Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 48 (2015) 507 – 512

Procedia Computer Science

International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2015)

Conference Organized by Interscience Institute of Management and Technology,

Bhubaneswar, Odisha, India

Part of speech tagging in odia using support vector machine

Bishwa Ranjan Das[a], Smrutirekha Sahoo[b], Chandra Sekhar Panda[c], Srikanta Patnaik[d]

[a,d]*Department of Computer Science & Information Technology.*
Institute of Technical Education and Research. SOA University, Bhubaneswar, India
[b]*Department of Computer Science & Application, North Odisha University, Baripada, India.*
[c]*Department of Computer Science & Application, Sambalpur University, Burla, India*

Abstract

Part of Speech (POS) Tagging is a challenging task to identify the meaning of each word in a sentence. This paper shows the task of identifying each word in an odia sentence using the technique of Support Vector Machine. The POS Tagger is developed using a very small tagset of five tags. Various features sets are taken for different contextual information is helpful in predicting the POS classes. An Odia corpus of 10,000 words has taken and tested it very carefully. The previous POS Tagger was done using Artificial Neural Network (ANN) had given the accuracy of 81%. But this SVM based POS Tagger for Odia gives the result with an accuracy of 82%. It is very helpful to use in many field of natural language process. The result of this system compares with POS tagger using ANN which was previously done.

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)

Keywords: Part of Speech: Support Vector Machine; Odia Corpus

* Bishwa Ranjan Das. Tel.:+91-8895007227;
*E-mail address:*biswadas.bulu@gmail.com

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)
doi:10.1016/j.procs.2015.04.127

1. Introduction

POS Tagging is the process of assigning a part of speech, like noun, verb, pronoun, adverb, adverb or other lexical class marker to each word in a sentence. The solving of ambiguity in POS tagging system is challenging task for all Natural Language Processing (NLP) researchers. The input to a tagging algorithm is a string of words of a natural language sentence and a specific tagset the output is a single POS Tag for each word. There are different machine learning approaches to the problem of assigning each word of a text with a parts of speech tag, which is known as POS tagging. In this paper the performance of a POS Tagger for Odia language is shown using SVM. Support Vector Machine is basically used for classification and recognizes the pattern. [1]SVMs have high generalization performance independent of dimension of feature vectors.

Asif Ekbal[1] shows labeling each words in the corpus using SVM with accuracy 16.84%. Cutting[2] described details about POS Tagging using Hidden Markov Model. Helmut Schmid[3], a new part of speech tagging method based on neural networks (net tagger) is presented and its performance is compared to that of a HMM-Tagger (Cutting et al 1992) and a trigram based tagger (Kempe, 1993). A part-of-speech tagger based on a multilayer perceptron network is presented. It similar to the network of Nakamura et al (1990) in so far as the same training procedure (Back propagation) is used; but it differs in the structure of network and also in its purpose (Disambiguation Vs Prediction). The performance of the tagger is measured and compared to that of two other taggers (Cutting et al. 1992; Kempe 1993).

2. POS Tagging in Odia

2.1 Various terminology uses in Odia

Various terminology uses in Odia language which are used for Odia Pos Tagging like Noun, Adjective, Verb, Pronoun, Adverb, Preposition etc. in Odia.

Noun -> Bisheshya, Adjective -> Bisheshana, Verb -> Kriya, Pronoun -> word use instead of Noun, Adverb -> Kriya bisheshana etc.

2.2 Morphological Analysis

To find the root or base word in Odia many suffixes are there, these suffixes are used in verb as well as noun also. These noun suffixes are come from inflection list (Bivokti) and some suffix list are use in verb, from these suffix we find out no of nouns and verbs. Here suffix list are mentioned in which is they are used in noun and verb.

3. Support Vector Machine

Support Vector Machines is machine learning approach, basically used for classification and regression. SVMs are well known for their good generalization performance and also used for pattern recognition. The role of SVM in NLP is applied to text categorization, and gives the high accuracy with a large number of texts taken as features. I am defining very simple case, a two class problem where the classes are linearly separable. Let the data set D be given as (X_i,y_i), (X_2,y_2)…..(X_D,y_D), where X_i is the set of training tuples with associated class labels y_i,. Each y_i can take one of two values, either +1 or -1(i.e., $y_i \in \{+1,-1\}$. I see this is a 2-D data are linearly separable because a straight line can be drawn to separate all of the tuples of class +1 from all of the tuples of class -1. There are an infinite number of separating lines that could be drawn. It is to find the "best", one, that is, will have the minimum classification error on previously unseen tuples. It uses the term "hyperplane" to refer to the decision boundary that is seeking, regardless of the input attributes.
An SVM approach this problem by searching for the Maximum Marginal Hyperplane. SVM searches for the hyperplane with the largest margin, that is, the Maximum Mariginal Hyperplane(MMH). The associated margin gives the largest separation between classes. A separating hyperplane can be written as $W.X + b = 0$, Where W is a weight vector, $W= \{w_1, w_2,\ldots,w_n\}$; n is the number of attributes, and b is a bias, Let's consider two input attributes,