IDENTIFICATION AND CLASSIFICATION OF NAMED ENTITIES IN INDIAN LANGUAGES

Sudha Morwal and Deepti Chopra

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

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

ABSTRACT

The process of identification of Named Entities (NEs) in a given document and then there classification into different categories of NEs is referred to as Named Entity Recognition (NER). We need to do a great effort in order to perform NER in Indian languages and achieve the same or higher accuracy as that obtained by English and the European languages. In this paper, we have presented the results that we have achieved by performing NER in Hindi, Bengali and Telugu using Hidden Markov Model (HMM) and Performance Metrics.

KEYWORDS

Accuracy, HMM, Named Entities, NER, Performance Metrics

1. INTRODUCTION

Named Entities (NEs) are the proper nouns or the entities that represent the Name of Person, Location, Organisation, River, Percentage, Quantity and Time etc. NER is used extensively in Natural Language Processing. NER is the task to categorize all the NEs or the proper nouns in a document into different NEs classes' e. g Person, Location, Organisation, City, State, and Country etc [1][2].

Consider an annotated sentence in Hindi:

"प्रधानमंत्री/OTHER अटलबिहारीवाजपेयी/PER ने/OTHER कहा/OTHER है/OTHER कि/OTHER वित्तमंत्री/OTHER यशवंतसिन्हा/PER द्वारा/OTHER प्रस्तुत/OTHER बजट/OTHER विकास/OTHER की/OTHER गति/OTHER को/OTHER तेज/OTHER करेगा/OTHER ।/OTHER"

In the above sentence, the NER based system identifies the NEs and classify them into different NEs classes. Here, अटलबिहारीवाजपेयी and यशवंतसिन्हा are the NEs and are Name of Persons so these are allotted Name of Person tag (PER).The rest of the tokens are given OTHER tag since these are not Named Entities.

Various applications of NER include – Information Retrieval, Information Extraction, Text Summarization, Machine Translation, question answering system etc. Although, a lot of work has been accomplished in NER in English, Chinese and Spanish etc. But, no significant amount of work has been accomplished in NER in Indian languages (IL). This is due to many reasons. Firstly, IL is free word order, inflectional as well as morphologically rich in nature. Secondly, Unlike in English, there is no concept of Capitalisation in IL, in which capital letter is used to

DOI : 10.5121/ijnlc.2013.2105

detect NEs in a document. Thirdly, web lack in the resources in the Indian languages. Fourthly, IL contains NEs which also lie in dictionary as common nouns. So, we need to resolve ambiguities that arise in IL.

In this paper, we have discussed about NER based system for IL particularly for Hindi, Telugu and Bengali using Hidden Markov Model (HMM).

2. LITERATURE REVIEW

Based on similar sounding property, a phonetic matching technique is developed to perform NER in English and Hindi. Precision and Average recall obtained for English are-81.40% and 81.3%. and for Hindi are-80.2% and 54.97% [4]. Annotated Corpora of Bengali and Hindi having 122,467 tokens and 502,974 tokens are used in performing NER using Support Vector Machine. The testing is done on 35K and 60K tokens of Bengali and Hindi. Average recall, precision and f-score values for Bengali are: 88.61%, 80.12% and 84.15%, whereas for Hindi, it is: 80.23%, 74.34% and 77.17% [12]. NER is performed in Telugu in two phases. In the first phase, Telugu dictionaries, Noun Morphological Stemmer and Noun Suffixes are used to identify the Nouns. In the second phase, transliterated gazetteer lists, various Named Entity suffix features, context features and morphological features are used to identify the Named Entities. The accuracy achieved using this approach is up to 95.37% [3]. Maximum Entropy (ME) approach and Contextual information along with different orthographic word level features have been used for performing NER in Telugu. Contextual word features refer to the preceding and the following words of a given word. The Corpus of Telugu has been taken from Telugu Wikipedia, Telugu Local dailies; iinaa Du. The Performance Metrics i.e. Precision, Recall and F-Measure obtained is 75.89%, 53.35% and 61.62% [2]. It has been shown that the performance of the HMM model of a hybrid system is better than the CRF model. The F-Measure values obtained using HMM Model for languages such as Bengali, Hindi, Oriya, Telugu and Urdu are: 39.77, 46.84, 45.84, 46.58 and 44.73 [14].

3. IMPLEMENTATION AND RESULTS

We have developed a NER based system based on the Hidden Markov Model (HMM) approach Fig1. The input to this NER based system is the raw text and its output is the Named Entity tags.

This NER based system performs in three phases. The first phase is referred to as ‘Annotation phase’ that assists in producing tagged or annotated text from the raw text. The second phase is referred to as ‘Train HMM’. In this phase, it computes the three most essential parameters of HMM i.e. Start Probability, Emission Probability (B) and the Transition Probability (A) [11][12][16]. The last phase is referred to as ‘TEST HMM’. 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.

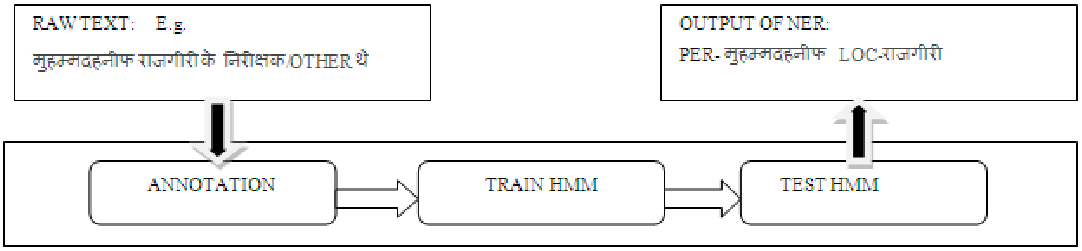
RAW TEXT: E.g.
मुहम्मदहनीफ राजगीरीके निरीक्षक/OTHER थे
OUTPUT OF NER:
PER- मुहम्मदहनीफ LOC-राजगीरी
ANNOTATION
TRAIN HMM
TEST HMM

Figure 1 NER in Indian languages 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, Bengali and Telugu. The details are mentioned in TABLE 1. For Hindi, we considered Tourism Domain Corpus which was developed at Banasthali Vidyapith. Also, we took Hindi, Bengali and Telugu Corpus from NLTK Indian Corpora. The Hindi text that we took from NLTK was related to Politics and Sports. Bengali and Telugu texts from NLTK were of general domain.

TABLE 1 Details of NER using HMM

S.NO.	LANGUAGE	SOURCE	DEVELOPED BY	WEBSITE	DOMAIN
1	HINDI		Banasthali Vidyapith		Tourism
2	HINDI	NLTK Indian Corpus	author: A Kumaran	http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml	Related to Politics and Sports News
3	BENGALI	NLTK Indian Corpus	author: A Kumaran	http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml	General domain and Related to Countries, Locations, Animals, languages etc.
4	TELUGU	NLTK Indian Corpus	author: A Kumaran	http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml	General

We have also done training on 8,623 words or 540 sentences in Hindi obtained from NLTK Indian Corpora. We considered 8 tags TABLE 2. Accuracy, Precision, Recall and F-Measure reported is 96%

TABLE 2 Tags used in NER in Hindi sentences from NLTK Indian Corpora

SNO	TAGS
1	PER (Name of Person)
2	LOC (Name of Location)
3	OTHER (Not a Named Entity)
4	CO (Name of Country)
5	MONTH
6	ORG (Name of Organization)
7	WEEK
8	PARTY (Name of Political Party)

We performed NER in Hindi. For training, we took 100 sentences or 2332 tokens from a Hindi tourism corpus, developed at Banasthali Vidyapith. We annotated it using 10 tags mentioned in TABLE 3. We obtained F-Measure of 93%.

TABLE 3.Tags used in NER in Hindi sentences from Tourism domain

SNO	TAGS
1	PER (Name of Person)
2	LOC (Name of Location)
3	OTHER (Not a Named Entity)
4	SPORT
5	TIME
6	MONTH
7	ORG (Name of Organization)
8	VEH (Name of Vehicle)
9	QTY(Name of Quantity)
10	RIVER

Also, we performed training on 9996 words or 994 Telugu sentences of NLTK Indian Corpora. F-Measure obtained is 98.6%. Table 4 shows the tags that we have used.

TABLE 4 Tags used in NER in Telugu sentences from NLTK Indian Corpora

SNO	TAGS
1	PER (Name of Person)
2	LOC (Name of Location)
3	OTHER (Not a Named Entity)
4	CO (Name of Country)
5	SUBJECT (Name of Subject)
6	LANGUAGE

Next, we performed training on 10,303 words or 899 sentences of Bengali obtained from NLTK Indian Corpora. F-Measure obtained is 98.5%. The tags used are shown in TABLE 5.

TABLE 5 Tags used in NER in Bengali sentences from NLTK Indian Corpora

SNO	TAGS
1	PER (Name of Person)
2	LOC (Name of Location)
3	OTHER (Not a Named Entity)
4	CO (Name of Country)
5	ANIMAL (Name of Animal)
6	RIGHT (Name of Fundamental Right)
7	DAIRY (Name of Dairy)
8	SONG
9	SPICES (Name of Spices)
10	MONTH
11	LANGUAGE
12	PANCHAYAT (Name of Panchayat)

So, finally the results we have obtained in terms of F-Measure are given below in TABLE 6

TABLE 6 Results obtained by NER using HMM

SNO	LANGUAGE	PRECISION	RECALL	F-MEASURE
1	Hindi(Tourism Corpus)	93%	93%	93%
	Hindi(General Domain)	96%	96%	96%
2	Bengali	98.5%	98.5%	98.5%
3	Telugu	98.6%	98.6%	98.6%

4. PERFORMANCE METRICS

Performance Metrics is measure to estimate the performance of a NER based system. Performance Metrics can be calculated in terms of 3 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 be termed 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 as follows: [2] [3] [9]
Precision (P): RA/R

Recall (R): RA/A
F-Measure: (2 * P * R) / (P + R)

5. CONCLUSION

In this paper, we have discussed about NER, some of the work that has been already been accomplished in NER using different approaches, Performance Metrics and the results that we have obtained by performing NER in Hindi, Bengali and Telugu using Hidden Markov Model. The F-Measure that we have obtained is 96%, 98% and 98.5% in Hindi, Bengali and Telugu. Also, it denotes that as the amount of training increases in Hidden Markov Model, then the accuracy and the F-Measure also increases.

ACKNOWLEDGEMENT

We would like to thank all those who helped us 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

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.