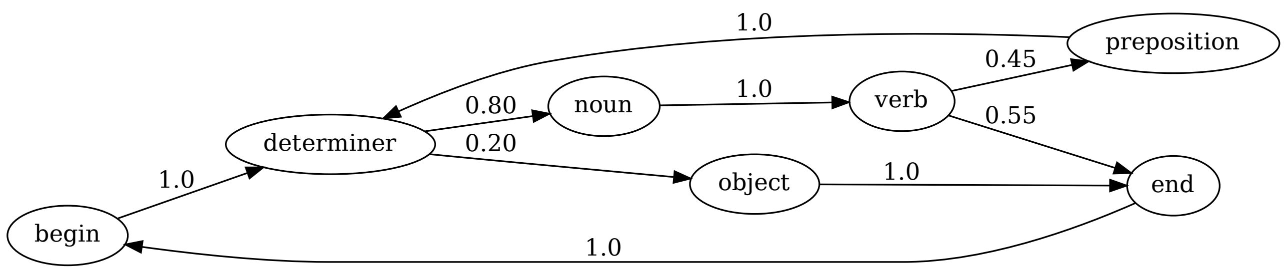
Homework 3 NLP: Written Assignment

CS461 2025-10-27

Imagine a language similar to English, but only containing the following:

1. class D, determiners: {the, that}
2. class N, nouns: {bird, cow}
3. class V, verbs: {jumped, flew}
4. class P, prepositions: {over, under}
5. class O, objects: {Sun, Moon}

A possible hidden Markov model to represent this language has seven hidden states, one for each of the five word classes and two more to mark the beginnings and endings of sentences. π_0, the initial state distribution, is $[1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]$, indicating that all sentence begin in the "begin" state. The state transitions and their probabilities are pictures below.



Part 1: Populate a state transition matrix, A, using the values in the graph above.

Part 2: Once the parameters of a Markov model have been determined, such as with the EM algorithm, the model can be used to make inferences about the language itself. You may have noticed that the transition matrix has loops, and thus there are an infinite number of sentences. Examination of the expected sentence length looks tedious, but is straightforward with a Markov model.

We want to solve for the *stationary distribution*. If π_0 is the initial state distribution, and $\pi_1 = \pi_0 A$ is the state distribution at time 1, and so on, then, at some point, we may reach $\pi = \pi A$, the stationary distribution (see Murphy, section 17.2.3.1). The existence of π is contingent upon a few preconditions; for example, our current state must not be solely determined by the current time (meaning that there is a lack of randomness in the system).

Solving for the stationary distribution is simply done by solving a system of linear equations. One equation comes from the rule of Markov models: $\sum_{i=1}^{k} \pi(i) = 1$. This just says that the sum of state probabilities must sum to 1. For the other equations, we get one equation per column in the transition matrix by expanding $\pi = \pi A$. For example, looking at the state transitions from part 1, class D is transitioned into from states P and bb with probability 1.0, so the expected fraction of time we are in state D is $\pi(D) = \pi(P) + \pi(bb)$. Likewise, the expectation for state P is $0.45\pi(V)$.

Solve the set of equations to determine the stationary distributions, π.

.

$\pi(bb) =$ ______

$\pi(D) =$ ______

$\pi(N) =$ ______

$\pi(V) =$ ______

$\pi(P) =$ ______

$\pi(O) =$ ______

$\pi(ee) =$ ______

Part 3: Use the fraction of time spent in the bb or ee states to deduce the average number of words in a sentence.

Part 4: The sentences created by this model tend may run on, e.g. "the cow jumped over a cat flew under the stars." Find a way to add another hidden state (that shares some words with other states and removes no words from any states) to solve this problem, preventing sentences from repeating the determiner->noun->verb pathway more than once.

Draw the new state transition diagram. You don't need to put in any probabilities.