Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages

Sudha Morwal
Department of Computer Science
Banasthali Vidyapith
Jaipur (Raj.), INDIA

Nusrat Jahan
Department of Computer Science
Banasthali Vidyapith
Jaipur (Raj.), INDIA

Abstract: ***Named Entity Recognition is the process to detect Named Entities (NEs) in a file , document or from a corpus and to categorize them into certain Named entity classes like name of city, State, Country, organization, person, location, sport, river, quantity etc. In this paper our main objective is to perform Named Entity Recognition in Natural languages using Hidden Markov Model (HMM) and provide ways to increase accuracy and the Performance Metrics (Precision, Recall, F-Measure). Named entity recognition (NER) is one of the applications of Natural Language Processing and is considered as the subtask of information retrieval. In this paper we discuss about NER, brief introduction about its approaches and some experimental result on Indian Languages like Hindi, Urdu and Marathi using HMM.***

Keywords: Named Entity recognition (NER), Hidden Markov Model (HMM), Accuracy, Named Entities (NEs), Indian Language (IL) .

I. INTRODUCTION

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. Numerous NER applications are found and observed in varied branches of knowledge and science such as Information Extraction, Question-Answering, Machine Translation, Automatic Indexing of documents, Cross-lingual Information retrieval, Text Summarization etc.

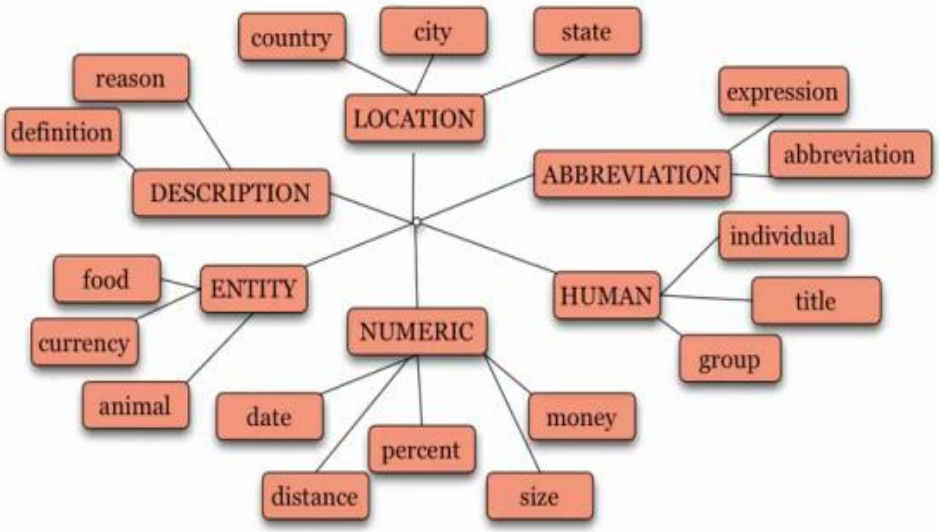

Fig1. A single Named Entity split into more specific Named Entities

For example consider the Hindi sentence like:

"मुहम्मद हनीफ राजगीरी के निरीक्षक थे।" after performing the named entities on these sentences the result is as follows "मुहम्मद/PER हनीफ/PER राजगीरी/LOC के/OTHER निरीक्षक/OTHER थे/OTHER ।/OTHER" means मुहम्मद and हनीफ

are name of person, राजगीरी specify the location and rest words are not named entities so assigned the tag OTHER. Similarly consider the example of an Urdu Language:
" جیسلمیر، جودھپور، اور بیکانیر ریت کے ٹیلوں کے لئے جانے جاتے ہیں لیکن ان میں سے سب سے زیادہ مشہور ہے جیسلمیر"

So after applying NER the result is as follows:
" جیسلمیر /CITY ،/OTHER جودھپور/CITY ،/OTHER اور/OTHER بیکانیر/CITY ریت/OTHER کے/OTHER ٹیلوں/OTHER کے/OTHER لئے/OTHER جانے/OTHER جاتے/OTHER ہیں/OTHER لیکن/OTHER ان/OTHER میں/OTHER سے/OTHER سب/OTHER سے/OTHER زیادہ/OTHER مشہور/OTHER ہے/OTHER جیسلمیر/CITY ./OTHER".
Generally NER can be treated as a two step process: Identification and classification

The named entities may be of any type. For example

Table1: Different named entities

NE tag	Definition
PER	Name of person
LOC	Name of Location
ORG	Name of Organisation
COUNTRY	Name of Country
STATE	Name of state
CITY	Name of City
RIVER	Name of River
MONTH	Name of Month
SPORT	Name of Any sport
QUANTITY	Any quantity like date, time etc.
OTHER	Not a named entity

II. APPROACHES FOR NER

There are basically two approaches that are used for Named Entity Recognition. These include:

A) Linguistic Approach

B) Machine learning based Approach.

Linguistic Approach/Rule based Approach:

The linguistic approach is the classical approach to NER. It uses rules manually written by linguists. Though it Requires a lot of work by particular domain experts, a NER system based on manual rules may provide very high accuracy. There are several rule-based NER systems, containing mainly lexicalized grammar, gazetteer lists, and list of trigger words.
The main disadvantages of these rule-based techniques are:

- They require huge experience and grammatical knowledge on the particular language or domain.
- The development is generally time-consuming.
- Changes in the system may be hard to accommodate.
- These systems are not transferable, which means that one rule-based NER system made for a particular language or domain cannot be used for other languages or domains.

Machine learning based Approach:

ML based techniques facilitate the development of recognizers in a very short time. Several ML techniques have been successfully used for the NER task. Here we discuss some NER systems that have used ML techniques. These are:

- Hidden Markov Models (HMM).
- Decision Trees.
- Maximum Entropy Models (ME).
- Support Vector Machines (SVM).
- Conditional Random Fields (CRF).

Combinations of different ML approaches are also used.

III. LITERATURE REVIEW

We have studied various papers by different researchers for identifying named entities from a text, document or from corpus by using different machine learning approach on different Indian Languages like Hindi, Urdu, Bengali, Telugu and Oriya.

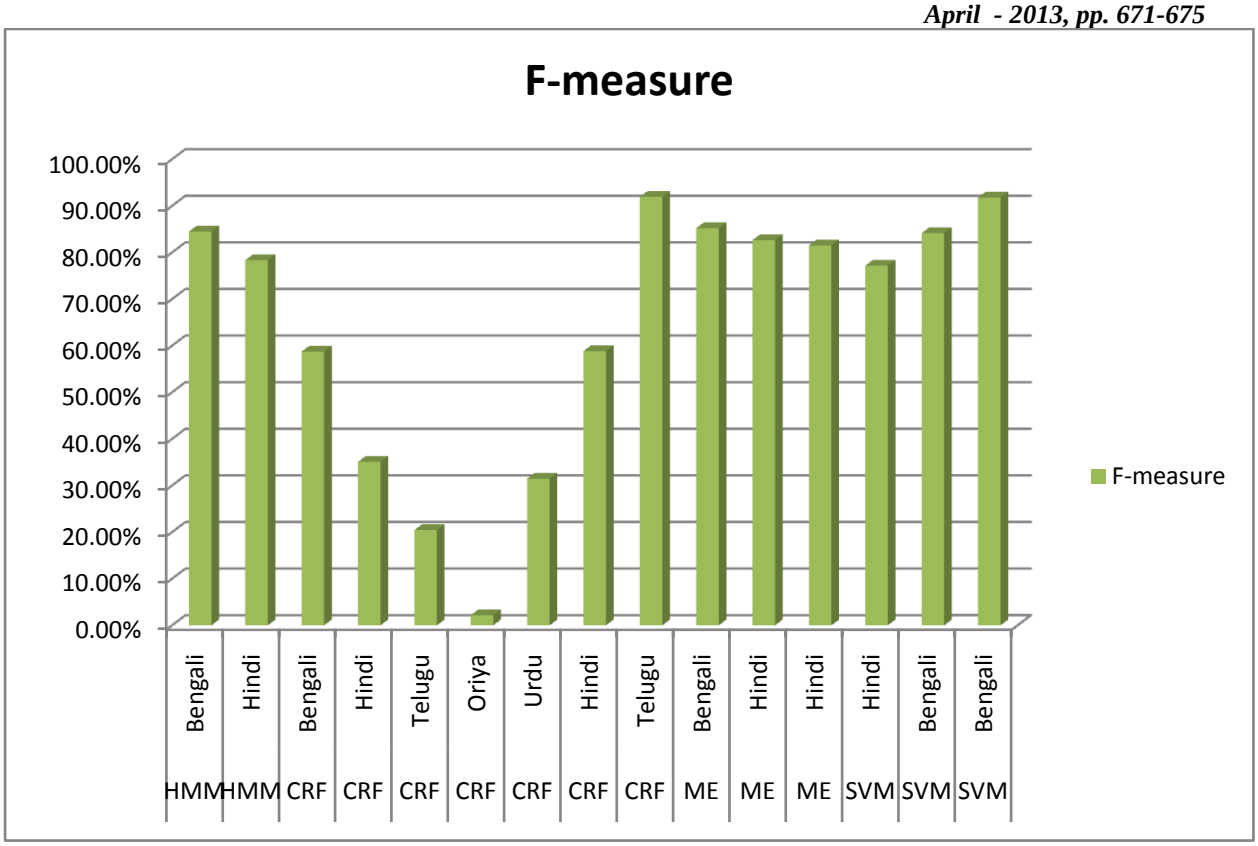

Fig2: Represents F-measure for different languages using different method.

IV. OUR PROPOSED METHOD

WE are using Hidden Markov Model based machine learning approach. Named Entity Recognition in Indian Languages is a current topic of research. The HMM based NER system 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'. In this phase, it computes the three parameters of HMM i.e. Start Probability, Emission Probability (B) and the Transition Probability (A) [11][12][7]. The last phase is the 'TESTING Phase'. In this phase, user gives certain test sentences to the system, and based on the HMM parameters computed in the previous state, Viterbi algorithm computes the optimal state sequence for the given test sentence.

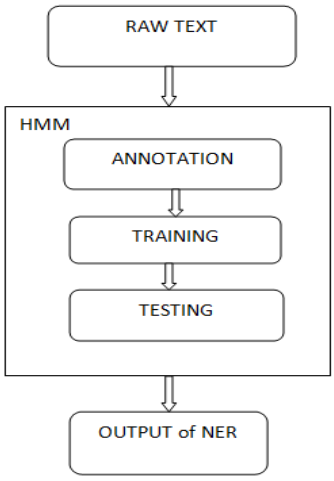

Fig3: Steps in NER using HMM

Mathematically, HMM parameters are given as follows:
A = aij = (Number of transitions from state si to sj) / (Number of transitions from state si).
B = bj (k) = (Number of times in state j and observing symbol k) / (expected number of times in state j).

We have performed NER in Hindi, Marathi and Urdu languages. For Hindi, we considered Tourism Domain Corpus and for Marathi language we consider Corpus from NLTK Indian Corpora. For Urdu language we have taken the Tourism domain corpus of Hindi language and translate it with the help of Google translator to convert it into Urdu language.

The experimental Result of each language is as follows:

A) For Hindi Language

We perform training and testing on 100 sentences and the tags are PER, LOC, COUNTRY, STATE, CITY, MONTH and OTHER. And we have seen that it gives 86% accuracy.

Sentences	Words	NE tag	PER	LOC	COUNTRY	STATE	CITY	MONTH	PLACE	Accuracy
100	2209	8	11	50	22	17	35	2	4	86%

Table2. Performance result of 100 Hindi sentences.

B) For Marathi Language

sentences	words	Total NE tag	PER	CITY	COUNTRY	STATE	ORG	LOC	Accuracy
100	1448	7	54	7	19	7	9	9	76%

Table3. Result of 100 Marathi sentences.

C) For Urdu Language

sentences	words	Total NE tag	PER	CITY	COUNTRY	STATE	RIVER	LOC	Accuracy
50	734	11	2	80	9	14	15	6	65%

Table4. Result of 50 Urdu sentences.

***Issues with Urdu language*:**

We are considering only 50 sentences for Urdu languages because of the following reason:

1. First major issue with Urdu language is those corpuses are not available.
2. Urdu language is written from right to left.
3. After tagging those sentences it automatically change the alignment.
4. It means forward space works like backward space and vice- versa.
5. Means we can read each word from left to right just like Hindi language. This is difficult for the person who does not have knowledge about Urdu.
6. For performing annotation we require lots of time as compare to other language.

V. FEATURES OF PROPOSED SYSTEM

Our Hidden Markov model based NER system has been trained and tested with different Indian Languages. We have performed training and testing on various corpuses and it gives better performance. The works reported in this paper differ from other previous work in terms of the following points:

A) language independent

This methodology works for any natural language European language also. This work tested for Hindi, Urdu and Marathi languages.

B) General Approach

This approach is not domain specific. This work tested for tourism corpus, general sentences and stories

C) High Accuracy

If rich corpus is developed it performs best. During testing we also get accuracy till 90 %.

D) Dynamic

All the parameters used by our system are of dynamic in nature means one can use according to their interest. This work is tested for Person, Location, river, Country tags in tourism corpus and Person, time, month, dry fruits, food items tags in story corpus.

E) Usefulness to other classification

Since the parameters are of dynamic in nature the same NER system can be used for other NLP classification like Part of speech tagging etc.

F) Fine grained tagging

Mostly systems allot location tag to name of place, river, palace etc. In this system you can set subclass of location tags according to your need. This system has been tested for country, river, and tree etc. tags.

G) Use of Annotated corpus

To use this system you have to design tagged corpus either with the help of proposed system or with other tools. This tagged corpus can be used in other natural language processing applications.

VI. PERFORMANCE METRICS

Performance Metrics is measure to estimate the performance of a NER based system.
Performance Metrics can be calculated in terms of three parameters: Precision, Accuracy and F-Measure [10] [6]. Consider the following terms:

Response (R): It may be defined as the output of a NER based system.

Answer Key (A): The interpretation of human may determined as Answer Key

Response –Answer key (RA): The output of a NER based system as well as the interpretation of Human. Hence, we define Precision, Recall and F-Measure [2] [3] [9] as follows:
Precision (P): RA/R
Recall (R): RA/A
F-Measure: (2 * P * R) / (P + R)

VII. CONCLUSION

A huge amount of work has already been done in Named Entity Recognition in English and in European Languages with high precision value. But, we have not performed much work in Indian languages. HMM is considered as one of the easiest Statistical approaches in Named Entity Recognition. If we perform Named Entity Recognition in HMM and also provide the ways to improve the accuracy and the performance metrics, then using same approach we can perform NER in the rest of the 21 Indian Languages and later a more efficient Language Independent approach can be used where a single NER based system can be used to perform Named Entity Recognition for all the Indian languages. Unknown words in Named Entity Recognition can be handled using transliteration approach, in which a Named Entity trained in one language, during testing same Named Entity in some other language can be handled.

ACKNOWLEDGEMENT

We would like to thank all those who helped us in accomplishing this task.

REFERENCES

[1] Sujan Kumar Saha, Partha Sarathi Ghosh, Sudeshna Sarkar, and Pabitra Mitra" *Named Entity Recognition in Hindi using Maximum Entropy and Transliteration".*

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

[3] 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.

[4] Asif Ekbal and Sivaji Bandyopadhyay 2008 *" Bengali Named Entity Recognition using Support Vector Machine"* Proceedings of the IJCNLP Languages, pages 51–58, Hyderabad, India, January 2008.

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

[6] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay "*Language Independent Named Entity Recognition 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.Available at: http://www.mt-archive.info/IJCNLP-2008-Ekbal.pdf

[7] 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.

[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] 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.

[10] 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

[11] 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

[12] 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. Available at: http://www.waset.org/journals/ijeee/v4/v4-2-19.pdf

