

Survey of Named Entity Recognition Techniques for Various Indian Regional Languages

Shrutika Kale

Department of Computer Engineering, Pillai
College of Engineering, University of Mumbai,
India

Sharvari Govilkar

Department of Computer Engineering, Pillai
College of Engineering, University of Mumbai,
India

ABSTRACT

Named entity recognition is a process and study of identification of entities that are proper nouns and classifying them to their appropriate pre-defined class, also called as tag. Named entity recognition is also called as entity chunking, entity identification and entity extraction. It is a sub task of information extraction, where structured text is extracted from unstructured text. Popular applications of NER are machine translation, text mining, data classification, question answering system. This paper presents a survey of different NERC techniques, approach, observations and features for Indian regional languages such as Hindi, Marathi, Bengali, Punjabi, Malayalam, Bengali, Kannada, Telugu, Tamil, Urdu, Oriya.

General Terms

Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.

Keywords

Named entity recognition, Named Entity, Machine learning techniques, Statistical Model, NERC

1. INTRODUCTION

In the world of exponentially growing web data, information is rapidly increasing. When user requests for any information, the response to the request is in an unstructured form. The conversion of unstructured information to structured information is called Information Extraction. The goal of information extraction system is to find and understand limited relevant parts text; normally we are going to run it into a context where system gathers information from across many pieces of text. The goal of doing this information gathering is to produce some sort of structural representation to the relevant information. Subtasks of IE are named entity recognition, noun phrase coreference resolution, semantic role recognition, entity relation recognition, and date and time line recognition [1]. Named Entity Recognition is a process where all the named entities are identified the proper nouns and classified into their pre-defined appropriate class. Very important applications of Named Entity Recognition (NER) are Information Retrieval, Information Extraction, Question Answering, Machine Translation, POS Tagging, Text Mining and Automatic Summarization etc. Various approaches are imposed for NER in various Indian and European languages. These approaches involve rule based, list lookup and machine learning.

This paper is organized into 6 sections. Section 1 presents the introduction. Section 2 gives the introduction about NER; section 3 presents the related work, the NER approach is explained in section 4, in section 5 NER techniques are

discussed, comparative analysis of all the techniques is done in section 6, and in section 7 it is concluded.

2. NAMED ENTITY RECOGNITION

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Named entity recognition (NER) or tagging is the task of finding names such as organizations, persons, locations, etc. in text. Since whether or not a word is name and the entity type of a name are determined mostly by the context of the word as well as by the entity type of its neighbors, NER is often posed as a sequence classification problem and solved by methods such as hidden Markov models (HMM) and conditional random fields (CRF).

In NER the aim is to distinguish between character sequence that represent noun phrases and character sequence that represents normal text [5]. Named entities are the pronouns which are the names of used for the referring persons, organizations, locations, etc. these named entities are classified into different categories as shown in Fig. 1

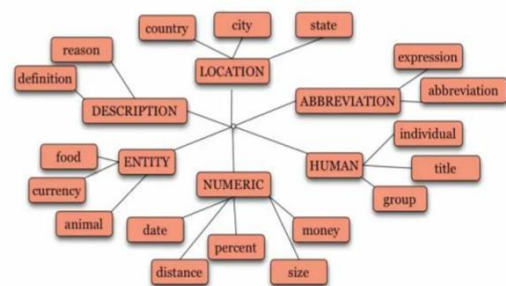


Figure 1. A single Named Entity split into more specific Named Entities [2]

The example given below shows the named entities and their classes.

Raman [PER] joined Symbiosis [ORG] in Pune [LOC] on 12th [DATE] Jan [MONTH] 2012 [YEAR] for a 3 [NUMB] course.

3. RELATED WORK

In this section we cite various Named Entity Recognition techniques applied on different Indian regional languages to identify the proper nouns and classify these nouns in their suited tagset.

Deepthi Chopra, Nisheeth Joshi and Iti Mathur [1] have presented a paper in which they have first described what is

NER, why it is difficult for Indian languages and then in the later part the authors have used Hidden Markov Model (HMM) Technique for identification on the entities in Hindi. The HMM technique enabled system proposed by the authors can identify 12 tags, the system performed training on 2343 tokens on tested on 105 tokens.

Nita Patil, Ajay S. Patil and B. V. Pawar [2], have elaborated the need of NE recognition for Marathi and discusses issues and challenges involved in NE recognition tasks for Marathi language. It also explores various methods and techniques that are useful for creation of learning resources and lexicons that are important for extraction of NEs from natural language unstructured text.

S. Amarappa and Dr. S. V. Sathyanarayana [3] have successfully implemented the system using CRF technique for NERC. They have also highlighted the drawbacks in identifying and classifying the entities for Indian regional languages as they are morphologically rich and have many affixes. Total of 14 main tags have been described by the authors for Kannada language, and these 14 tags are broken down into 23 tags have been described using CRF as it a supervised machine learning technique (SL).

Gowri Prasad, K.K. Fousiya, Dr. M. Anand Kumar and Dr. K.P. Soman [4], Malayalam being one of the oldest languages lacks an efficient NER system and so the growth in the field of information extraction for Malayalam language is very slow. The authors here have handled 10 tags and have identified and classified the beginning and ending of each tagged word. The proposed system does the process of NER with reasonable efficiency in terms of F-measure values.

Abinaya.N, Neethu John, Dr.M.Anand Kumar and Dr.K.P Soman [5], this paper describes Named Entity Recognition systems for English, Hindi, Tamil and Malayalam languages. This paper presents the working methodology and results on Named Entity Recognition (NER) Task of FIRE-2014. In this work, English NER system is implemented based on Conditional Random Fields (CRF) which makes use of the model learned from the training corpus and feature representation, for tagging. The other systems for Hindi, Tamil and Malayalam are based on Support Vector Machine (SVM).

Sudha Morwal and Deepti Chopra [6], authors in this paper have performed NER in Hind. The authors in this paper have performed NER in Hindi, Bengali and Telugu. For Hindi, they have considered Tourism Domain Corpus which was developed at Banasthali Vidyapith. For Hindi, Bengali and Telugu Corpus from NLTK Indian Corpora. The Hindi text that we took from NLTK was related to Politics and Sports. Bengali and Telugu texts from NLTK were of general domain.

Sudha Morwal and Deepti Chopra [7], the authors here have used HMM based NER system for Hindi, Urdu and Marathi, which works in three phases. The first phase is referred to as "Annotation phase" that produces tagged or annotated document from the given raw text, document or corpus. The second phase is referred to as "Training Phase". They have handled Per, Loc, Country, State, City, Month and other, tags.

S Amarappa and Dr. S V Sathyanarayana [8],building NERC for Kannada is challenging. It seeks to classify words which represent names in text into predefined categories like person name, location, organization, date, time etc. In this paper we have developed an algorithm based on supervised learning

techniques that include Hidden Markov Model (HMM). 12 tags have been handled in this paper.

Apurbalal Senapati, Arjun Das and Utpal Garain [9], this paper describes performance of two systems for Named Entity Recognition (NER) task of FIRE 2013 for Bengali language. The first system is a rule-based one whereas the second one is statistical (based on CRF) in nature. The systems vary in some other aspects too, for example, the first system works on untagged data (not even POS tag is done) to identify NER whereas the second system makes use of a POS tagger and a chunker.

Sudha Morwal, Nusrat Jahan and Deepti Chopra [10], in this the authors have created a system which is language independent and can be applied this system for any language domain. In this NER system the states are not fixed means it is of dynamic in nature one can use it according to their interest. The corpus used in this NER system is also not domain specific and authors have created and tested this HMM based NER system on three languages Hindi, Urdu and Punjabi.

Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa and Xuan Wang [11], in this paper we present a statistical Named Entity Recognition (NER) system for Urdu language using two basic n-gram models, namely unigram and bigram. Authors here have used gazetteer lists and some smoothing techniques with bigram NER tagger. This NER system is capable to recognize 5 classes of NEs using a training data containing 2313 NEs and test data containing 104 NEs.

Shilpi Srivastava, Mukund Sanglikar and D.C Kothari [12], the authors have introduced a hybrid approach for NER for Hindi language. We have experimented with Statistical approaches like Conditional Random Fields (CRF) & Maximum Entropy (MaxEnt) and Rule based approach based on the set of linguistic rules. In this paper they have used a combination of machine learning and Rule based approaches to classify named entities. The paper introduces a hybrid approach for NER, with a combination of rule based and machine learning approach.

B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu and Dr. A. Govardhan [13], identification of Named Entities using various features, gazetteer lists have been proposed in this system. The identification is done using rule based approaches for Telugu language, classified in two phases, first phase to identify and second phase to transliterate.

Padmaja Sharma, Utpal Sharma and Jugal Kalita [14],In this paper the authors have presented an overview of NER, the different approaches to NER and its existing work in Indian language. Finally they describe that the Assamese language and the issues faced by this language in NER and the rules of Assamese NER.

Vishal Gupta and Gurpreet Singh Lehal [15], this paper explains the Named Entity Recognition System for Punjabi language text summarization. A Condition based approach has been used for developing NER system for Punjabi language. Various rules have been developed like prefix rule, suffix rule, proper name rule, middle name rule and last name rule. For implementing NER, various resources in Punjabi, have been developed like a list of prefix names, a list of suffix names, a list of proper names, middle names and last names.

Kishorjit Nongmeikapam, Laishram Newton Singh, Tontang Shangkhunem, Bishworjit Salam, Ngariyanbam Mayekleima Chanu and Sivaji Bandyopadhyay [16], this paper deals about

the NER in the Manipuri Language. Manipuri uses two scripts; the first one is purely of its own origin (Meitei Mayek) while another one is a borrowed Bengali script. Feature selection is an important factor in recognition of Manipuri Name Entity using Conditional Random Field (CRF).

S. Biswas, M. K. Mishra, S. Acharya and S. Mohanty [17], the authors have developed a Named Entity Recognition system for Indian languages particularly for Hindi, Bengali, Oriya, Telugu using hybrid machine learning approach that used MaxEnt and HMM successively. They showed that with the preliminary data training through MaxEnt and appropriate classifier for error correction in the final recognition process through HMM, the performance of the proposed NER system can be greatly enhanced as compared to using only a single statistical model.

Sujan Kumar Saha, Partha Sarathi Ghosh, Sudeshna Sarkar, and Pabitra Mitra [18], in this paper, the authors have described a Maximum Entropy based NER system for Hindi. They have explored different features applicable for the Hindi NER task. They have incorporated some gazetteer lists in the system to increase the performance of the system. These lists are collected from the web and are in English. To make these English lists useful in the Hindi NER task, they have proposed a two-phase transliteration methodology.

P Srikant and Kavi Narayana Murthy [19], in this paper the authors have started with the experiments in building a CRF

(Conditional Random Fields) based Noun Tagger. Trained on a manually tagged data of 13,425 words and tested on a test data set of 6,223 words, this Noun Tagger has given an F-Measure of about 92%. Then developed a rule based NER system for Telugu. They have then developed a CRF based NER system for Telugu and tested it on several data sets from the Eenaadu and Andhra Prabha newspaper corpora developed by us here.

Asif Ekbal, Rejwanul Haque and Sivaji Bandyopadhyay [20], this paper reports about the development of a Named Entity Recognition (NER) system for Bengali using the statistical Conditional Random Fields (CRFs). The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various named entity (NE) classes. The training set consists of 150K words and has been manually annotated with a NE tagset of seventeen tags.

4. NAMED ENTITY RECOGNITION APPROACH

NER is sub divided into two stage problems first is Identification of proper nouns and Classification of these proper nouns into their respective classes. The process of NER involves few stages where it consists of pre-processing of text, data training, data testing, and lastly result evaluation. These steps are again broadly classified into pre-processing steps, feature extraction, NER algorithms and labeling. Fig. 2 depicts the whole NER process.

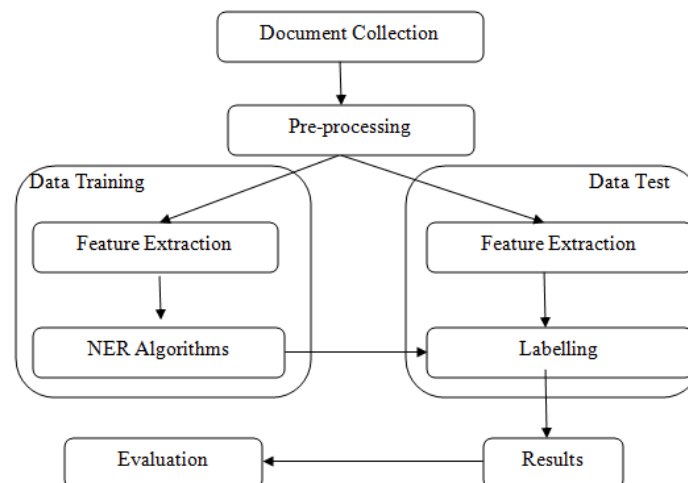


Fig 2: Named Entity Recognition Approach

Step 1: Document collection-Documents of varied formats such as .pdf, .html, .docx etc. from all sources will be collected. The languages of these documents will be all Indian languages like Marathi, Hindi, Punjabi, Bengali, Tamil, Telugu, Kannada, Oriya, etc. These documents will be inputs for the system.

Step 2: Pre-processing-Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure.

Step 2.1: Validation of input document-Validation is to check whether the given input text is in language for which the system is implemented. It also checks whether the input is syntactically correct, but does not check the semantic correctness.

Step 2.2: Tokenization-The aim of the tokenization is the exploration of the words in a Sentence where every word,

symbol, special character in the sentence is considered as a token.

Step 2.3: Stop word removal-In stop word removal, words that occur very frequently and does not contribute much to the context and content, and also have no impact on their existence are removed.

Step 2.4: Stemming-Trimming or cutting out the extraneous words to the stem is called stemming. Here inflections are removed using stemming algorithms.

Step 2.5: Morphological analysis-Morph analysis is the procedure to find out the root word. It is applied to recognize the inner structure of the word.

For example:

Input text - हॉकी भारताचा राष्ट्रीय खेळ आ

Tokenization - हॉकी | भारताचा | राष्ट्रीय | खेळ | आहे

Stemmer - हॉकी भारता राष्ट्री खेळ आहे

Morph Analysis - हॉकी भारत राष्ट्र खेळ आहे

POS tagging output- हॉकी/NNP भारत/NNP राष्ट्र/NN
खेळ/NN आहे/VM

Step 3: Data Training – This step is required to train the system. Training is done based on the feature extraction and the algorithm used. The output of this stage will be given to the testing stage.

Step 3.1: Feature extraction – In this process a small subset from the sentence is extracted and then a feature set is applied to the NER algorithms.

Step 3.2: NER algorithms – Various NER NLP algorithms include rule based, machine learning and hybrid approaches.

Step 4: Data Testing

Step 4.1: Feature extraction – This process is same as explained in the training data stage the only difference is this is done at testing stage. The extracted features are then tagged.

Step 4.2: Labeling (tagging) – In this process the entities are tagged using any of the algorithm.

Step 5: Result – the output of all the above stages will be then go through the evaluation stage using evaluation parameters.

Step 6: Evaluation – The accuracy level of NER is done by Precision (P), Recall (R) and F1-measure metrics.

5. NAMED ENTITY RECOGNITION METHODS AND TECHNIQUES

5.1 Rule based approach

Identifying and categorizing strings of text into different classes is a process defined as named entity recognition (NER). This process involves two main sub tasks firstly, identifying the proper nouns from the sequence of the text, secondly classifying these proper nouns into their predefined categories of Rule based approach. It is also called as hand crafted rules or linguistic approach. In this approach the rules are written manually by the researchers for the system and for the any particular language. Rule based systems parse the source text and produce an intermediate representation which may be a parse tree or some abstract representation [6]. Rule based are further classified into list lookup approach and linguistic approach [8].

5.1.1. List lookup Approach

In List Lookup a large corpus which is called as a bag of words are built for all the named entities and their classes. List lookup based approach works. This list is also called as gazetteer.

5.1.2. Linguistic Approach

In linguistic approach one should have a deep knowledge of the grammar of any specific language. The understanding and the knowledge of the language leads to more accurate rules so that the named entities will be identifies and classified very easily.

5.2. Machine Learning Approach

In Machine Learning-based NER system, the purpose of Named Entity Recognition approach is converting identification problem into a classification problem and employs a classification statistical model to solve it. Machine learning approaches are also named as corpus based approach. In this type of approach, the systems look for patterns and relationships into text to make a model using statistical models and machine learning algorithms. The systems identify and classify nouns into particular classes such as persons, locations, times, etc base on this model, using machine learning algorithms. There are three types of machine learning models that are used for NER that Supervised, semi-supervised and unsupervised machine learning model. Supervised learning utilizes only the labelled data to generate a model. Semi-supervised learning aims to combine both the labelled data as well as useful evidence from the unlabelled data in learning. Unsupervised learning is designed to be able to learn without or with very few labelled data [9]. These models will be broadly classified further.

5.2.1 Supervised Machine Learning Approach

Supervised machine approach is also called as statistical models. It has proved to be very effective. NER in statistical models usually treat recognising named entities as a sequence tagging problem in which each word is tagged to it entity if its present in the entity class. The process of learning is called supervised, human intervention is needed to train the system by giving the trained data examples to construct the statistical model, which cannot achieve good performance without large amount of training data. Different supervised machine approaches are as follows.

5.2.1.1 Hidden Markov Model (HMM)

It is a statistical language model that computes the likelihood of a sequence of words by employing a Markov chain, in which likelihood of next word is based on the current word. In this language model words are represented by states, NE classes are represented by regions and there is a probability associated with every transition from the current word to the next word. This model can predict the NE class for a next word if current word and its NE class pair is given. Among all approaches, the evaluation performance of HMM is higher than those of others. The main reason may be due to its better ability of capturing the locality of phenomena, which indicates names in text.

5.2.1.2. Maximum Entropy Markov Model (MEMM)

MEMM is also a statistical model which is very flexible. The assignment is output for each word or token is based on its future (f), history (h) and the features (g). Here an essential prerequisite is the selection of appropriate features, which is a tedious job especially for Indian languages because of its agglutinative and morphologically rich nature. A maximum entropy solution to this, or any other similar problem allows the computation of $p(f|h)$ for any f from the space of possible futures, F, for every h from the space of possible histories, H [9]. As the outputs of the models are defined by the futures, the solution for this is to compute $p(f|h)$ for any f from the space of possible futures, F and for any h from the possible histories, H. Thus in NER, finding the probability of f for any token with respect to its index (t) can be formulated as $p(f|ht)$.

Given a set of features and some training data, the maximum entropy estimation process produces a model in which every feature g_i has associated with it a parameter α_i [9].

$$p(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\alpha(h)} \quad (1)$$

$$Z_\alpha(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)} \quad (2)$$

The above formulas (1) and (2) tell us that given a set of features and training data, the conditional probability of the features and the history is the product of the weights for all the features on the active pair $\langle h, f \rangle$ that are normalized over the product of futures.

5.2.1.3. Conditional Random Field

Conditional random fields (CRFs) are a class of statistical modelling method often applied in pattern recognition and machine learning, where they are used for structured prediction. Where, the prediction of labels or tags for entities in any ordinary classifier is done without considering the context of neighbouring entity, CRF takes the context into account e.g. the linear chain CRF predicts sequences of labels for sequences of input samples. It is a discriminative undirected probabilistic graphical model (random fields) used to encode known relationships between observations and construct consistent interpretations.

CRF is used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $s = \langle s_1, s_2, \dots, s_T \rangle$ given an observation sequence $o = \langle o_1, o_2, \dots, o_T \rangle$ is calculated as:

$$P^{(s|o)} = \frac{1}{Z_o} \exp \left(\sum_{t=1}^T \sum_{k=1}^K \lambda_k \text{fk}(st-1, st, o, t) \right) \quad (1)$$

Where, $\text{fk}(st-1, st, o, t)$ is a feature function whose weight λ_k is to be learned via training. The values of feature function may range between $-\infty \dots \dots +\infty$ but typically they are binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor,

$$Z_o = \sum_s \exp \left(\sum_{t=1}^T \sum_{k=1}^K \lambda_k \text{fk}(st-1, st, o, t) \right) \quad (2)$$

which as in HMMs, can be obtained efficiently by dynamic programming. To train a CRF, the objective function to be maximized is the penalized log-likelihood of the state sequences given the observation sequences:

$$L = \sum_{i=1}^N \log(P^{(s_i|o_i)}) - \sum_{k=1}^K \frac{\lambda_k^2}{2\sigma^2}, \quad (3)$$

Where, $\{(s_i|o_i)\}$ is labelled as the training data.

5.2.1.4. Support Vector Machine (SVM)

This technique was introduced by Vapnik, gives high accuracy results for text categorization. Being machine learning is mainly used to solve two class pattern recognition problems. SVM is a classifier technique in which its goal is to find whether the specified entity or a vector belongs to the target class or not. During training in this approach, we

generate a hyper plane that is used to categorize the members into two classes (positive and negative classes) that exists on the opposite sides of a hyper plane. This approach also computes the distance of every vector from the hyper plane known as margin.

6. COMPARATIVE ANALYSIS OF NER TECHNIQUES

In this section, we compare listed NER techniques for Indian regional languages.

Table 1: Comparative analysis of NER techniques for Indian regional languages

Sr . no .	Name of the Algorithm	Advantages	Disadvantages	Implemented on Languages	Observations
1.	Rule Based	-Rules for any languages can be written -Language dependent	- They require huge experience and grammatical knowledge on the particular language or domain. -These systems are not transferable.	Bengali, Telugu, Urdu, Hindi, Assamese	Regular maintenance is required. High proficiency is required in a particular language.
2.	HMM	- Assumes features are independent - When labeling X_i future observations are taken into account (forward-backward) -Fast Computation	-Does not allow words in the input sentence to show dependency among each other. - Simple HMMs do not work well when small amount of labelled data are used to estimate the model parameters	Hindi, Marathi, Kannada, Bengali, Urdu, Telugu, Punjabi	- Generative Model - Find parameters to maximize $P(X, Y)$
3.	MaxEnt	-No longer assume that features are independent -We have exponential model for each state to tell us the conditional probability of the next states	-Sometimes this methods results in inadmissible assignment for tags belonging to the sequence that never happen -Do not take future observations into account (no forward-backward)	Hindi	Discriminative Model -Find parameters to maximize $P(Y X)$

4.	CRF	-Doesn't assume that features are independent -Can be interpreted in maximum entropy framework - Solves the problem of label bias.	- Slow Computation	Mani puri, Odia, Malayalam, Telugu, Bengali, Kannada	- Discriminative Model
5.	SVM	- Supports kernels, so you can model even non-linear relations	- SVMs do not consider state-to-state dependencies and feature-to-state dependencies	Hindi, Tamil, Malayalam	- It is one of the famous binary classifiers giving best results for fewer data sets and can be applied to multiclass problems by extending the algorithm.

7. CONCLUSION

Different techniques like Rule based Approach, Machine Learning Approach and Hybrid Approach form the major categories of the NLP NER algorithms. Out of these categories it has been observed that machine learning based approach is the best suited and most popular approach. In machine learning there are also many sub categories of techniques such as HMM, CRF, SVM, ME. Based on different evaluation techniques and result analysis and also according to the review of the literature survey of experiments conducted across India by different researches it has been proved the Rule based, CRF, HMM which are implemented mostly for Hindi, Marathi, Urdu, Punjabi, Bengali, Telugu. It has been observed that HMM, CRF gives the best results considering their limitations.

8. ACKNOWLEDGMENT

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks head of computer department and to the principal of Pillai College of Engineering, New Panvel for extending his support.

9. REFERENCES

- [1] Nita Patil, Ajay S., and B. V. Pawar, "Survey of Named Entity Recognition Systems with respect to Indian and Foreign Languages", January 2016
- [2] Sudha Morwal and Nusrat Jahan, "Named Entity Recognition Using Hidden Markov Model (HMM): An

Experimental Result on Hindi, Urdu and Marathi Languages", April 2013.

- [3] Ms. Jayshri Arjun Patil and Ms. Poonam Bhagwandas Godhwani, "Review of Name Entity Recognition in Marathi Language", June 2016
- [4] Yavrajdeep Kaur and Er. Rishamjot Kaur, "A Review of Name Entity Recognition (NER) in Hindi", July 2014
- [5] Nita Patil, Ajay S. Patil and B. V. Pawar, "Issues and Challenges in Marathi Named Entity Recognition", February 2016
- [6] Hinal Shah, Prachi Bhandari, Krunal Mistry, Shivani Thakor, Mishika Patel and Kamini Ahir, "Study of Named Entity Recognition for Indian Languages", March 2016
- [7] P Srikanth and Kavi Narayana Murthy, "Named Entity Recognition for Telugu", January 2008
- [8] Deepti Chopra, Nusrat Jahan and Sudha Morwal, "Hindi Named Entity Recognition by Aggregating Rule Based Heuristics and Hidden Markov Model", November 2012
- [9] Ms. Maithilee L. Patawar and Mrs. M. A. Potey, "Approaches to Named Entity Recognition: A Survey", December 2015
- [10] Shilpi Srivastava, Mukund Sanglikar and D.C Kothari, "Named Entity Recognition System for Hindi Language: A Hybrid Approach", 2011
- [11] Deepti Chopra, Nisheeth Joshi and Iti Mathur, "Named Entity Recognition in Hindi Using Hidden Markov Model", 2016
- [12] Sujan Kumar Saha, Partha Sarathi Ghosh, Sudeshna Sarkar, and Pabitra Mitra, "Named Entity Recognition in Hindi using Maximum Entropy and Transliteration", 2016
- [13] S. Amarappa and Dr. S. V. Sathyanarayana Kannada, "Named Entity Recognition and classification using Conditional Random Fields", 2015
- [14] Gowri Prasad, K.K. Fousiya, Dr. M. Anand Kumar and Dr. K.P. Soman, "Named Entity Recognition for Malayalam Language: A CRF based Approach", 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.16-19
- [15] Abinaya. N, Neethu John, Dr. M. Anand Kumar and Dr. K. P. Soman, "AMRITA@FIRE-2014 Named Entity Recognition for Indian Languages", 2014 in International Workshop: "NER shared Task" Forum for Information Retrieval Evaluation (FIRE-2014), Bengaluru, 2014.
- [16] Sudha Morwal and Deepti Chopra, "Identification and Classification of Named Entities in Indian Languages", 2013
- [17] S Amarappa and Dr. S V Sathyanarayana, "Named Entity Recognition and Classification in Kannada Language", 2013 International Journal of Electronics and Computer Science Engineering

- [18] Apurbalal Senapati, Arjun Das and Utpal Garain, "Named-Entity Recognition in Bengali@FIRE NER 2013", 2013
- [19] Sudha Morwal, Nusrat Jahan and Deepti Chopra, "Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach", 2012
- [20] Vishal Gupta and Gurpreet Singh Lehal,, "Named Entity Recognition for Punjabi Language Text Summarization", 2011 "A Highly Agglutinative Indian Language",2011
- [21] Kishorjit Nongmeikapam, Laishram Newton Singh, Tontang Shangkhunem, Bishworjit Salam, Ngariyanbam Mayekleima Chanu and Sivaji Bandyopadhyay, CRF Based Name Entity Recognition (NER) in Manipuri: "A Highly Agglutinative Indian Language",, 2011
- [22] Thoudam Doren Singh, Kishorjit Nongmeikapam, Asif Ekbal and Sivaji Bandyopadhyay, "Named Entity Recognition for Manipuri Using Support Vector Machine", 2009
- [23] Asif Ekbal Sivaji Bandyopadhyay, "A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi", Nov, 2009
- [24] S. Biswas, M. K. Mishra, S. Acharya and S. Mohanty, "A Two Stage Language Independent Named Entity Recognition for Indian Languages", 2008
- [25] Umrinder Pal Singh, Vishal Goyal and Gurpreet Singh Lehal, "Named Entity Recognition System for Urdu", 2008
- [26] Sudha Morwal, Nusrat Jahan and Deepti Chopra, "Named Entity Recognition using Hidden Markov Model (HMM)", 2008
- [27] Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa and XuanWang, "N- gram and Gazetteer List Based Named Entity Recognitionfor Urdu: Scarce Resourced Language", 2008
- [28] B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu and Dr. A. Govardhan, "Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach", 2008
- [29] Asif Ekbal and Sivaji Bandyopadhyay, "Bengali Named Entity Recognition using Support Vector Machine", 2008
- [30] Asif Ekbal, Rejwanul Haque and Sivaji Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field Approach", 2007