CSE 5525 Speech and Language Processing (Spring 2020)
In-class Exercise: Viterbi Algorithm and HMM

Prof. Wei Xu, The Ohio State University

February 12, 2020

This in-class exercise is dedicated to my PhD advisor, Ralph Grishman. He designed this lovely toy example with a two-word language, which namely consists of only two words: $fish$ and *sleep*. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden state. It can be used to solve Hidden Markov Models (HMMs) as well as many other problems.

Suppose we have a small training corpus. In this training corpus, word $fish$ appears 8 times as a noun (NN) and 5 times as a verb (VB); word *sleep* appears twice as a noun and 5 times as a verb.

(1) What are the emission probabilities?

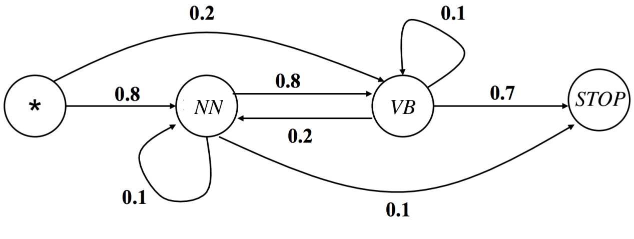
$e(fish|NN)$

$e(sleep|NN)$

$e(fish|VB)$

$e(sleep|VB)$

Also suppose we already have a simple Part-of-Speech HMM model, a bigram one:



The arrows in the diagram denote conditional probabilities, e.g. $q(NN|*) = 0.8$ and $q(NN|VB) = 0.2$.

(2) Use the Viterbi algorithm to find the most likely POS tag sequence for a test sentence *"Fish sleep."* For simplicity, we ignore the punctuation in the calculation for this toy example.

The Viterbi algorithm for a bigram HMM

Input: a sentence $x_1 \ldots x_n$, parameters $q(s|v)$ and $e(x|s)$
Initialization: Set $\pi(0, *) = 1$
Definition: $S_0 = *$, $S_k = S$ for $k \in \{1 \ldots n\}$
Algorithm:

```
For $k = 1 \ldots n$,
    For $v \in S_k$,

        $\pi(k, v) = \max_{w \in S_{k-1}} \big( \pi(k-1, w) \times q(v|w) \times e(x_k|v) \big)$

Return $\max_{v \in S_n} \big( \pi(n, v) \times q(STOP|v) \big)$
```