AREEFA 18K41A0505 ASSIGNMENT_2

In detailed 4th Question

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

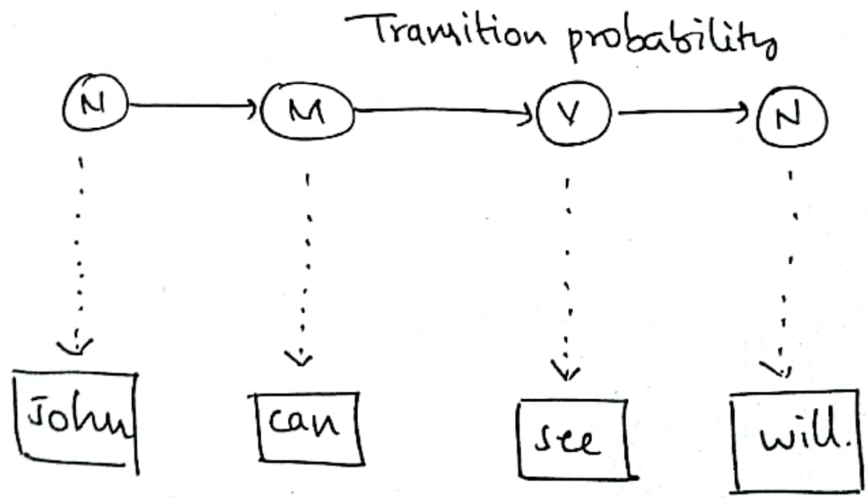
Ans. HMM (Hidden Markov model) is a stochastic technique for POS tagging.

→ Hidden markov model are known for their application to reinforcement learning and temporal pattern recognition such as speech, hand writing, gesture recognition, musical score following, partial discharge and bioinformatics.

POS tagging with Hidden Markov Model:

HMM (Hidden Markov Model) is a stochastic technique for POS tagging.

1) Let us consider an example proposed by Dr. Luis Serrano 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 of tags we require their transition probability and Emission probability.

1) The transition probability is the likelihood of a particular sequence for example how likely is that a noun is followed by a model and a model by a verb and a verb by a noun. This probability is known as transition probability. It should be high for a particular sequence to be correct.

2) Now what is the probability that the word Ted is noun, will is a model, spot is a verb and will is a noun. These set of probabilities are emission probabilities and should be high for our tagging to be likely.

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

i) Mary jane can see will.

ii) Spot will see Mary

iii) Will Jane spot Mary?

iv) Mary will pat spot?

N	N	M	V	N
Mary	Jane	can	see	will

N	M	V	N
spot	will	see	Mary

Scanned with CamScanner

N	N	V	N
Will	Jane	spot	Mary?
N	M	V	N
Mary	Will	Pat	spot

i) The above sentences, the word Mary appear for three 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 coloumn by the total number of their appearance's for example noun appears nine times in the above sentences so divide each term by 9 in the noun coloumn we get the following table after operation.

<S> N M V N <E>

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 tabe we infer that
The probability that Mary is 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, you can configure out the rest of probabilities, these are the emission probabilities.

Next, we have to calculate the transition probabilities to define two more tags <S> & <E>, <S> is placed at the begining of each sentence and <E> at the end as shown

<S> N M V N <E>
$\boxed{\text{Mary}}$ $\boxed{\text{Jane}}$ $\boxed{\text{can}}$ $\boxed{\text{see}}$ $\boxed{\text{will}}$

<S> N M V N <E>
$\boxed{\text{spot}}$ $\boxed{\text{will}}$ $\boxed{\text{see}}$ $\boxed{\text{Mary}}$

<S> M N V N <E>
$\boxed{\text{will}}$ $\boxed{\text{Jane}}$ $\boxed{\text{spot}}$ $\boxed{\text{Mary?}}$

<S> N M V N <E>

Mary will pat Spot

Let us again create a table and fill it with the co-occurence counts of the 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 one, thus the second entry is 1. In a similar manner, the rest of the table is filled.

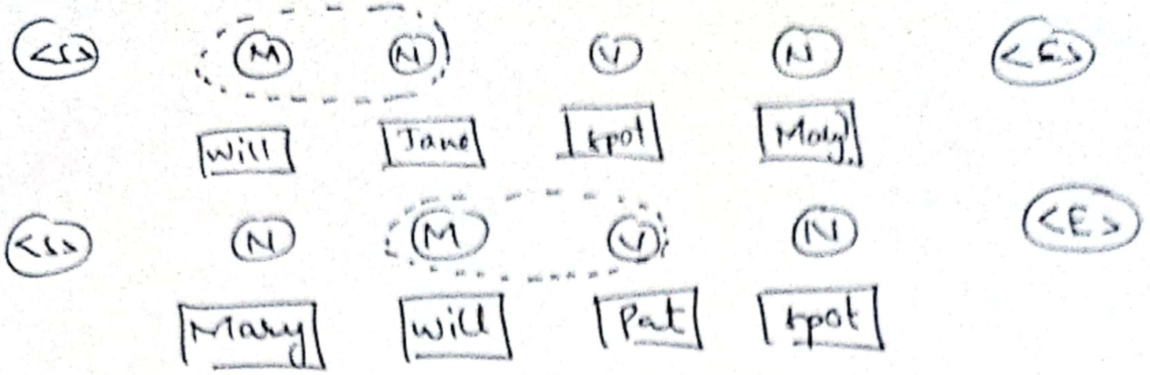
i) Next we divide each item term in a row of the table by the total number of co-occurences of the tag in consideration for example, The model tag is followed by any other tag four times as shown below. thus we devide each element in the third row by four.

<S> N N M V N <E>

Mary Jane can see will

<S> N M V N <E>

Spot will see Mary



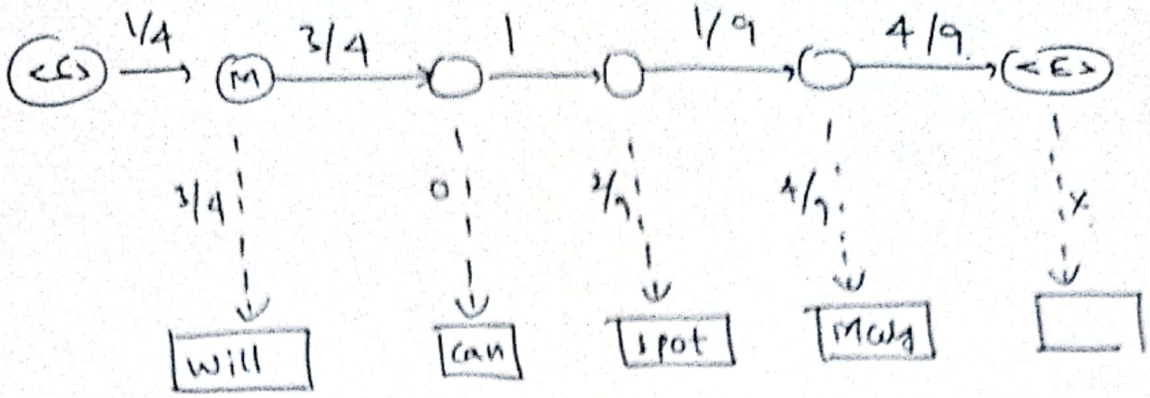
	N	M	V	<E>
<S>	3/4	1/4	0	0
N	1/9	3/4	1/9	4/9
M	1/4	0	3/4	0
V	4/4	0	0	0

i) There are the respective transition probabilities for the above four sentences. Now how does the HMM determine the appropriate sequence of tags for a particular sequence from the above tables? Let us find it out.

ii) Take a new sentence and tag them with wrong tags, Let the sentence are "Will can spot Mary" be tagged as

- Will is 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

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

$$= 0.00025720164$$

<S> → N → M → N → N → <E>

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

$$= 0.00000846154$$

<S> → N → M → N → V → <E> $= 3/4 * 1/9 * 3/9 * 1/4 * 3/4 * 1/4 * 4/9 * 4/9 * 1$

$$= 0.00025720164$$

Scanned with CamScanner

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.