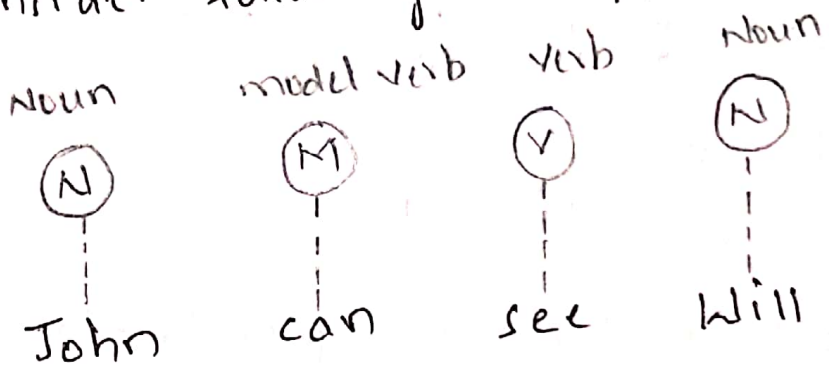
4Q POS tagging with HMM?

Ans) HMM (Hidden Markov Model) is a stochastic technique for POS tagging. Hidden Markov Models are known for their applications to reinforcement learning and temporal pattern recognisation such as speech, handwriting etc..

Let us consider following Example



To calculate the probability associated with the particular sequence of tags we require 1) Transition probability
2) Emission probability

Let us calculate above (2) prob to set of sentences.

1. Mary Jane can see will
2. Spot will see mary

3. Will Jane spot Mary ?

4. Mary will pat spot.

1. (N) Mary (N) Jane (M) can (V) see (N) Will

2. (N) spot (M) will (V) see (N) Mary

3. (M) Will (N) Jane (V) spot (N) Marry ?

4. (N) Mary (M) will (V) pat (N) spot

TO calculate Emission probability.

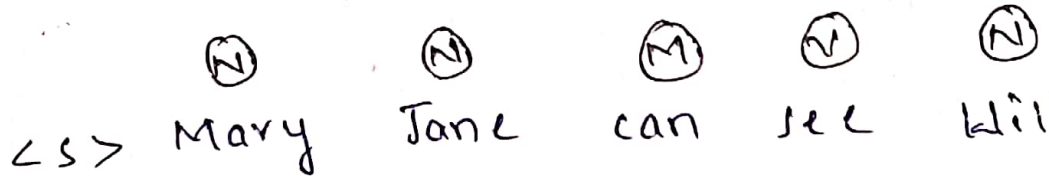
Words	Noun	Model	Verb
Mary	4	0	0
Jane	2	0	0
Will	1	3	0
spot	2	0	1
can	0	1	0
see	0	0	2
Pat	0	0	1

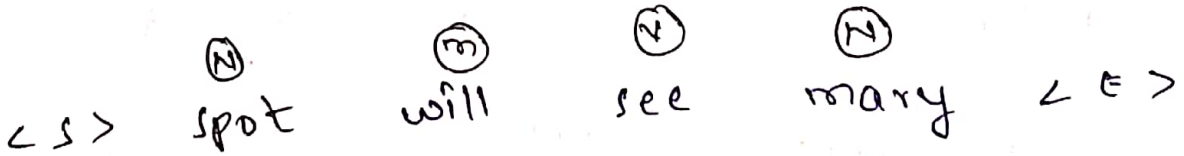
Now let us devide each colomn by total no. of apprearance

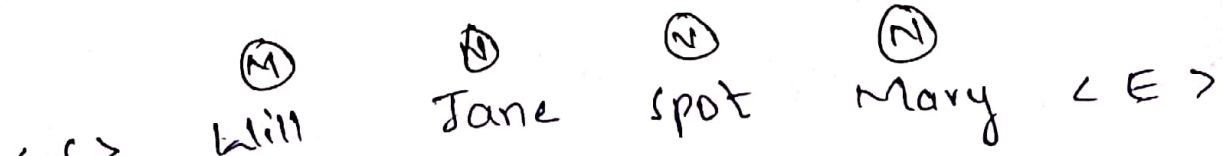
Words	Noun	model	verb
Mary	4/9	0	0
Jane	2/9	0	0
Will	1/9	3/4	0
spot	2/9	0	1/4
can	0	1/4	0
see	0	0	2/4
Pat	0	0	1

Next we have to calculate transition prob so define two more tags <S>, <E>

1.







	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 above figure <s> tag is followed by Noun tag 3 times thus first Entry is 3. in similar manner all.

Now Let us calculate probability

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

Now take a setence and tag it

1. Will can spot Mary

calculating prob

$$\frac{1}{4} * \frac{3}{4} * \frac{3}{4} * 0 * \frac{1}{2} * \frac{1}{9} * \frac{4}{9} = 0$$

* If the tag is correct probabity should be greater than '0'. so above tag is wrong.

Now correctly tagged.

(N) (M) (V) (N)

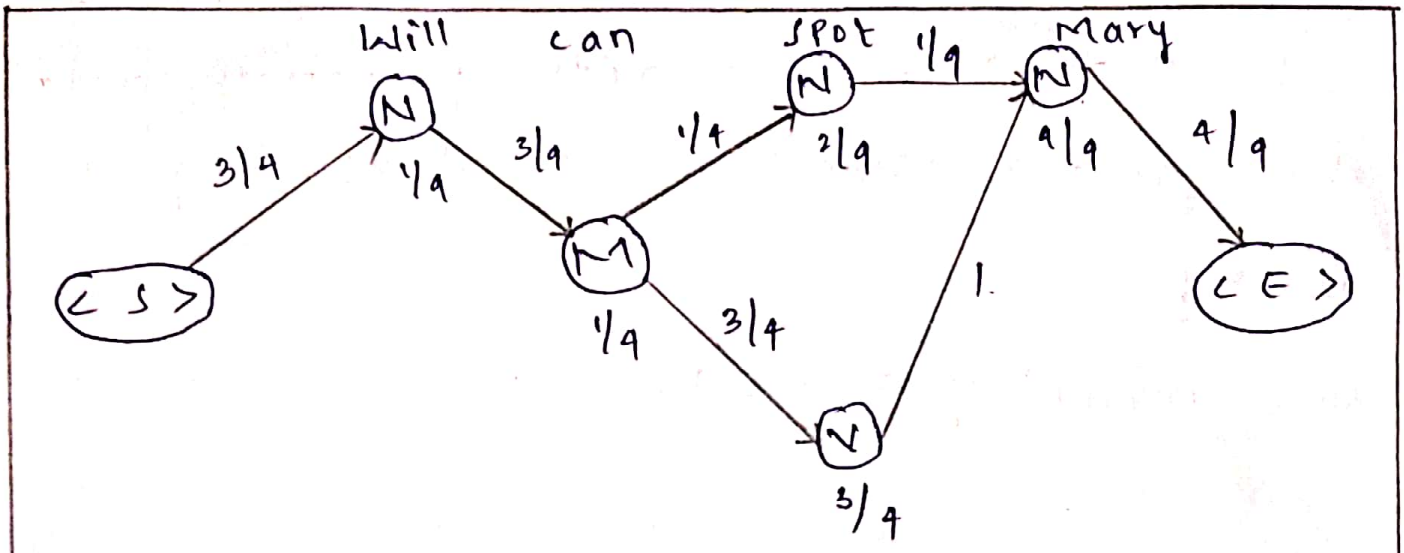
Will can spot Mary

$\frac{3}{4} * \frac{1}{9} * \frac{3}{9} * \frac{1}{4} * \frac{3}{4} * \frac{1}{4} * \frac{4}{9} * \frac{4}{9}$

$= 0.000 25 72$

* for above model we get 81 combinations calculating 81 seems larger so for these Penn tree bank project is taken under consideration

* After removing all the vertices & Edges with 0 probability we get 2 path that lead to End.



① $\langle S \rangle \rightarrow N \rightarrow M \rightarrow N \rightarrow N \rightarrow \langle E \rangle$

$= \frac{3}{4} * \frac{1}{9} * \frac{3}{9} * \frac{1}{4} * \frac{2}{9} * \frac{1}{9} * \frac{4}{9} * \frac{4}{9} = 0.000081$

② $\langle S \rangle \rightarrow N \rightarrow M \rightarrow N \rightarrow V \rightarrow \langle E \rangle$

$= \frac{3}{4} * \frac{1}{9} * \frac{3}{9} * \frac{1}{4} * \frac{3}{4} * \frac{1}{4} * \frac{1}{4} * 1 * \frac{4}{9} * \frac{1}{9}$

$= 0.000257 \approx 2$

clearly, the probability of ② sequence is much higher hence HMM is going to tag Each word in setence according to that sequence.