Named Entity Recognition in Punjabi Using Hidden Markov Model

Deepti Chopra[1], Sudha Morwal[2]
Department of Computer Science
Banasthali Vidyapith
Jaipur, INDIA
deeptichopra11@yahoo.co.in, sudha_morwal@yahoo.co.in

Abstract— **Named Entity Recognition (NER) is a task to discover the Named Entities (NEs) in a document and then categorize these NEs into diverse Named Entity classes such as Name of Person, Location, River, Organization etc. Since, huge amount of work in NER has been done in English; so, we now need to concentrate ourselves in performing NER in the Indian languages (IL). As, Punjabi is not only the Indian language but also it is the official language of Punjab, So we have developed NER based system for Punjabi. This paper discusses about NER, approaches of NER and the results achieved by us by performing NER in Punjabi using Hidden Markov Model (HMM).**

Keywords- Accuracy; HMM; Named Entities; NER; Performance Metrics

I. INTRODUCTION

Named Entity Recognition (NER) is considered as one of the key task in Natural language Processing and it forms the base for numerous applications such as Information Retrieval, Information Extraction, Question Answering, Text Summarization, Machine Translation etc.[1][2]

NER involves identification as well as the task of classification of Named Entities (NEs) in a given document. It may be defined as the procedure to search for the Named Entities (NEs) or proper nouns in a corpus and then classify them into different classes of NEs such as Name of Person, Organization, Location, City, River, Quantity, Percentage, Time etc.

Consider a sentence in Punjabi:

ਕਹਿਣ-'ਛੱਡ/OTHER ਪਰ੍ਹਾਂ/OTHER ਚਰਨਕੋਰੇ/PER ,/OTHER ਕਿਉਂ/OTHER ਕਲਪੀ/OTHER ਜਾਨੀ/OTHER ਆਂ/OTHER ।/OTHER

ਟੈਲਾਂ/OTHER ਲੁਆ/OTHER ਗੁਰਦੁਆਰੇ/LOC ।/OTHER

ਮਿਸਰੀ/DRYFRUIT ,/OTHER ਕਾਜੂ/DRYFRUIT ,ਬਦਾਮ/DRYFRUIT ,/OTHER ਖਰੋਟ/DRYFRUIT ਚੱਬਣ/OTHER ਨੂੰ/OTHER ।/OTHER

In the above tagged Punjabi text, NER based system identifies the NEs and its class and then allot an appropriate tag to it. In this sentence, various classes of NEs are: {'PER', 'LOC', DRYFRUIT'}. Here PER signifies Name of Person and LOC signifies Name of Location.

In the following paper, we have discussed about NER based system particularly for Punjabi language using Hidden Markov Model (HMM)

II. APPROACHES FOR NER

There are two methods used for performing NER i.e. Rule Based Approach and Machine Learning Based Approach.. [3][4][5]

The Rule Based Approach can either be List lookup Approach or a Linguistic Approach.

To perform NER using List lookup Approach or a Linguistic approach, a lot of human effort is required. In the List lookup Approach, firstly the Gazetteers are constructed, that contain collection of Named Entity classes. Then, we can perform a search operation to find out that a given word in a corpus is found under which category of a Named Entity class. In a Linguistic Approach, a Linguist frames certain set of rules to identify the NEs in a corpus and also to classify these NEs into different Named Entity classes. [1][6][7][8]

In a Machine learning based approach, less human effort is required. So, it is also known as automated approach. It is of the following types: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), Support Vector Machine (SVM) and Decision Tree. [1][9][10]

Among all these approaches, HMM is one of the easiest approach to implement but it requires large amount of training. There are basically 3 parameters that are employed in HMM, these include: Start Probability,

Transition Probability and Emission Probability. [11][12][16] A Viterbi Algorithm makes use of these parameters to obtain the optimal state sequence.

MEMM has higher recall and precision compared to HMM but it suffers from the label bias problem. [8][10] In CRF, if we know the conditional probability initially then we can estimate the conditional probability at a later point of time. [6] In SVM, we determine if a given vector belongs to a specific target class or not.[12][19] In Decision tree, we can classify the entities as positive entities or negative entities, depending on whether they are according to human interpretation or not.[13][14][15]

III. CHALLENGES IN INDIAN LANGUAGES

Following are the issues in performing NER in Indian languages (IL):

- Unlike in English, no Capitalisation concept is present in IL.
- IL are morphological and inflectionally rich and free word order languages
- Web lack in the resources in IL.
- NEs in IL exist as common nouns also.
- There lies a lack of proper standardisation in spellings in IL.[7][9][18]

IV.PERFORMANCE METRICS

Performance Metrics is a measure to describe the performance of a NER system in terms of Precision, Recall and F-Measure.

We can refer to the output of NER based system as “response” and NEs identified by the human beings may be referred to as “answer key”.[10] So, we describe the following definitions:

1. Correct-If response is identical to the answer key.
2. Incorrect-If response is not identical to the answer key.
3. Missing-If the answer key is tagged but the response is not tagged.
4. Spurious-If response is tagged but answer key is not tagged.

Following parameters are used to judge the performance of a NER based system: [2][3][9]

- Precision (P): It may be defined as follows:

  P = Correct/ (Correct + Incorrect + Missing)
- Recall (R): It may be defined as follows:

  R = Correct / (Correct + Incorrect + Spurious)
- F-Measure: It may be defined as follows:

  F-Measure = (2 * P * R) / (P + R)

V. OUR METHODOLOGY

We have developed a NER based system for Punjabi. We took Punjabi text from the web which are stories written in Punjabi. We then performed the task of Corpus Development or the Annotation task. The tags that we have considered for our corpus are shown in the TABLE I

TABLE I. Various Named Entity Tags. NE TAG: Named Entity Tags
PER: Name of Person, LOC-Location, CO-Country and QTY-Quantity

NE TAG	EXAMPLE
PER	ਚਰਨਕੋਰੇ, ਲਾਜੋ
LOC	ਛਾਰਪੁਰ, ਗੁਰਦੁਆਰੇ
CO	ਮਰੀਕੇ, ਕਨੇਡੇ
TIME	ਚਾਰਦਿਨ, ਸਵੇਰੇ, ਤਕਾਲੀਂ
DRUG	ਫੀਮ, ਭੁੱਕੀ, ਜਰਦਾ
DRYFRUIT	ਮਿਸਰੀ, ਕਾਜੂ, ਬਦਾਮ, ਖਰੇਟ
GARAM_MASALA	ਲੈਗਾਂ, ਲੈਚੀਆਂ
QTY	ਦੂਣੀ
FOOD	ਦੁੱਧਾਂ, ਘੇਏ, ਦੁੱਧ

The tags chosen for the annotation depends on the content of the document under consideration and the choice of tags may vary from one person to another.

The next step after Corpus development is HMM Training or HMM Parameter Estimation. Here, input is the annotated Punjabi text file and output is the three parameters of HMM i.e. Start Probability, Transition Probability and Emission Probability. Start Probability may be defined as the probability that a given Named Entity tag exists first in a sentence. Transition probability may be defined as the probability of existence of the next tag in a sentence given a particular tag exists at present. Emission Probability may be defined as the probability of occurrence of an output sequence for a given particular tag.

The next step after performing HMM Training is the HMM Testing. In this, we test the performance of the trained NER based system and run the viterbi algorithm whose input is the test sentences and output is optimal state sequence.

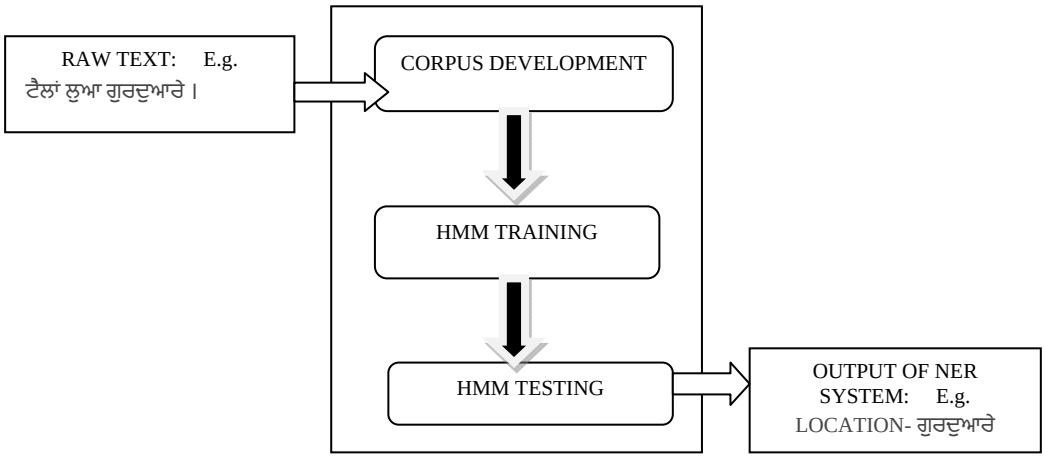

Figure 1 NER in Punjabi using HMM

Then, using Performance Metrics we can describe the performance of the NER based system. Our whole methodology of NER based system using HMM for performing NER in Punjabi is depicted in Fig1.

Consider an Example:

STEP 1: Consider the raw text in Punjabi:

ਟੈਲਾਂ ਲੁਆ ਗੁਰਦੁਆਰੇ ।

ਲਾਜੋ ਵੀ ਖ਼ਬਰਿਆ ਤਾਈਉਂ OTHER ਨੀ ਆਈ ।

STEP 2: Annotated text obtained is given as:

ਟੈਲਾਂ/OTHER ਲੁਆ/OTHER ਗੁਰਦੁਆਰੇ/LOC ।/OTHER

ਲਾਜੋ/PER ਵੀ/OTHER ਖ਼ਬਰਿਆ/OTHER ਤਾਈਉਂ/OTHER ਨੀ/OTHER

Sਆਈ/OTHER ।/OTHER

STEP 3: HMM training is performed i.e. HMM parameters (Start Probability, Transition Probability and Emission Probability) are computed:

- start probability {'OTHER': 0.5, 'PER': 0.5}
- transition probability: {'LOC': {'LOC': 0.0, 'OTHER': 1.0, 'PER': 0.0}, 'OTHER': {'LOC': 0.1111111111111111, 'OTHER': 0.6666666666666666, 'PER': 0.1111111111111111}, 'PER': {'LOC': 0.0, 'OTHER': 1.0, 'PER': 0.0}}
- emission probability{'LOC': {'ਲੁਆ': 0, 'ਲਾਜੋ': 0, 'ਗੁਰਦੁਆਰੇ': 1.0, '।': 0, 'ਟੈਲਾਂ': 0, 'ਆਈ': 0, 'ਨੀ': 0, 'ਖ਼ਬਰਿਆ': 0, 'ਵੀ': 0, 'ਤਾਈਉਂ': 0}, 'OTHER': {'ਲੁਆ': 0.1111111111111111, 'ਲਾਜੋ': 0, 'ਗੁਰਦੁਆਰੇ': 0, '।': 0.2222222222222222, 'ਟੈਲਾਂ': 0.1111111111111111, 'ਆਈ': 0.1111111111111111, 'ਨੀ': 0.1111111111111111, 'ਖ਼ਬਰਿਆ': 0.1111111111111111, 'ਵੀ': 0.1111111111111111, 'ਤਾਈਉਂ': 0.1111111111111111}, 'PER': {'ਲੁਆ': 0, 'ਲਾਜੋ': 1.0, 'ਗੁਰਦੁਆਰੇ': 0, '।': 0, 'ਟੈਲਾਂ': 0, 'ਆਈ': 0, 'ਨੀ': 0, 'ਖ਼ਬਰਿਆ': 0, 'ਵੀ': 0, 'ਤਾਈਉਂ': 0}}

STEP 4: We now run Viterbi Algorithm on trained HMM for our test sentences:

Test Sentence: ਲਾਜੇ ਵੀ ਖ਼ਬਰਿਆ ਤਾਈਉਂ OTHER ਨੀ ਆਈ ।

Output of Viterbi Algorithm : (2.477927800044668e-07, ['PER', 'OTHER', 'OTHER', 'OTHER', 'OTHER', 'OTHER', 'OTHER'])

VI RESULTS

We have generated a Punjabi Corpus from the Punjabi short stories available on the web. We have done annotation manually. The Named Entity tags used in this Punjabi Corpus are displayed in TABLE 1. We have developed a NER based system that employs HMM based approach to perform NER in Punjabi. We have performed training on 631 sentences or 3887 tokens of Punjabi text. Out of total 69 tokens of the testing sentences, 61 tokens could be identified correctly. So, Accuracy, Recall, Precision and F-Measure obtained are 88.4%. The results are shown in TABLE II. Fig 2 depicts that NER system has identified tags such as PER (Name of Person) and TIME with 100% Accuracy for the given test data.

TABLE II. Results of NER in Punjabi using HMM

TOTAL TRAINING SENTENCES	631		
TOTAL TESTING SENTENCES	10		
NO. OF TESTING SENTENCES GIVING WRONG RESULT	1		
	NO. OF TAGS IN TRAINING SENTENCES	NO. OF TAGS IN TESTING SENTENCES	NO. OF CORRECT TAGS IDENTIFIED
PER	13	6	6
LOC	4	0	0
OTHER	3846	57	50
CO	2	1	0
TIME	7	4	4
DRUG	4	0	0
DRYFRUIT	5	1	1
GARAM_MASALA	2	0	0
QTY	1	0	0
FOOD	3	0	0
TOTAL	3887	69	61
ACCURACY=(61/69)*100=88.4%	RECALL = 88.4%	PRECISION = 88.4%	F-MEASURE = 88.4%

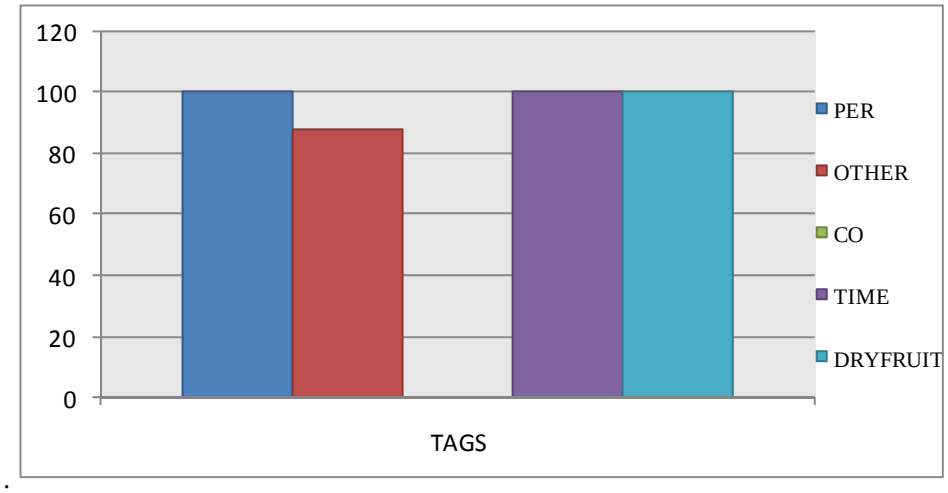

Figure 2 Results depicting accuracies of individual tags

VII. CONCLUSION

In this paper, we have discussed about NER, Challenges in NER in the Indian languages, Performance Metrics and finally our methodology and the results. We have obtained F-Measure and accuracy of about 88.4% by performing NER in Punjabi using Hidden Markov Model (HMM). HMM is known to be one of the easiest approaches of NER. So, now our challenge lies in developing a language independent based NER system that is used to perform NER in all the rest of the Indian languages and use the hybrid approach that would involve combination of HMM with the other approaches and this would lead to the increase in the accuracy of the NER based system. Also, we need to develop a system that would help in the Corpus Development and assist in generating the annotated text from a given raw text.

ACKNOWLEDGEMENT

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

REFERENCES

[1] 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

[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] 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

[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 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

[7] 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

[8] 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.

[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

[13] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos.”Learning Decision Trees for Named-Entity Recognition and Classification Available at: http://users.iit.demokritos.gr/~petasis/Publications/Papers/ECAI-2000.pdf

[14] 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

[15] Padmaja Sharma, Utpal Sharma, and 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

[16] 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

[17] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos.”Learning Decision Trees for Named-Entity Recognition and Classification”
Available at: http://users.iit.demokritos.gr/~petasis/Publications/Papers/ECAI-2000.pdf

[18] 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

[19] 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

[20] 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