A Survey on Various Approach used in Named Entity Recognition for Indian Languages

Dikshan N. Shah
Assistant Professor
S S Agrawal Institute of Compueter Science,
Navsari

Harshad Bhadka, PhD
Dean
Faculty of Computer Science,
C U Shah University,
Wadhwan

ABSTRACT
Named Entity Recognition (NER) is an application of Natural Language Processing (NLP). NER is a activity of Information Extraction. NER is a task used for automated text processing for various industries, key concept for academics, artificial intelligence, robotics, Bioinformatics and many more. NER is always essential when dealing with chief NLP activity such as machine translation, question-answering, document summarization etc. Most NER work has been done for other European languages. Among Indian constitutional languages, NER work has been done for few languages. Not enough work is possible due to some challenges such as lack of resources, ambiguity in language, morphologically rich and many more. In this paper, we found many challenges available in NER for Indian languages and compared by measuring standard evaluation metrics values of accuracy, precision, recall and F-measure.

Keywords
Natural Language, Named Entity Recognition, NER approach, Indian languages

1. INTRODUCTION
(NLP) Natural Language Processing is a very captivating method of human computer communication. Natural-language perceptive is sometimes referred to as an AI-complete problem, because it seems to require wide knowledge about the outside world and the aptitude to operate it.

The fundamentals of NLP lie in a number of disciplines, viz. mathematics, computer and information sciences, artificial intelligence and robotics etc. Named Entity Recognition (NER) is the application of NLP.

The task of NER is to identify all named entities from given document or paragraph and after that classify all named entities such as: Name of the person, Gender of the person, Location name (location can be city, state, country etc), Organization name, Date, E-mail address, Event etc.

1.1 Applications of NER
NER attains application in most of the NLP applications. The following list mentions few of its applications.

1) Mostly useful for Search engines.

2) In the context of Cross-Lingual Information Access Retrieval (CLIR), given a query word, it is very important to find if it is a named entity or not.

3) An amount of information can be examined using named entities, like plotting the popularity of entities over time and generating geospatial heat maps.

4) Mainly used in machine translation. Usually, entities identified as Named Entities and are transliterated as disparate to getting translated.

5) Most of the words indexed in the back index of a book are Named Entities.

2. EXISTING APPROACH
Named Entity Recognition basically classified into mainly two approaches as follows:

2.1 Rule Based or Hand crafted Named Entity Recognition System
Studies made in Named Entity Recognition primarily were based on handcrafted rules. Human made rules forms the main background of rule based Named Entity Recognition. Rule based approach can be further classified as List Look up Approach and Linguistic Approach.

2.2 Automated or Machine learning approach
Machine learning approaches are advantageous over rule based approaches as all these approaches are statistical in nature. Some machine learning approaches are Conditional Random Field (CRF), Hidden Markov Model (HMM), Decision Trees (DT), Maximum Entropy Markov Model (MEMM) and Support Vector Machine (SVM).

3. EVALUATION METRICS
The Named Entity Recognition performance is always measured in terms of Accuracy (A), Precision (P), Recall(R) and harmonic mean of precision and recall F-Measure (F).

$$Accuracy\ (\%) = \frac{Correct\ Words}{Total\ Named\ Entities} \times 100 \quad ---[36]$$

$$Precision(P) = \frac{number\ of\ Correct\ responses}{number\ of\ responses} \quad ---[36]$$

$$Recall\ (R) = \frac{number\ of\ Correct\ responses}{number\ of\ correct\ in\ key} \quad ---[36]$$

$$F - Measure\ (F) = \frac{2PR}{(P+R)} \quad ----[36]$$

4. LITERATURE SURVEY

They took the issue of handling the named entities as without it quality of translation would get affected. Only rule based approach was not sufficient, they have used a hybrid approach for it and collected 10000 sentences from news web sites as a corpus. And they used Standford's NER tool for name entity recognition. Out of total name entities of 9234, 9180 entities were generated from the system. So they achieved an accuracy of 83.65% Precision, 83.16% Recall and 83.40% as F-Measure value. **[1]**

In their paper, they found that among the Indian languages, Kannada language has no capitalization and lack of non-availability of larger gazetteer, lack of standardization and spelling. They found that there is a lack of annotated data and it is highly agglutinating and inflected language. They developed a Supervised Statistical Machine Learning system for Kannada Language using Multinomial Naïve Bayes classifiers. They have used 22 named entities and corpus of 95170 words. They recognized named entities from their developed model and got 83% Precision, 79% Recall and 81% of F-Measure value. **[2]**

They could achieved results to overcome the limitations due to the nature of Arabic Language and lack of available linguistic resources. Available corpora are not annotated with name entity and the relations do not include a sufficient number of annotated examples to be exploited for learning approaches. They combined the Machine Learning and Genetic Algorithm rules to enhance the overall performance of Machine learning method. Hybrid approach gives performance of Precision 84.8%, 67.6% Recall and 75.22% of F-Measure value. **[3]**

By using a combination of Rule Based Approach and List look approach, they found solution to overcome the limitations of availability of data corpus and need to resolve ambiguities of named entity recognition in Hindi Language. They found different values of Precision, Recall and F-Score for Location, Person, Organization, Date, Money, Direction, Transport etc. They achieved 95.77% accuracy in their system. **[4]**

They could achieve performance to overcome the limitations of ambiguous names, no capitalization, scarcity of resources and tools, lack of standardization and spelling, lack of labeled data, non availability of large gazetteer in Hindi language. For better presence of Hindi language, they have tested with different available approaches of NER and used voting method to improve the performance. By using CRF approach, they achieved 71.43% Precision, 30.86% Recall and 43.10% F-1 Measure for 5 testing files. For MaxEnt, they achieved 76.92% Precision, 19.8% Recall and 31.49% F-1 measure for 5 testing files. As per Rule based approach, they achieved 96.05% Precision, 86.90% Recall and 91.25% F-1 measure for 3 testing files. **[5]**

They found several challenges due to rich morphology of Kashmiri language. As compared to English language, Kashmiri language does not have capitalization. Based on 1000 Kashmiri words, they have conducted test for noun identification and got good performance by using dictionary gazetteer, lists, morphological suffix mapping techniques. They have identified nouns and achieved result as 93.32% and 07.75% errors. Using NE tags, ambiguity is also resolved using gazetteers lists and features. **[6]**

Among the constitutional Indian languages, they have found challenges in Manipuri language. No capitalization, redundant named entities available in dictionary with other specific meaning, highly inflectional language resulting in large complex word forms, free word order language which difficult to compared with others, resource constrained language. Using CRF approach evaluation has been done and achieved 81.12% Recall, 85.67% Precision and 83.33% F-Score value. Manipuri gazetteer list can be formed using the NER. **[7]**

They described named entity recognition for Mising language. It is a Tibeto-Burman language which is inhibit in Assam. It is a resource constrained language. They used 12 named entity tagset for feature extraction and in this language Roman Script. For classification of recognized entities they have used Support Vector Machine. As a limited availability of resources for the language, authors have to define their own corpus. Out of 34000 training data, 16000 data has been tested and achieved 90.58% Recall, 85.14% Precision and 87.77% F-Score value. **[8]**

In their paper they described that Urdu language challenges for Urdu Named entity recognition as No Capitalization, Scarce Resources, Agglutinative nature feature, free-word order, Complexity of spelling variations, borrow words, nested and compound named entities and many more. **[9]**

By using two machine learning approach in their paper as Hidden Markov Model and Entropy Markov Model for Punjabi named entity recognition, they have focused on general challenges for Punjabi language. Lack of spelling standardization, Non-availability of large gazetteer as Punjabi language is not much used on internet as compared to other languages. In all Indian languages number of common words which are also used as Named Entities. No capitalization feature as compared to English language. Scarcity of resources and tools for Punjabi language. 42k words for Hidden Markov Model and 61k words for Maximum Entropy Model of training corpus has been taken from various news articles and Punjabi newspapers. For HMM F-Score evaluation, in person named entity class 83.46%, 82.20% for Location NE class, 86.13% for Organization achieved. In MaxEnt model F-score was evaluated as 87.93% for Person NE class, 83.32% for Location NE class and 89.92% for Organization NE class. **[10]**

They found that among the 22 Indian languages, named entity recognition accuracy is not comparable with foreign languages which were deeply explored in NER. Large amount of information is available about Punjabi Language, but it is not in proper format so could not used for local users. Web sources for various gazetteer lists are not available in this language. They described that Machine learning approach is best suited with Punjabi language for NER by using 12 Named Entity tagset. They found that context window of word size 3, 5 and 7 gives the same results in F-Score value. Experiment has done by using Conditional Random Field approach. They achieved feature set comprising of word window size 5, digit features, Infrequent word has confirmed the highest F-Score value 87.46% and for feature set comprising of word window size 3, second highest F-Score value is 87.40%. According to them compare to all language independent features with word window size 5 gives highest results. **[11]**

They described that Indian languages are inflectional, free order and morphologically rich and lack in resources. Indian languages are ambiguous so recognition is too difficult. One major challenge they have found that one NER system built for one domain did not work well with other domains. 90% F-

Measure value achieved in Indian languages using various NER approach. In their research, they found some issues in Urdu language. They have used Hidden Markov Model for their research in NER. They achieved 100% performance results of tourism corpuses and 100% accuracy performance in seven sentences in Urdu of BBC news. The found that many tools for NER are available but they all are language dependent. No such tool is there, which is language independent. **[12]**

By using combined approaches as SVM and CRF, different categories of Geological Named Entities to recognize, to classify and identify various geological entities. They found that NER is a hard problem to identify proper names. Due to capitalization in English language up to some extent problem of identification is solved. They have developed new tag set for Geological corpus. For various locations using CRF approach they got 77.05% precision, 77.27% recall and 75.81% F-measure value. As per hybrid approach for various locations based NEs they achieved 81.99% Precision and 78.36% Recall result. **[13]**

They mainly focused on English to Hindi transliteration and they suggested a hybrid approach for the same. They found challenges as difficult to identify capitalized word in Hindi as there is no capitalization concept available, highly phonetic and inflectional language, places names are homographic as matched with person names, multiple transliterations possible for one word. They have defined approach based on knowledge and syllabification to identify named entities and achieved 84.23% accuracy. Their system was designed only to identify person named entity only. **[14]**

They described NER for Hindi, Marathi and Urdu language. For these mentioned Indian languages, they have taken Tourism Domain corpus for Hindi language, for Marathi language they considered NLTK Indian corpora and for Urdu language first they need to translate tourism domain corpus with the help of Google translator. They found major challenges for Urdu language as language corpus is not available, writing format is from right to left, annotation is time consuming. By using Hidden Markov model, they have perform training set and tested set for 100 sentences, 2209 words with 8 named entity tagset and achieved 86% of accuracy for Hindi language, For Marathi language, they have taken 100 sentences, 1448 words with 7 named entity tagset and achieved 76% result. As lack of Urdu language resources, they have tested 50 sentences, 734 words with 11 named entity tagset and got 65% performance. **[15]**

They have faced many challenges in NER while working with Indian languages. They found that there is a lack of resources for Indian languages even Indian languages are free order language, morphologically rich and inflectional and numerous entities exist as common nouns in dictionary. They discussed various approaches for NER and among them they used Hidden Markov Model for 9 named entity tagset for Hindi, Punjabi and Urdu languages and achieved F-measure value as 98.16%, 96.6%, 95.5% respectively. **[16]**

They discussed named entities from documents and categorized into proper nouns which will be useful component for NLP in English language. They have used 25 files of Treebank corpus, 6680 training words with 8 different name tags. Using Hidden Markov Model, performing NER in English language they obtained 73.8% F-measure value. With 70% of accuracy obtained in identifying named entity especially the names of Person. **[17]**

They mentioned linguistic rules to identify named entities in Oriya language using Hidden Markov Model and MaxEnt Model. They found that linguistic rules or Oriya language plays a crucial role in identifying NEs. Different 7 types of name tags were used and training data for this language contains more than 56k. Transliteration used to translate web resources into Oriya language. They secured F-measure value in science domain, 86.10%, 79.24%, 87.34% for Person, Location and Organization respectively. 88.23%, 83.33% and 77.98% for Arts domain. 82.12%, 86.65% and 85.98% for World affairs domain. In commerce F-measure values were 79.88%, 77.78% and 89.78%. **[18]**

In their paper they discussed several challenges they faced to identify named entity from Hindi language. They mentioned that system developed for English and European languages is not applicable for Indian languages. They found lack of language resources, No Capitalization, free-word language, rich in morphology and inflectional language. For 9 different tagset, according to shallow parsing technique, they obtained 47.5% accuracy for 325 detected NEs out of 687 NEs. In HMM, result was achieved as 89.78% accuracy for 325 NEs out of 362 NEs, while using Hybrid approach gives 94.61% accuracy for 650 identified NEs out of 687 total NEs. **[19]**

They mentioned that among the 22 Indian constitutional languages, Kannada is the seventh language spoken in state of Karnataka. Due to lack of available resources, annotated corpora, named dictionaries, Parts of Speech taggers, labeled data, security of resources and tools, they have taken corpus from Kannada news papers, and Kannada Wikipedia. They tested training set using Hidden Markov Model. **[20]**

They discussed challenges that Indian languages are free-word order, inflectional and rich in morphology. Unlike in English language, no Capitalization concept exists in other Indian languages. Lack of web resources for Indian languages. There is an ambiguity in Indian language NEs as some NEs are exist as common nouns in dictionaries. To resolve above challenges, they have used Hidden Markov Model for Hindi, Bengali and Telugu language. For Hindi language they have taken corpus from Tourism Domain which is developed by Banasthali Vidhyapith and Political and Sports news corpus taken from NLTK Indian Corpus. With 540 sentences, 8623 words from NLTK Indian corpora they reported accuracy level of Recall, Precision and F-Measure is 96% for Hindi. For training, they have taken 100 sentences or 2332 tokens from a Hindi tourism corpus, developed at Banasthali Vidyapith. They annotated it using 10 tags and obtained F-Measure of 93%. They performed training on 9996 words or 994 Telugu sentences of NLTK Indian Corpora and achieved F-Measure is 98.6%. Also they performed training on 10,303 words or 899 sentences of Bengali language taken from NLTK Indian Corpora with securing 98.5% F-Measure value. **[21]**

They mentioned that in Indian languages text prevails in undefined format from which entity extraction is a monotonous process. Entity extraction is crucial in Indian languages since language evolves with time and new words are added in vocabulary frequently. In this paper they deal with Malayalam language corpora of social media text and twitter resources with 33 different entities. They used SVM based classifier for supervised, unsupervised, known and unknown tokens. They achieved 93.33% Precision, 72.41% Recall and 81.55% F-measure standard results. **[22]**

In their paper they discussed for Indian languages that there is no concept of Capitalization. All words are written in similar

case. There is uncertainty and no particular standard has been defined to write spellings or words. Indian languages are having no specific order of words. Indian languages are considered to be poor languages as a good dictionary, web source and well-built language analyzer is not available yet. Still a lot of technological enhancement is required in Indian languages. While implementing in NER they faced that NER for one language cannot be implemented for the other language. Implementing same NER for different languages requires too much work and efforts. In some cases, Rule based Approach gives optimal solution with high accuracy. But, Language experts are required to generate rules for each language. As Indian Languages can be written in any order, new words can be easily created and thus maintaining the word list is a big task. But still a Gazetter method gives an acceptable output. To resolve all above mentioned challenges they suggested Hidden Markov Model for Gujarati language. **[23]**

They suggested CRF based identification of named entities from Social Media Text as corpus for Hindi and English language. They suggested that NER in Indian languages is still considered to be a budding topic of research in the domain of NLP and much of work is required to be performed in this regard. They have used 22 named entity tags. By using CRF approach they achieved 25.65% Precision, 16.14% Recall and 19.98% F-measure for Hindi language. Also 4.13% Precision, 3.39% Recall and 3.72% F-measure result evaluated for English language. **[24]**

Among the 22 Indian constitutional languages, Kannada is the Dravidian language spoken in the state of Karnataka. Basically Kannada is free-word order, inflected and agglutinating, rich heritage and large grammar. They found that extraction from Kannada language is challenging process. To handle noun inflections is major challenge and research work. By using Hidden Markov Model and hand crafted rule to identify and extract root entities from manually created database. From the different domains of Kannada language, total samples tested 130 sentences and more than 10000 words. They obtained 95.10% Precision, 94.61% Recall and 94.85% F-measure values. **[25]**

They described Statistical Hidden Markov model for different 7 Indian languages. They did not use gazetteer list because for all Indian languages gazetteer is not available. They evaluated traditional Precision, Recall and F-measure values. They obtained these values for Bengali language is 84.47%, 87.56% and 85.99%, for English language, 76.83%, 77.25% and 77.04%, for Hindi language 75.40%, 75%, 75.20%, for Marathi language 53.05%, 36%, 42.89%, for Punjabi language 54.97%, 54.13%, 54.55%, for Tamil language 32.21%, 72.79%, 44.66% and for Telugu language 37.73%, 42.63% and 40.03% correspondingly. **[26]**

They targeted the Nepali language in their research. They found that there is no capitalization, ambiguous names difficult to recognized, relatively free-word order language, non availability of large gazetteer, inflectional and rich in morphology. Training dataset consists of about 234k words collected from the newspaper "Dainik Jagaran" and manually tagged with 17 classes consists of 16482 NEs. Analysis of n-gram technique with gazetteer method on newspaper corpus which has about 1000 sentences. Total number of tags in corpus for person is 169, organization is 59 and for location are 31. Accuracy is obtained as 79.54% from 1000 sentences using n-gram and gazetteer method. **[27]**

They discussed key task of NLP as classification and identification of named entities. Loads of information is being shared by people in twitter on a daily basis. This information is unstructured and often includes important information about organizations, politics, disasters, promotional advertisements etc. They have identified 22 tags from the training data using Conditional Random Field approach. Twitter data for experiment was provided by FIRE 2015. They obtained 93.82% Precision for n-fold in English, 92.28% for Hindi and 86.94% for Tamil language. Recall value for n-fold in English is 80.53%, 76.23% for Hindi language and 73.87% for Tamil language. F-measure value for English language obtained as 86.66%, for Hindi language 83.49% and 79.87% for Tamil language. **[28]**

In their paper they mentioned that identifying the different entities in social media text is an important challenging task due to the informal nature of text present on social media. They faced a major challenge as code mixing in Indian social media text. The raw data files consist of 2700 tweets in Hindi-English corpus and 3200 tweets in Tamil-English corpus. 22 tags were present in the corpus. They proposed hybrid approach of a dictionary cum supervised classification approach for identifying different entities. They tested in proposed system for 3 runs and evaluated Precision for Hindi English language for Run 1 is 58.66%, 32.93% and 42.18% correspondingly. For Run 2 58.84%, 35.32% and 44.14% are evaluated. And for Run 3, they obtained 59.15%, 34.62% and 43.68% results. They also evaluated 3 Runs results for Tamil English. In Run 1, they achieved 55.86%, 10.87% and 18.20%, in Run 2 58.71%, 12.21% and 20.22% and for Run 3 they got 58.94%, 11.94% and 19.86% standard measure values. **[29]**

In this paper they found that among the 22 Indian languages, named entity recognition accuracy is not comparable with foreign languages which were deeply explored in NER. They described Machine learning approach is best suited with Punjabi language for NER by using 12 Named Entity tagset. They found that context window of word size 3, 5 and 7 gives the same results in F-Score value. Experiment has done by using Conditional Random Field approach. They achieved feature set comprising of word window size 5, digit features, Infrequent word has confirmed the highest F-Score value 87.46% and for feature set comprising of word window size 3, second highest F-Score value is 87.40%. According to them compare to all language independent features with word window size 5 gives highest results. **[30]**

In their paper, they found various NER approaches like CRF, ME, SVM used for various Indian Languages. They encountered some issues in Nested Entity, Agglutinative nature, spelling variations etc by using Rule Based Approach. As a result, they defined some rules for identification of named entities from the corpus of 5000 words of Assamese online articles and found 500 Person names and 250 location names. **[31]**

In their paper, they described some issues for NER in Indian Languages. 90% F-Measure value achieved in Indian languages using various NER approach. In their research, they found some issues in Urdu language. They have used Hidden Markov Model for their research in NER. They achieved 100% performance results of tourism corpuses and 100% accuracy performance in seven sentences in Urdu of BBC news. The found that many tools for NER are available but they all are language dependent. No such tool is there, which is language independent. **[32]**

5. COMPARATIVE STUDY TO IDENTIFY NAMED ENTITIES IN VARIOUS INDIAN LANGUAGES

This table shows the standard measure values of Precision, Recall and F-measure achieved by various authors to find named entities in various Indian languages.

Table 1. Standard Measure Values for Various Indian NER

No	Year	Language	Approach	Precision	Recall	F-Measure
1	2011	Hindi	CRF [5]	71.43%	30.86%	43.10%
2	2011	Hindi	MaxEnt [5]	76.92%	19.80%	31.49%
3	2011	Manipuri	CRF [7]	85.67%	81.12%	83.33%
4	2011	Hindi	Rule Based [5]	96.05%	86.90%	91.25%
5	2013	Tamil	Gazetteer [26]	32.21%	72.79%	44.66%
6	2013	Telugu	Gazetteer [26]	37.73%	42.63%	40.03%
7	2013	Marathi	Gazetteer [26]	53.05%	36%	42.89%
8	2013	Hindi	Gazetteer [26]	75.40%	75%	75.20%
9	2013	English	Gazetteer [26]	76.83%	77.25%	77.04%
10	2013	Bengali	Gazetteer [26]	84.47%	87.56%	85.99%
11	2013	Kannada	HMM [25]	95.10%	94.61%	94.85%
12	2013	Hindi	HMM [21]	96%	96%	96%
13	2013	Telugu	HMM [21]	98.50%	98.50%	98.50%
14	2013	Bengali	HMM [21]	98.60%	98.60%	98.60%
15	2014	Hindi	CRF [24]	25.65%	16.14%	19.98%
16	2014	English	Hybrid [1]	83.65%	83.16%	83.40%
17	2014	Arabic	Machine Learning , Genetic Algorithm [3]	84.80%	67.60%	75.22%
18	2014	Tamil	Dictionary/Supervised classification approach	86.94%	73.87%	79.87%
19	2014	Hindi	Dictionary/Supervised classification approach	92.28%	76.23%	83.49%
20	2014	English	Dictionary/Supervised classification approach	93.82%	80.53%	86.66%
21	2015	English	CRF [13]	77.05%	77.27%	75.81%
22	2015	English	Hybrid [13]	81.99%	78.36%	--
23	2015	Kannada	Machine Learning – Naïve Bayes [2]	83%	79%	81%
24	2016	Tibeto-Burman	SVM [8]	85.14%	90.58%	87.77%
25	2016	Malayalam	SVM [22]	93.33%	72.41%	81.55%

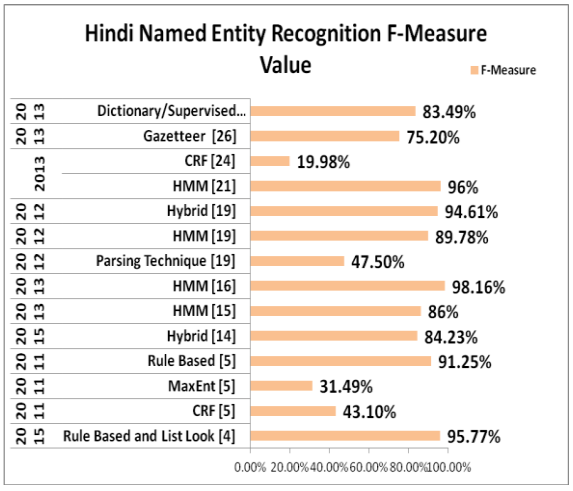

Figure 1 Hindi NER F-Measure

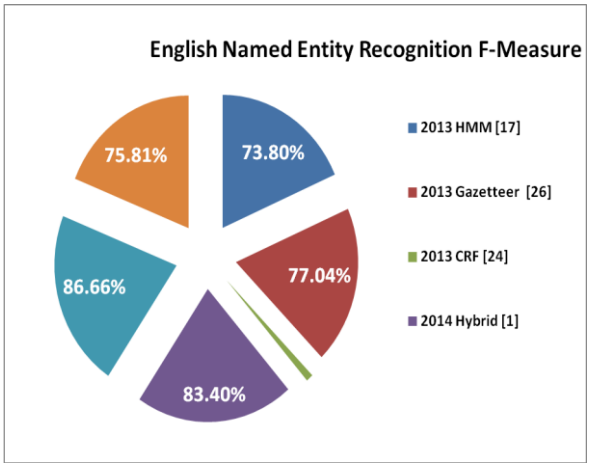

Figure 2 English NER F-Measure

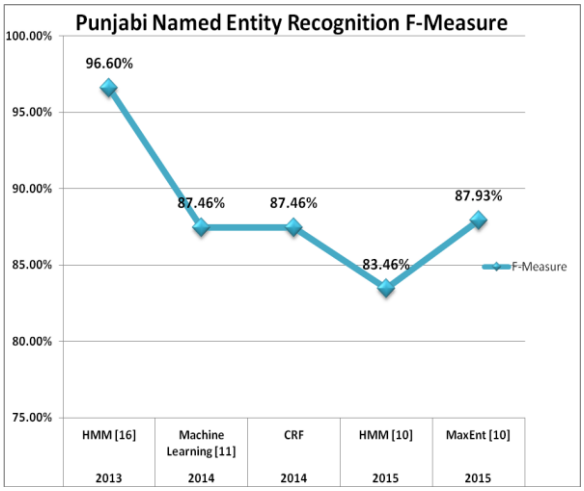

Figure 3 Punjabi NER F-Measure

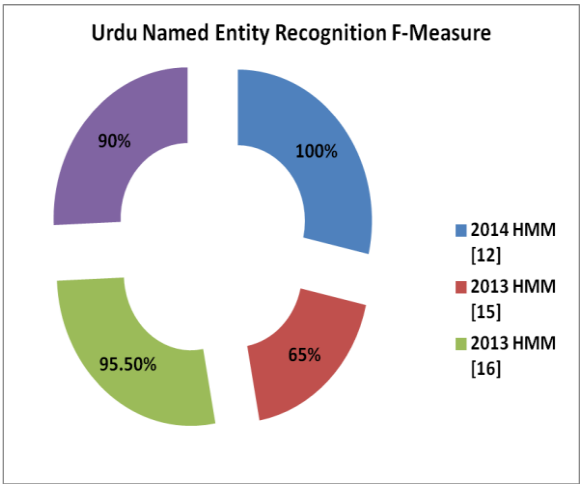

Figure 4 Urdu NER F-Measure

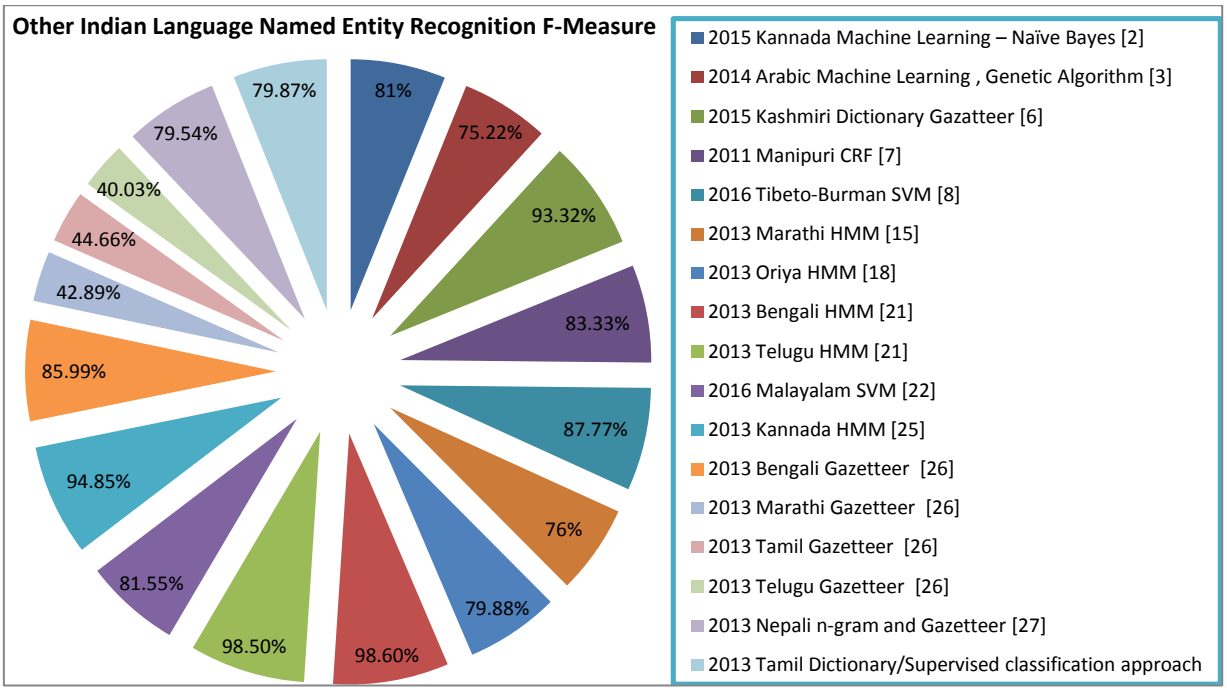

Figure 5 Other Indian Language NER F-Measure

6. PROBLEM DEFINITION

Precise named entity recognition systems are now available for European languages specially English, for South and South East Asian languages, the problem of NER is still not covered. Problem definition shows various challenges available in Indian language NER. They are as follows :

- No capitalization
- Morphologically rich
- Ambiguity
- Lack of standardization and Spell Variations
- Less Resources
- Lack of labeled data
- Agglutinative Nature
- Proper Name Ambiguity
- Lack of easy availability of annotated data

7. CONCLUSION

Indian languages suffer deeply from lack of available annotated corpora, agglutinative nature and diverse writing methodologies, demanding morphology and no concept of capitalization and many more. A massive amount of Named Entity Recognition work has already been done in European languages but not significant amount of work has been done for Indian languages. We conclude that Rules once define for one language could not be applied for other Indian languages, because every Indian language is different from other and they have their own language structure.

8. FUTURE WORK

In this research, we have compared various named entities from Indian Languages. There are so much work has been done for many Indian languages. Among the 22 Indian

constitutional languages, Gujarati is the language which is mainly spoken in Gujarat State. Not sufficient work has been done for Gujarati NER as so many challenges are there in Indian Languages. Not enough data sources are available. The future work will be to create a document for Gujarati language and then find various NEs and then classify them such as person name, location, date, time named entities from it.

9. REFERENCES

[1] Shruti Mathur, Varun Prakash Saxena [2014], "Hybrid Approach to English-Hindi Name Entity Transliteration"

[2] S Amarappa & Dr. S V Sathyanarayana [2015], "Kannada Named Entity Recognition And Classification (Nerc) Based On Multinomial Naïve Bayes (Mnb) Classifier", International Journal on Natural Language Computing (IJNLC) Vol. 4, No.4, August 2015

[3] Ines Boujelben, Salma Jamoussi, Abdelmajid Ben Hamadou [2014], "A hybrid method for extracting relations between Arabic named entities", Journal of King Saud University – Computer and Information Sciences (2014) 26, 425–440

[4] Yavrajdeep Kaur, Er.Rishamjot Kaur [2015], "Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look up Approach", International Journal of scientific research and management (IJSRM, Volume 3 Issue 3 Pages 2300-2306

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

[6] Amir Bashir Malik and Khushboo Bansal [2015], "Named Entity Recognition for Kashmiri Language using Noun Identification and NER Identification Algorithm", Volume-3, Issue-9

[7] Kishorjit Nongmeikapam * [2011], "CRF Based Name Entity Recognition (NER) in Manipuri: A Highly Agglutinative Indian Language", 978-1-4244-9581-8/11 IEEE

[8] Sadiq Hussain et al [2016], "The First Step towards Named Entity Recognition inMising Language", 978-1-4673-9939-5/16 IEEE

[9] Saeeda Naz et al [2014], "Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language", Res. J. App. Sci. Eng. Technol., 8(10): 1272-1278

[10] Jaspreet singh, Gurpreet singh lehal [2015], "Named entity recognition for Punjabi language using Hmm and Memm", IRF International Conference, 8th March 2015, Pune, India, ISBN: 978-93-82702-75-7

11] Amandeep Kaur, Gurpreet Singh Josan, [2015], "Evaluation of Named Entity features for Punjabi Language", Procedia Computer Science 46 (2015) 159 – 166

[12] Nusrat Jahan, Mohammad Alamgir Siddiqui [2014], "Urdu Named Entity Recognition Using Hidden Markov Model", IJACKD Journal of Research, Vol 3, Issuse 1

[13] Sobhana N V [2015], "A Hybrid Approach of Similarity Based SVM and CRF for Named Entity Recognition", International Journal of Advanced Research in Computer Science and Software Engineering 5(6), June- 2015, pp. 229-233

[14] Vaishnavi Singhal, Neha Tyagi [2015], "A Hybrid Approach Of English- Hindi Named-Entity Transliteration", International Journal of Advanced Technology in Engineering and Science, Volume No.03, Special Issue No. 02

[15] Sudha Morwal, Nusrat Jahan [2013], "Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages", International Journal of Advanced Research in Computer Science and Software Engineering 3(4), pp. 671-675

[16] Deepti Chopra and Sudha Morwal [2013], "Detection And Categorization Of Named Entities In Indian Languages Using Hidden Markov Model", International Journal of Computational Science and Information Technology (IJCSITY) Vol.1, No.1

[17] Deepti Chopra and Sudha Morwal, "Named Entity Recognition In English Using Hidden Markov Model", International Journal on Computational Sciences & Applications (IJCSA) Vo3, No.1

[18] Sitanath Biswas *, "A Hybrid Oriya Named Entity Recognition system: Harnessing the Power of Rule", International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (1): Issue (1)

[19] Deepti Chopra, Nusrat Jahan, Sudha Morwal [2012], "Hindi Named Entity Recognition By Aggregating Rule Based Heuristics And Hidden Markov Model", International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012

[20] S Amarappa, Dr. S V Sathyanarayana, "Named Entity Recognition and Classification inKannada Language", V2N1-281-289

[21] Sudha Morwal and Deepti Chopra, "Identification And Classification Of Named Entities In Indian Languages", International Journal on Natural Language Computing

[22] Remmiya Devi G, Veena P V, Anand Kumar M, Soman K P [2016], "Entity Extraction for Malayalam Social Media Text using Structured Skip-gram based Embedding Features from Unlabeled Data", Procedia Computer Science 93 (2016) 547 – 553

[23] Komil Vora, Dr. Avani Vasant, Rachit Adhvaryu, "Named Entity Recognition and Classification for Gujarati Language"

[24] Saatvik Shah et al, "Hierarchical classification for Multilingual Language Identification and Named Entity Recognition"

[25] Vira Bagiya et al, "Entity Extraction from Social Media Text Indian Languages (ESM-IL)"

[26] S Amarappa1and S V Sathyanarayana [2013], "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

[27] Vivekananda Gayen, Kamal Sarkar [2013], "An HMM Based Named Entity Recognition System for Indian Languages"

[28] Arindam Dey, Bipul Syam Prukayastha [2013], "Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach", International Journal of Computer Applications (0975 – 8887) Volume 84 – No 9, December 2013

[29] Abinaya.N, Neethu John, Dr.M.Anand Kumar and Dr.K.P Soman [2014], "Named Entity Recognition for Indian Languages"

[30] Pallavi K P., Srividhya K., Rexiline Ragini John Victor, Ramya M. M. [2015], "Twitter based Named Entity Recognizer for Indian Languages",

[31] Rupal Bhargava, Bapiraju Vamsi, Yashvardhan Sharma, "Named Entity Recognition for Code Mixing in Indian Languages using Hybrid Approach"

[32] Poonam Kumari, Dr. Mahesh Yadav [2015], "Study And Implementation Of Named Entity Recognition In Hindi Language Using Php", International Journal of Engineering Sciences & Research Technology

[33] Sadiq Hussain *[2016], "The First Step towards Named Entity Recognition in Mising Language", 978-1-4673-9939-5/16 IEEE

[34] Maithilee L. Patawar, M. A. Potey [2016], "Named Entity Recognition from Indian tweets using Conditional Random Fields based Approach", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 5

[35] Rayner Alfred, Leow Chin Leong, Chin Kim On, and Patricia Anthony [2014], "Malay Named Entity Recognition Based on Rule-Based Approach", International Journal of Machine Learning and Computing, Vol. 4, No. 3, June 2014

[36] Kamaldeep Kaur, Vishal Gupta [2012], "Name Entity Recognition for Punjabi Language", IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol. 2, No.3

[37] Mai Mohamed Oudah, Khaled Shaalan, "A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach", Proceedings of COLING 2012: Technical Papers, pages 2159–2176,

[38] Nusrat Jahan, Sudha Morwal and Deepti Chopra [2012], "Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 3

[39] Khaled Shaalan, Mai Oudah [2014], "A hybrid approach to Arabic named entity recognition", Journal of Information Science, Vol. 40(1) 67–87

[40] Kommaluri VIJAYANAND and R. P. Seenivasan [2011], "Named Entity Recognition and Transliteration for Telugu Language", Problems of Parsing in Indian Languages 64