SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 007
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – 19AD601 – Natural Language Processing

III Year / VI Semester

Unit 2 – WORD LEVEL ANALYSIS

Topic 8- Hidden Markov Model

Hidden Markov Model

An HMM is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), it computes a probability distribution over possible sequences of labels and chooses the best label sequence.

Markov Chains

The HMM is based on augmenting the Markov chain. A Markov chain is a model that tells us something about the probabilities of sequences of random variables, states, each of which can take on values from some set.

These sets can be words, or tags, or symbols representing anything, for example the weather.

A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state. All the states before the current state have no impact on the future except via the current state.

Hidden Markov Model

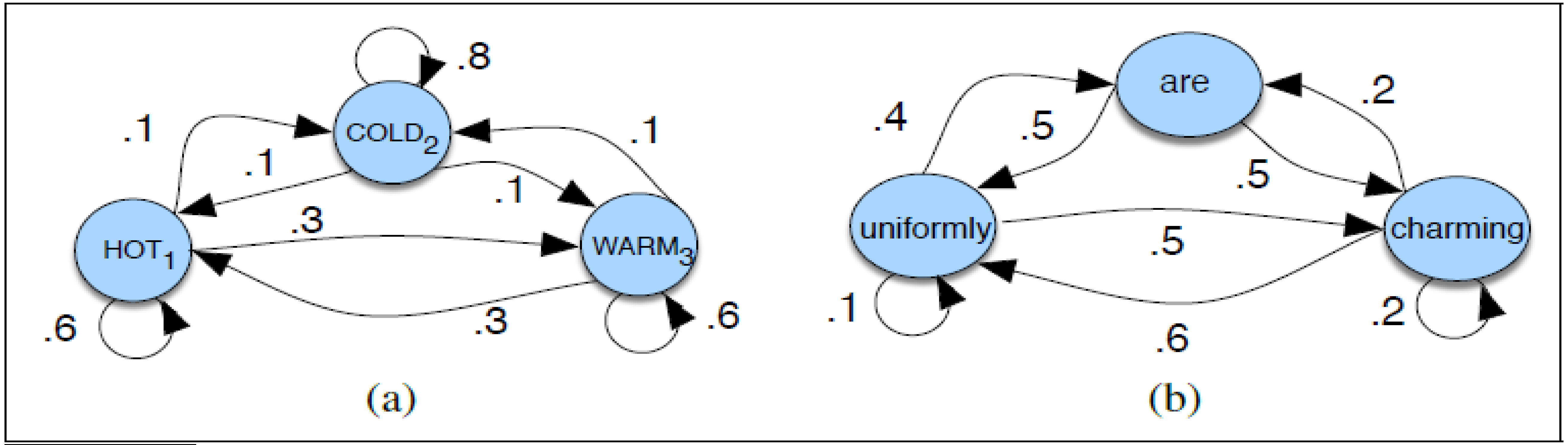

Figure 8.8 A Markov chain for weather (a) and one for words (b), showing states and transitions. A start distribution π is required; setting $\pi = [0.1, 0.7, 0.2]$ for (a) would mean a probability 0.7 of starting in state 2 (cold), probability 0.1 of starting in state 1 (hot), etc.

Hidden Markov Model

Formally, a Markov chain is specified by the following components:

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Hidden Markov Model

The Hidden Markov Model

Markov chain is useful to compute a probability for a sequence of observable events. In many cases, the events we are interested in are hidden events:

- We don’t observe hidden events directly.
- For example we don’t normally observe part-of-speech tags in a text. Rather, we see words, and must infer the tags from the word sequence.
- We call the tags hidden because they are not observed.

A hidden Markov model (HMM) allows us to talk about both observed events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model.

Hidden Markov Model

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state q_i
$\pi = \pi_1, \pi_2, ..., \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Hidden Markov Model

Example

For example, modal verbs like ***will*** are very likely to be followed by a verb in the base form, a VB, like ***race***, so we expect this probability to be high.

In the WSJ corpus, for example, MD occurs 131210 times of which it is followed by VB 101071, for an MLE estimate of

$$P(VB|MD) = \frac{C(MD,VB)}{C(MD)} = \frac{10471}{13124} = .80$$

Hidden Markov Model

Maximum Entropy models

HMM model has the following limitations,

- HMM – Tag and observed word both depend only on previous tag
- Need to account for dependency of tag on observed word
- Need to extract “features” from word & use

To overcome the limitations pf HMM, Maximum entropy model is used,

Maximum entropy classification is a method that generalizes logistic regression to multiclass problems. The Maximum Entropy model is a type of log-linear model.

Hidden Markov Model

Maximum Entropy models

HMM model has the following limitations,

- HMM – Tag and observed word both depend only on previous tag
- Need to account for dependency of tag on observed word
- Need to extract "features" from word & use

To overcome the limitations pf HMM, Maximum entropy model is used,

Maximum entropy classification is a method that generalizes logistic regression to multiclass problems. The Maximum Entropy model is a type of log-linear model.

THANK YOU