Tribhuvan University

Institute of Science and Technology

Support Vector Machines Based Part of Speech Tagging for Nepali Text

Dissertation

Submitted to

Central Department of Computer Science and Information Technology

Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements

for the Master's Degree in Computer Science and Information Technology

by

Tej Bahadur Shahi

March 18, 2012

Tribhuvan University

Institute of Science and Technology

Support Vector Machines Based Part of Speech Tagging for Nepali Text

Dissertation

Submitted to

Central Department of Computer Science and Information Technology

Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements

for the Master's Degree in Computer Science and Information Technology

by

Tej Bahadur Shahi

March 18, 2012

Supervisor

Assoc. Prof. Dr. Tanka Nath Dhamala

Co-supervisor

Mr. Bikash Balami

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Student’s Declaration

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

… … … … … … …
Tej Bahadur Shahi
March 18, 2012

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Supervisor's Recommendation

We hereby recommend that this dissertation prepared under our supervision by **Mr. Tej Bahadur Shahi** entitled **"Support Vector Machine Based Part of speech Tagging for Napali Text**" be accepted as partial fulfillment of the requirements for the degree of M. Sc. in Computer Science and Information Technology. In our best knowledge this is an original work in computer science.

……………………………..

Assoc. Prof. Dr. Tanka Nath Dhamala

Head of Department (HOD)

Central Department of Computer Science and
Information Technology(CDCSIT)

Tribhuvan University

Kritipur, Kathmandu, Nepal

(Supervisor)

……………………………..

Mr. Bikash Balami
Lecturer

Central Department of Computer Science and
Information Technology(CDCSIT)

Tribhuvan University

Kritipur, Kathmandu, Nepal

(Co-Supervisor)

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that we have read this dissertation and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirement of Masters Degree in Computer Science and Information Technology.

Evaluation Committee

……………………………..

Assoc. Prof. Dr. Tanka Nath Dhamala
Head of Department (HOD)
Central Department of Computer Science and
Information Technology(CDCSIT)
Tribhuvan University
Kritipur,

(Supervisor)

……………………………..

Assoc. Prof. Dr. Tanka Nath Dhamala
Head of Department (HOD)
Central Department of Computer Science and
Information Technology(CDCSIT)
Tribhuvan University
Kritipur

(External Examiner)

(Internal Examiner)

ACKNOWLEDGEMENT

It is a great pleasure for me to acknowledge the contributions of a large number of individuals to this work. I deeply extend my heartily acknowledgement to my respected teacher and dissertation advisor Assoc. Prof. Dr. Tanka Nath Dhamala, Head of Department, Central Department of Computer Science and Information Technology(CDCSIT), for giving me an opportunity to work under his supervision and for providing me guidance and support throughout this work. With this regard, I wish to extend my sincere appreciation to respected Co-supervisor Mr. Bikash Balami for his valuable suggestions.

I owe special debt of gratitude and deeply grateful to Prof. Dr. Yogendra Prasad Yadav (Head, Central Department of Linguistic), Mr. Balaram Prasain (Lecturer, Central Department of Linguistic), and Mr. Bal Krishna Bal (Madan Puraskar Pustakalaya) for their kind support during the study. They provided me ample time for discussion on issue related to the study and valuable suggestion in linguistic portion of my dissertation. I am very much grateful and thankful to the Madan Puraskar Pustakalaya for the linguistic support and data resources.

I would like to express my gratitude to the respected teachers Prof. Dr. Shashidhar Ram Joshi, Prof. Dr. Subarna Shakya, Prof. Sudarshan Karanjeet, Mr. Min Bahadur Khati, Mr. Nawaraj Paudel, Mr. Jagdish Bhatt, Mr. Arjun Singh Saud and others staffs of CDSCIT for granting me broad knowledge and inspirations within the time period of two years.

I can not remain without admiring the efforts put by my friends and others for their exceptional participation on this work. Last but not list, I would like to thank my family members for their constant support and encouragement.

ABSTRACT

Optimal part-of-speech tagging have great importance in various field of natural language processing such as machine translation, information extraction, word sense disambiguation, speech recognition and others. Due to the nature of the Nepali language, tagset used and size of the corpus (training data), getting accurate part-of-speech tagger is of challenging issue. This study is oriented to build an analytical machine learning model based on which it can be possible to determine the attainable accuracy. To complete this task, the support vector machine based part-of-speech tagger has been developed and tested for various instances of input to verify the accuracy level. The SVM tagger construct the feature vectors for each word in input and classify the word into one of two classes (One Vs Rest).

The performance analysis includes different components such as known words, unknown words and size of the training data. The present study of support vector machine based part of speech tagger is limited to use certain set of features and it use a small dictionary which affects its performance.

The learning performance of tagger is observed and found that it can learn well from the small set of training data and increases the rate of learning on the increment of training size.

TABLE OF CONTENTS

3. LITERATURE REVIEW

4. IMPLEMENATATION

5. TESTING AND ANALYSIS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

POS	Part-Of-Speech
HMM	Hidden Markov Model
CLAWS	Constituent Likelihood Automatic Word-tagging System
SOV	Subject Object Verb
SVM	Support Vector Machine