Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach

Nusrat Jahan [1], Sudha Morwal [2] and Deepti Chopra [3]
Department of computer science, Banasthali University, Jaipur-302001, Rajasthan, India
nusratkota@gmail.com
sudha_morwal@yahoo.co.in
deeptichopra11@yahoo.in

Abstract-**Named Entity Recognition (NER) is the task of processing text to identify and classify names, which is an important component in many Natural Language Processing (NLP) applications, enabling the extraction of useful information from documents. Basically NER is a two step process and used for many application like Machine Translation. Indian languages are free order, and highly inflectional and morphologically rich in nature. In this paper we describe the various approaches used for NER and summery on existing work done in different Indian Languages (ILs) using different approaches and also describe brief introduction about Hidden Markov Model And the Gazetteer method for NER. We also present some experimental result using Gazetteer method and HMM method that is a hybrid approach. Finally in the last the paper also describes the comparison between these two methods separately and then we combine these two methods so that performance of the system is increased.**

Keywords: Hidden Markov Model (HMM), Named Entities (NEs), Named Entity Recognition (NER), Indian Languages (ILs).

I. INTRODUCTION

Named Entities (NEs) such as person names, location names and organization names usually carry the core information of spoken documents, and are usually the key in understanding spoken documents. Therefore, Named Entity recognition (NER) has been the key technique in applications such as information retrieval, information extraction, question answering, and machine translation for spoken documents [14]. In the last decades, substantial efforts have been made and impressive achievements have been obtained in the area of Named Entity recognition (NER) for text documents.

Example- Consider a Hindi sentence as follows:

“मुहम्मद/PER हनीफ/PER राजगीरी/LOC के/OTHER निरीक्षक/OTHER थे/OTHER ।/OTHER”

In the above sentence, the NER based system first identifies the Named Entities and then categorize them into different Named Entity classes. In this sentence, first word मुहम्मद refers to the Person name, so it is allotted ‘PER’ tag. The second word हनीफ refers to the name of person. So, it is allotted ‘PER’ tag. The third word राजगीरी refers to the location. So it is assigned the tag LOC. Here ‘OTHER’ means not a Named Entity tag.

In the last decades, substantial efforts have been made and impressive achievements have been obtained in the area of Named Entity recognition (NER) for text documents.

Since NER is the current topic of research interest in India .A lot of work has been done for European language but for IL it has many challenges. So our aim is to develop a NER system for IL which gives accurate result.

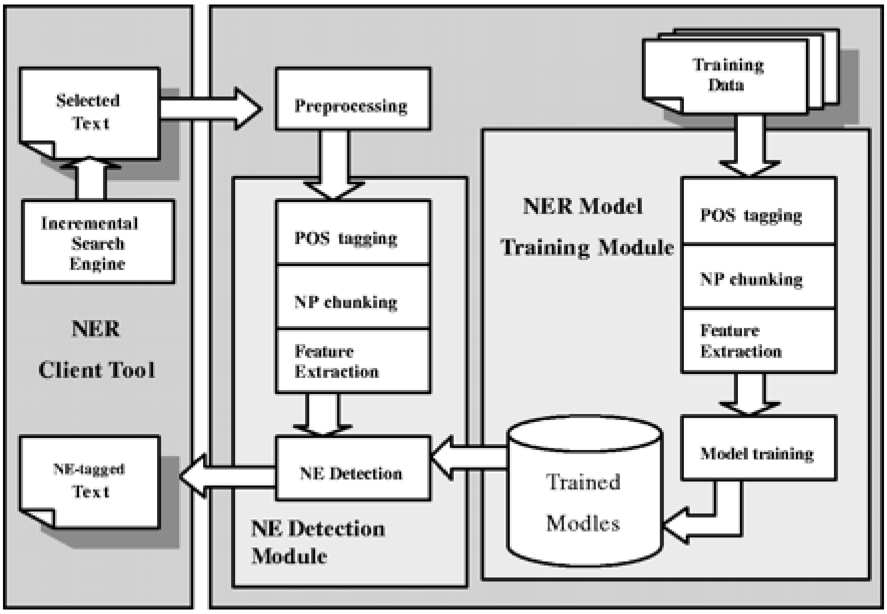

Fig.1 A Typical Named Entity Recognition Based System

NER can be treated as a two-step process - identification of proper nouns and its classification. The first step is the identification of proper nouns from the text and the second step is the classification of these proper nouns into any one of the classes like person name, organization name, location name and other classes. The main problem of NER is how to tag the words and what tag is assigned to the entities like person, organization and location etc. Sometimes ambiguities exist in the document and we have to resolve them in order to assign the correct tag.

II. APPROACHES TO NER

There are basically two methodologies that are employed in Named Entity Recognition. The major approaches to NER are:

A. Linguistic or Rule based approach.

B. Machine learning (ML) based approach.

C. Hybrid approach

A. Linguistic or Rule based approach

The linguistic approach mainly uses rules manually written by linguists. So there are many rule based NER system containing:

- Lexicalized grammar
- Gazetteer lists
- List of trigger words

B. Machine learning (ML) based approach

The most commonly used machine learning methods for NER which give accurate result up to extent are:

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

Each of these machines learning approach has advantages and disadvantages. Maximum entropy model does not solve the label biasing problem. Sequence labelling problem can be solved very efficiently with the help of Markov Models. The conditional probabilistic characteristic of CRF and MEMM are very useful for development of NER system. CRF is flexible to capture many correlated features, including overlapping and non-independent features [1].

C. *Hybrid approach*

The hybrid approach uses both rule based and machine learning methods. So in the hybrid approach we combine any of the two methods in order to improve the performance of the NER system. So the hybrid approach may be combination of HMM model and CRF model or CRF and MEMM approach.

In this paper we consider the hybrid approach i.e. Gazetteer method and Hidden Markov Model to increase the accuracy of the NER System.

Table 1: Comparison of Rule based and Machine learning approach

Rule Based Approach	**Machine Learning Approach**
This approach contains set of hand written rules. Rules are written by the language experts so for this approach human experts are required.	Developers do not need language expertise.
Require only small amount of training data.	Require large amounts of annotated training data.
These systems are not transferable to other languages or domains.	Once we build the machine learning based system may be used other language or domains.
Development can be very time consuming.	It requires less human effort.
Some changes may be hard to accommodate.	Some changes may require re-annotation of the entire training corpus.

III. CURRENT STATUS IN NER FOR INDIAN LANGUAGES

Although a lot of work has been done in English and other foreign languages like Spanish, Chinese etc with high accuracy but regarding research in Indian languages is at initial stage only. Accurate NER systems are now available for European Languages especially for English and for East Asian language. For south and South East Asian languages the problem of NER is still far from being solved. There are many issues which make the nature of the problem different for Indian languages.

For example:- The number of frequently used words (common nouns) which can also be used as names (Proper nouns) is very large for European language where a large proportion of the first names are not used as common words.

IV. ISSUES WITH HINDI LANGUAGE

Since for English Language lots of NER system has been built. But we can't use such NER system for Indian Language because of the following reason [3]:

- Unlike English and most of the European languages, Indian languages lack the capitalization information that plays a very important role to identify NEs in those languages.
- Indian names are ambiguous and this issue makes the recognition a very difficult task.
- Indian languages are also a resource poor language. Annotated corpora, name dictionaries, good morphological analyzers, POS taggers etc. are not yet available in the required quantity and quality [2].
- Lack of standardization and spelling [2].
- Web sources for name lists are available in English, but such lists are not available in Indian languages.
- Although Indian languages have a very old and rich literary history still technology development are recent [3].
- Non-availability of large gazetteer.
- Named entity recognition systems built in the context of one domain do not usually work well in other domains.
- Indian languages are relatively free-order languages [3].

V. GAZETTEER METHOD

The Gazetteer Method maintains the separate list for each Named entities and then applies lookup operation on the list to classify the names [7]. This method require as input a collection of gazetteers, one for each named entity class of interest and one for other class that gives examples of entities that we do not want to extract. For creating gazetteers list this method uses large corpus to create list of named entities. But it does not resolve ambiguity in a given document. Having list of entities in hand makes NER trivial. For example one can extract city name from a given document by searching in the document for each city name in a city list. But this strategy fails because of ambiguous words present in the documents or corpus.

For Example: - For example if in a document we have a name Ganga. That means when we prepare the gazetteer list then Ganga may be in the list of person name and in the list of river name. So there ambiguity exists. And it is difficult task for gazetteer method to correctly identify or tag the Ganga.

A. The gazetteer method work in two phases:

- In the first phase it creates large gazetteers of entities, such as list of cities, name of person, name of river etc and other list of entities of interest.
- In the second phase it uses simple heuristic to identify and classify entities in the context of a given document.

Without resolving ambiguity the system can't perform robust, accurate NER.

B. Advantage of gazetteer method

- The gazetteer based approach results in fast and high precision NER. Since one simple looks for occurrences of any entries in the gazetteer list is required.
- The accuracy of the gazette based method is dependent on the completeness of the gazette used.
- That means if the list is properly maintained and we made the list correctly then it gives very high performance.
- Creating the gazetteer manually is effort-intensive, error-prone and subjective.
- But the problem is how to automatically create a gazetteer with less effort, in less time and with high accuracy using a given document.

C. Disadvantage of gazetteer method

- Ambiguity resolution is difficult.
- Since the words are created repeatedly. So keeping a gazetteer list for these words up-to-date is challenging.
- Without ambiguity resolution the precision is low.

When the list is too large then the searching takes more time to find each word in the list. If we choose sequential search then it takes O (n) time to find a word in the list. Here n is the number of words in the list.

VI. HIDDEN MARKOV MODEL

Name recognition may be viewed as a classification problem, where every word is either part of some name or not part of any name. In recent years, hidden Markov models (HMM's) have enjoyed great success in other textual classification problems—most notably part-of-speech tagging.

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 [17]. Moreover, HMM seems more and more used in NE recognition because of the efficiency of the Viterbi algorithm [Viterbi67] used in decoding the NE-class state sequence. But the performance of a machine-learning system is always poorer than that of a rule-based system.

The Viterbi algorithm (Viterbi 1967) is implemented to find the most likely tag sequence in the state space of the possible tag distribution based on the state transition probabilities. The Viterbi algorithm allows us to find the best *T* in linear time. The idea behind the algorithm is that of all the state sequences, only the most probable of these sequences need to be considered. The trigram model has been used in the present work.

HMM consists of the following:

- Set of States, S where |S|=N. Here, N is the total number of states.
- Start State, S.
- Output Alphabet, O where |O|=k .Here, k is the number of Output Alphabets.
- Transition Probability, A
- Emission Probability B
- Initial State Probabilities π

HMM may be represented as: λ= (A, B, π)[6].

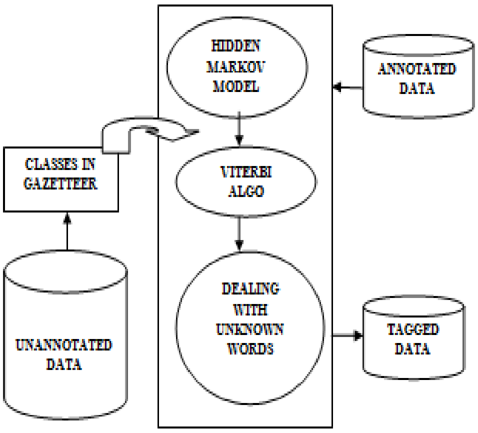

Fig. 2: Architecture of HMM used for NER

VII. EXISTING WORK ON DIFFERENT INDIAN LANGUAGES IN NER

Table 2: Different Approaches According to Their Accuracies.

Author	Language	Approach	Words	Class	Accuracy
[4]	Telugu	CRF	13,425	-	91.95%.
[4]	Telugu	ME	-	-	50.00% Approx
[5]	Tamil	CRF	94K	106	80.44%
[9]	Hindi	ME	25K	4	81.52%
[10]	Hindi	CRF	-	-	60.00%
[12]	Hindi	CRF	-	-	-
[13]	Bengali	CRF	150 K	17	90.7% Approx
[14]	Hindi	SVM	502,974	12	77.17% Approx
[14]	Bengali	SVM	122,467	12	84.00% Approx
[16]	Hindi	ME	-	-	75.89%
[16]	Bengali	SVM	150K	17	90.00% Approx
[16]	Bengali	ME	-	12	80.00% Approx
[17]	Bengali	HMM	150K	16	83.00% Approx

VIII. RESULT ANALYSIS

When we perform gazetteer method on tourism corpus which has 100 sentences. The size of the list increases drastically and for each named entities we have to search entire list from starting which take much time. In our case we have consider four list namely Person (PER), location (LOC), temple, River and rest are assign other tag.

Table 3: Total number of tags in the corpus

	Person(PER)	Location(LOC)	Temple	River
Total	49	250	3	5

To reduce the list size we maintain the separate list for prefix and suffix of these tags. And then find the accuracy of Gazetteer method which is as follows:

Table 4: So the overall accuracy is 40.13% for 100 sentences using Gazetteer method

	Person tag		Location tag	
	Total PER tag	Correctly observed tag	Total LOC tag	Correctly observed tag
	49	28	250	92
Accuracy	57%		37%	

Now we apply Hidden Markov Model on these sentences which are the machine learning approach to identify the named entities. After performing the training on the viterbi algorithm for each sentence we observe the following accuracy:

Table 5: So accuracy is 97.3% for training 100 sentences using HMM.

	Person tag		Location tag	
	Total PER tag	Correctly observed tag	Total LOC tag	Correctly observed tag
	49	46	250	245
Accuracy	95.90%		98%	

When we perform testing on 40 sentences the result is as follows:

Table 6: So accuracy is 93.8% for testing 40 sentences using HMM.

	Person tag		Location tag	
	Total PER tag	Correctly observed tag	Total LOC tag	Correctly observed tag
	14	12	67	64
Accuracy	85.70%		95.50%	

Now we combine these two approaches and perform NER in order to improve accuracy and the result is as follows: In this hybrid approach we first apply Gazetteer method which correctly classifies 28 tags out of 49 PER tag and 92 location entities out of 250 location tags. After that for identifying the remaining tags we apply HMM and the result obtained is as follows:

Table 7: Overall accuracy is 98.375%.

Method	Person tag		Location tag	
	Total PER tag	Correctly observed tag	Total LOC tag	Correctly observed tag
Gazetteer	49	28	250	92
HMM	21	20	158	155
Accuracy	97.95%		98.80%	

IX. Conclusion

Building a NER based system in Hindi using HMM is a very conducive and helpful in many significant applications. We have studied various approaches of NER and compared these approaches on the basis of their accuracies. India is a multilingual country. It has 22 Indian Languages. So, there is lot of scope in NER in Indian languages. Once, this NER based system with high accuracy is build, then this will give way to NER in all the Indian Languages and further an efficient language independent based approach can be used to perform NER on a single system for all the Indian Languages. We perform some experiment using Gazetteer method and

HMM method and get accuracy as 40.13% and 97.30%.Then we combine both the approach to improve the performance and get accuracy as 98.37%.

REFERENCES

[1] P. K. Gupta and S. Arora, "An Approach for Named Entity Recognition System for Hindi: An Experimental Study," in *Proceedings of ASCNT-2009*, CDAC, Noida, India, pp. 103–108.

[2] "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.)Available at: http://www.languageinindia.com/may2011/v11i5may2011.pdf

[3] 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

[4] B. Sasidhar#1, P. M. Yohan*2, Dr. A. Vinaya Babu3, Dr. A. Govardhan4," A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu**"**, http://www.ijcsi.org/papers/IJCSI-8-2-438-443.pdf

[5] Asif Ekbal, Rajewanul Hague, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay 2008 "Language Independent Named Entity Recognition in Indian Languages" *Proceedings of the IJNLP-08 Workshop on NER for South and South East Asian Languages*, Hyderabad, India.

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

[7] Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra "Gazetteer Preparation for Named Entity Recognition in Indian Languages". Available at: http://www.aclweb.org/anthology-new/I/I08/I08-7002.pdf

[8] Sachin Pawar, Rajiv Srivastava and Girish Keshav Palshikar "Automatic Gazette Creation for Named Entity Recognition and Application to Resume Processing "in Tata Research Development and Design Centre, Pune, India.Available at: http://www. pawar_agcfneraatrp_2012.pdf.

[9] S. K. Saha, S. Sarkar, and P. Mitra, "A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition," in *Proceedings of the 3rd International Joint Conference on NLP*, Hyderabad, India, January 2008, pp. 343–349.

[10] A. Goyal, "Named Entity Recognition for South Asian Languages," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South- East Asian Languages*, Hyderabad, India, Jan 2008, pp. 89–96.

[11] S. K. Saha, P. S. Ghosh, S. Sarkar, and P. Mitra, "Named Entity Recognition in Hindi using Maximum Entropy and Transliteration," *Research journal on Computer Science and Computer Engineering with Applications*, pp. 33–41, 2008.

[12] W. Li and A. McCallum, "Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper)," *ACM Transactions on Computational Logic*, pp. 290–294, Sept 2003.

[13] A. Ekbal, R. Hague, and S. Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field," in *Proceedings of ICON*, India, pp. 123–128.

[14] A. Ekbal and S. Bandyopadhyay, "Named Entity Recognition using Support Vector Machine: A Language Independent Approach," *International Journal of Computer, Systems Sciences and Engg (IJCSSE)*, vol. 4, pp.155–170, 2008.

[15] A. Ekbal and S. Bandyopadhyay, "Bengali Named Entity Recognition using Support Vector Machine," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages*, Hyderabad, India, January 2008, pp. 51–58.

[16] M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay, "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi, "*International Journal of Recent Trends in Engineering*, vol. 1, May 2009.

[17] A. Ekbal and S. Bandyopadhyay, "A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies," in *Proceedings of 2nd International conference in Pattern Recognition and Machine Intelligence*, Kolkata, India, 2007, pp. 545–552.

Authors

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.

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.