International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 7 Issue 11, November-2017, pp. 13~21
ISSN (O): 2249-3905, ISSN(P): 2349-6525 | Impact Factor: 7.196

Named Entity Identification and Classification in Telugu Using Knowledge Base Approach

G. V. Subba Raju[1] ,K.Venkata Rao[2]

Department of Computer Science and Systems Engineering
Andhra University,
A.P,India

Abstract—Telugu textual information becomes available more and more through the Web in homes and businesses, via Internet and Intranet services, there is an urgent need for technologies and tools to process the relevant information. Named Entity Recognition (NER) is an Information Extraction task that has become an integral part of many other Natural Language Processing (NLP) tasks, such as Machine Translation and Information Retrieval Part of speech tagging. The characteristics and peculiarities of Telugu, a member of the agglutinating and suffix orientation languages family, make dealing with NER a challenge. In this paper we proposed a knowledge base approach for named Entity identification and Classification for Telugu language. The paper is divided as follows. First, we present introduction about Named entity in section I and II. Then we present simple literature survey in section III. Next we present a system architecture and module description in section IV. Next we showed our system performance in section V.

Keywords—*NE, Rule-based, gazetteers, Telugu*

INTRODUCTION

The term "Named Entity" [1]. (which might also be called as proper name) now widely used in Natural Language Processing. The Named Entity information in a document is crucial for many language processing tasks. Identifying references to named entities in text was recognized as one of the important sub-tasks of information extraction and was called as "Named Entity Identification and Classification (NEIC)". Named Entity Recognition (NER)(which might also be called as proper name classification) is a computational linguistic task in which we seek to classify every word in a document into some predefined categories like person name, location name, and organization name, miscellaneous name (date, time, percentage and monetary expressions) and "none-of-the above".

Named Entity Recognition, is much simpler than either of the tasks described above and it is a necessary precursor to them. Clearly, before we can determine the relationship between **BJP** andAmith Shah , we must first properly categorize them respectively as an organization and a person. Similarly **Sriharikota**must first be identified as a location before we can identify it as a Satellite Launching Station of Indian Space Research Organization site.

Ambiguity is the major challenge in Telugu named entities. Many words can appear as Named Entities of correct category, sometimes the same words may appear as Incorrect category and sometimes as common noun dependent on various contexts [2]. Therefore identification of the correct category is very difficult. The correct category depends on the context. The identification and classification of names often involve challenging ambiguity. One of the

properties of the named entities in these languages is the high overlap between common names and proper names. In cases where an entity can have two valid tags, the more appropriate one is to be used. The annotator has to make the decision in such cases.

ABOUT TELUGU NAMED ENTIY(NE)

I. Characteristics of NE

- Named entities are not typically available in general purpose lexical resources.
- Named entities, generic terms and PRO terms are used interchangeably and form chains of co-referring items. (E.g. Tony Blair visited ..., The Prime Minister emphasized...)
- The surface of named entity can vary. (e.g. Narendra Damodardas Modi, Narendra Modi, Modi, etc)
- Ambiguity of named entity types:
    1. For example, Nagarjuna(నాగార్జున), Satyam(సత్యం) may be a person name or a company name.
    2. When the word Tirupati(తిరుపతి) or Annavaram(అన్నవరం) being used as a name of person and as name of city?
    3. When the word kavita(కవిత) or giita (గీత) being used as a name of person and as common nouns?
    4. Swathi/rohini(స్వాతి/ రోహిణి) (person name Vs birth star name)

Proper nouns are identified in singular forms and take an altogether statistical distribution of 'case marker' incidence having no effect of modifiers viz. demonstratives quantifying and qualifying adjectives, possessive nouns and pronouns, relative participles. These modifiers behave as such in limited situations but do not have regular features.

We have used part of news articles from Telugu local dailies and Telugu wikipedia for all our experiments. For all of our experiments we are using the roman translitration form of these articles. Ambiguity with common words for example ``raajuరాజు(king) and raaNiరాణి(queen)'' can be a person name as well as a common word.

II. Telugu Language and It's Complexity

Telugu, a language of Dravidian family, is spoken mainly in southern part of India and ranks third among Indian languages in terms of number of speakers. Telugu is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and statistical features. Telugu is one of the languages which is characterized by a rich system of inflectional morphology and a productive system of derivation, saMdhi and compounding[3].

RELATED WORK ON NE IN INDIAN LANGUAGES

NLP research around the world has taken giant leaps in the last decade with the advent of efficient machine learning algorithms and the creation of large annotated corpora for various languages. However NLP research in India started with the development of rule based systems due to lack of annotated corpora. Statistical NLP research can only be given a push by the creation of annotated corpus for Indian languages.

There is not much work done on named entity recognition in Telugu language, compared to English. Ideas and features that are used for English cannot be borrowed directly to Telugu language. For example, capitalization feature is not there in Telugu language. Dravidian languages in case of Telugu are rich in morphology. In Indian languages Hindi has simplest morphology, as we go from north to south the complexity increases. Dravidian languages have complex morphology. Indian names are more diverse

in nature, i.e. there are a lot of variations for a given named entity. For example "telugudeeshaMpaaThii" is written as Ti. Di. pi, TiDipi, tedeepaa, te. Dee. paa, etc. And inflections are also added to named entities.

Bengali is the seventh popular language in the world, second in India and the national language of Bangladesh. In the year 2009 [4], the development of NER in the language was reported by Ekbal and Bandyopadhyay. He tried to get through the task successfully by combining the output of the classifier like ME, CRF and SVM respectively. The training set comprises of 150k word formation for tracing the 4 NE Tags Viz. person, location, organization and miscellaneous objects. In order to enhance the performance of the classifier, about three million word forms were used, extracted from lexical context pattern generated from an un-labelled Bengali corpus. Evaluation results of 30k word forms have shown that altogether, precision and f-score values as 87.11%, 83.61% and 85.32%. This indicates an improvement of 4.66% in f-score over the best performing SVM based system and 95% in f-score over the ME based system.

A report on the development of Bengali news corpus from the web comprising of 34 million word forms was propounded by Ekbal and Bandyopadhyay in 2008 [5]. Part of it, about 150k word forms, is manually tagged with 16 NE and with one non-NE tag, besides, 30k word forms are tagged up with a tag set of 12 NE tags explained and defined for the IJCNLP–08 NER shared task for SSEAL. A change in tag (conversion) routine has been fared to convert the 16 NE tagged corpus of 150k word forms to the corpus tagged with IJCNLP– 08, 12 NE tags where the former has been used to develop the Bengali NER system suing HMM ME, CRF, SVM respectively. The output (evaluation) results of 10 fold cross validation experiments give the F-score of 84.5% for HMM, 87.4% for ME and 90.7% for CRF and 91.8% for SVM.

Ekbal and Bandhopadhyay in 2008 [5] reported on the development of NER system in Bengali combining the outputs of the classifier like ME, CRF, and SVM. The corpus consisting of 250k word forms is manually tagged with four NEs namely person, location, organization, and miscellaneous. The system makes use of different contextual information of words along with a variety of features that help in identifying the NES experimental results and indicates the effectiveness of the proposed approach with overall average recall, precision and f-score values of 90.78%, 87.35% and 89.03% respectively. This shows an improvement of 11.8% in f-score over the best performing SVM based baseline system and an improvement of 15.116 in f-score over the least performing ME based system.

In the year 2008 [6]Vijayakrishna and Sobha L brought out "Domain Focused–Named Entity Recognition for Tamil using conditional Random fields", developed a model titled "Domain focused NE Recognizer for tourism Domain conditional Random Fields Approach on Tamil language". They used 106 tag sets for tourism domain and five feature templates. About Ninety four thousand words corpus was collected in Tamil for this domain. NE annotations NP Chunking, POS tagging, Morph analysis are presented as to their performance manually on the corpus. It comprised of roughly 20,000 titled entities divided into two sets. Whereas the fore most formed the training data while the other the test data, constituting 80% and 20% of the total data respectively. A total of 4059 entities where taken on testing for experiment and got overall F–measure 80.44%.

Development of Hindi NER using ME approach was elucidated by Saha et al. (2008) [7, 8]. About 234 k words were stated to have comprised as training data, collected from the news papers "DainikJagaran" which were manually tagged with 17 classes, with 16,482 NEs.

The development of a module was also reported in the paper about the semi–automatic learning of context pattern, using a blind test corpus of 25k the

system was evaluated as having 4 classes and achieved an F-measure of 81.52%.

A detailed observation was made out by Gupta and Arova in 2009 [9] and the experiment conducted on CRF models for developing Hindi NER. It indicates some features making the development of NER system more complex. It narrates the different approaches for NER. The information used for the training of the model was taken from tourism domain which is manually tagged in 10B format.

Using the SVM system, in the year 2008 [5], Ekbal and Bandyopadhyay developed NER system for Bengali.

Further to the usage of appropriate unlabelled data in 2009, [5] Ekbal and Bandyopadyay briefed about a voted NER system. This above procedure locates the basis in supervised classifier, namely ME, SVM, CRF where SVM makes use of two different systems known as forward parsing and backward parsing. It was tested for Bengali comprising 35,143 news document and 10 million word forms and make use of language independent features along with different contextual information of the words. A the end, the models were combined into an ultimate system with an arranged voting technique and the test results extended the effectiveness of the proposed approach with the recall precision and f–score values of 93.81%,92.18% and 92.98% respectively.

A language independent NER in Indian languages [4] was developed by Asif Ekbal in 2008, using the statistical Conditional Random Fields (CRF).

The system utilized variety of contextual information of the words along with different features that was supportive in forecasting (predicting) the various NE classes in both the language dependent and language independent areas.

The latter was applied to Hindi, Bangali Oriya Telugu and Urdu and language dependent features were applied to only Bengali and Hindi. The system was experimented with Bengali. (1,22,467 tokens), Hindi (5,02,974 tokens) Telugu (64,026 tokens), Oriya (93,173 tokens) and Urdu (35,447 tokens) and tested with Bengali (30,505 tokens), Hindi (38708 tokens), Telugu (6, 356 tokens), Oriya (24,640 tokens) and Urdu (3,782 tokens), and found the maximal F-measure of 53.46% for Bengali whereas for Telugu F-measure was found as a very performer.

PROPOSED APPROACH

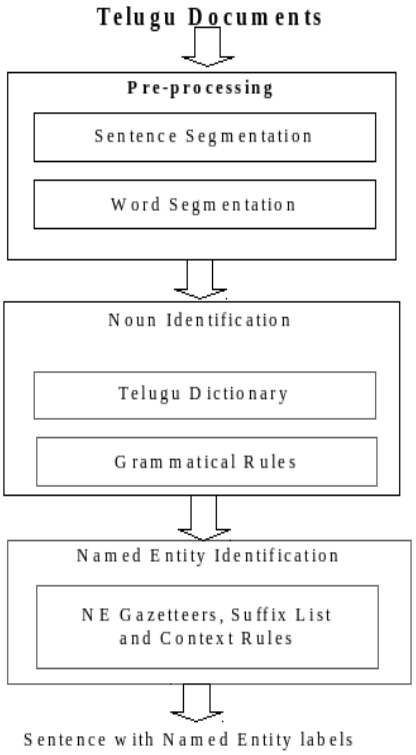

Fig 1. NEIC system block diagram

Fig. 1 shows data flow in Named Entity identification system. In this paper we proposed a knowledge base or rule based approach for named entity identification and classification processes. It is a two step process. First step is identification of common nouns and second step identification of named entity identification. In this system has three main modules. First module is preprocessing, second module is noun identification module and final

module is named entity identification and classification.

Steps for pre-processing:

1. Open file (document) and dividing it into sentences (sentence segmentation using knowledge base).
2. Each sentence, split it into tokens.(Tokenization or word segmentation)

Steps for Noun identification

1. First check each token in Telugu dictionary. If the word is found in a Telugu dictionary, then assign category of a word (token).
2. If it is not found, then apply all Telugu grammatical suffix features for identification of nouns. If the word is matching with any one of the suffix, then assign the appropriate suffix category.
3. if token not found in step1 and step2, then check token(word) position, if token is end of the sentence and it's length has more than 3 akshara's (syllables) then it is marked as Verb. (Telugu is verb-final language in general. It is generally observed that most of the sentences end with (90%) verbs.
4. If it is not found in any one then assign unknown category for that token.

Steps for Named Entity Identification:

1. Select noun and unknown words.
2. For noun and unknown tokens are checked in the Telugu gazetteers list (Person, Location, and Organization). If it is found then assign appropriate category.
3. If it is not found, then apply all named entity suffix and named entity context features for identification. If the word is matching with any one of the feature, then assign the appropriate suffix category.

In this processes we need Telugu Dictionary, gazetteers and suffix list , grammatical features context features. We collected Telugu dictionary from online resources. Next section we will discuss how to prepare gazettes list using raw corpus.

Gazetteers Preparation

Gazetteers, or named entity dictionaries, are important for performing named entity recognition (NER) accurately in knowledge base system. Since building and maintaining high-quality gazetteers by hand is very expensive, many methods have been proposed for automatic extraction of gazetteers from text[10].

In Telugu we have prepared very large raw corpus (20000K) from different sources like Telugu online new papers and Wikipedia. After that we apply suffix and context pattern methods for gazetteers extraction from corpus.

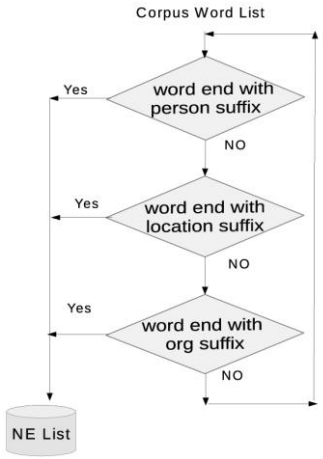

Fig. 2 A flow diagram for gazetteers preparation

This method seeks some high precision suffix and context patterns by using some seed entities and hits the patterns to the raw corpus to prepare the named entity gazetteers.

Grammatical suffix

In this processes we use language dependent grammatical features (it is available in Telugu

grammar books) like noun suffixes and verb suffixes [11, 12] etc.. .

Context Features Description

The features that we have identified for the Telugu NEIC task are:

Surrounding Words

As the surrounding words are very important to recognize a NE, previous and next words of a particular word are used as features. During experiment different combinations of previous two to four words to next two to for words are used as features. These features are multi valued. For a particular word wi, its previous word wi-1 can be any word in the vocabulary, which makes the feature space very high. For example aaMdhravisvavidhyaalayaM(andhra university(AU) or aadikavinannayyavisvavidhyaalayaM(adikavinannaya University).

Named Entity Context Lists

Context words are defined as the frequent words present in a word window for a particular class. In our experiment we have listed all the frequent words present anywhere in wi-2 … wi+2 window for a particular class. Then this list is manually edited to prepare the context word list for a class. For example, location context list contains rooD (road), nagaraM (nagar) raajdhaani (capital), daggara (located in) etc. The feature is defined as, for a word wi, if any of its surrounding words (wi-2 … wi+2) is in a class context list then the corresponding class context feature is 1.

Named Entity Tags of Previous Words

Named entity (NE) tags of the previous words (ti-m...ti-1) are used as feature. This feature is dynamic. The value of the feature for wi is available after obtaining the NE tag of wi-1.

Final akshara or syllable

If the word end with consonant that present word is not a Telugu word. All Telugu words are end with vowel, It is a named word or other language word, then this feature is set to 1. Otherwise, it is set to 0.

Numerical Word

If a word is a numerical word, i.e. it is a word denoting a number (e.g. muuDu (three), pradhama (first) etc.) then the feature NumWord is set to 1.

Word Suffix

Suffix information is useful to identify the named entities. This feature can be used in two ways. The first and naive one is that a fixed length word suffix of current and surrounding words can be treated as feature. During evaluation, it was observed that this feature is useful and able to increase the accuracy by a considerable amount. Still, better approach is to use suffix based binary feature. Variable length suffixes of a word can be matched with predefined lists of useful suffixes for different classes of NEs. Suffix list of locations is very useful since most of the location names in India end with a specific list of suffixes. Suffix list of locations endings contains 116 suffixes like, peeTa, puraM, puur, nagaraM, palli, vaaDa etc.

Word Prefix

Prefix information of a word is also useful. A fixed length word prefix of current and surrounding words can be treated as feature.

Using above features we extracted gazetteers from raw corpus.

Table 1 LIST OF CONTEXTUAL RULES

Features	Example
person names	raavu, raaju, naayuDu, muurti,
person names	
person names	Srii, sriimati,adhyakshuDu
location name	vaaDa, paTnaM, puraM,
location name	Jilla, maMDalam
Grammatical	loo, looki,loona,paina,paiki,

The overall methodology of identification of named entity from test corpus is summarized as follows:

Step1: Finding the test corpus. It must be machine readable format.

Step2: Do preprocessing on test corpus

Step3: Apply noun classification on test corpus. This classification techniquedivides each word in a sentence either noun or not noun.

Step4: Extraction named entities from nouns using seed patterns and identified features.

Step5: Validation was done by extracted named entities randomly.

Table 2 TAG Description of Named Entity suffixes

TAG	Description	# suffix
PER-SUF	Person affix	190
PER-BEG	Person -starting suffix	1,714
PER-CON	Person context sufffix	90
PER-END	Person endingsufiix	301
LOC-SUF	Location affix	120
LOC-BEG	Location beginning suffix	1,347
LOC-END	Location ending suffix	116
ORG-SUF	Organization affix	27
ORG-CON	Org contextsuffix	35
N-PP	Grammatical suffix	400

RESULTS and DISCUSSIONS

In this paper, system performance tested on Telugu new paper corpus files. The tables below show preliminary named entity identification and classification results based on Telugu dictionary, named entity gazetteers, suffix and context feature and finally we use context based dis-ambiguity rules.

After gathering all useful requirements, we tested our system using newspaper corpus. We demonstrate few expermental result bellow.

Table 3 Named Entity Identification in word level

No. of words tested	No. of NE's Identified by manually				No. of NE's Identified by our system			
	NE	PER	LOC	ORG	NE	PER	LOC	ORG
457	30	04	07	19	29	03	07	19
417	39	28	11	0	37	26	11	0
489	36	16	19	01	35	15	19	01
426	17	08	08	01	15	07	08	00
1181	45	20	19	07	41	18	16	07
1224	77	69	07	01	77	69	07	01

Table 4 Named Entity Identification in file level

	#words	Actual #NE's in file	#NE's Identified	%Identification
News-1	944	77	66	85.7
News-2	426	17	15	88.3

News-3	6809	150	138	92.0
News-4	10059	620	587	94.7
News-5	62159	4986	4986	100

Performance Metrics

Precision (P): Precision is the fraction of the documents retrieved that are relevant to the user's information need.

Precision (P) = correct answers/answers produced

Recall (R): Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

Recall (R) = correct answers/total possible correct answers

F-measure: F-measure is the weighted harmonic mean of precision and recall.

F-measure (F) = (β2+1) PR / (β2R+P)

β2 is the weighting between precision and recall. The typical value of β2 is 1. When recall and precision are evenly weighted i.e. β2 = 1, F-measure is called F1 measure.

F1-measure (F1) = 2PR / (P + R)

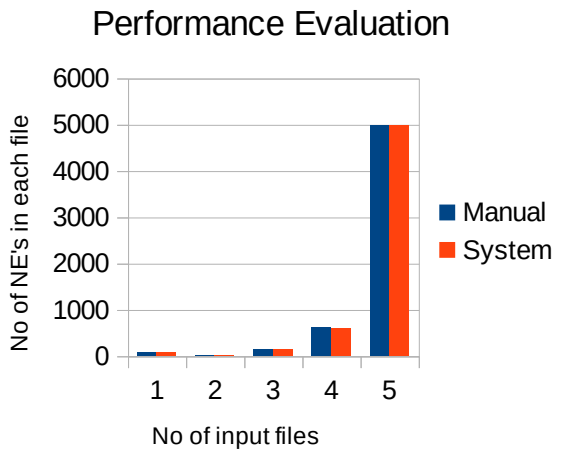

Fig 3. system performance evaluation

CONCLUSION

we observed that rule based approaches may give satisfactory results with sufficient gazetteers list and language dependent rules. Language dependent rules are specific for each language. Named entities are open class words, every day new words added to languages and gazetteers list is long. To store all words in gazetteers is a practically difficult. In this paper we proposed new approach for NEIC. According to this approach, gazetteers are needed to divide into finite lists like suffix and context words etc., All Rule based approaches are language dependent. We intend to implement language independent NER system for Telugu languages. Our main aim is to minimize manual effort, with less resource, obtaining good result. In the future work we have to collect more suffix and context features and try to improve the system performance and use this system out as train data for machine learning system.

REFERENCES

[1] Oudah, Mai, and Khaled Shaalan. "Person name recognition using the hybrid approach." *18th International Conference on Applications of Natural Language to Information Systems*, Springer Berlin Heidelberg, 2013. 237-248.

[2] Florian, R., luycheriah, A., Jing, H., Zhang, T. 2003. Named Entity Recognition through Classifier Combination. *In Proceedings of the International Conference on Natural Language Learning* (CoNLL-2003), Edmonton, Canada ,168-171.

[3] Murthy, Kavi Narayana, and G. Bharadwaja Kumar. "*Language identification from small text samples*." *Journal of Quantitative Linguistics* 13, no. 01 (2006): 57-80.

[4] Ekbal, A. and Bandyopadhyay, S. 2008. "Named Entity Recognition using Support Vector Machine: A Language Independent Approach",

International Journal of Computer, Systems Sciences and Engg.(IJCSSE), vol. 4, pp. 155–170.

[5] Asif Ekbal, and Bandyopadhyay, S. 2008. "Bengali Named Entity Recognition using Support Vector Machine". In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, January. pp. 51–58.

[6] Krishna. V. R., and Sobha. L. 2008. "Domain focused Named Entity Recognizer for Tamil using Conditional Random Fields". In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, Hyderabad, India, pp. 59-66.

[7] Saha, S. K., Sarkar, S., and Mitra, P. January 2008 "A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition". In Proceedings of the 3rd International Joint Conference on NLP, Hyderabad, India, pp. 343–349.

[8] SujanKumar Saha, Sanjay Chatterji, SandipanDantapat, Sudeshna Sarkar and PabitraMitra 2008. "A Hybrid Approach for Named Entity Recognition in Indian Languages". Proceedings of the IJNLP-08 workshop on NER for South and South East Asian Languages Hyderabad, India

[9] Gupta, P. K., and Arora S. 2009. "An Approach for Named Entity Recognition System for Hindi: An Experimental Study". In Proceedings of ASCNT-2009, CDAC, Noida, India, pp. 103–108.

[10] Kazama, J.I. and Torisawa, K., 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. *Proceedings of ACL-08: HLT*, pp.407-415.

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

[12] Brown,C.P.TheGrammaroftheTeluguLanguage.1991,NewDelhi:LaurierBooksLtd.