LING 570 – HW6

Q1

The states in the bigram output HMM represent the POS tags that are collected from the training data. Each unique POS tag collected from the training data is represented by a state. All possible combinations of state transitions, from one tag to another are listed in the output file. These also include the BOS and EOS tags. In a bigram model, if we have a tag sequence of WP VBD DT, they will be represented by:

From State (WP) => To State (VBD)
From State (VBD) => To State (DT)

From all possible combinations of state transition, the transition probabilities are calculated. The transition probability is a measure of how likely we would have a tag (T_i) given the previous tag (T_{i-1}). The likelihood of tag sequence can be mapped to the transition of $state_i$ to $state_j$ and this is represented by $a_{ij} = P(S_j \mid S_i)$, where a_{ij} is the transition probability.

Taking an example from the bigram output HMM, since the tag # is followed by the tag CD in the training data. We can map # to $state_i$ and CD to $state_j$. Since the tag pair sequence of # followed by CD occurred exactly once and at the same time # on its own occurred exactly once, we have

Count of (#, CD) = 1
Count of (#) = 1

P(CD | #) = 1.

The emission probability is the likelihood of a word being produced by a state. Since each state is a representation of a tag, it is also the likelihood of a word given the tag. The emission probabilities in the output HMM were calculated using this formula: $b_{jk} = P(W_k \mid S_j)$.

Using the same example, word # is tagged with tag # in the training data. The #/# pair occurred exactly once and similarly the tag # also occurred only once, we would have

Count of (#, #) = 1
Count of (#) = 1

P(# | #) = 1.

Q2

The states in the trigram output HMM represent the POS tag pairs that are collected from the training data. Each unique POS tag pair collected from the training data is represented by a state. In a trigram model, if we have the tag sequence of IN NN RB VBN gathered from the training data, we would have:

From State (IN, NN) => To State (NN, RB)
From State (NN, RB) => To State (RB, VBN)

In other words, the occurrence of tag RB depends on the previous tags of IN and NN, while the occurrence of tag VBN would then depends on previous tags of NN and RB. The transition probability is a measure of how likely we would have a tag of (T_i) given the two previous tags of (T_{i-2}) and (T_{i-1}). Each each tag pair sequence can then be mapped to the transition of $state_i$ to $state_j$ and similarly this is represented by $a_{ij} = P(S_j \mid S_i)$, where a_{ij} is the transition probability, $S_j = (T_{i-1}, T_i)$ and $S_i = (T_{i-2}, T_{i-1})$.

The transition probabilities in the trigram output HMM were calculated using this formula:

$$P(T_3 \mid T_1,T_2) = L3*P(T_3 \mid T_1,T_2) + L2*P(T_3 \mid T_2) + L1*P(T_3).$$

Smoothing was used to account for unseen tag sequences and L1, L2 and L3 are Lamda values that were used for interpolation. To achieve the desire effect of smoothing, all possible sequences of unseen trigram tags would have to be generated and added to the calculation of probability values. As such, the trigram output HMM included both tag sequences observed from the training data and as well as unseen tags.

The emission probability in the trigram output HMM is the likelihood of a word being associated with a given tag at a given state. Since each state is a representation of a tag pair (T_{i-1}, T_i), it is the likelihood of a word given the tag (T_i). The emission probabilities in the trigram output HMM were calculated using this formula: $b_{jk} = P(W_k \mid S_j)$, where S_j is the state that contains the tag that generates the word. In addition, unknown words are handled through the introduction of an output symbol <unk> along with a given probability associated to each tag. Based on the given probability of the <unk> symbol, the probabilities for all the known words associated to each given tag are renormalized. That is to say for a given tag, the sum of probabilities for all known words and probability of <unk> equals 1.

End of HW6 – submitted by Wee Teck Tan

Student ID: 0937003

Course Name: LING 570