NLP ASSIGNMENT-II

A.Sreechandana

18K41A0565

4. Explain POS (Parts-of-speech) with HMM?

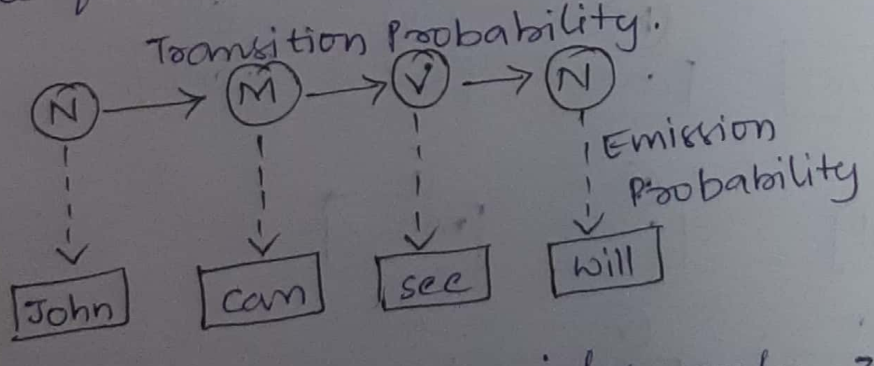
A. HMM (Hidden Markov Model) is a stochastic technique for POS tagging.

* Hidden Markov models are known for their applications to reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, musical score following, partial discharge and bio-informatics.

POS tagging with Hidden Markov Model:-

HMMC is a stochastic technique for POS tagging.

* Let us consider an example proposed by Dr. Luissera no 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.

* 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.

* Now, what is the probability that the word 'Ted' is a noun, 'will' is a model, 'spot' is a verb and 'will' is a noun.

* Let us calculate the above 2 probabilities for the set of sentences below.

* Mary Jane can see will
* spot will see mary
* will Jane spot Mary?
* Mary will pat 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 appears 4 times as a noun.

Next, we have to calculate transition probabilities so define 2 more tags <S> and <E>. <S> is placed at the beginning of each sentence and <E> at the end

	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>

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

* Next, we divide each item in a row table by the total no. of co-occurrences of the tag in consideration.

	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>

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 9 times in the above sentences so divide each term by 9 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.

Probability mary is Noun = 4/9

Probability Mary is Model = 0

Probability will is Noun = 1/9

Probability will is model = 3/4

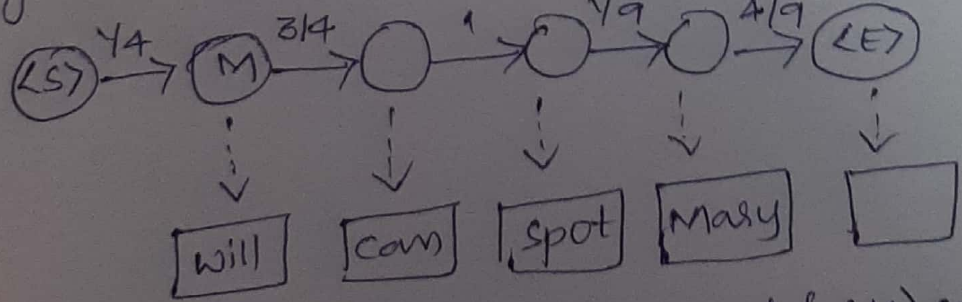
	N	M	V	<E>
<S>	3/4	1/4	0	0
N	1/9	3/9	1/9	4/9
M	1/4	0	3/4	0
V	4/4	0	0	0

* These are the respective transition probabilities for the above 4 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.

* Take a new sentence and tag them with wrong tags, Let the sentence '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

$3/4 * 1/4 * 3/9 * 1/4 * 3/4 * 1/4 * 1 * 4/9 * 4/9 = 0.00025720$

$\langle S \rangle \to N \to M \to N \to V \to \langle E \rangle = 3/4 * 1/9 * 3/9 * 1/4 * 3/4 * 1/4 * 4/9$
$* 4/9 * 1 = 0.00025720164$

* 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.