

CS4248 Assignment 2

U096883L Shawn Tan

1 Introduction

This assignment requires us to build a Part-of-Speech (POS) tagger, using training data from part of the Penn Treebank. The method used in our approach has to employ a Hidden Markov Model (HMM). This entails learning from the training data a set of parameters required for the HMM. Table 1 shows the various sets of data we have to collect for this particular assignment.

In the following sections, we will explain in detail how individual aspects of the HMM was created, and experiment with two simple smoothing techniques, Laplace and Witten-Bell smoothing. The two techniques will be evaluated according to their precision, recall and F1 measures.

Name	Description
V	all unique words
S	all unique POS tags
$p(s_i s_{i-1})$	transition probability from one POS tag to another
$p(w s)$	probability of seeing a word given a POS tag

Table 1: Parameters for a POS tagger HMM