DETECTION AND CATEGORIZATION OF NAMED ENTITIES IN INDIAN LANGUAGES USING HIDDEN MARKOV MODEL

Deepti Chopra[1] and Sudha Morwal[2]

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

ABSTRACT

Named Entity Recognition (NER) is the task in which proper nouns in a given document are discovered and then categorized into respective classes. The various classes of proper nouns may be name of location, name of person, Organization, River, Quantity, Time, Percentage etc. Today, there is a great need to perform NER in the Indian Languages, since not much work has been done in the field of Information retrieval in the Indian languages. In this paper, we have tried to explain NER, different approaches of NER and finally some results of NER in natural languages.

KEYWORDS

HMM; NER; Performance Metrics; Accuracy

1. INTRODUCTION

Named Entity Recognition (NER) is the process that involves finding the NEs in a corpus and then be able to distinguish them into various classes of NEs such as person, location, organization, time, river, sport, vehicle, country, state, quantity, number, time etc. The various applications of NER are: Question Answering, Information Extraction, Automatic Summarization, Machine Translation, Information Retrieval etc.

Consider a Hindi sentence:

नैनीताल/LOC कुआऊँ/LOC पहाड़िया/OTHER म/OTHER स्थित/OTHER है/OTHER ।/OTHER यह/OTHER नैनीताल/LOC हा/OTHER था/OTHER जहाँ/OTHER कनलजम्सएडवडकाबट/PER का/OTHER जन्म/OTHER हुआ/OTHER था/OTHER ।/OTHER

In the above sentence, we have tagged नैनीताल and कुआऊँ as LOC, since these are locations. Here, कनलजम्सएडवडकाबट is the name of person, so it is shown by a PER tag. The OTHER tag signifies not a named entity tag. There are many challenges that have to be dealt with while performing Named Entity Recognition in Indian languages. Indian languages lack in proper resources, so before performing Named Entity Recognition in Indian languages, we have to carry out the task of Corpus development which include doing annotation on the raw text, preparing Gazetteer etc. Indian languages are free word order, inflectional and morphologically rich in nature. In Indian languages, there are numerous named entities that also exist as common nouns in the dictionary.

2. APPROACHES OF NER

For performing Named Entity Recognition in natural languages, two approaches are used. These include: [5] [1] [18] Rule Based Approach and Machine learning based Approach [11] [6] [16].

2.1 Rule based Approach

It is one of the primitive approaches of NER and is referred to as handcrafted approach. Rule based approaches are generally not performed alone. These are coupled with the Machine learning based approaches in order to increase the accuracy of a NER based system. It is of two types:

2.1.1 List Lookup Approach

This approach makes use of Gazetteers that contain collection of various lists of Named Entity classes and a list look up procedure is performed to find whether a given word is Named Entity or not. The Named Entity tag allotted to a word depends on the list in which it is found. Web contains collection of Named Entities which are not in the Indian languages but are in English. So, we can prepare Gazetteers by using transliteration approach in the Indian languages. In a domain specific document, seed values are used to remember the context patterns and then by using the bootstrapping approach, the Named Entities are retrieved.[17]List Lookup Approach is very easy to implement. The major drawback of List Lookup Approach is its inability to solve the problem of ambiguity.

2.1.2 Linguistic Approach

In Linguistic approach, a linguist frames certain language dependent rules that help in the identification of Named Entities in a document. [3][20][19]Such rules cannot be used to perform Named Entity Recognition in the other languages. [11]

2.2 Machine Learning Based Approach

This approach is often employed in performing Named Entity Recognition in natural languages and is referred to as Statistical or automated. These are of following types:

2.2.1 Hidden Markov Model (HMM)

HMM is a Machine learning based approach that consists of hidden set of states. The output of HMM is the sequence of tokens which is optimal state sequence. HMM makes a use of the Markov Chain Property i.e. the probability of the happening of the next state is dependent on the occurrence of just previous state. HMM can be implemented very easily.Fig1 But, the drawback of HMM is that it needs lot of training, to get good results and it cannot solve large dependencies. [12]

2.2.2 Maximum Entropy Markov Model (MEMM)

It involves combination of both the theories of Hidden Markov Model as well as Maximum Entropy Model. While training in MEMM, the unknown values are never conditionally independent to each other but are rather connected together. MEMM has the advantages that it is able to resolve the large dependency problem of HMM. Also, as compared to HMM, MEMM has higher precision and recall.

The disadvantage of MEMM approach is that it suffers from the label bias problem i.e. the transition probabilities from a given state must always sum to one. So, at all times, MEMM favor the states through which fewer transitions occur. [16]

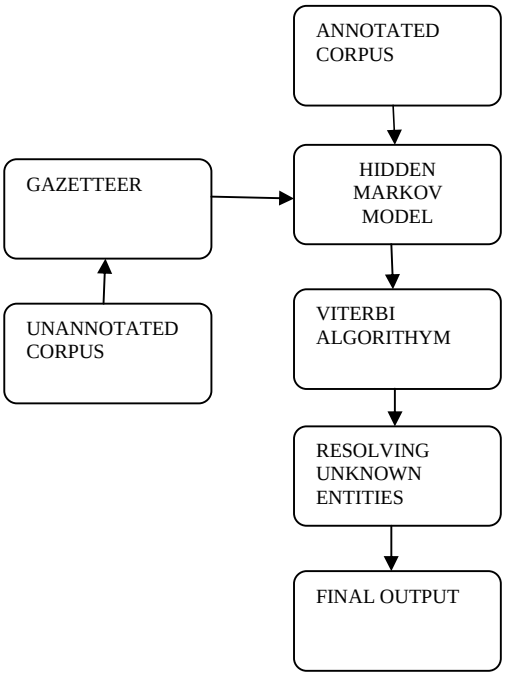

Figure 1: Architecture description of NER using HMM

2.2.3 Conditional Random Field (CRF)

Conditional Random Field is referred to as the undirected graphical model. A different approach as compared to the other approaches, CRF also takes into account the neighboring samples or the context information. CRF is called as Random field because it estimates the conditional probability on the subsequent node given the current node probability values. Like MEMM, CRF also can resolve the large dependency problem and has high precision and recall. The advantage of CRF over MEMM is that the CRF can solve the label bias problem which is faced by MEMM. [3]

2.2.4 Support Vector Machine (SVM)

Vapnik introduced this methodology. SVM is a supervised machine learning based approach. The main aim of this methodology is to locate whether a particular vector is a part of a specific target class or not.[2] In this approach, both the training and the testing information belong to the single dimension vector space.

While training, a hyper plane is created that is used to classify the elements into two classes: the positive and the negative classes. These two classes are present on both the opposite faces of a hyper plane. This approach also calculates margin, which is the displacement of each vector from the hyper plane. The advantage of this methodology is that it provides high precision for the text categorization problem. [4]

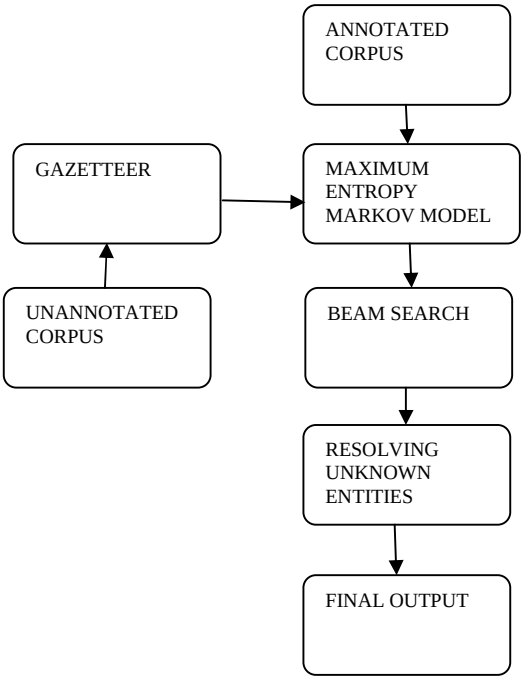

Figure 2: Architecture description of NER using MEMM

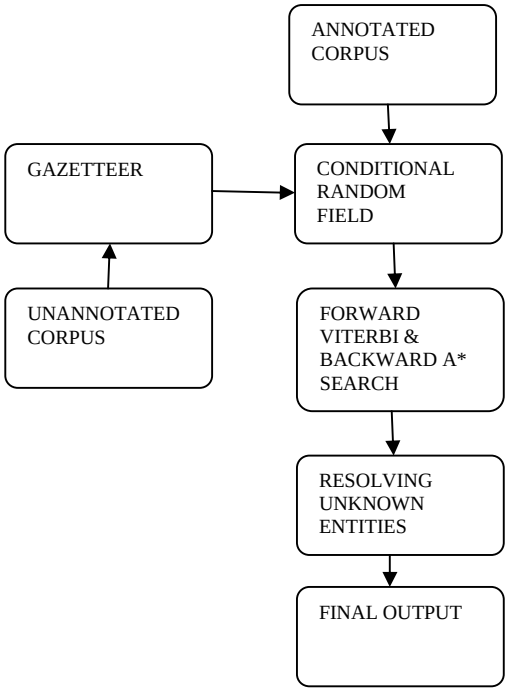

Figure 3: Architecture description of NER using CRF

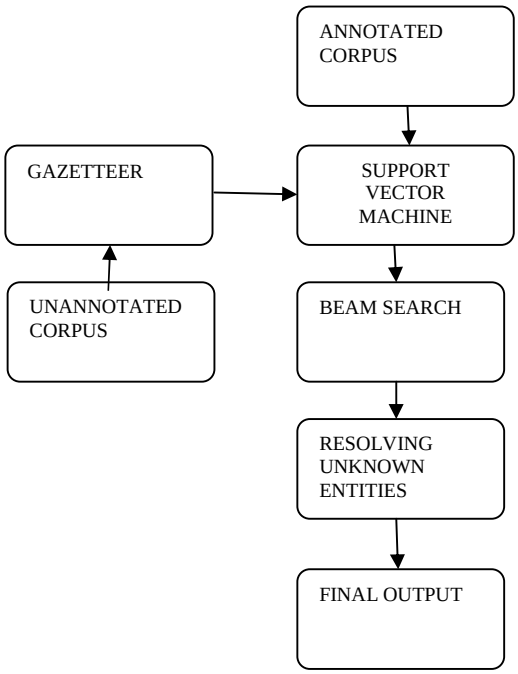

Figure 4: Architecture description of NER using SVM

2.2.5 Decision Tree

It is a famous approach whose aim is to detect and classify the Named Entities in a given document. In this methodology, some identification rules are applied to the unannotated training document so that all the Named Entities are recovered. After this, we match the Named Entities with the actual Named Entities given by the human expert. If the Named Entity is same as the one given by the human, then it is known as the positive example otherwise it is referred to as the negative example. [7]. A decision tree is therefore constructed that categorizes the Named Entities present in the testing document. [9] The leaf node of a decision tree gives the final result of test.

3. RESULTS

We have considered different corpus of Hindi, Punjabi and Urdu. The Named Entities that we have used are: LOC (Name of Location), PER (Name of Person), ORG (Name of Organization), TIME, MONTH, RIVER, SPORT, VEH (Name of Vehicle) and QTY (Quantity).This is shown in TABLE 1. Figure 5 depicts that we have obtained different accuracies for different tags in Hindi, Punjabi and Urdu corpus. We have obtained following F-Measure by performing NER in Hindi, Punjabi and Urdu as: 98.16%, 96.6% and 95.5%.

TABLE 1 Named Entities used in NER Using HMM

SNO	TAGS
1	LOC (Name of Location)
2	PER (Name of Person)
3	QTY (Name of Quantity)
4	TIME
5	ORG (Name of Organization)
6	SPORT
7	RIVER
8	VEH (Name of Vehicle)
9	MONTH

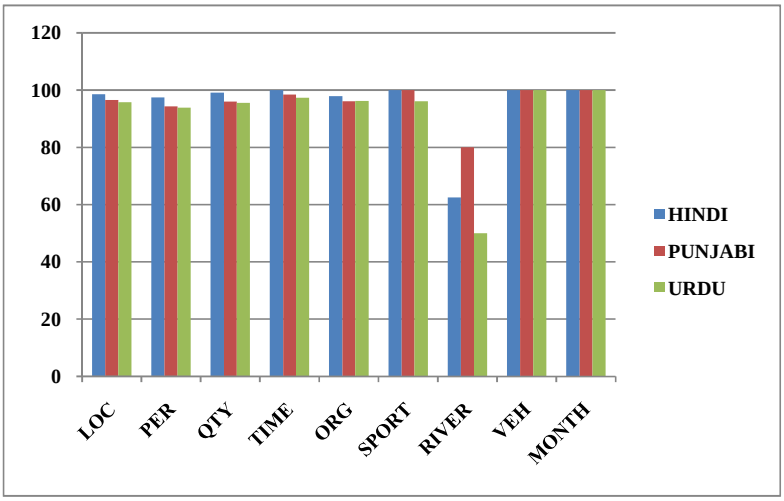

Figure 5 Accuracy obtained in different tags in Hindi, Punjabi and Urdu

4. CONCLUSION

HMM is considered as one of the simplest and efficient approaches of Named Entity Recognition. NER is a very important subtask of information extraction. At Present, we have performed NER in Hindi, Punjabi and Urdu and the F-Measure obtained in these languages are: 98.16%, 96.6% and 95.5%.There lays a lot of scope in NER in many Indian languages on which not much work has been performed as yet.

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. Available at: http://www.aclweb.org/anthology-new/I/I08/I08-5014.pdf

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

[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.Available at: http://www.mt-archive.info/IJCNLP-2008-Ekbal.pdf

[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..Available at: http://www.aclweb.org/anthology-new/I/I08/I08-5008.pdf

[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 2011Available at: http://www.ijcsi.org/papers/IJCSI-8-2-438-443.pdf

[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.Available at: http://ijcsi.org/papers/7-6-239-245.pdf

[7] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine .Spyropoulos."Learning Decision Trees for Named-Entity Recognition and Classification"Available at: http://users.iit.demokritos.gr/~petasis/Publications/Papers/ECAI-2000.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" Available at: http://www.jatit.org/volumes/research-papers/Vol13No2/4Vol13No2.pdf

[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.)Available at: 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". Available at:http://www.aclweb.org/anthology-new/I/I08/I08-7002.pdf

[18] Sujan Kumar Saha Sanjay Chatterji Sandipan Dandapat . "A Hybrid Approach for Named Entity Recognition in Indian Languages"Available at: http://aclweb.org/anthology-new/I/I08/I08-5004.pdf

[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.
Available at: http://www.ijcsit.com/docs/vol1issue4/ijcsit2010010416.pdf

[20] Vishal Gupta, Gurpreet Singh Lehal "Named Entity Recognition for Punjabi Language Text Summarization" International Journal of Computer Applications (0975 – 8887) Volume 33– No.3, November 2011.
Available at:http://www.advancedcentrepunjabi.org/pdf/NER%20for%20Summarization.pdf

About 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. She has published many papers in international journals and conferences.

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. She has published many papers in International Conferences and Journals.