NERHMM: A TOOL FOR NAMED ENTITY RECOGNITION BASED ON HIDDEN MARKOV MODEL

Sudha Morwal and Deepti Chopra

[1]Department of Computer Engineering, Banasthali Vidyapith, Jaipur (Raj.), INDIA

sudha_morwal@yahoo.co.in
deeptichopra11@yahoo.co.in

ABSTRACT

Named Entity Recognition (NER) is considered as one of the key task in the field of Information Retrieval. NER is the method of recognizing Named Entities (NEs) in a corpus and then organizing these NEs into diverse classes of NEs e.g. Name of Location, Person, Organization, Quantity, Time, Percentage etc. Today, there is a great need to develop a tool for NER, since the existing tools are of limited scope. In this paper, we would discuss the functionality and features of our tool of NER with some experimental results.

KEYWORDS

HMM, NER, F-Measure, Accuracy, NEs

1. INTRODUCTION

Named Entity Recognition 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. [8][11]

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. RELATED WORK

Natural language toolkit (NLTK) is a free and open source computational linguistic tool. Apart from Named Entity Recognition, this tool can be used for performing tokenization, classification, stemming, parsing, tagging etc. NLTK provides support in carrying out research in areas like linguistic, artificial intelligence, machine learning, information retrieval etc.[21]

Scikit-learn also known as Scikits.learn is an open source machine learning tool. It efficiently implements the algorithms of Hidden Markov Model. [22]

Stanford Named Entity Recognizer (NER) is a java based NER toolkit that uniquely tags Named Entities such as Name of Person, Company, gene, proteins etc. Stanford NER is also known as CRF Classifier. [23]

3. PROPOSED TOOL-NERHMM

Hidden Markov Model (HMM) is Statistical approach that was initially used for Speech Recognition but it can now be used to perform Named Entity Recognition also. HMM has three parameters: Start Probability (π), Transition probability (A = aij) and Emission Probability (B = {bj(O)}), represented as λ = (A, B, π). [12]

Start Probability (π) is the probability that a given tag occurs first in a sentence.

Transition probability (A = aij) is the probability of occurrence of the next tag j in a sentence given the occurrence of particular tag i at present.

Emission Probability (B = {bj(O)}) is the probability of occurrence of output sequence given a state j.

For performing NER using HMM, we need to perform two tasks i.e. HMM Training and HMM Testing. Before, performing HMM Training, we need to perform Annotation that accepts raw data as an input and generates the annotated data or tagged data as an output.

Consider a raw text in Hindi:

geeta badminton kheti hai |

The above sentence is a raw text and is not tagged. We need to perform annotation on this raw text to obtain the annotated data.

Output of Annotation is:

geeta/PERSON badminton/SPORT khelti/O hai/O |/O

In the above sentence, geeta is a name of person, so it is tagged with a PERSON tag, badminton is a name of sport, so we have tagged it with SPORT tag. ‘O’ signifies not a named entity tag or not a proper noun.

The input to the HMM Training process is the annotated data and the output is the three parameters of HMM. The next step is the HMM Testing that accepts sentences as an input and generates optimal state sequence and Named Entities as an output.

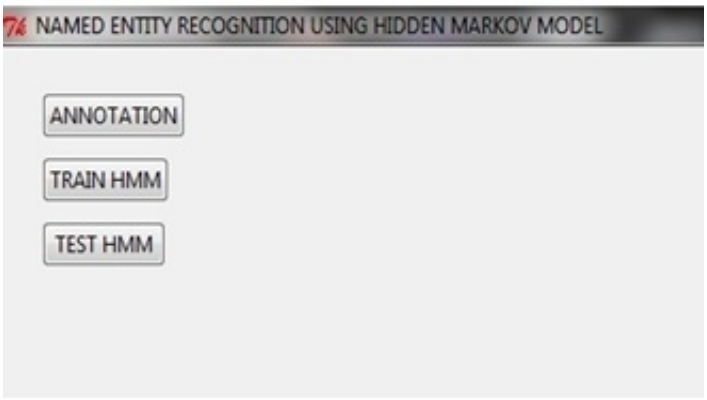

Figure 1: NER tool using HMM

We have made a tool NERHMM that perform all the above mentioned tasks. Initially, our aim was to perform NER in the Indian languages. But, the tool we have developed is able to perform NER in all the natural languages. Figure1 displays the first screen of our tool.

When we click on the 'ANNOTATION' button, we have an option to either write the raw text or select the unannotated text or raw text using a browse button. We can then choose appropriate tag from the generated list to tag each token in a sentence to obtain annotated data. This process of converting the raw text into annotated data is known as 'corpus development phase'. Using NERHMM, we can annotate all the natural languages text by using any number and any kind of tags to obtain the annotated text. Still there are many languages for which annotated data do not exist on web. So, using NERHMM tool we can obtain annotated data that can be further be utilised to perform various Natural language processing task such as NER.

When we click on 'TRAIN HMM' button, we have an option to either write the annotated data or select from the existing using browse button. The output of TRAIN HMM is shown in Figure 2. In Figure 2, we have states= {'OTHER', 'LOC', 'PER', 'TIME', 'SPORT', 'MONTH'}.Here OTHER means not a Named Entity tag, PER is Name of Person tag and LOC is a location tag. Observation is a test sentence or set of test sentences on which we wish to perform NER.

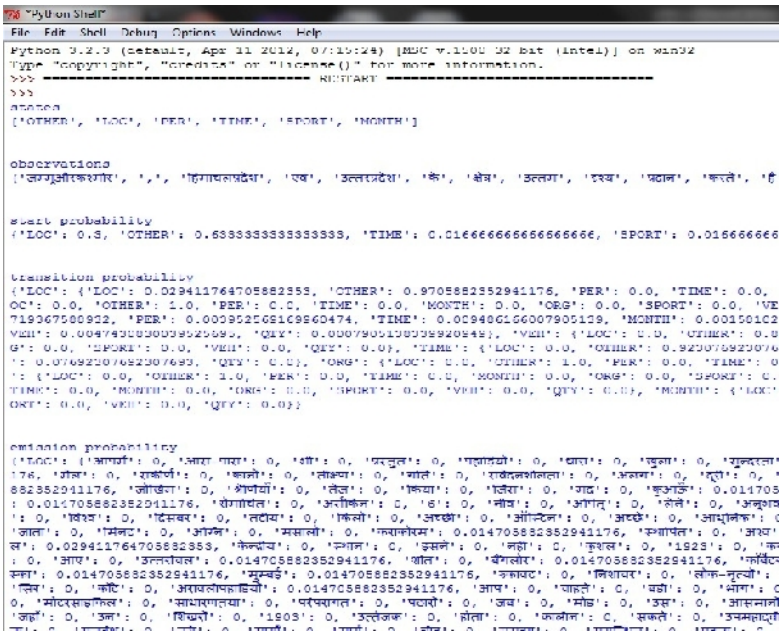

Figure 2: Start Probability, Transition Probability and Emission Probability

HMM parameters are calculated by the tool, NERHMM automatically and are displayed as shown in Figure 2. Finally, on clicking 'TEST HMM', we can either write test sentence(s) or select from a file using browse button. Viterbi algorithm is made to run that accepts all the HMM parameters computed by the tool and displays optimal state sequence as shown in Figure 3. Thus, in Figure 3, Jammu Kashmir, Himachal Pradesh and Uttar Pradesh are the Names of locations, so the output of NERHMM is shown by a LOC tag. 'OTHER' signifies that the rest of the tokens are not Named Entities.

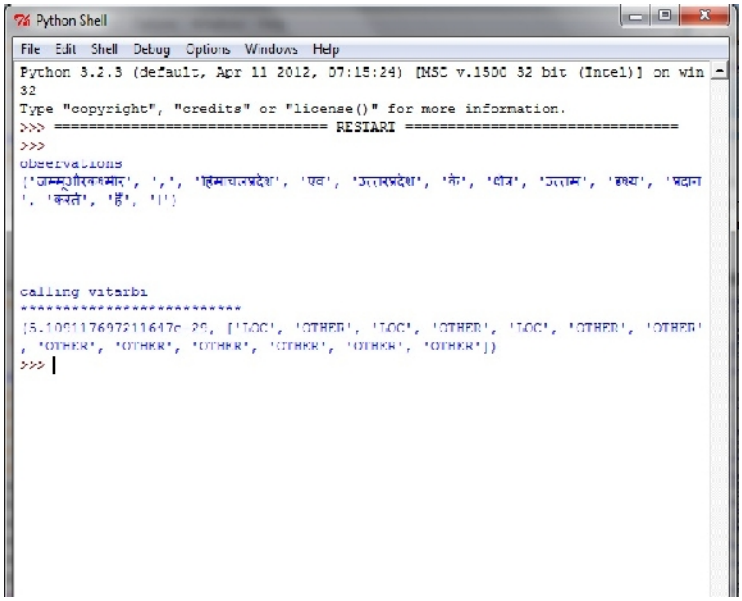

Figure 3: Output of tool showing optimal state sequence

4. FEATURES OF OUR PROPOSED TOOL

Some of the characteristic features of our NER tool are listed below:
1. Our tool works for all the Natural languages. This tool has been tested for languages such as English, Hindi, Bengali, Urdu, Punjabi, Marathi and Telugu.
2. This tool is not domain specific in nature. It has been tested for documents from tourism domain, general sentences and short stories.
3. If we perform large amount of training, then we obtain high accuracy. We have been able to achieve more than 90% of accuracy on testing.
4. The tags used in a document are not fixed. They can be even modified according to the individual desire. Hence, the tags used are of dynamic nature.
5. This tool is also suited to perform part-of-speech tagging, in which standard tags may be used such as NNP, VB, and JJ etc.
6. This tool can include rich tag set. E.g. the location tag may further get split into state, city, country, street, town, palace, temple etc. tags.
7. This proposed tool also facilitates in annotation of raw text to obtain annotated text. This tagged text can further be utilized for other NLP applications.
8. This tool can handle multilingual task i.e. it can perform Named Entity Recognition on document containing multiple languages. This has been tested for a document containing languages such as English, Hindi, Telugu, Bengali and Punjabi.
9. This tool is very user friendly. It solves the major problem of parameter estimation of Hidden Markov Model and it also assist in achieving annotated document from the raw text.

5. RESULTS

Figure 4 shows analysis in terms of F-Measure for files of different sizes. It depicts that in a file having 12 tokens, we have achieved 15% F-Measure and in file having 29 tokens, 17.94% F-Measure is achieved. Till now, we have performed training and testing on multilingual data i.e. data from Hindi, Bengali, Urdu, English, Punjabi and Telugu are combined. We have done training on 42,784 tokens and we have observed that as the amount of training increases, the F-Measure also increases henceforth the performance of a NER based system is determined by the amount of training performed on it.

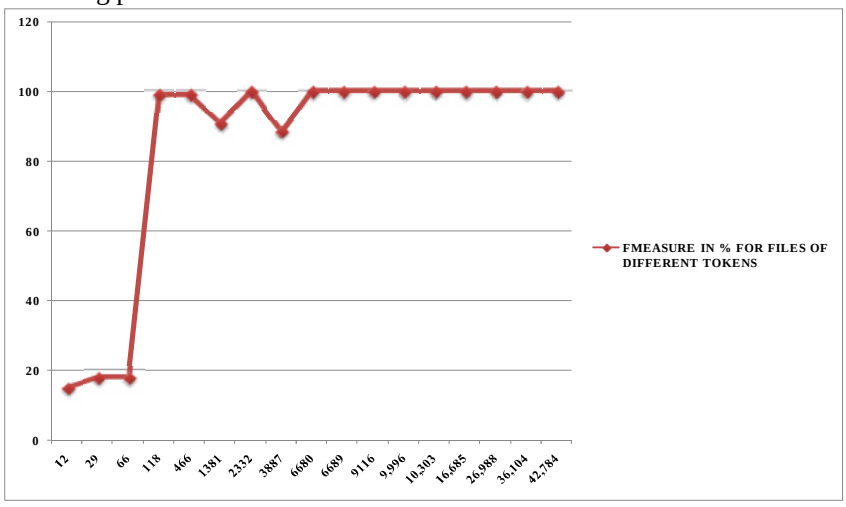

Figure 4: F-Measure in % for files of different sizes

6. CONCLUSION

HMM is considered as one of the simplest and efficient approaches of Named Entity Recognition. We have introduced a tool that provides an easiest way to perform NER in all the natural languages using HMM. There are many natural languages that are resource poor in nature. So, this tool also facilitates in annotation on the raw corpus to obtain the annotated or tagged corpus. In other words, this tool helps in the Corpus Development. In this tool, we have the facility to use the tags of our own choice according to the context of the corpus that we are referring to. At Present, we have performed NER in Hindi, Punjabi, Urdu, English, Marathi, Telugu and Bengali. We have used document of various sizes for NER and through analysis we arrive at a conclusion that as the amount of training increases, the Performance of a NER based system also improves.

ACKNOWLEDGEMENT

I would like to thank all those who have 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

[21] NLTK Toolkit. Available at: http://nltk.org/

[22] Scikit-learn tool. Available at: http://scikit-learn.org/stable/

[23] Stanford Named Entity Recognizer. Available at: http://nlp.stanford.edu/software/CRF-NER.shtml

AUTHORS

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.

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.