# HINDI NAMED ENTITY RECOGNITION BY AGGREGATING RULE BASED HEURISTICS AND HIDDEN MARKOV MODEL

Deepti Chopra, Nusrat Jahan, Sudha Morwal

Department of Computer Engineering, Banasthali Vidyapith Jaipur (Raj.), INDIA
deeptichopra11@yahoo.co.in
nusratkota@gmail.com
sudha_morwal@yahoo.co.in

***ABSTRACT***

*Named entity recognition (NER) is one of the applications of Natural Language Processing and is regarded as the subtask of information retrieval. NER is the process to detect Named Entities (NEs) in a document and to categorize them into certain Named entity classes such as the name of organization, person, location, sport, river, city, country, quantity etc. In English, we have accomplished lot of work related to NER. But, at present, still we have not been able to achieve much of the success pertaining to NER in the Indian languages. The following paper discusses about NER, the various approaches of NER, Performance Metrics, the challenges in NER in the Indian languages and finally some of the results that have been achieved by performing NER in Hindi by aggregating approaches such as Rule based heuristics and Hidden Markov Model (HMM).*

***KEYWORDS***

*HMM, Accuracy, NER, Performance Metrics, Named Entities*

## 1. INTRODUCTION

There are numerous applications of Named Entity Recognition (NER).Some of these include: Information Extraction, Question Answering, Information Retrieval, Automatic Summarization, Machine Translation etc. The Named Entities can be known to us, if we perform computations on the natural language. The task of extracting necessary details and retrieving important information can be made easier and faster, if the Named entities are already known to us. NER is the process in which Named Entities are detected in a document and are classified into their respective Named Entity classes using any of the NER based approaches. According to the 8th schedule, India is known to have 22 official Indian languages. NER in Indian languages is still considered to be a budding topic of research in the field of NLP and much of work is needed to be performed in this regard.

Consider an example of NER in Hindi as follows:

"Mohit/PER ne/O mi road/LOC se/O kitab/O khareedi/O ।/O

In the above sentence, the task of a NER based system is to extract and then classify the named entities into certain classes. Here, we have considered 'Mohit' as the name of a person, so it is

shown by a PER tag. 'mi road' is the name of a location, so we have allotted a LOC tag to it. The named entity tags that we choose may vary every time. It depends on the individual choice and the contents that we have considered for the Named Entity Recognition. TABLE I lists some of the Named Entity Tags. Named Entity tags may be of the general type or may further be divided into sub tags which are of specific types. E.g. location tag (LOC) may further be classified into continent tag, country tag, city tag, state tag, town tag, street tag etc.

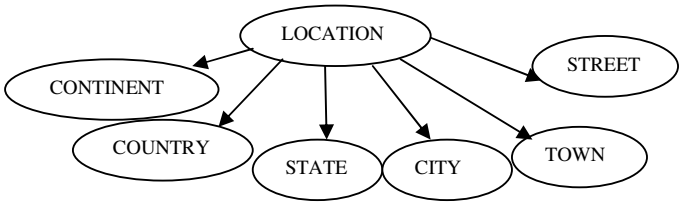


Figure1. A single Named Entity tag split into more specific Named Entity tags

Table 1
Various Named Entity Tags. NE Tags: Named Entity Tags
PER: Name of Person, CO-Country, ORG-Organization, VEH-vehicle and QTY-Quantity

| NE TAG | EXAMPLE |
|---|---|
| PER | Deepti, Sudha, Rohit |
| CITY | Jaipur, Mumbai, Kolkata |
| CO | India, China, Pakistan |
| STATE | Rajasthan, Maharashtra |
| SPORT | Hockey, Badminton |
| ORG | TCS, Infosys, Accenture |
| RIVER | Ganga, Krishna, kaveri |
| DATE | 27-04-2012, 31/01/1989 |
| TIME | 10:10 |
| PERCENT | 100% |

## 2. METHODOLOGIES FOR NER

There are basically two approaches that are employed in Named Entity Recognition. [5] [1] [18] These include: Rule Based Approach and Machine learning based Approach [11] [6] [16].

### 2.1. Rule based Approach

It is also known as handcrafted approach. It is of two types:

### 2.1.1 List Lookup Approach

In this approach, Gazetteers are used that consists of different lists of Named Entity classes and a simple look up operation is performed to conclude whether a word is a Named Entity or not. If a particular word is found in a Named Entity class, then a Named Entity tag is allotted to that word according to the Named Entity class in which it is found. Indian languages lack in resources.

We can prepare Gazetteers in Indian languages using transliteration that would convert English Named Entities into Indian languages. Some seed values of a domain specific corpus can be used that would learn the context patterns and then Named Entities are produced by the concept of bootstrapping.[17]This methodology is easy and fast .The disadvantage of this approach is that it cannot overcome the problem of ambiguities.

E.g In a sentence:-““Ganga/PER Ne/O Ganga/RIVER nadi/O mein/O dupki/O Lagayi/O |/O””. In this sentence, Ganga is a Named Entity .But, it can be a person name or a river name .The ambiguity cannot be resolved by this methodology.

### 2.1.2. Linguistic Approach

In this approach, a linguist, who has an in depth knowledge about the grammar of specific language constructs some rules, so that the Named Entities can be recognized as well as classified easily. [3][20][19]The rules that are constructed are language independent and cannot be used to identify Named Entities in some other language. [11]

## 2.2. Machine Learning Based Approach

This approach is also known as automated approach or Statistical approach. Machine learning based approach is more efficiently and frequently used as compared to the Rule based approach.

### 2.2.1. Hidden Markov Model (HMM)

HMM is a statistical based approach in which states are hidden or unobserved .The HMM produces sequence of tokens that are nothing but optimal state sequence.
It is based on the Markov Chain Property i.e. the probability of occurrence of the next state is dependent on the just previous state. HMM is easy to implement. The disadvantage of this approach is that it requires lot of training in order to get better results and it cannot be used for large dependencies. [12]

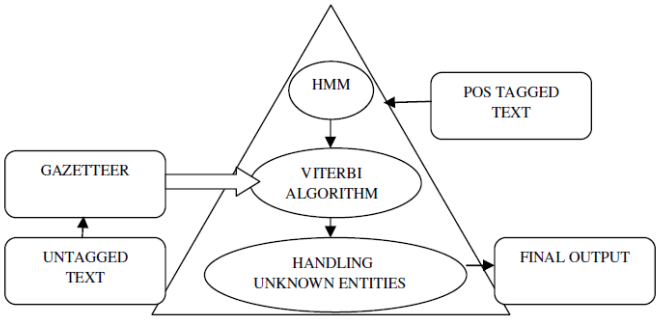


Figure 2: Diagrammatic description of HMM

### 2.2.2. Maximum Entropy Markov Model (MEMM)

It combines the concept of Hidden Markov Model and Maximum Entropy Model. While training, this model makes sure that the unknown values in a Markov Chain are connected and are not conditionally independent of each other.

The large dependency problem of HMM is resolved by this model. Also, it has higher recall and precision as compared to HMM. The disadvantage of this approach is the label bias problem. The probabilities of transition from a particular state must sum to one. MEMM favours those states through which less number of transitions occurs. [16]

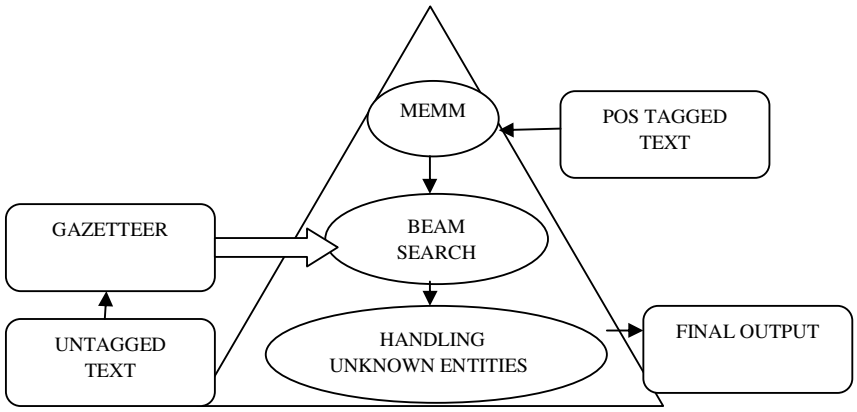


Figure 3: Diagrammatic description of MEMM

### 2.2.3. Conditional Random Field (CRF)

It is graphical undirected model .Unlike other classifiers, it also takes into consideration the context information or the neighbouring samples. It is known as Random field since it computes the conditional probability on the following node given the present node values.

This methodology has advantages same as that of MEMM. Also it resolves the label bias problem faced by MEMM. [3]

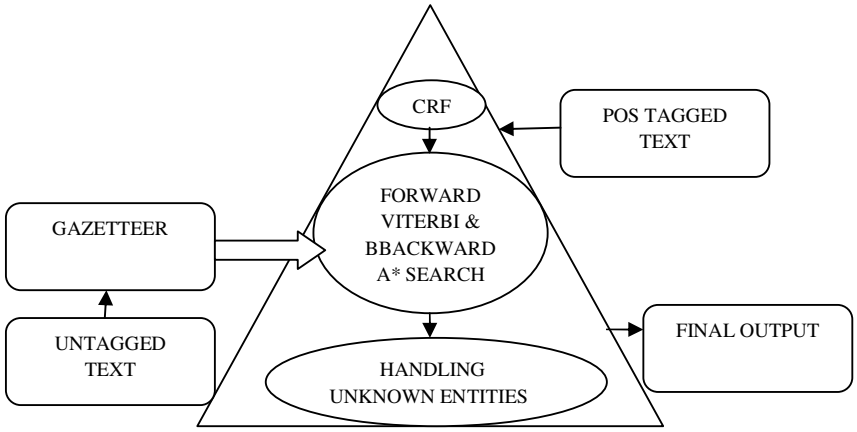


Figure 4.Diagrammatic description of CRF

### 2.2.4. Support Vector Machine (SVM)

This methodology was introduced by Vapnik. SVM is a supervised statistical approach. The main objective of this approach is to find whether a specific vector belongs to a particular target class or not. [2] In this approach, the training as well as the testing data belongs to the single dimension vector space.

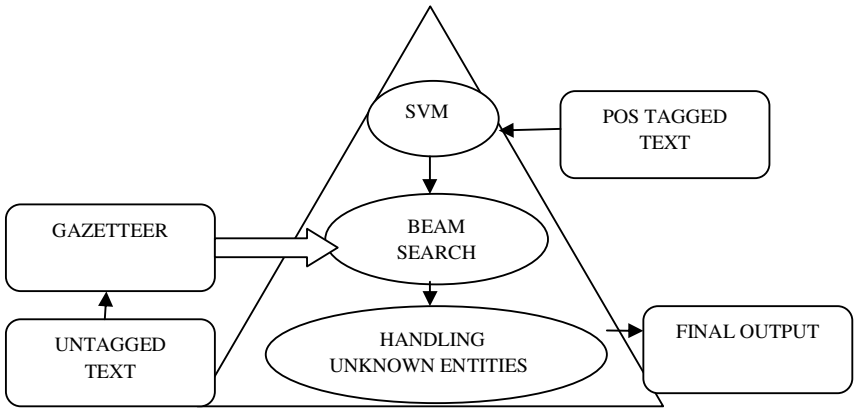


Figure 5: Diagrammatic description of SVM

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. The main advantage of this approach is that it gives high accuracy for the text categorization problem. [4]

### 2.2.5 Decision Tree

It is a well known methodology that is used to extract and categorize Named Entities in a given corpus .In this approach, some recognition rules are applied to the untagged training corpus so that Named Entities are retrieved. Now, we match these Named Entities obtained with the actual answer key provided by the humans. If the Named Entity is same as the answer key, then it is referred to as the positive example else it is known as negative example. [7]. A decision tree is build that classifies the Named Entities in the testing document.[9] The leaf node of decision tree depicts the resultant value of test .

## 3. PERFORMANCE METRICS

Performance Metrics is very important since it reveals the performance of a Named Entity Recognition based system in terms of Precision, Accuracy and F-Measure. The output of a NER system may be termed as “response” and the interpretation of human as the “answer key”. We consider the following terms:

1. Correct-If the response is same as the answer key.
2. Incorrect-If the response is not same as the answer key.
3. Missing-If answer key is found to be tagged but response is not tagged.
4. Spurious-If response is found to be tagged but answer key is not tagged.[6]

Hence, we define Precision, Recall and F-Measure as follows:

Precision (P): Correct / (Correct + Incorrect + Missing)
Recall (R): Correct/ (Correct + Incorrect + Spurious)
F-Measure: (2*P*R)/(P+R) [5][8]

## 4. ISSUES IN NER IN INDIAN LANGUAGES

We still have not performed much of the work in NER in the Indian languages. This is mainly due to the fact that Indian languages lack in resources such as annotated corpora and lexical resources. There are many challenges related to the Named Entity Recognition in the Indian languages .Some of them include the following:[6]

1. Lack of Capitalization: In Indian languages, the Capitalization concept is absent. Whereas, in English and in many of the European languages, the word in which first alphabet is capital is a proper noun. The NER based systems that are developed for the English and the European languages, henceforth cannot be used to perform named entity recognition in the Indian languages .Thus there is a need to develop an efficient NER based system for the Indian languages. [15]

2. Indian languages are inflectional and morphologically rich and are free word order.

3. Indian languages lack in resources .This problem is due to the fact that web mostly have lists of Named Entities which are in English and not in the Indian languages.[17].

4. In dictionary of the Indian languages, many common nouns also exists as proper nouns. E.g. Lata, Suraj, Aakash , Tara etc. are the Name of persons and common nouns as well. So, we need to resolve ambiguities, which is also one of the issues in NER in the Indian languages

## 5. RESULTS

We have prepared a general corpus from the Hindi newspapers on the web. We have annotated it manually. The Named Entity tags that we have used are: PER (Name of Person), LOC (Name of Location), TIME, MONTH, SPORT, ORG (Name of Organization), VEH (Name of Vehicle), RIVER and QTY (Quantity).In the first phase, we have applied the Rule based heuristics or the shallow parsing technique over the Corpus, in which some of the helping words are used to detect the Named Entities, that occur just after or before the Named Entities to be identified. In the second phase, we apply Hidden Markov Model (HMM) to detect the rest of the Named Entities.

Table 2 Results of Rule based heuristics or shallow parsing technique

| NAMED ENTITIES | TOTAL NAMED ENTITIES(NEs) | NAMED ENTITIES (NEs) IDENTIFIED | ACCURACY |
|---|---|---|---|
| LOC | 247 | 125 | 50.60% |
| PER | 56 | 29 | 51.79% |
| QTY | 79 | 40 | 50.63% |
| TIME | 67 | 34 | 50.75% |
| ORG | 135 | 68 | 50.37% |
| SPORT | 45 | 23 | 51.11% |
| RIVER | 11 | 6 | 54.54% |
| VEH | 25 | 0 | 0% |
| MONTH | 22 | 0 | 0% |
| | TOTAL NEs = 687 | TOTAL NEs DETECTED = 325 | TOTAL ACCURACY = 47.5% |

Table 3 Results of Hidden Markov Model (HMM)

| NAMED ENTITIES | TOTAL NAMED ENTITIES (NEs) UNDETECTED | NAMED ENTITIES (NEs) IDENTIFIED | ACCURACY |
|---|---|---|---|
| LOC | 122 | 107 | 87.70% |
| PER | 27 | 24 | 88.89% |
| QTY | 39 | 34 | 87.18% |
| TIME | 33 | 29 | 87.88% |
| ORG | 67 | 59 | 88.06% |
| SPORT | 22 | 20 | 90.90% |
| RIVER | 5 | 5 | 100% |
| VEH | 25 | 25 | 100% |
| MONTH | 22 | 22 | 100% |
| | TOTAL NEs = 362 | TOTAL NEs DETECTED = 325 | TOTAL ACCURACY = 89.78% |

Table 4 Results of Combination of Approaches or Hybrid Approach

| TOTAL NAMED ENTITIES (NEs) | NAMED ENTITIES (NEs) IDENTIFIED | ACCURACY |
|---|---|---|
| 687 | 650 | 94.61% |

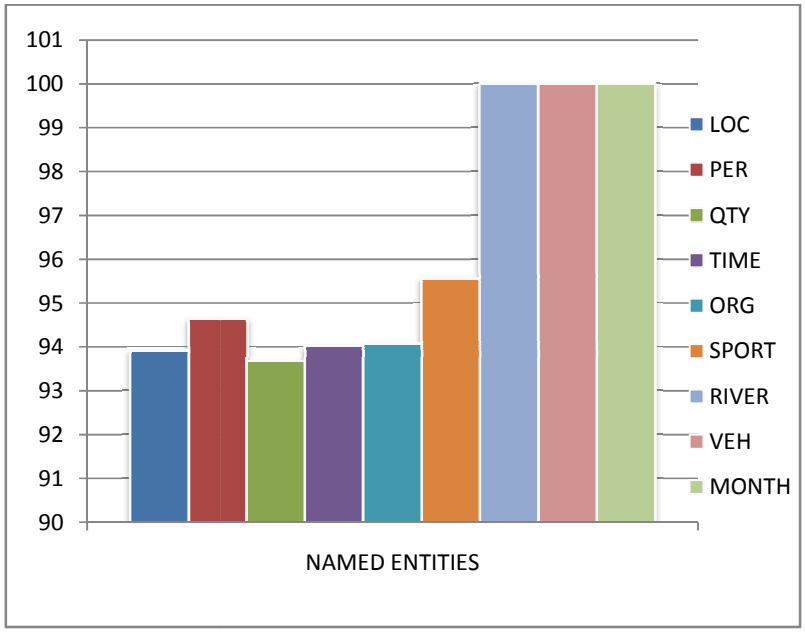


Figure 6 Results of using Combined Approach

## 6. CONCLUSIONS

We have obtained accuracy of about 94.61% by aggregating the rule based heuristics and the HMM, as shown in Table 4. Table 2 depicts that if we applied only Rule Based Heuristics, then it performed very poorly, and the accuracy obtained by this approach was 47.5%. Similarly, Table 3 depicts that if we applied only HMM, then its performance was average, and the accuracy obtained by this approach was 89.78%. This shows that if we apply hybrid approach or the combined approach, then it gives very good results in a Named Entity Recognition based system.

## ACKNOWLEDGEMENT

I would like to thank all those who helped me in accomplishing this task.

## REFERENCES

[1] Animesh Nayan,, B. Ravi Kiran Rao, Pawandeep Singh,Sudip Sanyal and Ratna Sanya "Named Entity Recognition for Indian Languages" .In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages ,Hyderabad (India) pp. 97–104, 2008.

[2] Asif Ekbal and Sivaji Bandyopadhyay. "Named Entity Recognition using Support Vector Machine: A Language Independent Approach" International Journal of Electrical and Electronics Engineering 4:2 2010.

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

[4] Asif Ekbal and Sivaji Bandyopadhyay 2008 " Bengali Named Entity Recognition using Support Vector Machine" Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 51–58, Hyderabad, India, January 2008..

[5] B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu3, Dr. A. Govardhan. "A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[6] Darvinder kaur, Vishal Gupta. "A survey of Named Entity Recognition in English and other Indian Languages" . IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.

[7] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos."Learning Decision Trees for Named-Entity Recognition and Classification"

[8] G.V.S.RAJU, B.SRINIVASU, Dr.S.VISWANADHA RAJU, 4K.S.M.V.KUMAR "Named Entity Recognition for Telugu Using Maximum Entropy Model"

[9] Hideki Isozaki "Japanese Named Entity Recognition based on a Simple Rule Generator and Decision Tree Learning" .Available at:http://acl.ldc.upenn.edu/acl2001/MAIN/ISOZAKI.PDF

[10] James Mayfield and Paul McNamee and Christine Piatko "Named Entity Recognition using Hundreds of Thousands of Features" .Available at: http://acl.ldc.upenn.edu/W/W03/W03-0429.pdf

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

[12] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", In Proceedings of the IEEE, 77 (2), p. 257-286February 1989.Available at: http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf

[13] "Padmaja Sharma, Utpal Sharma, Jugal Kalita."Named Entity Recognition: A Survey for the Indian Languages. " . (LANGUAGE IN INDIA. Strength for Today and Bright Hope for Tomorrow .Volume 11: 5 May 2011 ISSN 1930-2940)AvailableAt:http://www.languageinindia.com/may2011/v11i5may2011.pdf

[14] Praveen Kumar P and Ravi Kiran V" A Hybrid Named Entity Recognition System for South Asian Languages". Available at-http://www.aclweb.org/anthology-new/I/I08/I08-5012.pdf

[15] S. Pandian, K. A. Pavithra, and T. Geetha, "Hybrid Three-stage Named Entity Recognizer for Tamil," INFOS2008, March Cairo-Egypt. Available at: http://infos2008.fci.cu.edu.eg/infos/NLP_08_P045-052.pdf

[16] Shilpi Srivastava, Mukund Sanglikar & D.C Kothari. "Named Entity Recognition System for Hindi Language: A Hybrid Approach" International Journal of Computational Linguistics (IJCL), Volume (2) : Issue (1) : 2011.Available at: http://cscjournals.org/csc/manuscript/Journals/IJCL/volume2/Issue1/IJCL-19.pdf

[17] Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra "Gazetteer Preparation for Named Entity Recognition in Indian Languages".

[18] Sujan Kumar Saha Sanjay Chatterji Sandipan Dandapat. "A Hybrid Approach for Named Entity Recognition in Indian Languages"

[19] S. Biswas, M. K. Mishra, Sitanath_biswas, S. Acharya, S. Mohanty "A Two Stage Language Independent Named Entity Recognition for Indian Languages" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (4), 2010, 285-289.

[20] Vishal Gupta, Gurpreet Singh Lehal "Named Entity Recognition for Punjabi Language Text Summarization" International Journal of Computer Applications (0975 – 8887) Vpl.33 No.3, Nov. 2011

## Authors

**Deepti Chopra** received B.Tech degree in Computer Science and Engineering from Rajasthan College of Engineering for Women, Jaipur, Rajasthan in 2011.Currently she is pursuing her M.Tech degree in Computer Science and Engineering from Banasthali University, Rajasthan. Her research interests include Artificial Intelligence, Natural Language Processing, and Information Retrieval.

**Nusrat Jahan** received B.Tech degree in Computer Science and Engineering from R.N. Modi Engineering College, Kota, Rajasthan in 2010.Currently she is pursuing her M.Tech degree in Computer Science and Engineering from Banasthali University, Rajasthan. Her research interests include Artificial Intelligence, Natural Language Processing, and Information Retrieval.

**Sudha Morwal** is an active researcher in the field of Natural Language Processing. Currently working as Associate Professor in the Department of Computer Science at Banasthali University (Rajasthan), India. She has done M.Tech (Computer Science) , NET, M.Sc (Computer Science) and her PhD is in progress from Banasthali University (Rajasthan), India.