ISSN:0975-9646

A Two Stage Language Independent Named Entity Recognition for Indian Languages

S. Biswas
ITER, SOA University, Bhubaneswar
Sitanath_biswas2006@yahoo.com

M. K. Mishra
ITER, SOA University, Bhubaneswar
mkmishra_iter@yahoo.com

S. Acharya
AIET, Bhubaneswar
Sweta_acharya20@yahoo.co.in

S. Mohanty
Utkal University
sangham1@rediffmail.com

Abstract

This paper describes about the development of a two stage hybrid Named Entity Recognition (NER) system for Indian Languages particularly for Hindi, Oriya, Bengali and Telugu. We have used both statistical Maximum Entropy Model (MaxEnt) and Hidden Markov Model (HMM) in this system. We have used variety of features and contextual information for predicting the various Named Entity (NE) classes. The system uses both language dependent and language independent rules. We have also tried to identify the nested named Entities (NES) by giving some linguistic rules and the rules are purely language independent. We have also used gazetteer list in addition to the rules for Oriya, Bengali and Hindi for better accuracy. The system has been trained with Hindi (450, 150 tokens), Oriya (150, 100 tokens), Bengali (93, 023 tokens), and Telugu (50, 250 tokens). The system has been tested with 35,018 tokens of Hindi 45,100 tokens of Oriya, 28,123 tokens of Bengali and 4,320 tokens of Telugu.

1. Introduction

Named Entity Recognition (NER) is an important tool in natural language processing pertaining to information extraction (IE), Machine Translation (MT), Information Retrieval (IR), Question Answering etc. NER is the task of identifying and classifying all proper nouns in a document as person name, location name, organization name, number time etc.

In this paper we have presented a Named Entity recognition system for the Indian languages, particularly for Hindi, Oriya, Bengali and Telugu. Hindi is the most popular and National language in India and other languages are also equally important and popular in India. The two statistical model namely Maximum Entropy Model (MaxEnt) and Hidden Markov Model (HMM) has been used to develop the system. The system also makes use of the different contextual information of the words along with the variety of orthographic word level features that are helpful in predicting the various named entity classes. In our system we have considered language independent as well as the language dependent features. The different language independent features include the contextual words, prefix and suffix information of all the words in the training corpus, several digit features. The system includes linguistic features for Hindi, Bengali and Oriya. Linguistic features for Bengali and Oriya includes the set of known suffixes and some clue words which may appear with the named entities. Which helps to identify person names, location names and organization name has been used for Hindi, Bengali and Oriya. No linguistic rules have been considered for Telugu.

2. Previous Work

There are several classification methods which are successful to be applied on this task. Chieu and Ng and Bender et al.used Maximum Entropy approach as the classifier. Conditional Random Filed (CRF) was explored by McCallum and Li to NER. Mayfield et al. applied Support Vector Machine (SVM) to classify each name entity. Florian et al. even combined Maximum Entropy and Hidden Markov Model (HMM) under different conditions. Some other researchers are focused more on extracting some efficient and effective features for NER. Chieu and Ng successfully used local features, which are near the word,

and global features, which are in the whole document together. Klein et al. and Whitelaw et al.reports that character-based features are useful for recognizing some special structure for the name entity. Linguistic approach uses hand-crafted rules, which needs skilled linguists. Some recent approaches try to learn context patterns through ML which reduce amount of manual labour. Talukder et al.(2006) combined grammatical and statistical techniques to create high precision patterns specific for NE extraction.

In rule-based approaches, a set of rules or patterns is defined to identify the named entities in a text. These rules or patterns consist of distinctive word format, such as particular preposition prior to a named entity. For instance, a string of words behind titles such as 'sri', 'srimati', etc will be identified as name of a person, whereas a word after a preposition such as , 'deikeri', 'pakhare', etc is most likely to be a location. By implementing a finite set of carefully predefined pattern matching rules, the named entities within a text could be found systematically.

3. Maximum Entropy Model

(MaxEnt)

For the development of our Oriya NER system, we have used MaxEnt model which is the Java based open-nlp MaxEnt toolkit and freely available at `www.maxent.sourceforge.net.` It gives the probability values of a word belonging to each class. That is, given a sequence of words, the probability of each class is obtained for each word. To find the most probable tag corresponding to each word of a sequence, we can choose the tag having the highest class conditional probability value.

A Maximum Entropy approach models a random process by making the distribution satisfy a given set of constraints, and making as few other assumptions as possible. The constraints are specified as real-valued feature functions over the data points. The expected value of each feature function under the ME distribution must equal the empirically expected value of function as found in the training dataset. In all other respects, the target distribution should be as uniform as possible, which means it must have the highest entropy.

Let X be the set of conditions, usually very big, and Y the set of possible outcomes. We assume that there is a true joint distribution P(x,y), but we are interested only in modeling the conditional P(y|x). For this purpose we can use a training set $\{(x_k,y_k)\}k=1..N$ generated by the true distribution, and a set of features fi :X×Y→R. Typically, the features are binary and test for specific conditions. It can be shown that the unique most uniform distribution that satisfies all feature constraints has the form:

$$(*)\ \ p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_i \lambda_i f_i(x, y)\right]$$

where λi –s are the parameters chosen to maximize the likelihood of the training data, and Z(x) is a normalization constant, which ensures that for every x the sum of probabilities of all possible outcomes is 1. The most common procedure for parameter estimation is the Generalized Iterative Scaling algorithm.

3.1 Maximum Entropy Markov Models

A MaxEnt consists of |Y| conditional ME models py'(y|x) = p(y|x,y'), one for each y'. The model py'(y|x) estimates the probability of appearance of the label y immediately after the label y' in the context x. The probability of a whole label sequence y = y1 y2… ym, given the sentence $x = x_1\ x_2 \ldots\ x_m$, is the product

$$P(y \mid x) = P_0(y_1 \mid x_1) \prod_{i=1}^{m-1} p_{y_i}(y_{i+1} \mid x_{i+1})$$

The best tagging can be found using Dynamic Programming similar to Vitterbi algorithm. The model p_0 (y|x) used at the beginning of a sentence is separate.

4. Hidden Markov Model (HMM)

After the MaxEnt walkthrough, all the tagged named entities in the testing corpus are used as training data for HMM to make the final tagging. We are confident that there will be sufficient training after parsing through the corpus using MaxEnt. In our system, HMM is used mainly for global context checking, that is to check the occurrences of the same named entity in different sections of the same text document. We believe that checking the

context from the whole document is important as this will ensure the consistency of the tagged named entities and resolve some ambiguous cases. For instance, an organization's name is often abbreviated especially when it has already been mentioned somewhere in a document. By checking the global information, we are able to identify the abbreviation as an organization. Besides that, we often encounter some entities that are highly ambiguous, and their categories cannot be determined without taking the global context into consideration. The phrase 'Honda City' in sentences such as "Honda City is nice" or "Promotion for Honda City" could easily be misinterpreted as a location based on the local contextual evidence, unless we found another sentence that sounds like "I am driving Honda City".

Similar to the previously used MaxEnt, we use HMM to compute the likelihood of words occurring within a given category of named entity. Every tokenized word is now considered to be in ordered pairs. By using a Markov chain, the likelihood of the words is calculated simply based on the previous word. For classifying the named entities, our system finds the most likely tag t for a given sequence of words w that maximizes P (t|w). The occurrences of the given events are counted throughout the whole text based on the calculation below:

$$P(y|y_{-1}x_{-1}) = \frac{count(y, y_{-1}, x_{-1})}{count(y_{-1}, x_{-1})}$$

Finally, we use a classifier to correct the errors in the results derived from MaxEnt to perform the final tagging process using HMM.

5. Named Entity Recognition in Indian Languages

Named Entity Recognition in Indian Languages (IL) is difficult and challenging as capitalization is not present as clue in ILs like English. Indian languages are very highly inflective in nature and have non-independent and diverse overlapping features.

6. Training Data Preparation

The training data in all the languages were annotated with the twelve NE tages using the shakti standard standard format (SSF) (http://shiva.iiit.ac.in/SPSAL 2007/ssf.html). For example

((NP <ne=NEL>

((NP < ne = NEP>

1.1.1 ((NP <ne = NETP>

1.1.1.1 Biju

))

Pattnaik

))

1.2 Road

))

Here Biju Pattnaik road was annotated as location and assigned the tag "NEL" EVEN IF Biju, Pattnaik are named entity title person (NETP) and person name (NEP) respectively. The training data is searched for the multiword NES. Each component of a multiword NE is also checked whether the component is made up of digits only and if it is then assigned the tag "NEN". For preparing the training data, the list of gazetteers, which have been used given in Table 1.

Gazetteer List	**No. of entries**
First person name in Oriya	20,801
Last person name in Oriya	6,280
Middle name in Oriya	1,302
Person name designation in Oriya	732
Location name in Oriya	5,880
First person name in Hindi	80,963
Last person name in Hindi	5,630
Middle Name in Hindi	370
Month names in Hindi, Bengal, Oriya	14
Words that denote measurement in Oriya, Bengali and Hindi	65

Table 1. Gazetteer lists used during training data preparation.

7. Named Entity Features

Experiments were carried out to find out most suitable features for NE tagging because feature selection plays an important role in statistical framework. In addition, various gazetteer lists have been developed to use in our model particularly for Hindi, Oriya and Bengali.

Following is the details of the set of features that were applied to the NER system.

Word suffix and prefix:

To identify NES, suffix and prefix information plays an crucial role. We have taken a list of common suffixes and prefixes of person name, location names in Hindi, Bengali and Oriya. Some location suffixes are " Vihar", "Nagar", "pur". Some person names suffixes are "ku", "ra" "re", "babu", "da"). A fixed length word prefix of current and surrounding words are treated as features. We have considered the world window size three.

Context word feature : previous and next words of a particular word might be used as a feature. We have considered the word window of size five.
Digit : Digit can be many type. It can be single or multiple based on the above fact we have defined digits as some form like if a token contains four digits then it may be helpful in identifying the time (e.g. 2010 barsa or 2010) expressions. If a token contains two digits then it can be also helpful in identifying time expressions (7 am, 8 p.m). If n digit is followed by slash or hyphen then it may be helpful for identifying time expression (02/04/2010 or 02-04-2010). For identifying numerical quantities we took a token which contains a digit and period (e.g. 2503.50). If a token contains digits and percentage then it may be useful for predicting measurements (e.g. S20%).

8. Evaluation

We have used precision, recall and F-measure for the evaluation purpose for all the four languages. The measure are calculated in different ways –

Maximal matches : the largest possible named entities are matched with reference data.

Nested matches : The largest possible as well as nested named entities are matched.
Maximal lexieal item matches : the lexical items inside the largest possible named entities are matched.
Nested lexical item matches: the lexical items inside the largest possible as well as nested named entities are matched.

9. Experimental Results

The MaxEnt and HMM based NER system has been trained and tested with four different Indian languages namely, Hindi, Bengali, Oriya and Telegu data. The training and test sets statistics are presented in table 2. Result of evaluation is explained in table 3 in terms of precision, recall and F-measure.

Experimental result in table 3 shows promising performances for Hindi, Bengali and Oriya. For Telegu, the problem we faced is the lack of enough data and the identification of linguistic rules.

Language	Number of tokens in the training set	Number of tokens in the test set
Hindi	450, 150	35, 018
Bengali	93,023	28,123
Oriya	150, 100	45, 100
Telugu	50,250	4,320

Table-2: Training and Test Sets statistics

Conclusion

In this paper, we have developed a Named Entity Recognition system for Indian languages particularly for Hindi, Bengali, Oriya, Telugu using hybrid machine learning approach that used MaxEnt and HMM successively. We showed that with the preliminary data training through MaxEnt and appropriate classifier for error correction in the final recognition process through HMM, the performance of our proposed NER system can be greatly enhanced as compared to using only a single statistical model. Moreover, our system is also able to adapt to different

Measure → / Language ↓	Precision			Recall			F-measure		
	Pm	Pn	Pl	Rm	Rn	Rl	Fm	Fn	Fl
Bengali	54.55	51.11	53.89	59.98	62.28	71.03	57.13	56.14	61.28
Hindi	78.33	71.12	76.12	66.28	71.72	68.23	71.80	71.41	71.95
Oriya	62.11	69.24	76.38	69.23	72.13	73.24	65.47	70.65	74.77
Telugu	12.10	11.09	18.78	2.81	3.28	1.29	4.56	5.06	2.41
M: maximal, n:Nested, l:Lexical									

Table-3: Evalution of the four languages

domains without human intervention, and maintain desirable performance regardless of the size of the training corpus.

While our experimental results have been quite positive, we reckon that our proposed approach is still fairly immature. Much work needs to be done to make the performance of our system more robust also we need to add parts of speech (POS) information of the current word and surrounding word(s)

References

Hai Leong Chieu and Hwee Tou Ng, Named Entity Recognition with a Maximum Entropy Approach. In: Proceedings of CoNLL-2003, Edmonton, Canada, 2003, pp.160-163.

Oliver Bender, Franz Josef Och and Hermann Ney, Maximum Entropy Models for Named Entity Recognition In: Proceedings of CoNLL- 2003, Edmonton, Canada, 2003 pp.148-151.

Bikel Daniel M., Miller Scott, Schwartz Richard and Weischedel Ralph. 1997. Nymble: A High Performance Learning Name-finder. In Proceedings of the Fifth Conference on Applied Natural Language Processing, 194–201.

Borthwick Andrew. 1999. A Maximum Entropy Approach to Named Entity Recognition. Ph.D.thesis, Computer Science Department, New York University.

Cucerzan Silviu and Yarowsky David. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999, 90–99.

Kumarn. and Bhattacharyya Pushpak. 2006. Named Entity Recognition in Hindi using MEMM. In Technical Report, IIT Bombay,India..

Li Wei and McCallum Andrew. 2004. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper).In ACM Transactions on Computational Logic.

McDonald R., Crammer K. and Pereira F. 2005. Flexible text segmentation with structured multilabel classification. In Proceedings of EMNLP05.

Srihari R., Niu C. and Li W. 2000. A Hybrid Approach for Named Entity and Sub-Type Tagging. In Proceedings of the sixth conference on Applied natural language processing.

Ekbal, Asif, and S. Bandyopadhyay. 2007a. pattern Based Bootstrapping Method for Named Entity Recognition. In Proceedings of 6th International Conference on Advances in Pattern Recognition, Kolkata, India, 349-355.

Ekbal, Asif, and S. Bandyopadhyay. 2007 b. Lexical pattern Learning from Corpus Data For Named Entity Recognition. In Proceedings of the 5th International Conference on Natural Language Processing, Hyderabad, India, 12-128.

Ekbal, Asif, Naskar, Sudip and S. Bandyopadhyay. 2007c. Named Entity Recognition and Transliteration in Bengali. Named Entities : Recognition, Classification and use, Special issue of Lingvisticae Investigations Journal, 30:1 (2007), 95-114.