CS4248Assignment 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, outlining some of the technical difficulties faced. We also experiment with two simple smoothing techniques, Laplace (add-one) and Witten-Bell smoothing. The two techniques will be evaluated according to their precision, recall and F1 measures.

2 Learning from `sents.train`

The `sents.train` dataset contains 39,832 lines, each word annotated with POS tags.

In order to extract the relevant information we count.

In order to obtain $p(s_i \mid s_{i-1})$, we need to count the different number of times one tag is followed by another. For each line, we extract the POS for each token, leaving a list of POS

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

Table 1: Parameters for a POS tagger HMM

tags. We then prepend a ‘’□ at the beginning, and a ‘$’ character at the end. This ensures that the probabilities for a given POS starting a sentence are also taken into account