HMM based POS Tagger for a Relatively Free Word Order Language

Arulmozhi Palanisamy and Sobha Lalitha Devi

AU-KBC Research Centre, MIT Campus,
Chromepet, Chennai - 600044, India.
{p.arulmozhi,sobha}@au-kbc.org

Abstract. We present an implementation of a part-of-speech tagger based on hidden markov model for Tamil, a relatively free word order, morphologically productive and agglutinative language. In HMM we assume that probability of an item in a sequence depends on its immediate predecessor. That is the tag for the current word depends up on the previous word and its tag. Here in the state sequence the tags are considered as states and the transition from one state to another state has a transition probability. The emission probability is the probability of observing a symbol in a particular state. In achieving this, we use viterbi algorithm. The basic tag set including the inflection is 53. Tamil being an agglutinative language, each word has different combinations of tags. Compound words are also used very often. So, the tagset increases to 350, as the combinations become high. The training corpus is tagged with the combination of basic tags and tags for inflection of the word. The evaluation gives encouraging result.

1 Introduction

Part of speech (POS) tagging is the task of labeling each word in a sentence with its appropriate syntactic category called part of speech. It is an essential task for all the language processing activities. There are two factors determining the syntactic category of a word (a) the words lexical probability (e.g. without context, *bank* is more probably a noun than a verb) (b) the words contextual probability (e.g. after a noun or a pronoun, bank is more a verb than a noun, as in "I bank with HSBC "). Hence it disambiguates the part of speech of a word when it occurs in different contexts. For any POS work the tagset of the language has to be developed. It should contain the major tags and the morpho-syntactic features called the sub tags.

In this paper we present a POS tagger developed for Tamil using HMM and Viterbi transition. Tamil is a South Dravidian language, which is relatively free word order, morphologically productive and agglutinative in nature. It is an old language and most of the words are built up from roots by following certain patterns and adding suffixes. Due to the agglutination many combination of words and suffixes occur to form a single word whose POS is a combination of all the suffixed words or suffixes. Hence the total number of tagset increases as the combination increases. In this work we confine to a fixed set of tagset, which gives the most commonly needed tags for any

Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 37-48