Named Entity Recognition for Odia Text Using Machine Learning Algorithm

[1]Bishwa Ranjan Das, [2]Hima Bindu Maringanti, [3]Niladri Sekhar Dash
[1,2]Department of Computer Application.
Maharaja Sriram Chandra Bhanjadeo University, Baripada, India
[3]Linguistic Research Unit, Indian Statistical Institute, Kolkata, India
[1]biswadas.bulu@gmail.com, [2]profhbnou2012@gmail.com, [3]ns_dash@yahoo.com

ABSTRACT

This paper presents a novel approach to recognize named entities for Odia newspaper text. The development of a NER system for Odia newspaper text using Support Vector Machine is a challenging task in the field of intelligent computing. Named Entity Recognition aims at classifying each word in a piece of document into predefined target named entity classes in a linear as well as non-linear fashion. Starting with named entity annotated corpora and a set of features it requires to develop a base-line NER System. Some language specific rules are added to the system to recognize some specific NE classes. Moreover, some gazetteers and context patterns are added to the system to increase its performance level as it is observed that identification of rules and context patterns requires language-based knowledge to make the system work better. A lexical database is used to prepare the rules as well as to identify the context patterns for Odia text. A very large corpus including one lakhs sentences both training and test set is taken for experimental test and results show that our approach achieves much higher accuracy than previous approaches.

Keywords: Support Vector Machine, Name Entity Recognition, Part of Speech Tagging, Root word

Introduction

Named Entity Recognition (NER) is a technique to identify and classify named entities for particular domain of a piece of text. It is an important task as it is directly related to applications like Information Extraction, Question Answering, and Machine Translation, Data Mining, and other NLP (Natural Language Processing) applications. This paper proposes a novel NER system for Odia, one of the Indian national languages. It performs NER act on three types of named entity - person names, location names, and organization names. These named entities are addressed because identification of these is the most challenging task in the whole scheme of NER. For our task, suitable set features are first identified for the named entities in Odia. The feature list includes orthography features, suffix and prefix information, morphological information, part-of-speech information as well as information about the neighboring words and their POS tags, which are combined together to develop the Support Vector Machine (SVM) based NER System of the language. Some rules are defined for classification of person, location, organization names based on certain criteria, which are made available to the system through gazetteers-based identification for person, location, and organization names.

There are several named entity classification methods which may be successfully applied on this task. Taku Kudo, et.al.[1] have used the Support Vector Machine in chunking which may also help in our proposed work. Biswas, et.al.[2] have used the Max Entropy model for hybrid NER

for classification. Their approach can achieve higher precision and recall, if it is provided with enough training data and appropriate error correction mechanism. Ekbal, et. al.[3] have used the SVM for classification of Bengali named entities with 91.8% accuracy. Saha, et. al.[4] have described the development of Hindi NER system by using ME approach with 81.51% accuracy. Their system is tested with a lexical database of 25k words having 4 classes of named entities. Goyal[5] has also developed a system for NER for South Asian Language. Saha et.al (2008) [4] have identified suitable features for Hindi NER task that are used to develop an ME based Hindi NER system. Two-phase transliteration methodology has been used to make the English lists useful in the Hindi NER task. This system gives the accuracy with 81.2%. Various approaches that are used in NER system include Rule Based, Handcrafted Approach, Machine Learning, Statistical Approach, and Hybrid Model[6]. In Rule-Based approaches, a set of rules or patterns is defined to identify the named entities in a text. For instance, while pre-tags like 'sri', 'sriman', 'srimati' etc. are used to identify person names, forms like 'nagar', 'sahara', 'vihar' etc. are used to identify place names, and the forms like 'vidyalaya', 'karjyalaya' etc. are used to identify organization names. Hai and Hwee[7] have used Maximum Entropy Model to find NE (Global information) with just one classifier. In another work ("A survey of named entity recognition and classification"), they have presented a survey of fifteen years of research (1991 to 2006) in NERC field. The introduction of this paper describes in some details the early works on NER system development in other languages; Section 2 describes the composition and content of the Odia newspaper text corpus; Section 3 describes the Support Vector Machine which is used for classification of named entities; Section 4 describes the training data, how it is specially used, and data mapping with the test dataset; Section 5 presents the evaluation results to show how our proposed system works; and Section 6 describes the conclusion part of the paper.

THE ODIA TEXT CORPUS

An Odia newspaper text corpus is recently developed to describe in details the form and texture of the Odia language used in the present data Odia newspapers. Following some well-defined strategies and methods this Odia corpus is designed and developed in a digital with texts obtained from Odia newspapers. The corpus is developed with sample news reports produced and published by some major Odia newspapers published from Bhubaneswar and neighboring places [15]. We have followed several issues relating to text corpus design, development and management, such as, size of the corpus with regard to number of sentences and words, coverage of domains and sub-domains of news texts, text representation, question of nativity, determination of target users, selection of time-span, selection of texts, amount of sample for each text types, method of data sampling, manner of data input, corpus sanitation, corpus file management, problem of copy-right, etc. Since this corpus is very much rich with data relating to named entities of various types, we have been using it to perform linear and nonlinear classification of named entities in which we prepared our own digital corpus from various Odia newspapers. In essence, we are using this corpus to identify and classify Odia person names, place names and organization names along with some miscellaneous named entities.

SUPPORT VECTOR MACHINE

The Support Vector Machines is a binary learning machine with some highly elegant properties that are used for classification and regression. It is a well-known system for good generalization performance and it is used for pattern analysis. In NLP, it is applied to categorize the text, as it gives high accuracy with a large number of features set. We have used this machine to defining a very simple case – a two class problem where the classes are nonlinearly separable. Let the data

set D be given as (X1,y1), (X2,y2)…..(XD,yD), where Xi is the set of training tuples with associated class labels yi,. Each yi can take one of two values, either +1 or -1(i.e., yi ∈ {+1,-1}.

Fig. 1: Classification of textual data

A separating hyperplane equation can be written as wx + b = 0, where x is an input vector, w is the adjustable weight, and b is the bias. Training tuples are 2-D, e.g. x={x1, x2}, where x1, x2 are the values of attributes A1 and A2 respectively for x. It finds an optimal hyperplane which separates the training data as well as the test data into two classes. It find separating hyperplane which maximizes its margin. Two parallel lines and margin M can be expressed as wx + b = +1, M=2/||w||. To maximizes this margins r = 1/ ||w|| and Minimize ||w|| = ||w||2/2, Subject to di (w. xi + b) ≥ 1, where i = 1, 2, 3,…, l) . Any training tuples that falls on either side of the margins are called support vector. It has strength to carry out the nonlinear classification. The optimization problem can be written usual form, where all feature vectors appear in their dot products. By simple substituting every dot product of xi and xj in dual form with a certain Kernel function K(xi, xj). SVM can handle nonlinear hypotheses. Among these many kinds of Kernel function available. We shall focus on the polynomial kernel function with degree d such as K (xi, xj) = (xi * xj + 1)d. Here d degree polynomial kernel function helps us to find the optimal separating hyperplane from all combination of features up to d. The hypothesis space under consideration is the set of functions. The linear separable case is almost done. The nonlinear SVM classifier gives a decision making function f(x).

$$f(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b, \quad g(x) = \text{sign}\big(f(x)\big) \qquad \ldots\ldots\ldots (1)$$

If g(x) is +1, x is classified as class C1 and -1 x is classified as class C2. zi are called support vectors and representative of training examples, m is the number of support vectors is a kernel that implicitly maps vectors into a higher dimensional space and can be evaluated efficiently. The polynomial kernel K(x, zi) = (x.zi)d .

TRAINING DATA

We used our own training data set that was developed by ourselves in Odia. It gives the 100% correct result for our system.

$T = \{x_k, d_k\}_{k=1}^{Q}$ Where xk € IRn, dk € {-1, +1} (2)

A. Features

It is mentioned the following set of features that have been applied to the NER task.

i. After POS tagging, the nominal word or surrounding word is set to be +1 otherwise it is set to -1. This binary value used to all POS feature.
ii. Person prefix word, if the prefix belongs to 'sriman', 'srimati' etc. then set to +1.
iii. If middle names like 'kumar', 'ranjan', 'prasad' etc. appear inside the person name, then it is set to be +1.
iv. If surnames like 'Das', 'Mishra', 'Sahoo' etc. appear set to be +1.
v. Location name with suffix 'nagar', 'sahara', 'podaa', 'vihar' etc. is set to be +1.
vi. Organization name with suffix 'mahabidyalaya', 'karjyalaya', 'bidyalaya' etc. is set to be +1, otherwise set to be -1.

All positive words used in the training set are considered as +1 and rest of the words are considered as -1.

It is identified that various features may be considered to find out NE in Odia language as mentioned below. Following the features many place names, person names, and organization names are identified. Also some rules are mentioned in this paper that is used for such purpose, as summarized below.

(a) A Odia word which is associated with its prefix or suffix word and its surrounding words i.e., *desha* "country", *rajya* "state", *anchala* "area", *jilla* "district", *sadar mahakumaa* "dist. head quarter", *grama* "village", *panchayata* "panchayata", *pradesh* "state", *sahara* "town", are treated as place names. Some other words which belong to *nagara, vihara, pura, podaa* also used to identify place name.
(b) An Odia word which is associated with *sriman, srimati, kumara, kumari, ranjan*, etc. are used to identify person names. Some of the bivokti or markers are also used in Odia to identify person names, e.g., *-ku, -re, -ro*.
(c) An Odia word which is associated with forms like *bidyalaya* "school", *mahabidyalaya* "college", *vishwabidyalaya* "university", *karjyalaya* "office", is used to identify organization name.

The following flowchart to find Named Entity (Fig. 2).

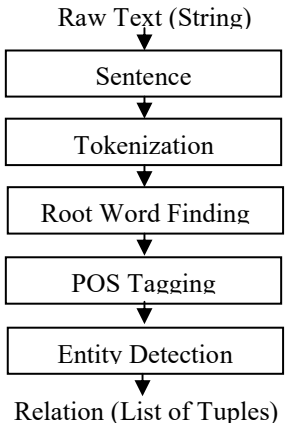

Fig. 2: Flowchart of finding NER

B. Suffix & Prefix

Some suffix and prefix alphabets are used to identify NE, which are mentioned in the features. Firstly a fixed length word suffix of the current and surrounding words are used as features.

C. Part of Speech Tagging

POS tagging is used to find out noun and verb as POS information of the current word and the surrounding words are useful features for NER. For this purpose an Odia POS tagger using ANN is used here. The tagset of the tagger contains 28 tags. The POS values of the current and surrounding tokens as features is used here.

D. Root Word

Morphological analyzer is used to find the root words by stripping suffix-prefix from a word.

E. Algorithm used

The proposed algorithm is used for finding the NE in the Odia corpus data. First, the entire Odia text corpus is entered by user in our proposed system, and then the process of NER is divided into seven steps which are described in the following algorithm.

Step 1 : Enter a text.
Step 2 : Convert entire text into token by tokenization.
Step 3 : Find root word using morphological analysis.
Step 4 : Compare each word with our valid features.
Step 5 : Extract the features from each and every word.
Step 6 : Compare each word with the training data set.
Step 7 : Find the exact Name Entity.

RESULT ANALYSIS

Odia news corpus is used to identify the test set for NER experiment. Out of one lakh word forms, a set of one thousand word forms has been manually annotated with the 10 tags initially. In our system we have used several important features to find NE and these are already described in the earlier sections. The general result obtained from our experiment is presented below (Fig. 3).

For classification of NE, thereby SVM technique is used. A baseline model is defined where the NE tag probabilities depend only the current word.

$$P(t_1, t_2, t_3 \ldots\ldots t_n | w_1, w_2, w_3 \ldots\ldots w_n) = \prod_{i=1..n} P(t_i, w_i) \qquad \ldots\ldots\ldots (3)$$

The test data is assigned to a particular NE tags POS tags that occur in the training data after some empirical analysis. The combination of words from a set ‘F’ gives the best features for Odia NER. The given set ‘F’ mentioned below.

F= {wi-4, wi-3, wi-2, wi-1, wi, wi+1, wi+2, wi+3, |prefix|<= 3, |suffix|<=3, NE information, POS information of current word, digit features}

Some experimental notations are used in this work as follows: pw (previous word), cw (current word), nw (next word), pp (POS tag of previous word), cp (POS tag of the current word), np

(POS tag of the next word). The cardinality of the prefix, suffix length is measured up to 3 characters.

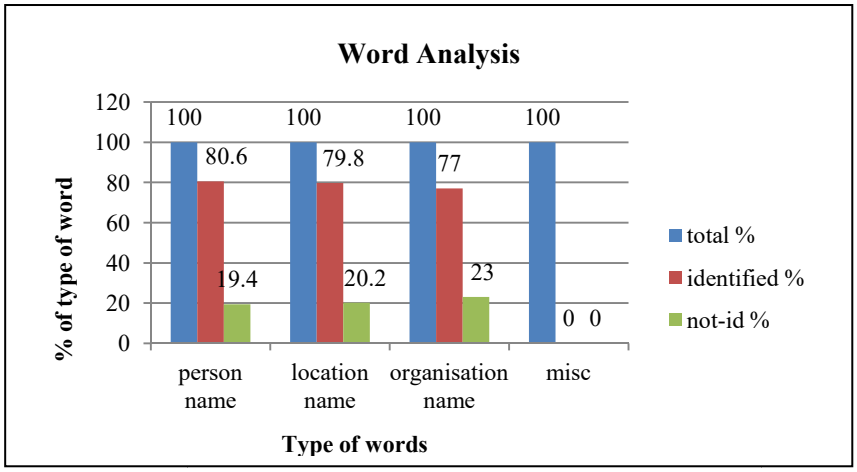

Fig. 3: Analysis of the proposed system

The Precision, Recall, and F_Score formula are used for measuring the level of accuracy of results. Mathematical equations, which get from SVM, are giving proper classification. Construction of SVM, taking training set in the equation (2) Minimize,

Φ(w) = ½‖w‖2, subject to the constraints di (wTxi + b) – 1 ≥ 0, i= 0, 1, 2…N. The objective was to maximize the margin 1/‖W‖. Since the square root is monotonic function, one can switch to ‖w‖2 instead of ‖w‖, and in order to minimize ½ ‖w‖2. To solve this optimization problem, the technique of language multiplier is used to turn here. It is used because it is easy to handle. Also to find the accuracy, we use the mathematical formula of precision, recall, F_score. POS information helps to fine the accuracy. Most of the words are tagged with appropriate tagset. From the tagged word, named entities can find easily.

Precision = |ANE ∩ ONE | / |ONE|

Recall = |ANE ∩ ONE| / |ANE|

F_Score = 2*(Recall * Precision) / (Recall + Precision)

Here ANE - Actual named entity, ONE - Obtained named entity. Precision means how many correct entities from whatever has been obtained are. Recall means out of the correct once how many have been obtained named entities. Here accuracy is calculated through F_Score in percentage. With the help of harmonic mean (HM) more accurate result also calculated.

Let us consider some instances to know how it works. For instance, let us consider a sentence in Odia: - Sriman Hariprasad jone volo gayaka,(“Sriman Hariprasad is a good singer”). Here the term “Hariprasad” is Person Name Entity, because it contains the middle name ‘prasad’. Similarly, consider this sentence Hariprasadnko ghara Bhubaneswar “Hariprasad’s home is at Bhubaneswar”. Here the term Bhubaneswar is a Location Name Entity. Similarly, in the sentence Ravenshaw mahabidyalayare se patho podhithile “He was studying at Ravenshaw College”. The term “Raveshaw” is an Organization Name Entity.

CONCLUSION

Our proposed system tries to identify NE nearly accurately with a success rate of 86% without any error. Although this system worked fine on the Odia newspaper text, we are not sure if this will work equally well in other types of Odia text. Since Odia is a resource-poor as well as less-researched language, it is obvious that we need more exhaustive research in this direction before we can claim appreciable success in recognition and identification of named entities used in Odia written texts. The performance of this system has been compared with the existing one Odia NER[2] system and one Bengali NER[3] system. There are many linguistic and stylistic issues (e.g., agglutinative nature and different writing style, etc) that also need careful attention for developing NER system for the Odia language. Definitely, the availability of an Odia text corpus of only five lakh words collected from Odia newspapers cannot be the benchmark trial database for systems like this, even if SVM system works fine on our database. With this limited success we propose to move further as application relevance of NER is approved in many domains of NLP: parsing, word sense disambiguation, information retrieval, question answering, machine learning – to mention a few.

REFERENCES

[1] Taku kudo, Yuji Matsumoto, *Chunking with Support Vector Machine*, Proceedings of NAACL-2001, pp 192-199.

[2] Sitanath Biswas, S.P. Mishra, S. Acharya, and S. Mohanty, *A Hybrid Oriya Named Entity Recognition system: Harnessing the Power of Rule*, International Journal of Artificial Intelligence and Expert Systems,2010, Volume 1, Issue 1, pp 639- 643.

[3] Asif Ekbal, and Sivaji Bandyopadhyay, *Bengali Named Entity Recognition using Support Vector Machine*, Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp 51–58.

[4] S.K. Saha, S. Sarkar, and P. Mitra, *A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition*, Proceedings of the 3rd International Joint Conference on NLP, Hyderabad, India, January 2008, pp. 343–349.

[5] A. Goyal, *Named Entity Recognition for South Asian Languages*, Proceedings of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages, Hyderabad, India, Jan 2008, pp. 89–96.

[6] B. Sasidhar, P.M. Yohan, A. Vinaya Babu, and A. Govardhan, *A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu*, International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011, ISSN: 1694-0814 www.IJCSI.org

[7] Hai, Leong Chieu, Hwee Tou Ng, *Named Entity Recognition: A Maximum Entropy Approach Using Global Information*, 19th International Conference on Computational Linguistics, COLING 2002, August 24 - September 1, 2002.

[8] Padmaja Sharma, Utpal Sharma, and Jugal Kalita, *Named Entity Recognition: A Survey for the Indian Languages*, Language in India, www.languageinindia.com Volume 11, No. 5, May 2011 Special Volume: Problems of Parsing in Indian Languages.

[9] Asif Ekbal and Sivaji Bandyopadhyay, *Named Entity Recognition using Support Vector Machine: A Language Independent Approach*, International Journal of Electrical and Electronics Engineering, Volume 4, No. 2, 2010, pp. 155-170.

[10] Saha, S.K., P.S. Ghosh, S. Sarkar, and P. Mitra, *Named Entity Recognition in Hindi using Maximum Entropy and Transliteration*, Research journal on Computer Science and Computer Engineering with Applications, pp. 33–41, 2008.

[11] Akshar Bharati, Rajeev Sangal and Veenit Chaitnya, *Natural Language Processing – A Paninian Perspective*, 1995, New Delhi, Prentice Hall-India.

[12] Pradipta Ranjan Ray, V. Harish, Sudeshna Sarkar, and Anupam Basu, *Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi*, Proceedings of the International Conference on Natural Language Processing. ICON-2003, pp.118-125.

[13] Kumar, Satish. Neural Network Book: *A Classroom Approach*, 10th edition, 2010, TMH publication, New Delhi.

[14] Mahapatra, Dhaneswar, Adhunika *Odia Byakarana* (Modern Odia Grammar), 5th Edition, 2010, Cuttack, Kitab Mahal.

[15] Das, Bishwa Ranjan, Srikanta Patnaik, Niladri Sekhar Dash, *Development of Odia Language Corpus from Modern News Paper Texts: Some Problems and Issues*, Proceedings of the International Conference On Intelligent Computing, Communication & Devices (ICCD 2014), 18-19 Apr 2014, SOA University, Bhubaneswar, India, Springer Book Series on AISC, Pp. 88-94.

[16] Dash, Niladri Sekhar, *Indian scenario in language corpus generation*, in, Dash, Niladri Sekhar Dash, Probal Dasgupta, and Pabitra Sarkar (eds.) Rainbow of Linguistics: Vol. I., 2007, pp. 129-162, Kolkata: T. Media Publication.