STUDY OF NAMED ENTITY RECOGNITION FOR INDIAN LANGUAGES

Hinal Shah, Prachi Bhandari, Krunal Mistry, Shivani Thakor, Mishika Patel and Kamini Ahir

Shrimad Rajchandra Institute of Management & Computer Application,
Uka Tarsadia University, Bardoli, India

ABSTRACT

Named Entity Recognition is a prior task in Natural Language Processing. Named Entity Recognition is a sub task of information extraction and it identifies and classifies proper nouns in to its predefined categories such as person, location, organization, time, date etc. In this document the major focus is given on NER approaches and the work done till now for various languages to identify Named Entities is been discussed. Author have done comparative study to recognize named entity and identified that CRF approach proven best for Indian languages to identify named entity.

KEYWORDS

Named Entity Recognition, Hidden Markov Model, Support Vector Machine, Conditional Random Field, Maximum Entropy

1. INTRODUCTION

Language is a necessary entity for human communication. To make machine understand such kind of natural languages Natural Language Processing (NLP) is used. There has been growing interest in this field of research since the early 1990s. Named-entity recognition is an automated approach and it (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.[11]. Named Entity Recognition is a two-step process: Identification and Classification. In identification stage, proper nouns or NEs are identified. And those NEs are separated out in their classes using classification [12].

Example:

Consider a statement: “Rohit joined IBM as a PG student in Pune on 14th January, 2014”.

Here “Rohit” is classified in PERSON entity, “IBM” is classified in ORGANIZATION entity, “Pune” is classified in LOCATION entity and “14th January, 2014” is classified in DATE entity.

The NER task was first appear in Sixth Message Understanding Conference (MUC-6) Sundheim(1995) [9]. It has provided the benchmark for named entity systems that performed a

DOI : 10.5121/ijist.2016.6202

variety of information extraction tasks [34]. MUC-7 classifies name entity into following categories and sub categories:

- ENAMEX: person, organization, location.
- NUMEX: money, percentage.
- TIMEX: date, time etc. [2][9].

Named Entity recognition is the primary technique in application such as information retrieval, extraction of information, question answering, Machine translation, text summarization [3] [12] [14]. For instance, the key to a question processor is to identify the asking point (who, what, when, where, etc...), so in many cases the asking point corresponds to a NE. In biology text data, the named entity system, can automatically extract the predefined names (like protein and DNA names) from raw documents [4].

Although over the years there has been considerable work done for NER in English and other European languages, the interest in Indian language has been quite low until recently. Indian languages have rich morphology compared to other European languages.

Rule based approach provide better performance and accuracy if the set of rules is under control. The Indian languages are morphologically rich in features and agglutinative in nature, hence rule-based approaches may fail in situations where we need full-fledge NLP based system. To make full-fledge NLP based system we need high number of rules and linguistic expertise which is difficult hence many researchers are now working on statistical and hybrid approaches [6].

Over the past decade Indian language content on various media types such as websites, blogs, email, chats has increased significantly. Content growth is driven by people from non-metros and small cities need to process this huge data automatically especially companies are interested to ascertain public view on their products and processes. This requires natural language processing software systems which identify entities, identification of associations or relation between entities. Hence an automatic Named Entity recognizer is required.

2. LITERATURE REVIEW

In paper [2], authors semantically investigate features and machine learning algorithms for NER in Inflectional Language. About 1000 sentences are collected randomly from different domains of an inflectional language. One third of 1000 sentences were used to train the NER systems and one third for testing. Authors investigated the effects of different types of features including bag-o-characters, word segmentation, part-of-speech, and section information, and different machine learning algorithms including conditional random fields(CRF), support vector machines(SVM), maximum entropy(ME), and structural SVM(SSVM) on the Inflectional language NER task. All classifiers were trained on the training dataset and evaluated on the test, and micro-averaged precision, recall, and F-measure were reported.

In paper [9], authors have explained Named Entity Recognition techniques and how they find named entities from given input. Translation model check the accuracy of target sentences given the source sentence and decoder maximizes the probability of translated text of target language. Authors found that there are some character exists in English which are double meaning like "you" is also written as "u" and the major inaccuracies in the transliteration are due to poor word selection. They proposed a framework for machine learning. In which they have made a NER model to find proper entities and this framework enhances the capability of machine translation. And in it they classified name and entity which is input of language model. Therefore language model perform efficiently.

In paper [10], authors have done surveys of various approaches used for identification of Named Entities for Indian languages. They reveal that hybrid models which combine of both rules and a machine leaning algorithm perform better for Indian languages. And they also found that work done for NER in Indian languages is very less compared to what all need to be done.

In paper [11], authors have reviewed about Naming Entity Recognition and various proposed approaches for NER. They found that existing methods have not a proper corpus for Hindi language. And they said that they will prepare a Hybrid approach for NER in Hindi for enhancing the accuracy. In future authors will create a new system for NER in Hindi to improve the accuracy of existing system and produce high accurate result.

In paper [15], authors depict Named Entity Recognition system for Hindi language. And they have proposed NER system for Hindi language using Hybrid approach which is the combination of rule based approach and list look up approach. And in this paper they have identifies three new named entities that is money value, direction values and animal/bird entities. Authors claimed that proposed system works on new rule that is "no name entity rule". No name entity rule is used which improves or modifies the existing rules. This rule analyses the various NER systems and results are compared with the existing approaches. Different tables have been created in database for Hindi language and named entities have been extracted from these tables in list look up approach. And the accuracy of proposed system is 95.77%.

In paper [27], authors have use Support Vector Machine approach for NER for Nepali text. They have made a system which is efficient feature extraction and the comprehensive recognition techniques. They found that the system can learn well from the small set of training data and NE recognition is increased when the size of training data is increased. They also recognize that Nepali NER is very complex as it's a free order language and due to lack of resources.

In paper [34], authors have proposed a named-entity recognition system that combines named entity extraction with a simple form of named-entity disambiguation. They have used some simple yet highly effective to perform named-entity disambiguation. They compare the performance of their unsupervised system with that of a basic supervised system, using the MUC 7 NER corpus. And also show that their technique is general enough to be applied to other named-entity types, such as car brands, or bridge names. To support this claim, they include an experiment with car brands. And in future they will create a system that can recognize named entities in a given document without training (supervised learning) or manually constructed gazetteers.

In paper [37], 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.

In paper [38], 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.

3. APPROACHES OF NER [2]

3.1. Rule based/linguistic Approach
3.2. Statistical Approach
3.3. Hybrid Approach

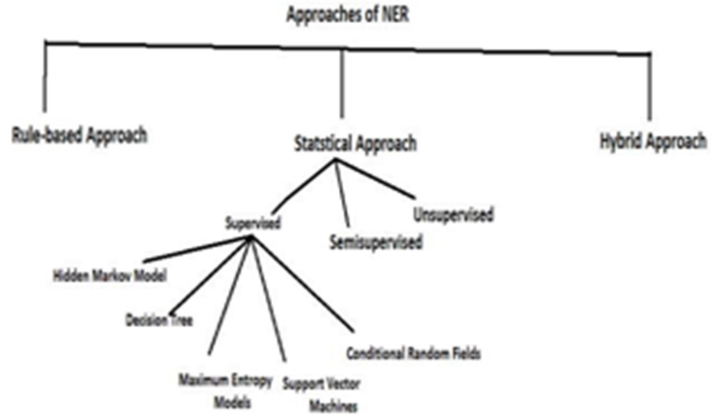


Figure 1. Approaches of NER

3.1. Rule based/linguistic Approach

Ralph Grishman in 1995 developed rule-based NER systems.

The Rule based approach is a classical approach to NER. It uses rules manually written by linguists. Rule based systems parse the source text and produce an intermediate representation which may be a parse tree or some abstract representation. To use this, a large gazetteer list has to be built for different Named Entity classes.

3.2. Statistical Approach

It is automated approach in which mathematical logic and formulas are used for classifying the named entity. In statistical approach a corpus is initially studied .Corpus is a large collection of a texts .Based on the corpus a training module is made where the system is trained to identify the named entity and then on their occurrences in the corpus with particular context and class a probability value is counted. Every time when text is given based on the probability value the result is fetched. For better effectiveness large amount of annotated training data are required [1]. For statistical approach very less amount of human labor is required as the system is more dependent on corpus developed for it.

This approach has following types:

3.2.1. Supervised Approach

This approach uses pre-annotated corpus to train the system. The System here reads the annotated corpus, memorizes it and used the same to identify the entities from the input text. The supervised learning approach requires separating labeled training data to construct a statistical model, but it can't achieve a good performance without a large amount of training data [8].

Example:

Learn how to predict if a patient will get cancer based on past data of labeled patients?

Here first, all the data from past patients who were suffering from cancer will be gathered. Means what symptoms is common between all patient is identified and make it as response variable. And when the next patient will come and if that patient have same symptoms, then it is said that, patient is suffering from cancer.

Different techniques used in Supervised Method

3.2.1.1. Hidden Markov Model

HMM is the earliest model applied for solving NER problem by Bikel et al.(1999) for English [7].

HMM is a generative model in which states are hidden or unobserved [9][11][22]. This model assigns a join probability to paired observation and label sequence. In a regular markov model the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters [29]. In HMM states are not directly visible, but output depends on the state and is visible. Each state has a probability distribution over the possible outcome tokens so, that's why the sequence of tokens generated by HMM gives some information about the sequence of states [29]. It is based on the Markov Chain Property i.e. the probability of occurrences of the next state is depending on the just previous state [28]. HMM model is used to sequentially process where the observed sequence is actually depend on a separate process.

Example

Based on the past analysis of weather data user can predict the weather of tomorrow? Assuming only four possible weather states

- Sunny
- Cloudy

- Rainy
- Snowing

In past the determination of the 4 states of weather and from those stored predictable state user can determine the weather of tomorrow.

3.2.1.2. Decision Trees

The first representative of decision tree is Sekine in who implemented it in New York University system in 1998 [7].

Decision tree is a popular and powerful tool for categorizing and forecast [1]. Decision tree is a classifier in the form of a tree structure where each node represents a leaf node. Indicates the values of the output attribute of languages, decisions, require some text to be supported out on a single attribute value with one branch and sub-tree for each possible come the output of the text. In training the decision tree, a series of training sets is used, each with differing number of sentences, in order to examine the effect of training size on overall performance [33]. In this approach some reorganization tools are applied to the untagged training purpose so that named entities are retrieved. Now, match this named entities obtained with the actual answer key provided by the humans. If the named entity is same as answer key it is referred to as positive example else it is known as negative example [28][30]. It is an inductive approach to acquire knowledge on classification.

3.2.1.3. Maximum Entropy Models (MaxEnt)

MaxEnt model was first applied by Andrew Borthwick in 1999. He developed MENE system [8][19].

Maximum Entropy Model is discriminative model. The MaxEnt framework is a powerful learning model, which has been successfully employed for many natural languages processing task. MaxEnt principle seeks the distribution that maximizes the entropy of the distribution subject to the known constraints [4]. It has multiple features of a word which are extracted from one word and handle this dependency for the extended term [1][11]. It is a model for list bias which considers all known fact is a one which maximizes entropy. Each source has a model of exponential that takes the inspection feature as input and distribution over possible next state as a output. The output result labels are related with that states [1]. The main advantage of MaxEnt is robust and statistically efficient [19], while still allowing for easy representation and incorporation of different features [4].For easiness of training data of MaxEnt model bootstrapping is used to help in supervised learning [23].

3.2.1.4. Support Vector Machine

SVM was first introduced by Cortes and Vapnik in 1995[2][9].

SVM is a popular machine learning approach based on the structural risk minimization of statistical learning theory [4]. The main objective of this approach is to find whether a specific vector belongs to a particular target class or not [28]. SVM is a kind of binary classifiers that search for an optimal separating hyper-plane between positive and negative samples/classes and make decisions based on support vectors which are selected as the only effective examples in the training sets[4][9].This approach also computes the distance of every vector from the hyper-plane known as Margin. Large margin suggest that the distance between the hyper-plane and the point from either instances is maximum. The points closest to hyper-plane on either side are known as support vectors. It is probably the most widely used kernel learning algorithm. The SVMs are known to robustly manage large feature sets and to develop models that maximize their generalizability [1].

3.2.1.5. Conditional Random Fields (CRF)

CRF was introduced by Lafferty et al in 2001 as a statistical modeling tool for pattern recognition and machine learning using structured prediction. McCallum and Li in 2003 proposed a feature induction method for CRF[9].

CRF is a type of discriminative probability model [13][17]. It has all the advantage of Maximum Entropy instead the label bias problem [1]. CRFs are undirected graphical models and also called random fields which are used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes[22][28]. In CRF each vertex represents a random variable whose distribution is to be inferred, edges represents dependencies between two random variables [13]. CRF were these models are commonly used for decoding test instances where only the context is observed. The approach is achieved empirical success in many NLP problems. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences [10]. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference [10].

3.2.2. Semi-Supervised Approach

Semi-supervised learning use both labeled and unlabeled corpus to create their own hypothesis [9]. It is falls between unsupervised learning and supervised learning. In Semi-Supervised learning, some initial entities called seeds are trained into the system. The system then searches for this seeds and identified them. Then the system tries to identify other entities that appear in similar context where it identifies the seeds. The learning process is then again applied using this new context [13]. The main technique for SSL is called bootstrapping and involves a small degree of supervision [1], such as a set of seeds, for starting the learning process.

3.2.3. Unsupervised Approach

Collins et al. discusses an unsupervised model for named entity classification by use of unlabeled examples of data, Koim et.al. Proposes an unsupervised named entity classification models and their ensembles that uses a small-scale named entity dictionary and an unlabeled corpus for classifying named entities [31][32].

Unsupervised learning method is another type of machine learning model, where an unsupervised model learns without any feedback or past data [25].Unsupervised learning approaches doesn’t expect any implicit or structural information about the data they are processing [1]. Many languages don’t have large annotated corpus available at their disposal. To deal with lack of annotated text across domains and languages, unsupervised techniques for NER have been proposed. Labels are not given and trying to extract information in general out of data. In unsupervised learning the goal is to give representations from data. These representations are then be useful for data comparison, classifying, decision making and other purposes [8][22]. Unsupervised learning technique that promises fast deployment for many entities types without the prerequisite of an annotated corpus.

The main technique behind unsupervised learning is clustering. In clustering large number of entities occurring in similar context is grouped into one unit and the system is made to learn this cluster. Whenever it is implemented it looks for the clusters and entities resembling similar contexts are identified as ones in the trained cluster [13].

3.3. Hybrid Approach

In Hybrid approach, two approaches can be merged together and make a new methods using strongest points from each method [8].It improves the performance of NER Systems. It can be the combination of linguistic and statistical models like Gazetteer list and HMM, HMM and CRF, etc. Although this type of approach can get better result than some other approaches, but the weaknesses of handcraft Rule-based NER remains the same that is when there is a need to change the domain of data.

4. EVALUATION MATRICS

Evaluation Metric mathematically defines the measurement of the system's performance against human-annotated, gold standard [18]. Here for every experiment before checking the performance of the system, a human tagged test data is prepared to evaluate the system [20].The system's performance is measured in terms of Accuracy (A), Precision (P), Recall (R) and harmonic mean of precision and recall F-Measure (F).

• ACCURACY

Accuracy is the measure of correct words found from the total name entities found. And the quality output can be found using following formula,

$$\text{Accuracy (\%)} = \frac{\text{Correct Words}}{\text{Total name entities}} * 100$$

• Precision

The precision measures the number of correct NEs, obtained by NER system, over the total number of NEs extracted by NER system [22].

$$\text{Precision (P)} = \frac{\text{number of correct responses}}{\text{number of responses}}$$

• Recall

The recall measures the number of correct NEs, obtained by NER system over the total number of NEs in a text that has been used for testing [22].

$$\text{Recall (R)} = \frac{\text{number of correct responses}}{\text{number of correct in key}}$$

• F – Measure

The F-measure represents harmonic mean of precision and recall [22].

$$\text{F-Measure} = \frac{RP}{1/2(R+P)}$$

The term response is used to denote “answer delivered by a name-finder”, the term key or key file is used to denote “an annotated file containing correct answers”.

5. PROBLEM FACED IN INDIAN LANGUAGES [8][12][20][24][26][27][34]

• No capitalization

In English language, capitalization plays a vital role, as it helps to recognize proper noun. But in Indian languages no capitalization concept is there

• Morphologically rich

Indian languages are very rich morphologically, so they identification of root is very difficult.

• Ambiguity

Ambiguity between common and proper noun.

• Lack of standardization and Spell Variations

One of the main problems with Indian languages is that in India that different people spells same entity differently.

• Less Resources

As not much work had been done in NER with Indian languages, so there exist so less resources and tools required for pre-processing such as part-of -speech tagging and chunking that helps to recognize NE. Or tools which are available have poor performance.

• Indian languages are relatively free order language

• Lack of labeled data

For Indian languages not more corpus or training data is available.

- **Agglutinative Nature**

  Agglutinative means that some additional features can be added to the word to add more complex meaning.

Example:

Let us consider the root word as Rup and suffix as Ali(GOD) then combination these two words becomes Rupali (a name of person) as new word.

- **Proper Name Ambiguity**

  Ambiguity in proper name present in Indian language as in English. The names like White are ambiguous in English-name or color. Indian Person Names are more diverse compared to the other languages and a lot of these words can be found in the dictionary with some other specific meanings. There is a surprising amount of ambiguity even among proper names.

Example:

People vs. Companies: Vimal etc.

People vs. Locations: Gandhinagar (person vs. city) People vs. Organizations: Nirma (person vs. university).

Acronyms vs. Organizations: MRI (Magnetic Resonance Imaging vs. Mental Research Institute).

People vs. Months: Shrawan (month of Indian Calendar).

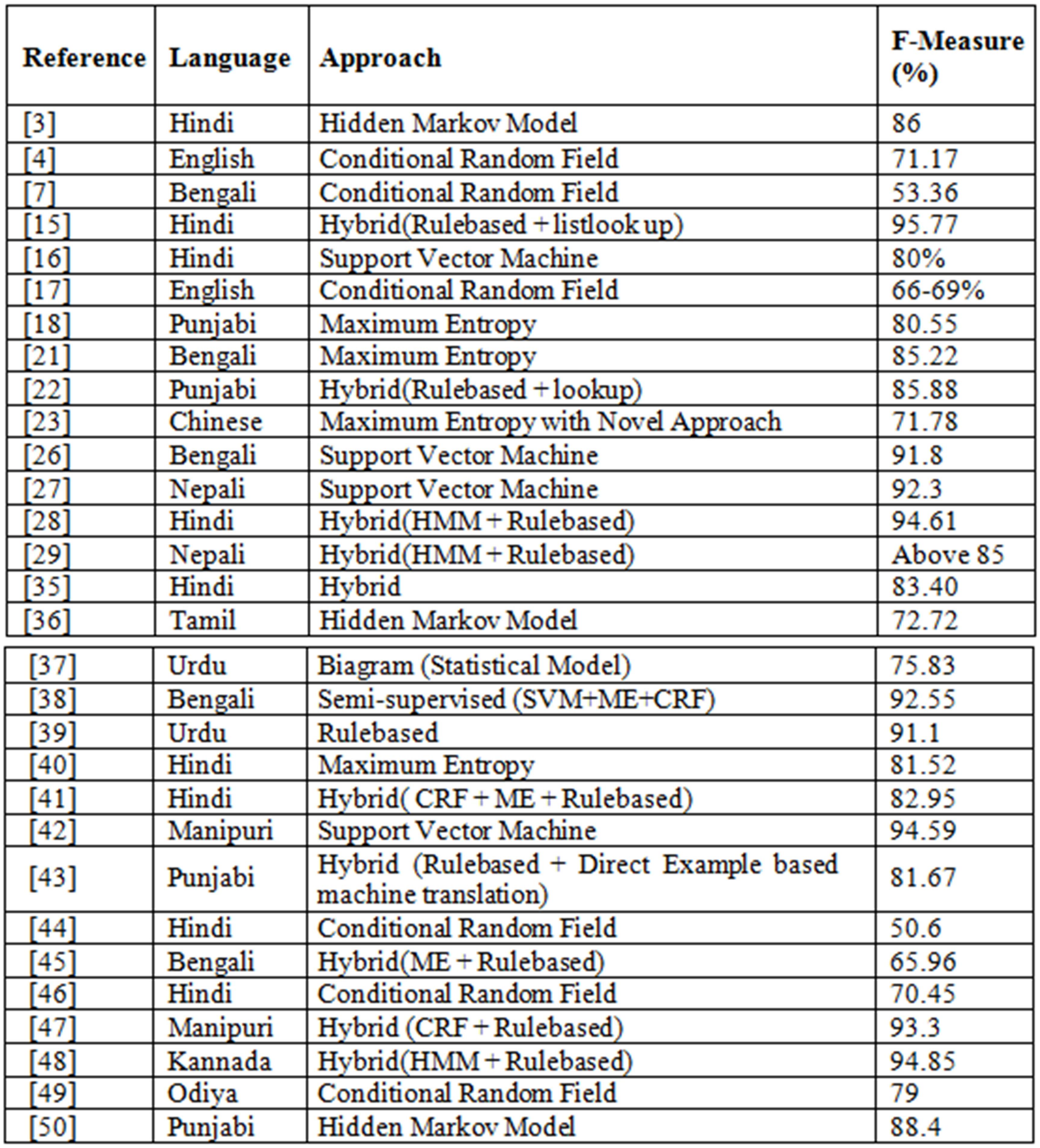
- **Lack of easy availability of annotated data**

  As less work is done with NER in Indian languages, it gets hard to get annotated data and corpus in the community.

6. COMPARATIVE STUDY TO IDENTIFY NAMED ENTITIES IN VARIOUS LANGUAGES

This table shows the F-measure achieved by various authors to find named entities in different languages

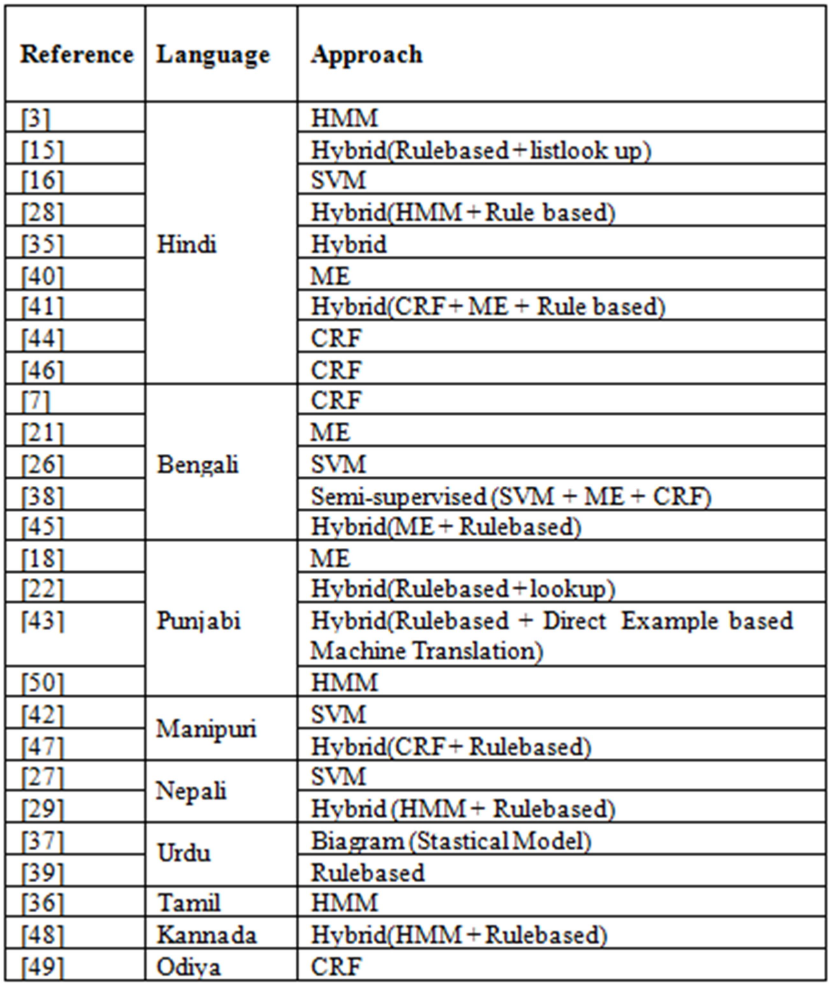
Table 1. F-Measure of various languages



Reference	Language	Approach	F-Measure (%)
[3]	Hindi	Hidden Markov Model	86
[4]	English	Conditional Random Field	71.17
[7]	Bengali	Conditional Random Field	53.36
[15]	Hindi	Hybrid(Rulebased + listlook up)	95.77
[16]	Hindi	Support Vector Machine	80%
[17]	English	Conditional Random Field	66-69%
[18]	Punjabi	Maximum Entropy	80.55
[21]	Bengali	Maximum Entropy	85.22
[22]	Punjabi	Hybrid(Rulebased + lookup)	85.88
[23]	Chinese	Maximum Entropy with Novel Approach	71.78
[26]	Bengali	Support Vector Machine	91.8
[27]	Nepali	Support Vector Machine	92.3
[28]	Hindi	Hybrid(HMM + Rulebased)	94.61
[29]	Nepali	Hybrid(HMM + Rulebased)	Above 85
[35]	Hindi	Hybrid	83.40
[36]	Tamil	Hidden Markov Model	72.72
[37]	Urdu	Biagram (Statistical Model)	75.83
[38]	Bengali	Semi-supervised (SVM+ME+CRF)	92.55
[39]	Urdu	Rulebased	91.1
[40]	Hindi	Maximum Entropy	81.52
[41]	Hindi	Hybrid(CRF + ME + Rulebased)	82.95
[42]	Manipuri	Support Vector Machine	94.59
[43]	Punjabi	Hybrid (Rulebased + Direct Example based machine translation)	81.67
[44]	Hindi	Conditional Random Field	50.6
[45]	Bengali	Hybrid(ME + Rulebased)	65.96
[46]	Hindi	Conditional Random Field	70.45
[47]	Manipuri	Hybrid (CRF + Rulebased)	93.3
[48]	Kannada	Hybrid(HMM + Rulebased)	94.85
[49]	Odiya	Conditional Random Field	79
[50]	Punjabi	Hidden Markov Model	88.4

This table depicts the work till now done for Indian languages using various approaches.

Table 2. Approaches used for Indian Languages



Reference	Language	Approach
[3]	Hindi	HMM
[15]		Hybrid(Rulebased+listlook up)
[16]		SVM
[28]		Hybrid(HMM+Rule based)
[35]		Hybrid
[40]		ME
[41]		Hybrid(CRF+ME + Rule based)
[44]		CRF
[46]		CRF
[7]	Bengali	CRF
[21]		ME
[26]		SVM
[38]		Semi-supervised (SVM + ME + CRF)
[45]		Hybrid(ME+Rulebased)
[18]	Punjabi	ME
[22]		Hybrid(Rulebased+lookup)
[43]		Hybrid(Rulebased + Direct Example based Machine Translation)
[50]		HMM
[42]	Manipuri	SVM
[47]		Hybrid(CRF+Rulebased)
[27]	Nepali	SVM
[29]		Hybrid (HMM+ Rulebased)
[37]	Urdu	Biagram (Stastical Model)
[39]		Rulebased
[36]	Tamil	HMM
[48]	Kannada	Hybrid(HMM+Rulebased)
[49]	Odiya	CRF

Author studied 50 different papers to recognize named entity in different languages which uses various approaches and conveys that,

To identify named entity, Most of the researchers have used Hybrid or CRF Approach. The results of CRF can be improved if used with ME & Rule based.

7. CONCLUSION

A huge amount of work has already been done in Named Entity Recognition in English and other European languages but not considerable amount of work has been done for Indian languages. CRF approach is performed best for Indian languages to find named entities. HMM is not much used to identify named entities. If CRF is used with POS (Part of Speech) and prefix and suffix feature then the performance of NER improves.

REFERENCES

[1] N. Kaur & Y. Kaur, (2014) "Optimized name entity recognition of machine translation", IJRASET, Vol. 2, pp24-30.

[2] A. Dey, J. Abedinand & B. Purkayastha, (2014) "A Comprehensive Study Of Named Entity Recognition On Inflectional Languages", Int. J. of Advanced Research in Computer Science and Software Engineering, Vol. 4, pp696-701.

[3] S. Morwal & N.Jahan, (2013) "NER using HMM: an experimental results on hindi, urdu and marathi languages", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3.

[4] H. Wang, T. Zhao, H. Tan & S. Zhang, "Biomedical Named Entity Recognition based on classifiers ensemble", International Journal of Computer Science and applications, Vol.5, pp1-11.

[5] D. Nadeau, P. Turney & S. Matwin, "Unsupervised Named-Entity Recognition: generating gazetteers and resolving ambiguity", unpublished.

[6] G.Garje and G.Kharate, (2013) "Survey of machine translation systems in India", International Journal On Natural Language Computing(IJNLC), Vol. 2, pp47-67.

[7] A. Ekbal, R.Haque, A.Das, V. Poka & S. Bandyopadhyay, (2008) "Language independent Named Entity Recognition in Indian languages", IJCNLP-08, pp33-40.

[8] D. Kaur, V. Gupta, (2010) "A survey of Named Entity Recognition in English and other Indian languages", International Journal of Computer Science Vol. 7, pp239-245.

[9] R. Sharnagat, "Named Entity Recognition: A Literature Survey", unpublished.

[10] P. Hiremath & B. R. Shambhavi, (2014) "Approaches to Named Entity Recognition in Indian Languages: A Study", International Journal of Engineering and Advanced Technology, Vol. 3, pp191-194.

[11] Y.Kaur & R.Kaur, (2014) "A review Name Entity Recognition in Hindi", International Journal of Computer Engineering and Application, Vol. 7, pp1-8.

[12] P. Sharma, (2011) "Named Entity Recognition: A Survey for the Indian Languages", Problems of parsing in Indian Languages.

[13] H. Mahanta, (2013) "A study on the approaches of developing a Named Entity Recognition tool", International Journal of Research in Engineering and Technology, Vol. 2, pp58-61.

[14] "A Hybrid Approach for Named Entity Recognition in Indian Languages", unpublished.

[15] Y. Kaur & R. Kaur, (2015) "Named Entity Recognition system for Hindi Language using combination of rule based approach and list look up approach", International Journal of scientific research and management, Vol. 3, pp2300-2306.

[16] N. Kaur & Y. Kaur, (2014) "Optimized machine translation from English to Hindi by improving Name Entity Recognition in NLP", International Journal of scientific research Engineering & Technology, Vol. 3, No. 4.

[17] N.Ponomareva, P.Rosso, F. Pla & A. Molina, "Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task", unpublished.

[18] G. V. Raju, K.Kumar & B.Srinivasu, "Named Entity Recognition for Telugu using MaxEnt model", Journal of Theoritical and Applied Informational Technology.

[19] A. Borthwick, (1999) "A Maximum Entropy Approach to Named Entity Recognition", Degree of Doctor of Philosophy Computer Science Department New York University.

[20] S Amarppa & S.Sathyanarayana, "Named entity recognition and classification in kannada language", International Journal of Electronics and classification in Kannada Language.

[21] M. Hasanuzzaman, A.Ekbal & S. Bandyopadhyay, (2009) "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi", International Journal of Recent Trends in Engineering, Vol. 1.

[22] K. Kaur & V. Gupta, (2012) “Name Entity Recognition for Punjabi Language”, International Journal of Computer Science and Information Technology & Security, Vol. 2, No.3.

[23] D. Feng, Y. Lv & M. Zhou, “A New Approach for English-Chinese Named Entity Alignment”, unpublished.

[24] “Named Entity Recognition for Indian Languages”, unpublished.

[25] A. Mansouri, (2008) “Named Entity Recognition approaches”, International Journal Of Computer Science and Network Security, Vol. 8, pp339 – 344.

[26] A. Ekbal & S. Bandyopadhyay, “Bengali Named Entity Recognition using Support Vector Machine”, unpublished.

[27] S. Bam & T. B. Shahi, (2014) “Named Entity Recognition for Nepali text using Support Vector Machine”, Intelligent Information Management, pp21-29.

[28] D.Chopra, N.Jahan & S. Morwal, (2012) “Hindi Named Entity Recognition By aggregating rule based heuristic and Hidden Markov Model”, International Journal of Information Sciences and Techniques, Vol. 2.

[29] A. Dey, A. Paul & B. Purkayastha, (2014) “Named Entity Recognition for Nepali language: A Semi Hybrid Approach”, International Journal of Engineering and Innovative Technology, Vol. 3.

[30] A. Pillai & Pallavi, (2013) “Named Entity Recognition for Indian Languages: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, pp1215-1218.

[31] D. Collins & Y. Singer, “Unsupervised models for Named Entity Classification”, unpublished.

[32] J. Kim, I. Kang & K. Choi, “Unsupervised Named Entity Classification models and their ensembles”, unpublished.

[33] M. Fleischman, (2001) “Automated Subcategorization of named entities”, Conference of the European Chapter of Association for Computational Linguistic, pp25-30.

[34] S. Kulkarni, (2014) “A survey on Named Entity Recognition for South Indian Languages”, National Conference on Indian Language Computing.

[35] S. Mathur & V. Saxena, “Hybrid Approach to English-Hindi Name Entity Transliteration”, unpublished.

[36] S. Pandian, T. Geetha & K. Pavithra, (2008) “Hybrid Three-stage Named Entity Recognizer for Tamil”, INF02008.

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

[38] A. Ekbal & S. Banyopadhyay, (2009) “Named Entity Recognition Using Appropriate Unlabeled Data, Post-Processing and Voting”.

[39] K. Riaz, “Rule-based Named Entity Recognition in Urdu”, unpublished.

[40] S. Saha, S. Sarkar & P. Mitra, “A Hybrid feature set based Maximum Entropy Hindi Named Entity Recognition”, International Journal of Advanced Research in Computer Science and Software Engineering.

[41] S. Srivastava, M. Sanglikar & D. Kothari, (2011) “Named Entity Recognition system for Hindi Language: A Hybrid Approach”, International Journal of Computational Linguistics, Vol. 2.

[42] T. Singh, K. Nongmeikapam, A. Ekbal & S. Bandyopadhyay, “Named Entity Recognition for Manipuri using Support Vector Machine”, 23rd Pacific Asia Conference on Language, Information and Computation, pp811–818.

[43] H. Kaur & Dr. V. Laxmi, (2013) “A web based English to Punjabi MT system for News Headlines”, A Web Based English to Punjabi MT System for news headlines, Vol. 3.

[44] K. Gali, H. Surana, A. Vaidya, P. Shishtla & D. Sharma, “Aggregating machine learning and rule based heuristics for Named Entity Recognition”, unpublished.

[45] S. Saha, S. Chatterji, S. Dandapat, S. Sarkar & P. Mitra, “A Hybrid Approach for Named Entity Recognition in Indian Languages”, IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp17–24.

[46] R. Sharma & V. Goyal, (2011) “Name Entity Recognition Systems for Hindi using CRF approach”, International Conference on Information Systems for Indian Languages 2011, pp31-35.

[47] L. Jimmy & Darvinder Kaur, "Named Entity Recognition in Manipuri: A Hybrid Approach", Springer International Publishing AG, Part of Springer Science+Business Media, Vol. 8105, pp 104-110.

[48] S Amarappa & S. Sathyanarayana, "A Hybrid approach for Named Entity Recognition, classification and extraction in Kannada Documents", unpublished.

[49] R. Balabantaray, S Lenka & D. Sahoo, (2013) "Name Entity Recognizer for Odia using Conditional Random Fields", Indian Journal of Science and Technology, Vol. 6.

[50] D. Chopra & S. Morwal, (2012) "Named Entity Recognition in Punjabi using Hidden Markov Model", International Journal of Computer Science & Engineering Technology, Vol. 3.

AUTHORS

Hinal Shah received BCA(Bachelor of Computer Application degree from B. V. Patel Inst. Of BMC & IT, Bardoli, Gujarat. Currently she is pursuing her MCA from SRIMCA, UTU, Bardoli, Gujarat. Her Subject of interest includes Natural Language Processing.

Prachi Bhandari received BCA(Bachelor of Computer Application degree from B. V. Patel Inst. Of BMC & IT, Bardoli, Gujarat. Currently she is pursuing her MCA from SRIMCA, UTU, Bardoli, Gujarat. Her Subject of interest includes Natural Language Processing.

Krunal Mistry received BCA(Bachelor of Computer Application degree from B. V. Patel Inst. Of BMC & IT, Bardoli, Gujarat. Currently He is pursuing his MCA from SRIMCA, UTU, Bardoli, Gujarat. His Subject of interest includes Natural Language Processing and Artificial Intelligence.

Shivani Thakor received BCA(Bachelor of Computer Application degree from B. V. Patel Inst. Of BMC & IT, Bardoli, Gujarat. Currently she is pursuing her MCA from SRIMCA, UTU, Bardoli, Gujarat. Her Subject of interest includes Natural Language Processing.

Mishika Patel received BCA(Bachelor of Computer Application degree from B. V. Patel Inst. Of BMC & IT, Bardoli, Gujarat. Currently she is pursuing her MCA from SRIMCA, UTU, Bardoli, Gujarat. Her Subject of interest includes Natural Language Processing.

Kamini Ahir received BCA(Bachelor of Computer Application degree from Vivekanand College, Surat, Gujarat. Currently she is pursuing her MCA from SRIMCA, UTU, Bardoli, Gujarat. Her Subject of interest includes Natural Language Processing.