**Hidden Markov Model**

**POS Tagging**

**Project Report**

# Natural Language Processing

*To the*

# The University Of Texas at Dallas



**`** *Submitted by*:

**Prashant Prakash**

**(Net id: pxp141730)**

## The University of Texas at Dallas Dallas – 75252

December, 2015

1. **INTRODUCTION**

In Corpus linguistics , part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Research on part-of-speech tagging has been closely tied to corpus linguistics. The first major corpus of English for computer analysis was the Brown Corpus developed at Brown University by Henry Kučera and W. Nelson Francis, in the mid-1960s. It consists of about 1,000,000 words of running English prose text, made up of 500 samples from randomly chosen publications. Each sample is 2,000 or more words (ending at the first sentence-end after 2,000 words, so that the corpus contains only complete sentences).

The Brown Corpus was painstakingly "tagged" with part-of-speech markers over many years. A first approximation was done with a program by Greene and Rubin, which consisted of a huge handmade list of what categories could co-occur at all. For example, article then noun can occur, but article verb (arguably) cannot. The program got about 70% correct. Its results were repeatedly reviewed and corrected by hand, and later users sent in errata, so that by the late 70s the tagging was nearly perfect (allowing for some cases on which even human speakers might not agree).

For some time, part-of-speech tagging was considered an inseparable part of natural language processing, because there are certain cases where the correct part of speech cannot be decided without understanding the semantics or even the pragmatics of the context. This is extremely expensive, especially because analyzing the higher levels is much harder when multiple part-of-speech possibilities must be considered for each word.

1. **POS TAGGING**
   1. **What is it?**

POS tagging of assigning a part-of-speech to each word in a sentence. (automatically or manually). It is useful in resolving lexical ambiguity.

**Tags**

**WORDS**

* 1. **Why useful?**

POS tagging is useful in

* Information Retrieval
* Text to Speech: object(N) vs. object(V)
* Word Sense Disambiguation

POS tagging is useful as a preprocessing step of parsing.

* Unique tag to each word reduces the number of parses.
  1. **Why POS tagging is hard?**

POS tagging is hard mainly because of “**Ambiguity”** . The major ambiguity found in tagging a word to its part of speech is “Noun” gets tagged to Verb most of the times and vice-versa. Few examples are shown below:

1. “Plants/N need light and water.”
2. “Each one plant/V one.”
3. “Flies like a flower”
   1. Flies: noun or verb?
   2. like: preposition, adverb, conjunction, noun, or verb?
   3. a: article, noun, or preposition?
   4. flower: noun or verb?
   5. **Methods for POS tagging**

There are different approaches taken to do POS tagging.

* Stochastic tagging: Maximum likelihood, Hidden Markov model tagging.

Pr(Det-N) > Pr (Det-Det)

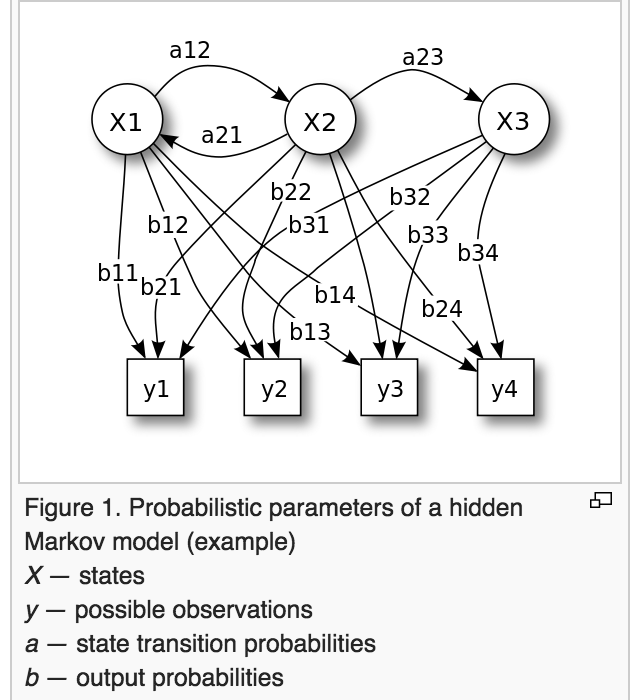
* Rule based tagging: if <some pattern> then … <some part of speech > (allows for several passes)
* Constrain Grammar (CG) tagging: The cautious approach: ‘Don’t guess, just eliminate the impossible!’
* Transmation-based (TB) tagging: The whimsical approach: ‘ Guess first , then change your mind is necessary!’
  1. **Asdsd**

1. **HIDDEN MARKOV MODEL**
   1. **What is it?**

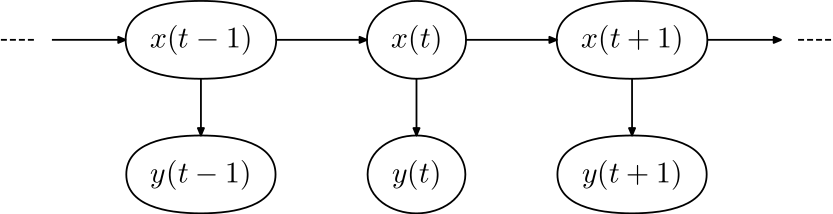
A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.

In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. The adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

* 1. **Architecture**

****

The diagram below shows the general architecture of an instantiated HMM. Each oval shape represents a random variable that can adopt any of a number of values. The random variable x(t) is the hidden state at time t (with the model from the above diagram, x(t) ∈ { x1, x2, x3 }). The random variable y(t) is the observation at time t (with y(t) ∈ { y1, y2, y3, y4 }). The arrows in the diagram (often called a trellis diagram) denote conditional dependencies.



From the diagram, it is clear that the conditional probability distribution of the hidden variable x(t) at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden variable x(t − 1); the values at time t − 2 and before have no influence. This is called the Markov property. Similarly, the value of the observed variable y(t) only depends on the value of the hidden variable x(t) (both at time t).

* 1. **asadasd**