International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 03 - Issue 01 || January 2018 || PP. 50-58

Knowledge Base Approach for Named Entity Identification and Classification in Telugu

Dr. Srinivasu Badugu, J. Adarshavathi
[1](*Computer Science & Engineering , Stanley College of Engineering for Women / Osmania University, India)*
[2](*Computer Science & Engineering, Gandhi Institute Of Technology / GITAM University, 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 rule based system 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: Gazetteers ,NE, Rule-based, Telugu

1. 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 **Microsoft** and **Satya Narayana Nadella**, 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 NERC. 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.

1.1 About Named Entity (NE)

1) 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. naaraa caMdrabaabu naayuDu, caMdrabaabu naayuDu, caMdrabaabu etc)
- Ambiguity of named entity types:For example, Daa. ReDDi డా . రెడ్డి (dr. Reddy) may be a person name or a company name.
- kavita (Person Vs common)
- When the word Tirupati being used as a name of person and as name of the city?
- Ambiguity between named entities and common words. When the word "viina or giita or kavita" being used as a name of person and as common word?

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 ``raaju రాజు (king) and raaNi రాణి (queen)" can be a person name as well as a common word.

2) 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].

2. 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 Indian languages, compared to English. Ideas and features that are used for English cannot be borrowed directly to Indian languages. For example, capitalization feature is not there in Indian languages. Indian languages are rich in morphology. In Indian languages Hindi has simplest morphology, as we go from north to south the complexity increases. Dravidian languages have more morphology. Indian names are more diverse in nature, i.e. there are a lot of variations for a given named entity. For example "telugudeeshaM paaThii" 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 "Dainik Jagaran" 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".

"Parveen Kumar et al. (2008) [13] reports about the development of a NER system for 5 languages (Hindi, Bengali, Oriya, Telugu, Urdu) by using Hidden Markov Language. They have used POS-Tag and Chunk information. They obtained a decent F-measure of 39.77%, 46.84%, 45.84%, 46.58% and 44.78% respectively for all 5 languages".

"Praneeth M Shishtla et al. (2008) [14] presents the hypothesis by making experiments on NER system in Telugu language by adopting CRF approach. Recall, Precision and F-score are claimed to be 64.70%, 34.57% and 44.91% respectively".

"In paper [15], authors proposed a method that is a combination of Maximum Entropy (MaxEnt) and Conditional Random Field (CRF) and Support Vector Machine (SVM) for NER in Bengali. They take approximately 272k word forms of training set for testing. And they have developed semi-supervised learning technique that uses the unlabeled data during training of system. Authors describe that use of large corpora is not enough but system should measures to automatically select effective documents and sentences from the unlabeled data. They have finally used an approach that is weighted voting approach to combine the models. And the average experimental result of recall, precision, and f-score values is 93.79%, 91.34%, and 92.55% respectively".

"In paper [16], authors have also made use of gazetteer lists with both techniques as well as some smoothing techniques with bigram NER tagger. This NER system is capable to recognize 5 classes of NEs i.e. Person, Location, Organization, Date and Time. They have used gazetteer lists to improve the results of n-gram statistical models. The unigram tagger trained with training data and combined with gazetteers produced up to 65.21% precision, 88.63% recall and 75.14% f-measure. A bigram NER tagger is trained with training data, combined with gazetteers and back off smoothing produced up to 66.20% precision, 88.18% recall and 75.83% f-measure".

"Hasanuzzaman et al. [17] describe the development of NER system in Bengali and Hindi using ME framework with 12 NE tags.The average recall, precision and f-measure were 88.01%, 82.63%, 85.22%, respectively for Bengali and 86.4%, 79.23% and 82.66%, respectively for Hindi".

"Li and McCallum [18] describe the application of CRF with feature induction to a Hindi NER task. The experimental results for validation and test sets found out to be 82.55% and 71.50% respectively".

"Goyal [19] focused on building Hindi NER using CRFs. He used the NLPAI Machine Learning Contest 2007 data for experiments. This method was evaluated on two different test sets and attained a maximum F1-measure of around 49.2% and nested F1-measure of around 50.1% for test set1, maximum F1-measure around 44.97% and nested F1-measure around 43.70% for test set2 and F-measure of 58.85% on the development set. The author also compared the results on Hindi data with English data of CONLL shared task of 2003. They trained this system on English data of CoNLL-2003 shared task, considering only contextual features since they give the maximum accuracy. They obtained an overall F-measure of 84.09% and 75.81% on both the test sets".

"B.Srinivasu et al. [20] developed a Telugu NER system by using the ME approach.Evaluation results came out with a an F-measure of 72.07% for person, 60.76%, 68.40% and 45.28% for organization, location and others, respectively".

"VijayKrishna and Sobha [21] developed a domain specific Tamil NER for tourism by using CRF. The system obtained an F-measure of 80.44%".

"Srikanth and Murthy [22] used CRF approach on a part of the Language Engineering Research Centre at University of Hyderabad (LERC-UoH) Telugu corpus consisting of a variety of books and articles, and two popular newspapers They obtained an F-measure of 91.95%. Then they developed a rule-based NER system using a corpus of 72,152 words including 6,268 Named Entities. Finally they developed a CRF based NER system. They achieved an overall F-measures between 80% and 97% in various experiments. Shishtla et al. [23] conducted an experiment on the development data released as a part of NER for South and South East Asian Languages (NERSSEAL) Competition using CRF.The best performing model gave an F-1 measure of 44.91%".

3. Proposed Approach

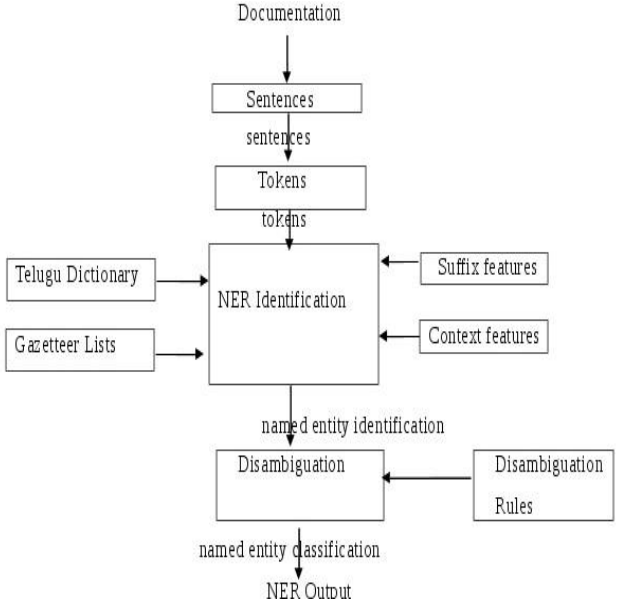

Fig 1. NERC system block diagram

Fig. 1 shows simple data flow in our proposed system. In this system We proposed new approach for named entity identification and classification for resources are not available languages.

3.1 Preprocessing

Sentence: A set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses.

Tokens: This tokens an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.

3.2 NER Identification

Telugu dictionary: A book or electronic resource that lists the words of a language (typically in alphabetical order) and gives their meaning, or gives the equivalent words in a different language, often also providing information about pronunciation, origin, and usage.

Gazetteer List: The Gazetteer list consist of Proper Nouns for instants Person, Location, Organization.

Suffix Features: Suffixes are word endings that add a certain meaning to the word.

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.

Disambiguation: Disambiguation refers to the removal of ambiguity by making something clear. Disambiguation narrows down the meaning of words and it's a good thing. This word makes sense if you break it down. Dis means"not," ambiguous means"unclear," and the ending -tion makes it a noun.

Steps for Named Entity Recognition:

1. Open file (document) and dividing it into sentences (sentence segmentation using Rules).
2. Divide each sentence and split it into tokens.(Tokenization)
3. If last token (word) in a sentence has long and ending with PNG(person number gender) marker, 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. Next we check each token other than final word with Telugu dictionary. If the word is found in a Telugu dictionary, then assign category of a word (token).
5. If not found in dictionary we check in the Gazetteers list (Person, Location, and Organization). If it is found then assign appropriate category.
6. If it is not found, we apply all suffix and grammatical suffix features for identification. If the word is matching with any one of the suffix, then assign the appropriate suffix category.
7. If there is any ambiguity exits between words then we call disambiguation technique and remove ambiguity.

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

3. 3 IMPLEMENTATION

Corpus Preparation: Collection of machine readable text is called "CORPUS". We have collected the corpus from ANDHRA JYOTHI, ANDHRA PRABHA, EENADU, and SAAKSHI which are popular Telugu local daily newspapers. All the news papers are available with UTF or font encoding formats. Some web tools automatically convert the font to Unicode. But this available data is not in machine readable format. So the text is converted into machine readable format using our own Transliteration format.

Gazetteers Collection : The NE gazetteer list consists of various categories of information, such as, Person-names, Location-names and Organization names. Person names are divided into four parts including beginning of a person, ending of a person, contexts of a person, and suffixes of a person. The person context list contains(26), names of a person(2561), beginning of a person (1714), ending of a person (301), suffix of a person(98), Location names are divided into four parts which include location names(20634), beginning of a location(1347), ending of a location(82), suffixes of a location(49), and suffixes of the organization (26). Month names(24), Week names(07). All these entries are collected from various sources and transliterated[10].

In this approach we have collected various lists like suffix list, features list and other lists related to Person, Location, and Organizations. We have collected these gazetteers using various pattern matching techniques. Using these techniques we are getting sure words (if any word is ending with some pattern like, raavu in a corpus, the word is compulsory for the name of the person).

We have collected sure patterns and using them. We are also collecting various lists. There may be a possibility of missing words. We have collected names of persons, organizations and locations for "clues information". Names are unique and infinite we do not have lists enough names.

Now here each word from the input is compared with Gazetteers list for Person, Location and Organization. If the word is matched with this Gazetteers list, then assign the tag according to its category. Otherwise check the word suffix matching with any one of the name suffix list. If the suffix is found, then assign the tag according to the category to which the NE belongs. If the word suffix is not found, then the word is again searched for context features list. If there is any match, assign the relevant tag.

If there is any ambiguity, that is, if the word contains more than one tag for the same name, then, check the context features list for a match. According to the context, the features are useful to assign the category. This way we finally identify Named Entities.

In this processes we use language dependent suffix features and grammatical features (it is available in Telugu grammar books like noun suffixes and verb suffixes [11, 12]), context features.

3.4 Suffix lists for NER Identification

NER systems have been developed for resource rich languages like English with very high accuracies. But constructing an NER for a resource-poor language is very challenging due to unavailability of proper resources. Name-dictionaries or gazetteers are very helpful NER resources and in Telugu language there is no publicly available list. The web contains lots of such resources, which can be used for Telugu language NER development. Most of the web resources are in English. But direct transliteration from English to Telugu language is not easy.

3. 4.1 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. 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.

Eg.

(Person)	Naagaraaju , chaMdrabaabunaayuDu
(Location)	vijayavaaDa, maciliipaTnaM
(Organization)	prajaaraajyaMpaarTii, aaMdhrabyaaMk

Here manually we gathered suffix list

Person suffix list - 119
Location suffix list - 37
Organization suffix list - 27

Here we gathered the 1 lakh words from the newspaper

Using this suffix list we extracted the proper nouns within the corpus. 1st we are comparing the suffix list with the corpus. If any suffixes are found in the corpus then assign the category.

Using the person suffix list we extracted 3,762 words. Using Location suffix list we extracted 3,762 words. Using Location suffix list we extracted 877 words and the Organisation suffix list words are 29 with in the corpus. Out of 119 person suffix list 46 suffixes gives the NE's. These suffixes surely gives the named entities. As well as out of the 37 Location suffix list only 21 suffixes gives the good result. That means these suffixes are 100% sure about Named Entities. Finally out of the 27 Organisation suffix features only 7 suffixes gives the good result. These 7 suffixes surly gives the N.E's.

	Suffix list	Extracted words
Person suffix list	119	3,762
Location suffix list	37	877
Organisation suffix list	27	29

3. 4. 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. 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.

E.g.

(Person)	**s`ree** naagees`vararavu gaaru, **DaakTar** varaprasaad
(location)	anaMtapuraM **jilla**, puliveMdula **graamaM**
(organization)	aaMdraprades` rooDDu ravaaNaa **saMsta,** s`raamika vidyaa **peeThamu**

Grammatical 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.

Eg. Common noun	aavulatoo aavu + lu + too
	(with cows) root word. + number + case marker
(Person)	iMdiraadeevitoo iMdiraadeevi + too
	(with raamaaraavutoo) root + case marker
(Location)	haidaraabaadunuMdi haidaraabaadu + nuMdi

(from haidaraabaadu) root + case markers

Here also manually we gathered the grammatical feature. Using this feature I gathered 38 grammatical features. Using theses grammatical features we extracted the word list.
Grammatical features are 38 and we extracted the 49 words.
Using the suffix features, context features and grammatical features for identification of proper nouns gave good result.

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. The overall methodology of extracting gazetteers from a raw corpus is summarized as follows:
Step1: Finding the useful corpus. It must be machine readable format.
Step2: Extraction gazetteers from corpus using seed patterns and identified features.
Step3: Validation was done by extracted gazetteers randomly.

The tables below shows preliminary named entity identification and classification results based on Telugu dictionary, named entity gazetteers, suffix and context feature and finally we use dis-ambiguity rules.

Table 1 TAG Description of Named Entity suffixes for Gazetteers

TAG Description	Size
PER-SUF - Person suffix	98
PER-BEG - Person -begging	1,714
PER-CON - Person context	26
PER-END - Person ending	301
LOC-SUF - Location suffix	49
LOC-BEG - Location beginning	1,347
LOC-END - Location ending	116
ORG-SUF - Organization suffix	27
ORG-CON- Org context	35
Noun suffix (case marks)	43
PSLF	16

After gathering all useful requirements we tested our system using news paper corpus. We demonstrate few experimental result bellow.

Table 2 Sentences, Words count in Files

Files	**Sentences**	**Words**
File1	03	20
File2	04	44
File3	08	135
File4	72	1035
File5	113	1969

No. of words tested	No. of NE's Identified by manually				No. of NE's Identified by our system			
	NE	PER	LOC	O RG	NE	PER	LOC	ORG
20	01	01	01	0	20	01	01	0
44	01	01	02	0	44	01	02	0
135	06	06	04	01	135	04	03	01
1035	15	15	18	08	1035	12	11	06

1969	30	30	22	10	1969	21	14	12
2503	46	44	41	19	2503	36	27	21

Table 4 Named Entity Identification in files

	#words tested	#NE's Identified manually	#NE's Identified by our system	%Identification
File 1	20	02	02	100
File 2	44	03	03	100
File 3	135	11	08	72.7
File 4	1035	41	29	70.7
File 5	1969	62	47	75.8

4. 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)

5. 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]. Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dantapat, Sudeshna Sarkar and Pabitra Mitra 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. The Grammar of the Telugu Language. 1991, New Delhi: Laurier Books Ltd.

[13]. Praveen Kumar, P., Ravi Kiran, V.: A Hybrid Named Entity Recognition System for South Asian Languages. In: Proceedings of the IJCNLP 2008 Workshop on NER for South and South East Asian Languages, Hyderabad, India, pp. 83–88 (January 2008)

[14]. Shishtla, P.M., Pingali, P., Varma, V.: Character n-gram Based Approach for Improved Recall in Indian Language NER. In: Proceedings of the IJCNLP 2008 Workshop on NER for South and South East Asian Languages, Hyderabad, India, pp. 67–74 (2008)

a. Ekbal & S. Banyopadhyay, (2009) "Named Entity Recognition Using Appropriate Unlabeled Data, Post-Processing and Voting".

[15]. F. Jahangir, W. Anwar, U. Bajwa1 & X. Wang, "N-Gram and gazetteer list based Named Entity Recognition for Urdu: A scarce resourced Language", unpublished.

[16]. Mohammad Hasanuzzaman, Asif Ekbal, and Sivaji Bandyopadhyay. Maximum EntropyApproach for Named Entity Recognition in Bengali and Hindi.International Journal of Recent Trends in Engineering, 1(1):408–412, May 2009.

[17]. Wei Li and Andrew McCallum. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper). ACM Transactions on Computational Logic , pages 290–294, Sept 2003.

[18]. Amit Goyal. Named Entity Recognition for South Asian Languages. In Proceedings of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages, pages 89–96, Hyderabad, India, Jan 2008.

[19]. G.V.S Raju, B.Srinivasu, Dr. S. Viswanadha Raju, and K.S.M.V Kumar. Named Entity Recognition for Telegu using Maximum Entropy Model. Journal of Theoretical and Applied Information Technology ,3(2):125–130,2010

[20]. Adam L.Berger, Stephen A.Della Pietra, and Vincent J.Della Pietra. A Maximum Entropy approach to Natural Language Processing. Computational Linguistic 22:39–71,1996.

[21]. P.Srikanth and Kavi Narayana Murthy. Named Entity Recognition for Telegu. In Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages, pages 41–50,Hyderabad, India, Jan 2008.

[22]. Praneeth M Shishtla, Karthik Gali, Prasad Pingali, and Vasudeva Varma. Experiments in Telegu NER: A Conditional Random Field Approach. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages, pages 105–110, Hyderabad,India, January 2008.