, W. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons. In Proc. Conference on Computational Natural Language Learning.

[17] Palmer, David D.; Day, D. S. 1997. A Statistical Profile of the Named Entity Task. In Proc. ACL Conference for Applied Natural Language Processing

[18] Mikheev, Andrei. 1999. A Knowledge-free Method for Capitalized Word Disambiguation. In Proc. Conference of Association for Computational Linguistics.

[19] Whitelaw, Casey; Patrick, J. 2003. Evaluating Corpora for Named Entity Recognition Using Character-Level Features. In Proc. Australian Conference on Artificial Intelligence.

[20] Alfonseca, Enrique; Manandhar, S. 2002. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In Proc. International Conference on General WordNet.

[21] Evans, Richard. 2003. A Framework for Named Entity Recognition in the Open Domain. In Proc. Recent Advances in Natural Language Processing.

[22] Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proc. International Conference on Computational Linguistics.

[23] Cimiano, Philipp; Völker, J. 2005. Towards Large-Scale, Open-Domain and Ontology Based Named Entity Classification. In Proc. Conference on Recent Advances in Natural Language Processing.

[24] Shinyama, Yusuke; Sekine, S. 2004. Named Entity Discovery Using Comparable News Articles. In Proc. International Conference on Computational Linguistics.

[25] Etzioni, Oren; Cafarella, M.; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; Yates, A. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study.

[26] Turney, Peter. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Proc. European Conference on Machine Learning.

[27] Kavi Narayana Murthy and Srinivasu Badugu, Roman Transliteration of Indic Scripts, 10th International Conference on Computer Applications, 28-29, February,2012, University of Computer Studies, Yangon, Myanmar

[28] Sasidhar, B., Govardhan, A., Vinaya Babu, A.2011. "Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach" in International Journal of Computer Applications, ISSN-0975-8887,Vol.22, No.8, May 2011. pp. 30-34.

[29] BH. Krishnamurthi and J.P.L. Gwynn. A Grammar of Modern Telugu". Oxford University Press, New Delhi, 1985.

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

[31] Uma Maheshwar Rao G., Amba Kulkarni. P., and Christopher Mala "A TELUGU MORPHOLOGICAL ANALYZER ".Center for Applied Linguistics and Translation Studies University of Hyderabad, Hyderabad, India.