A Hybrid Oriya Named Entity Recognition system: Harnessing the Power of Rule

Sitanath Biswas sitanath_biswas2006@yahoo.com
ITER, SOA University, Bhubaneswar

S. P. Mishra smitaprava@yahoo.com
ITER, SOA University, Bhubaneswar

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

S Mohanty sangham1@rediffmail.com
Utkal University, Bhubaneswar

Abstract

This paper describes a hybrid system that applies maximum entropy (MaxEnt) model with Hidden Markov model (HMM) and some linguistic rules to recognize name entities in Oriya language. The main advantage of our system is, we are using both HMM and MaxEnt model successively with some manually developed linguistic rules. First we are using MaxEnt to identify name entities in Oria corpus, then tagging them temporary as reference. The tagged corpus of MaxEnt now regarded as a training process for HMM. Now we use HMM for final tagging. Our approach can achieve higher precision and recall, when providing enough training data and appropriate error correction mechanism.

1 INTRODUCTION

Name Entity Recognition (NER) is an important activity in the Natural Language Processing pertaining to Information Extraction (IE), Machine Translation (MT), Information Retrieval (IR) 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.
This paper presents an Hybrid NER system for Oriya Language and the goal of the system is to recognize different types of NEs- person, designation, title-Person, organization, abbreviation, location, time, number, and measure.

To develop a MaxEnt and HMM based Oriya NER system, we have identified suitable features like Orthography features, suffix and prefix information, morphology information, part-of-speech information as well as information about the surrounding words and their tags in Oriya language. We have used gazetteers for identification of designation, title, of the person names etc. We have also used person and location name gazetteers in our system for better identification of NEs. We have discovered that linguistic rule also plays a crucial role in identifying NEs so we have used a no. of linguistic rules of Oriya language in our system like the rule to recognize time, number etc.
According to the specifications defined by MUC, the NER tasks generally work on seven types of named entities as listed below with their respective markup:

- PERSON (ENAMEX)
- ORGANISATION (ENAMEX)
- LOCATION (ENAMEX)
- DATE (TIMEX)
- TIME (TIMEX)
- MONEY (NUMEX)
- PERCENT (NUMEX)

The paper is organized as follows. A brief survey of different techniques used for the NER task in different languages and domains are presented in Section 2. A discussion on the training data is given in Section 3. The MaxEnt and HMM based NER system is described in Section 4 and 5. Various features used in NER are then discussed. Next we present the experimental results in Section 8. Finally Section 9 concludes the paper.

2 PREVIOUS WORKS

There are several classification methods which are successful to be applied on this task. Chieu and Ng [1] and Bender et al.[2] used Maximum Entropy approach as the classifier. Conditional Random Filed (CRF) was explored by McCallum and Li [3] to NER. Mayfield et al.[4] applied Support Vector Machine (SVM) to classify each name entity. Florian et al. [5] 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 [1] successfully used local features, which are near the word, and global features, which are in the whole document together. Klein et al. [6] and Whitelaw et al.[7] 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 TRAINING DATA

The annotated data used in our system is in the IOB formatted text in which a *B - XXX* tag indicates the first word of an entity type *XXX* and *I -XXX* is used for subsequent words of an entity. The tag O indicates the word is outside of a NE. The training data for Oriya contains more than 56K.

4 MAXIMUM ENTROPY MODEL

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 empirical 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 $f_i : X\times Y\rightarrow \boldsymbol{R}$. 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:

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

where λ_i –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.

4.1 MAXIMUM ENTROPY MARKOV MODELS

A MaxEnt consists of |*Y*| conditional ME models $p_{y'}(y|x) = p(y|x,y')$, 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** = y_1 y_2… y_m, given the sentence **x** = x_1 x_2… x_m, is the product

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

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.

4.2 FEATURES

Features play an important role when building any MaxEnt model based system. The different features are Orthographic features (like capitalization, decimal, digits), affixes, left and right context (like previous and next words), NE specific trigger words, gazetteer features, POS and morphological features etc. In English and some other languages, capitalization features play an important role but In Indian languages there is no capitalization of letters for distinguishing proper nouns from other category of words and no such database is available from which one can search the proper names like other nouns. The Indian languages are also morphologically rich in nature. The word reordering inside a sentence is also a common feature of these languages. In the following we have discussed about the features we have identified and used to develop the Indian language NER systems.

SUFFIX AND PREFIX:

To identify NEs in Oriya language, suffix and prefix information plays an important role. We have taken a list of common suffixes of person and location names in Oriya. Some location suffixes are *"vihar", "nagar", "pur".* Some person name suffixes are *"ku", "ra","re".* A fixed length word prefix of current and surrounding words are treated as features.

PARTS-OF-SPEECH (POS) INFORMATION:

Since NEs are noun phrases; the noun tag is very relevant, no need to give detail POS tags. We have taken the POS of the current word and the surrounding words as features.

ROOT WORDS:

We have used morphological analyzer to check the root words. As we know the Oriya language is morphologically very rich and words are inflected in different forms on its number, tense, case etc.

Apart from taking POS information, root words, Suffix and prefix as feature, we have also taken first word, digits, numerical word as feature.

5 HIDDEN MARKOV MODEL

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.

6 RULE FORMATIONS

We have developed 32 rules to identify numbers, measures, time etc. The rules are manually developed and required a huge knowledge about Oriya language. The result we are getting, much more accurate than MaxEnt and HMM.

7 GAZETTEER LISTS

We have prepared a list of specialized names from different web resources and transliterated those into Oriya Language as the resources were in English. Using transliteration we have constructed several lists. Which are month name and days of the week, list of common locations, location names list, first names list, middle names list, and surnames list.

8 EVALUATIONS

The accuracies of the system are measured in terms of the F-measure, which is the weighted harmonic mean of precision and recall. The test data for Oriya languages is provided. The size of the Oriya test data is 35,112.

	Correct	**In correct**
Selected	Cs	Is
Not selected	Cn	In

Table1. Components of F-measure

Precision: P= $\frac{Cs}{Cs + Is}$

Recall: R= $\frac{Cs}{Cs + Cn}$,

F1= $\frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$

Domain	**Category**		
	PER	**LOC**	**ORG**
Science	86.10	79.24	87.34
Arts	88.23	83.33	77.98
World affairs	82.12	86.65	85.98
Commerce	79.88	77.76	89.78

Table 2: F-measure score in percentage

9 CONCLUSIONS

In this paper, we presented a 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 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.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

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

[9] 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.