SURVEY ON NAMED ENTITY RECOGNITION FOR TELUGU

G. V. S Raju
Dept of Computer Science & Systems Engineering, Andhra University, India.
letter2raju@gmail.com

Prof.M.S.Prasad Babu
Dept of Computer Science & Systems Engineering, Andhra University India,

Prof.K.Venkata Rao
Dept of Computer Science & Systems Engineering, Andhra University, India

Abstract: *In this paper we present here a survey of various approaches used to recognize name entity in Telugu Language. First, we give a brief introduction on NER then we present various approaches used to recognize the NE in Telugu Language. Telugu language is resource poor language and agglutinative in nature, morphologically very complex. Available named entity gazetteers are insufficient.*

Key words: *Named Entity, Named Entity Recognition, Telugu, morphology, agglutinative*.

I. INTRODUCTION

Named Entity Recognition (NER) is one of the major tasks in Natural Language Processing (NLP). NER is an active field of research for past twenty years. NLP is a component of Artificial Intelligence (AI). A lot of progress has been made in detecting named entities, but NER still remains a big problem at large. NER involves the identification of named entities such as person names, location names, names of organizations, monetary expressions, dates, numerical expressions, etc. In the taxonomy of Computational Linguistics, NER falls within the category of Information Extraction which deals with the extraction of specific information from given documents. NER emerged as one of the sub-tasks of the DARPA-sponsored Message Understanding Conference (MUCs). The task has important significance in the Internet search engines and is an important task in many of the Language Engineering applications such as Machine Translation, Question-Answering systems, indexing for Information Retrieval and Automatic Summarization.

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 labelled data. It lacks capitalization feature which plays a major role in identification of named entities in English.

Telugu is most popular language in southern part of India. Telugu language occupied 15th position in the world and 2nd position in India. Telugu belongs to the Dravidian family of
languages. Telugu is a highly inflectional and agglutinating language providing one of the richestand most challenging sets of linguistic and statistical features resulting in long and complex word forms (G. Bharadwaja kumar et al, 2007) [19]. Each word in Telugu is inflected for a very large number of word forms. Telugu is primarily suffixing language, in which several suffixes added to the right. Telugu is a verb final language (in general) and free word-order language (Krishnamurthy and Gwynn, 1985) [20].

Issues and challenges in Telugu NERC:

1. No capitalization.
2. Agglutinative nature of the Telugu language. Each word is inflected for a very large number of word forms.
3. NE Ambiguities:
    1. Person name Vs Organization name:

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

2. Person name Vs Place name:

Ex: prakaashaM (Prakasham) Vs prakaashaM jillaa (Prakasham District)
raMgaareDDi (Rangareddy) Vs raMgaareDDI jillaa (Rangareddy District)

3. Place Vs Organization

Ex: usmaaniyaa yuunivarsiTy (Osmania

university) a location or an organization.

4. Person name Vs Common nouns:

Common noun sometimes occurs as a person name such as "suurya" (Surya) which means sun, thus creating ambiguities between common noun and proper noun.

4. Lack of Standardization and spelling variations:

Ex: aarsiipuramu, aarsipuraM, aar.si.puraM, Ti.Di.Pi., TiDiPi, tedeepaa, te.dee.paa

5. Non availability of large gazetteer.
6. Lack of labelled data.
7. Free word-order nature.

II. AII PPROACHES TO NER

There are several approaches to NER. They can be categorized into two broad classes:

A. Rule based (Linguistic) approaches: Rule based approaches rely on handcrafted rules which contain gazetteer lists, list of triggered words etc, which focuses on extracting names.

B. Machine learning (Statistical) approaches: Machine learning approaches rely on statistical models to make predictions about NEs in the given text. Large amounts of annotated training data are required .There are three main machine learning approaches: Unsupervised, Supervised and Semi-supervised.

Commonly used supervised statistical approaches are:

1. Hidden Markov Models: Hidden Markov model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states. Here, the state is not directly visible, but output, which depends on the state, is visible. HMMs have difficulty modeling overlapping, non-independent features. Conditional Random Fields (CRF) solves this problem.

2. Conditional Random Fields: These are undirected graphical models used to calculate the conditional probability of values on designated output nodes, given values assigned to other designated input nodes. They are conditionally trained probabilistic finite state automata. Since they are conditionally trained, these CRFs can easily incorporate a large number of arbitrary, non-independent features, while still having efficient procedures for non-greedy finite-state inference and training.

3. Maximum Entropy model: Maximum Entropy models are conditional probabilistic sequence models. The maximum entropy framework estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. Such constraints are derived from training data, expressing some relationship between the features and outcomes. The probability distribution that satisfies this property is the one with the highest entropy. 4) Support Vector Machines: SVM solves two-class pattern recognition problem. It takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. It gives best results when the data set is small, and with extended algorithms, it can be used in multi-class problems.

5. Decision Tree: Decision Tree is a classifier in the form of a tree structure where each node is either a leaf node-indicates the value of the target attributes (class) of expressions, or a decision node that specifies some text to be carried out on a single attribute value with one branch and sub-tree for each possible outcome of the text. It is an inductive approach to acquire knowledge on classification.

C. Semi-Supervised Methods: Semi supervised learning algorithms use both labeled and unlabeled corpora to create

their own hypothesis. Algorithms typically start with a modest quantity of seed data set and create more hypothesis' using large amount of unlabeled corpus. In this section, we will have a look at some of the semi-supervised NER system.

D. Unsupervised Methods: A major problem with supervised setting is a requirement of specifying large number of features. For learning a good model, a robust set of features and large annotated corpus is needed. Many languages don't have a large annotated corpus available at their disposal. To deal with the lack of annotated text across domains and languages, unsupervised techniques for NER have been proposed.

E. Hybrid Model Approach: In this approach Rule Based approach and Machine Learning approaches are mixed for more accuracy to identify NERs.

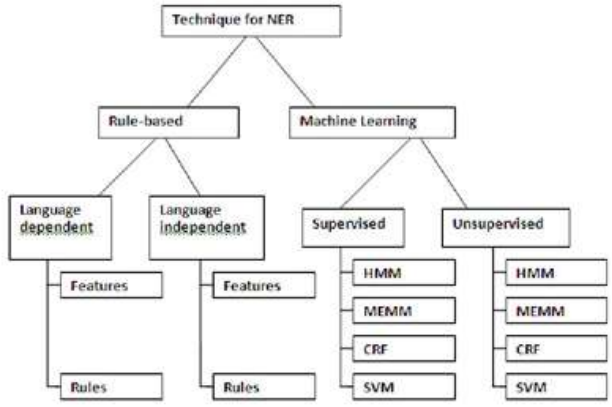

Fig 1: Approaches

III OBSERVATIONS AND DISCUSSIONS

P Srikanth**,** et. al[7] proposed a CRF based Noun Tagger. This system was built on three stages .In the first phase Trained on a manually tagged data is used to identify the nouns. In the second phase the nouns are now used to identify the named entities like person, place and organization names. In the next phase to handle the suffix and prefixes of named entities they have manually prepared a list of such suffixes gazetteer for both persons and locations, as also a list of prefixes for person names consisting of 1346 location names, 221 organization names. This process of semi-automatic tagging is continued for several iterations. This way they have developed a named entity annotated database of 72,157 words, including 6,268 named entities. By using the above data CRF based NER system for Telugu developed and tested it on several data sets from the Eenadu and Andhra Pradesh newspaper corpora. Good performance has been obtained using the majority tag concept. Obtained overall F-measures between 80% and 97% in various experiments. The Performance of the heuristic-based NER system is tested over two test data sets (AP-1 and AP-2). These test data sets are from the AP corpus.

Asif Ekbal, et. al[10] proposed a language independent NER by using the statistical Conditional Random Fields (CRFs) for South and South East Asian languages, particularly for Bengali, Hindi, Telugu, Oriya and Urdu as part of the IJCNLP-08 NER Shared Task 1 .They make use of the different contextual information about the words along with the variety of features that are helpful in predicting the various named entities (NE) classes. They used both the language independent as well as language dependent features. The language independent features are applicable for all the languages. The language dependent features have been used for Bengali and Hindi only. The training data were provided in five different Indian languages, namely Bengali, Hindi, Telugu, Oriya and Urdu in Shakti Standard Format. The training data in all the languages were annotated with the twelve NE tags, as defined for the IJCNLP-08 NER shared task target. The system has been trained with Bengali (122,467 tokens), Hindi (502,974 tokens), Telugu (64,026 tokens), Oriya (93,173 tokens) and Urdu (35,447 tokens) data. The system has been tested with the 30,505 tokens of Bengali,

38,708 tokens of Hindi, 6,356 tokens of Telugu, 24,640 tokens of Oriya and 3,782 tokens of Urdu. Evaluation results have demonstrated the highest maximal F-measure of 53.36%, nested F-measure of 53.46% and lexical F- measure of 59.39% for Bengali.

KashifRiaz,et. al [1]proposed a system which uses a hand crafted rule-based NER system for Urdu NER. These rules form a finite state automata (FSA) based on lexical cues. Some cues are at the start of the state, some are at the end of the state, sometimes the cues are found in the middle of the finite state machine. These rules are corpus-based, heuristic-based, and grammar-based. The rules are implicitly weighted in the order they are applied. The rule sets were created from 200 documents of Becker-Riaz corpus and the experiment was run on 2,262 documents. The algorithm execution resulted in 187 named entities and 171 of those were true named entities. The results show the recall of 90.7% and precision of 91.5%. This gives the $f1 - measure$ value of 91.1%.

Umrinder,et. al [3] proposed a Rule based NER for Urdu. The system first normalizes the input data, then the text is tokenized word by word. Next these words look up against the Gazetteer to recognize the NES. The rules are used to find person name, date, time, abbreviations, organizations which are not found in look up. The accuracy of system is 88.1%. If we consider all the thirteen tags of IJCNLP-08 the accuracy of the system is 74.09%.

Saha, et. al [15] proposed a hybrid NER system that applies the Maximum Entropy model (Max Ent), language specific rules and gazetteers to the task of Named Entity Recognition (NER) in Indian languages designed for the IJCNLP NERSSEAL shared task. Two approaches used here are the Linguistic approach where the typical rules written by linguists and the Machine Learning (ML) approach Where the system is trained using tags. Features identified in this NER system are: 1. Static word feature, 2. Context list, 3. Dynamic NE tag, 4. First word, 5. Contains digits, 6. Numerical word, 7. Word suffix, 8. Word prefix, 9. Root information of a word and Parts of speech information. Parts Of Speech(POS) tagging used the Coarse-grained tag set with only three tags - nominal (Nom), post position (PSP) and other (O). They also worked on Nested entities and Nominal Postpositions (Nom SPS). This paper reported that they received poor accuracy for Oriya, Telugu and Urdu languages compared to the other two languages due to lack of POS information, morphological information, language specific rules and gazetteer lists. Finally, the system was able to recognize 12 classes of NEs with 65.13% f-value in Hindi, 65.96% f-value in Bengali and 44.65%, 18.74%, and 35.47% f-value in Oriya, Telugu and Urdu respectively.

Shilpi Srivastava, et. al[16] proposed a Hybrid approach combination of rule based CRF and Maximum Entropy for named entity recognition system for Hindi with a result of 96% precision, 86.96 recall and 91% F-measure. Corpus used: LERC-UOH Telugu corpus, developed at the Language Engineering Research Centre at the Department of Computer and Information Sciences, University of Hyderabad.

Research work on NER for Telugu is (Praneeth M Shishtla et al, 2008),[18] "A Character n-gram Based Approach for Improved Recall in Indian Language NER" used Conditional Random Fields with Character based n-gram technique on two languages Telugu and Hindi with annotated Telugu corpus containing 45,714 tokens out of which 4709 were named entities, English corpus contained
45,870 tokens out of which 4287 were named entities and Hindi corpus contained 45,380 tokens out of which 3140 were named entities. A total of Nine features were used in training and testing and not used any of the language dependent

resources and used POS taggers, Chunkers, morphological analyzers... etc and also included some regular expressions and gazetteer information. Gram n=3
gave better F-measure up to 24.2% for 10k words, 35.38% for 20k words, 44.48% for 30k words and 48.93% for 35k words for Telugu, Gram n=2 gave better F-measure up to 52.92% for 10k words, 65.59% for 20k words, 67.49% for 30k words and 68.46% for 35k words for English and Gram n=4 gave better F-measure up to 40.96% for 10k words, 36.26% for 20k words, 42.36% for 30k words and 45.18% for 35k words for Hindi. The evaluation achieved an overall F-measure of 49.62% for Telugu and 45.07% for Hindi. More number of tested words giving a maximum F-
measure.

Sasidhar, et. al[21] proposed a 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

In the year 2008 [22] 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%

A language independent NER in Indian languages [23] 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.

IV CONCLUSION

Applications of Natural Language Processing are many like machine translation, text processing, information retrieval, speech recognition and so on. Named Entity Recognition is a critical task in all of these NLP applications. By using the rule based approach we can achieve better results. But here implementing the rules is the toughest task and also we must have full depth knowledge in the Language and large gazetteers lists are needed, which is not possible for resource poor languages such as Indian language. Coming to statistical approaches they also need large annotated Corpus for training and testing to get the bench mark results. Which are not

available for Indian Languages. This study brings out that hybrid models which combine both rules and a machine learning algorithm perform better for Indian languages. Finally, we conclude that the work done in NER for Indian languages is really less and need more research work is required for better system result.

V REFERENCES

[1] *KashifRiaz,"Rule-based Named Entity Recognition in Urdu", Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 126–135,Uppsala,Sweden,16 July 2010,Association for Computational Linguistics.*

[2] *Anup Patel, Ganesh Ramakrishnan, and Pushpak Bhattacharya, "Incorporating Linguistic Expertise using ILP for Named Entity Recognition in Data Hungry Indian Languages" Presented at the 19th International Conference on Inductive Logic Programming, Springer Berlin Heidelberg, 2009. 178-185.*

[3] *Umrinder Pal Singh, Vishal Goyal, Gurpreet and Singh Lehal, "Named Entity Recognition System for Urdu", Proceedings of COLING 2012: Technical Papers, pages 2507–2518, COLING 2012, Mumbai, December 2012.*

[4] *Sobha Lalitha Devi, Malarkodi C S, and Marimuthu K, "Named Entity Recognizer for Indian Languages", ICON NLPTool Contest2013.*

[5] *Sujan Kumar Saha,Shashi Narayan, Sudeshna Sarkar, Pabitra Mitra, "A composite kernel for named entity recognition", Pattern Recognition Lett, Published by Elsevier B.V, 31.12 (2010): 1591-1597.*

[6] *Praneeth M Shishtla, Karthik Gali, Prasad Pingali and VasudevaVarma,"Experiments in Telugu NER: A Conditional Random Field Approach", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 105–110, Hyderabad, India, January 2008, Asian Federation of Natural Language Processing.*

[7] *P Srikanth and Kavi Narayana Murthy, "Named Entity Recognition for Telugu" Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 41–50,Hyderabad, India, January 2008.*

[8] *Asif Ekbal and Sivaji Bandyopadhyay, "Named Entity Recognition using Support Vector Machine: A Language Independent Approach", International Journal of Electrical, Computer, and Systems Engineering 4.2 (2010): 155-170..*

[9] *Vivekananda gayan, Kama Sarkar, "A HMM based Named Entity Recognition System for Indian Languages", The JU System at ICON 2013.arXiv preprint arXiv:1405.7397 (2014).*

[10] *Asif Ekbal, Rejwanul Haque, Sivaji Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field Approach". Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33–40, Hyderabad, India, January 2008.*

[11] *Asif Ekbal,Rejwanul Haque,Amitava Das, Venkateswarlu Poka and Sivaji Banyopadyay, "Language Independent Named Entity Recognition in Indian Languages", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33–40, Hyderabad, India, January 2008.*

[12] *Malarkodi, C S. Pattabhi, RK Rao and Sobha and Lalitha Devi, "Tamil NER – Coping with Real Time Challenges", Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 23–38, COLING 2012, Mumbai, December 2012.*

[13] *Arjun Das and Utpal Garain, "CRF-based Named Entity Recognition, ICON 2013". arXiv preprint arXiv:1409.8008 (2014).*

[14] *Jisha P Jayan, Rajeev R R, Elizabeth Sherly, "A Hybrid Statistical Approach for Named Entity Recognition for Malayalam Language", International Joint Conference on Natural Language Processing, pages 58–63, Nagoya, Japan, 14-18 October 2013.*

[15] *Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar and Pabitra Mitra,"A Hybrid Approach for Named Entity Recognition in Indian Languages", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 41–50, Hyderabad, India, January 2008.*

[16] *Shilpi Srivastava, Mukund Sanglikar and D.C Kothari, "Named Entity Recognition System for Hindi Language: A Hybrid Approach", International Journal of Computational Linguistics (IJCL), Volume (2) : Issue (1) : 2011.*

[17] *S Amarappa and S Satyanarayana, "A Hybrid approach for Named Entity Recognition, Classification and Extraction (NERCE) in Kannada"Documents", Proc. of Int. Conf. on Multimedia Processing, Communication and Info. Tech., MPCIT, at Association of Computer Electronics and Electrical Engineers, 2013.*

[18] *Praneeth M Shishtla, Prasad Pingali, and Vasudeva Varma 2008 "ACharacter n-gram Based Approach for Improved Recall in Indian Language NER s" Proceedings of the IJNLP-08 Sorkshop on Ner for South and South East Asian Languages Hyderabad, India.*

[19] *Bh.Krishna Murthy and J.P.L.Gywnn. 1985. A Grammar of Modern Telugu. Oxford University Press, Delhi.*

[20] *G. Bharadwaja Kumar, Kavi Narayana Murthy, and B.B.Chaudhari. June 2007. Statistical Analysis of Telugu Text Corpora. IJDL,Vol 36, No 2, pages 71– 99.*

[21] *B Sasidhar, P M Yohan, Dr. Vinaya A Babu and Dr. A Govardhan. Article: Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach. International Journal of Computer Applications 22(8):30–34, May 2011.*

[22] *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.*

[23] *Asif Ekbal, Rajewanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay 2008 "Language Independent Named Entity Recognition in Indian Languages", Proceedings of the IJNLP-08 Workshop on NER for South and South East Asian Languages Hyderabad, India*