Natural Language Processing

Lab 6: Hidden Markov Models

---

1 Introduction

In this assignment, we are going to explore the use of Hidden Markov Models for predictive text entry, as used for text messaging on pre-smartphone mobile phones. This is an exercise in mathematical modeling and reasoning, so you will not be asked to do any programming or empirical evaluation this time.

2 The problem

Assume we have a keypad that looks like this:

When used to produce text messages on a mobile phone, this device produces a sequence of numbers that can represent multiple sequences of characters. For example, to type the word *box*, you would produce the key sequence 269. But the same sequence would be produced also for the word *bow*. The predictive text entry problem can be seen as the problem of predicting the most probable word given a sequence of numbers (keys).

3 Build a model

The predictive text entry problem can be seen as a sequence tagging problem, where we are given observable sequences of numbers and are required to find hidden sequences of characters (words). It is therefore a good fit for a Hidden Markov Model. Your first task is to define the model structure.

- What are the (hidden) states of the model?
- What are the (emitted) signals?

4 Estimate probabilities

Once you have defined the model structure (states and signals), there are two probability distributions that need to be estimated.

- How would you estimate the emission probabilities $P(\text{signal}|\text{state})$?
- How would you estimate the transition probabilities $P(\text{new state}|\text{previous state})$?

5 Use model for decoding

Given estimates of the probability distribution, the model can be used to decode key sequences by finding the most probable word.

- How would you compute the probability that 269 corresponds to *box* in your model?
- How would you compute the most probable word corresponding to 269 in your model?

Note: You are not supposed to carry out any concrete computations here, only think as a matter of principle how to do the computations.