Chapter 54
Development of Part of Speech Tagger for Assamese Using HMM

Surjya Kanta Daimary
Punjabi University, India

Madhumita Barbora
Tezpur University, India

Vishal Goyal
Punjabi University, India

Umrinderpal Singh
Punjabi University, India

ABSTRACT

This article presents the work on the Part-of-Speech Tagger for Assamese based on Hidden Markov Model (HMM). Over the years, a lot of language processing tasks have been done for Western and South-Asian languages. However, very little work is done for Assamese language. So, with this point of view, the POS Tagger for Assamese using Stochastic Approach is being developed. Assamese is a free word-order, highly agglutinate and morphological rich language, thus developing POS Tagger with good accuracy will help in development of other NLP task for Assamese. For this work, an annotated corpus of 271,890 words with a BIS tagset consisting of 38 tag labels is used. The model is trained on 256,690 words and the remaining words are used in testing. The system obtained an accuracy of 89.21% and it is being compared with other existing stochastic models.

1. INTRODUCTION

Part-of-Speech (POS) tagging is the process where every word in a natural language sentence is marked with its corresponding part of speech category like noun, verb, adjective, adverb, etc. based on both its definition and context. Besides words, punctuation characters and symbols are also labeled accordingly. It is a very important process because it resolves the ambiguity of words in a sentence by assigning accurate POS label to a word depending on the context. As Assamese is morphologically rich and agglutinative language, several words have more than one POS category that makes the word ambiguous. There is an inflection of noun and verb in a sentence in accordance with the grammatical characteristics

DOI: 10.4018/978-1-7998-0951-7.ch054

as well. Therefore, POS tagging becomes a challenging task for Assamese. POS Tagger tries to assign the accurate POS labels to ambiguous words in a sentence according to the context and it has a vital role in various NLP applications as because the POS tagged data is used in many other NLP tasks (Jurafsky & Martin, 2000), e.g., in Parsing, the tagged data helps in finding out noun and verb groups, in Named Entity Recognition, it helps in determining the proper names like the name of a person, place or a thing, in Information Retrieval, it helps in selecting the proper nouns or other important word classes from a given text, in Speech Recognition, it helps in modeling a language, in Machine Translation, it helps in generating the probability for word translation of the source language into the target language, as well as it is useful for many other NLP applications. Thus, it is considered as an initial step of the language processing task. As POS Tagger has a great impact on other NLP systems, a tagging result with high accuracy is always encouraging.

There are several methods of POS tagging and basically there are three main approaches which are Rule Based Approach, Stochastic Approach and Hybrid Approach. Rule Based POS tagging is the most primitive approach where hand-written linguistic rules are used for tagging. These rules identify the appropriate tag for an ambiguous word. This method is dependent on dictionary or lexicon to generate the possible POS tags for every word in input text. The Stochastic Approach is based on the probabilities of words that occur for a particular tag. The tag which occurs most repeatedly in the training data is assigned to unknown or ambiguous word. The probability of a given sequence of tags is calculated from the frequency of words from the annotated training corpus. Hybrid Approach is the combination of more than one method which usually contains rule- based and statistical methods. This model uses the essential feature of statistical approaches and uses the rules for better efficiency. The developed POS Tagger for Assamese follows the Stochastic Approach. A bigram Hidden Markov Model (HMM) is used which is one of the processes in this technique. It is a probabilistic model that uses an annotated training corpus. The tagging process is done by computing the tag sequence probability and the word likelihood probability of the corpus. This method is called supervised learning method. Therefore, HMM requires a large amount of annotated corpus to obtain high accuracy. On the other hand, unsupervised learning method does not use the annotated corpus and it calculates the probabilities by using automatic word groupings.

This paper is further divided into five more sections in which second section provides the related work and next section shows the morphological characteristics of Assamese. Fourth section describes the approach then fifth section gives the evaluation of the system. Finally, the paper is concluded in sixth section.

2. RELATED WORK

Till recent years, many research works have been accomplished on Part of Speech tagging for different languages using different approaches. In case of the Indian languages like Hindi, Malayalam, Bengali, etc. which are morphologically rich in nature, a good number of POS taggers have been developed using Stochastic Approach with varied accuracies. This approach takes less effort in implementation and involves very little knowledge about the language but requires large training data to obtain high accuracy. (Dalal et al., 2006), developed POS Tagger using Maximum Entropy Markov Model for Hindi with a corpus of 15562 words and using 27 POS tags and obtained an accuracy of 94.81%. (Joshi et al., 2013), developed HMM based POS Tagger for Hindi using IL POS tag set and obtained an accuracy of 92%.