NLP Assignment.

K. Pragathi
18K41A0528
IV yr CSE-A

4) Explain pos (parts of speech) with HMM?

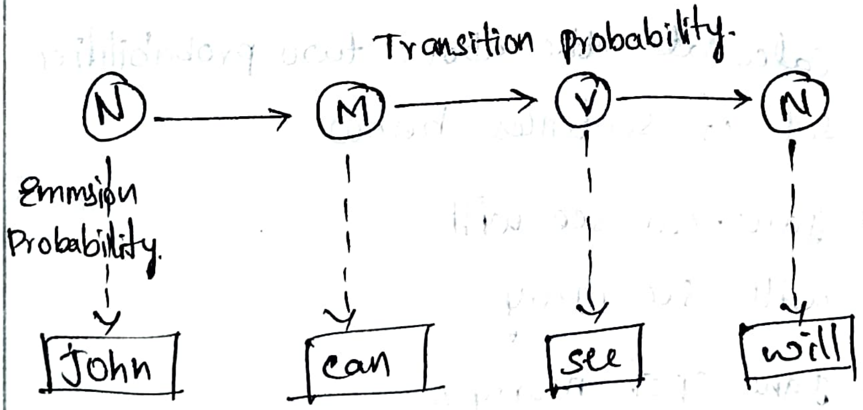
Ans HMM (Hidden Markov Model) is a stochastric technique for pos tagging

→ Hidden Markov models are known as their application to reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, musical score following, partial discharge and bioinformatics.

Pos tagging with Hidden markor model

HMM (Hidden markor model) is a stachostic technique for Pos togging.

→ Let us consider an example proposed by Dr. Lusserano and find out how HMM selects an appropriate tag sequence for a sentence.



In this example we consider only 3 Pos tags that are noun, model and verb

Let the sentence "Ted will spot will" be tagged as noun, model verb and a noun and to calculate the Probability associated with this particular sequence at tags we require their transition probability and Emmission probability.

→ The transition probability is the likelihood of a particular sequance for example how likely is that a noun is followed by a noun. This Probability is known as Transition Probability that the word. It should be high for a particular sequence to be corred.

→ Now, what is the Probability that the word led is a noun these set of probabilities are emission probobabilites and should be high for our tagging to be likely.

Let us calculate the above two probabilities for the set of sentences below.

* mary Jane can see will
* spot will see mary
* will Jane spot marry?
* Mary will Bat spot?

N	N	M	V	N
Mary	Jane	can	see	will

N	M	V	N
Spot	will	see	Mary?

M	N	V	N
will	Jane	spot	Mary?

N	M	V	N
Mary	will	pat	spot

→ The above sentences, the word mary appear four times as a noun to calculate the emission probabilities.

let us create a counting table in a similar manner.

words	Noun	model	verb.
mary	4	0	0
Jane	2	0	0
will	1	3	0
spot	2	0	1
can	0	1	0
see	0	0	2
Pat	0	0	1

Now let us divide each column by the total no. of their appearances for example noun appears nine times in the above sentences so divide each term by a in the noun column we get the following table after this operation.

Words	Noun	model	verb
mary	4/9	0	0
Jane	2/9	0	0
will	1/9	3/4	0
Spot	2/9	0	1/4
can	0	1/4	0
see	0	0	2/4
Pat	0	0	1

from the above table we inter that

The probability that mary noun = 4/9

The probability that mary is model = 0

the Probability that will is noun = 1/9

The probability that will is model = 3/4

In a similar manner, we can configure out the rest of the probabilities, these are the emission probabilities.

Next we have to calculate the transition Probabilities so define two more tags <S> and <E>, <S> is placed at the bigining of each sentence and <E> at the end as shown in the figure below.

	N	N	M	V	N	
<S>	mary	Jane	can	see	will	<E>

	N	M	V	N	
<S>	spot	will	see	mary	<E>

	M	N	V	N	
<S>	will	Jane	spot	mary	<E>

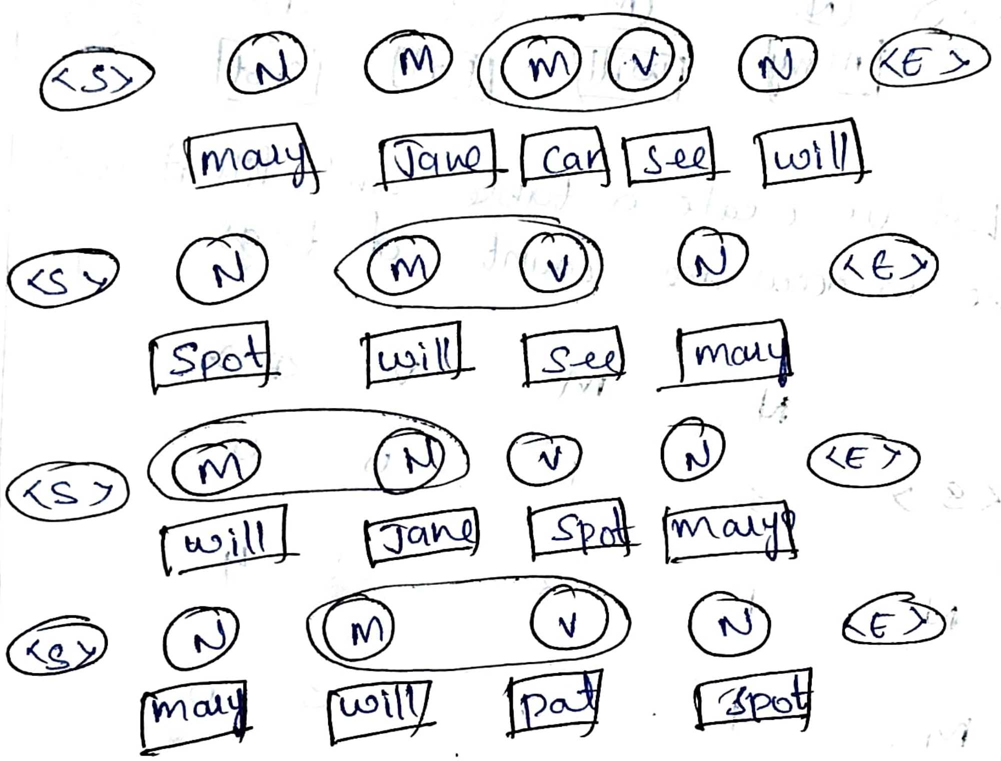
	N	M	V	N	
<S>	mary	will	Pat	spot	<E>

Let us create a table and fill it with the co-occurance counts of tags.

	N	M	V	<E>
<S>	3	1	0	0
N	1	3	1	4
M	1	0	3	0
V	4	0	0	0

→ In the above figure, we can see that the <S> tag is followed by the N tag three times, thus the first entry is 3. The model tag follows the <S> just once, thus the second entry is 1. In Similar manner, the rest of the table is filled.

→ Next we divide each item term in a row of the table by the total no. of co-occurance of the tag in consideration.
for example: The model tag is followed by any other tag four times as shown below, thus we divide each element in the third row by four.



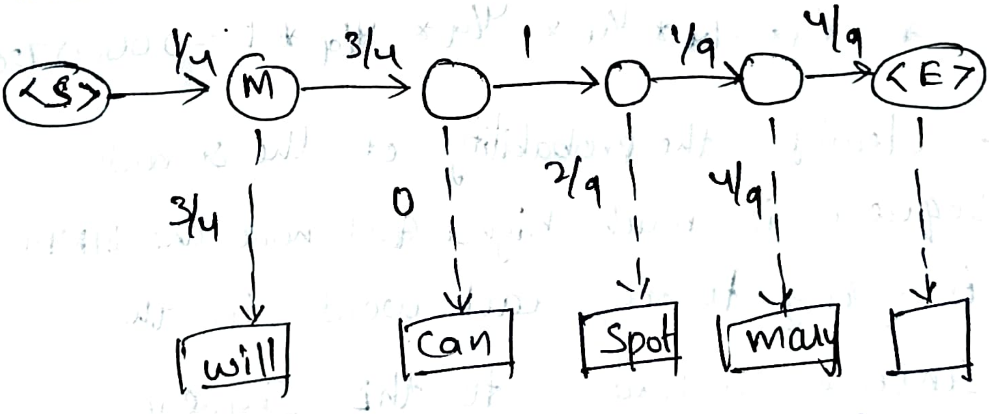
	N	M	V	<E>
<S>	3/4	1/4	0	0
N	1/9	3/9	1/9	4/9
M	1/4	0	3/4	0
V	4/4	0	0	0

→ These are the respective transition probabilities for the above four sentences. Now how does the HMM determine the appropriate sequences of tags for a particular sequance from the above tables ? let us find it out

→ Take a new sentence and tag them with wrong tags, let the sentences 'I will can spot mary' be tagged as

- will as a model
- can as a verb
- Spot as a noun
- mary as a noun

Now we calculate the probability of this sequence being correct in the following manner.



→ The probability of the tag model (M) comes after the tag <S> is 1/4 as seen in the table, Also the probability that the word will is a model is 3/4

→ Since the tags are not correct, the Product is zero.

$1/4 * 3/4 * 3/4 * 0 * 1 * 2/9 * 1/9 * 4/9 * 4/9 = 0$

when these words are correctly tagged we get a probability greater than zero as shown below.

Calculating the product of these terms we get

$3/4 * 1/9 * 3/9 * 1/4 * 3/4 * 1/4 * 1 * 4/9 * 4/9 =$

0.00025720164

$\langle S \rangle \rightarrow N \longrightarrow M \longrightarrow N \longrightarrow N \longrightarrow \langle E \rangle$

$= 3/4 * 1/9 * 3/9 * 1/4 * 1/4 * 2/9 * 1/9 * 4/9 * 4/9$

$= 0.00000846754$

$\langle S \rangle \longrightarrow N \longrightarrow M \rightarrow N \longrightarrow V \longrightarrow \langle E \rangle = 3/4 * 1/9 * 3/9$

$* 1/4 * 3/4 * 1/4 * 4/9 * 4/9 * 1 = 0.00025720164$

→ clearly, the probability of the second sequence is much higher and hence the HMM is going to tag each word in the sentence according to this Sequence.