N-Gram.

Advantage

* Easy to understand, implement
* Can be easily convert to any gram

Disadv:

* Underflow due to multiplication of probabilities
  Sol: Use log. Add probabilities
* Zero probability Problem.
  Sol: Use Laplace smoothing $\sum_{i=1}^{n} \boxed{\frac{value+1}{value + (uniquewords)}}$
  except <s>

Given Corpus:

<s> I am Henry </s>
<s> I like College </s>
<s> Do Henry like college </s>
<s> Henry I am </s>
<s> Do I like Henry </s>
<s> Do I like College </s>
<s> I do like Henry </s>.

Q) <s> I like college </s>

$= P(I|\langle s\rangle) \times P(like|I) \times P(college|like) \times P(\langle /s\rangle|college)$

$= 3/7 \times 3/6 \times 3/5 \times 3/3 = \frac{9}{70} = 0.13.$

$= \log(3/7) + \log(\frac{3}{6}) + \log(\frac{3}{5}) + \log(3/3) = \cancel{20} -2.0513$

Perplexity

It is the inverse probability of the test data which is normalized by the number of words.

$$PP(W) = P(w_1, w_2, w_3 \ldots w_n)^{-1/N}$$

$$PP(W) = \left(\prod_{i=1}^{n} \frac{1}{P(w_i \mid w_{i-1})} \right)^{1/N}$$

Ex: <s> I like College </s>

Perplexity for Bigram

$= P(I \mid \langle s \rangle) \times P(like \mid I) \times P(College \mid like) \times P(\langle /s \rangle \mid college)$

$= 3/7 \times 3/6 \times 3/5 \times 3/3$

$= 9/70$

$= 0.13$

$$PP(W) = (1/0.13)^{1/4} = \underline{1.67}$$

Perplexity for Trigram

$P(W) = P(like \mid \langle s \rangle I) \times P(college \mid I\ like) \times P(\langle /s \rangle \mid like\ college)$

$P(W) = 1/3 \times 2/3 \times 3/3$

$= 2/9$

$= 0.22$

$$PP(W) = \left(\frac{1}{0.22}\right)^{1/3}$$

$= \boxed{1.66}$ ← less probability so ~~more~~ best Predicting Model.

[Which language Model will have less perplexity is the best model for your data]

Parts of Speech Tagging (HMM).

Noun: john, car, India..

Verb: run, swim ...

Modal verbs must, will, would, can, may.

Mary Jane will
N N N

Jane saw will
N V N

Collect data — Create lookuptable — Tag our target sentence. — done

Lookuptable

Ex1: Jane saw will
N V N

Mary saw Jane
N V N

	N	V
Mary	1	–
saw	–	2
Jane	2	–
will	1	–

Mary Jane will
N N N

Mary will see will
N? M? V? ??

Ex2:

Mary will see Jane
N M V N

will will see Mary
N M V N

Jane will see will
N M V N

lookup table

	N	V	M
Mary	2	0	0
see	0	3	0
Jane	2	0	0
will	2	0	3

* So should consider "context" into consideration (2) using Bigrams.

	N-M	M-V	V-N
Mary-will	1	-	-
will-see	-	3	-
see-jane	-	-	1
will-will	1	-	-
~~will-see~~	~~0~~	~~0~~	~~1~~
see-mary	0	0	1
Jane-will	1	0	0
see-will	0	0	1

Mary will see will
? ? ? ?
Noun M V N

Ex: 3
N N N N
Mary Jane will Spot
Mary Jane Can See will
Spot will See Mary
Will Jane spot Mary?
Mary will ~~spot~~ pat Spot.

Lookup table

	N-M	M-V	V-N
Mary-Jane			
Jane-can			
Can-see			
⋮			

~~Quest~~ I/P stm is $\boxed{\text{Jane will spot will.}}$

If the occurance is not there in data? → Soln. HMM

2 types of Probabilities

1. Transition Probability → How likely ~~to~~ tags probability after another tag
2. Emission Probability → How likely ~~to~~ tag will allocate for a word.

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

lookup table

	N	M	V
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 1/2
pat	0	0	1/4

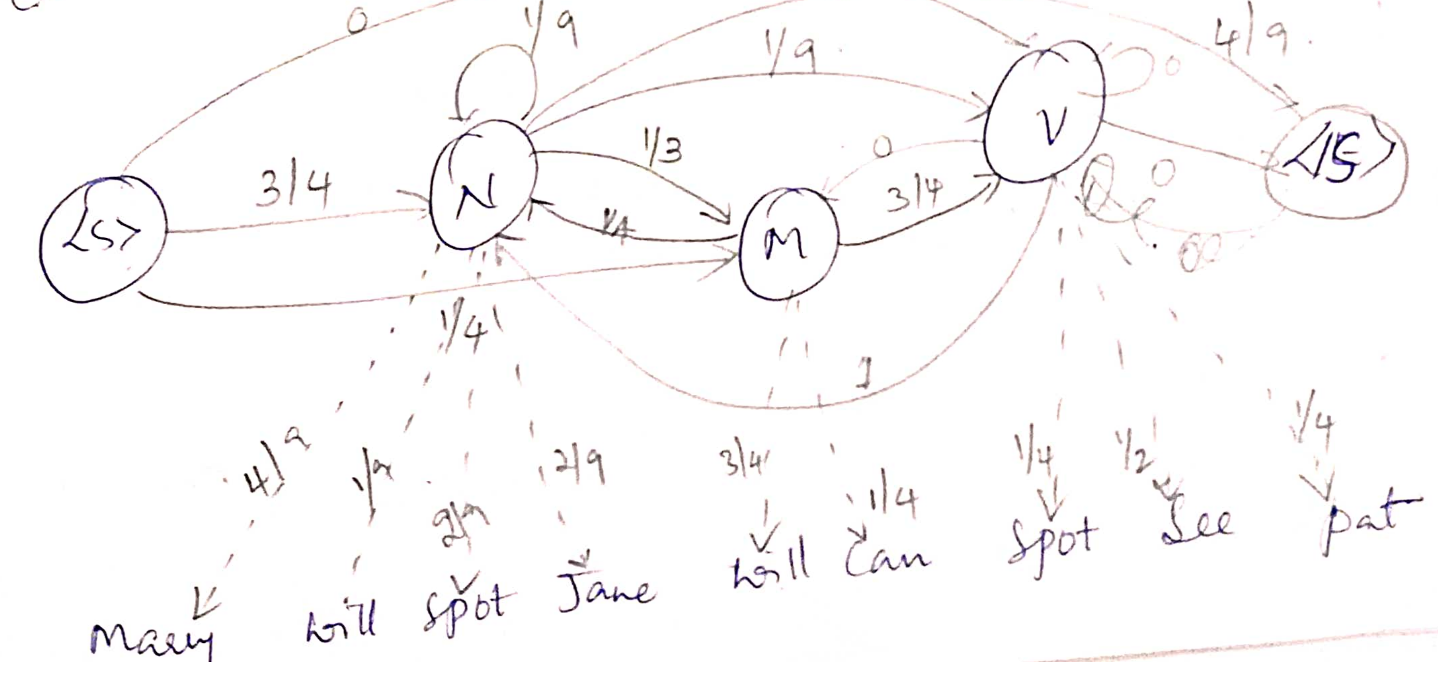
Emission Probability

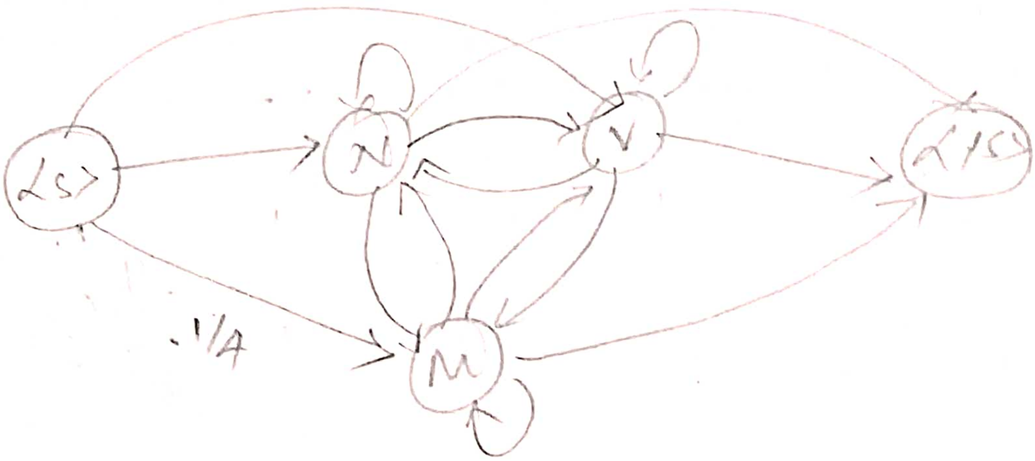
Transition Probability

add <S> & </S> to the stmts.

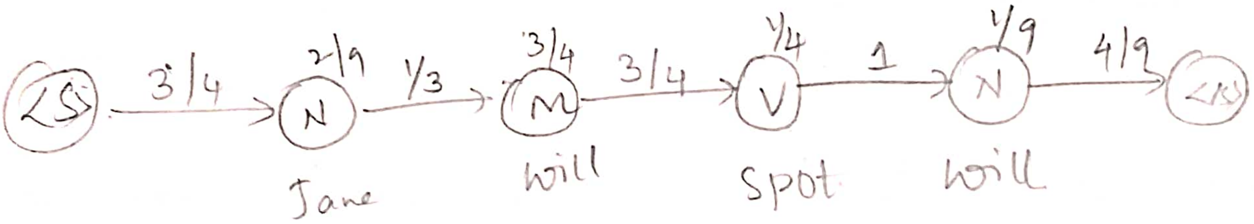
	N	M	V	</S>
<S>	3/4	1/4	–	–
N	1/9	1/3 3/9	1/9	4/9
M	1/4	0	3/4	–
V	1 4/4	–	–	–

Combine 2 Probabilities to create HMM.

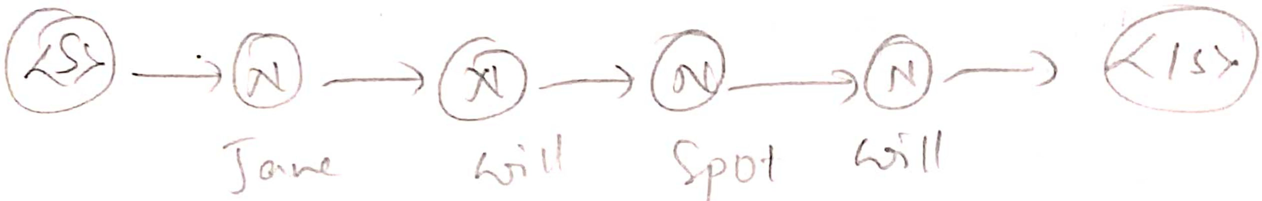




Jane will Spot Will



= $\boxed{0.003858}$ Multiply all the values.



= 0.0000002788

How many Combinations to compare:

3^4 => 81 (Hidden states) ^ no. of words in the stmt

* Computationally expensive.
* As data grows the no. of observations & comparing Probabilities also grow.

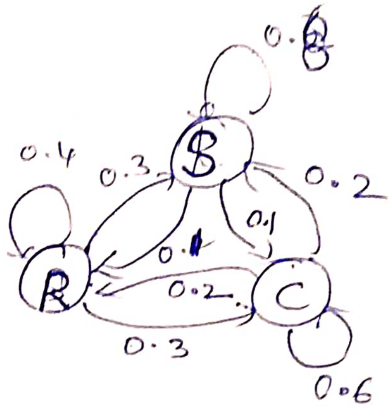
Weather observation

S1 : Rainy S2: Cloudy S3: Sunny.

The state transition probabilities are

$$A = \begin{array}{c|ccc} & R & C & S \\ \hline R & 0.4 & 0.3 & 0.3 \\ C & 0.2 & 0.6 & 0.2 \\ S & 0.1 & 0.1 & 0.8 \end{array}$$

, sum = 1



Question:

① Given that the weather on day 1 is sunny, what is the probability that the weather for the next 7 day will be "Sun-sun-rain-rain-sun-cloudy-sun?

$O = \{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$

$P(O|\text{model}) = P(S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3 | \text{Model})$

$= P(S_3)P(S_3|S_3)P(S_3|S_3)\,P(S_1|S_3)\,P(S_1|S_1)$
$P(S_3|S_1)\,P(S_2|S_3)\,P(S_3|S_2)$

$= 1 * 0.8 * 0.8 * 0.1 * 0.4 * 0.3 * 0.1 * 0.2$

$= \underline{1.536 \times 10^{-4}}$.

② ~~Probo~~ today is Sunny, what is the probability that Tomorrow is sunny & next day is Rainy?

P(Sunny | sunny) * P(Rainy | Sunny)

= 0.8 * 0.1

= 0.08//

CYK [Cocke-kasami-Younger] Parsing

Sentence
↓
[Set of rules (CNF form)] ⇄ [Parsing]
↓
Parsetree.

CNF rules [Chomskey Normal form]

1) should have one Starting Symbol
2) should have LHS & RHS
3) LHS will have only non-terminals.
4) RHS can contain a) only single terminal.
   b) Two non terminals.
5) RHS Cannot contain
   a) ↳ terminal & Nonterminal Combination
   b) ↳ Single non-terminal

Ex:

$A \to a$ ✓ 4(a)
$A \to BC$. ✓ 4(b)
$NP \to N\ PP$ ✓ 4(b)
$NP \to the\ N$
$NP \to the\ NOM$
$NP \to Det\ \cancel{NOM}^{noun}\ PP$ } ×

Ex. ① $NP \to the\ NOM$.
⇓ convert to CNF
Introduce dummy-non terminal
$NP \to Det\ NOM$.
$Det \to the$.

② $NP \to Det\ \cancel{NOM}^{un}\ PP$.
$NOM \to Noun\ PP$
$NP \to Det\ NOM$.

CKY - tree

CNF rules:

S → NP VP [0.80]

NP → Det N [0.30]

VP → V NP [0.20]

V → includes [0.05]

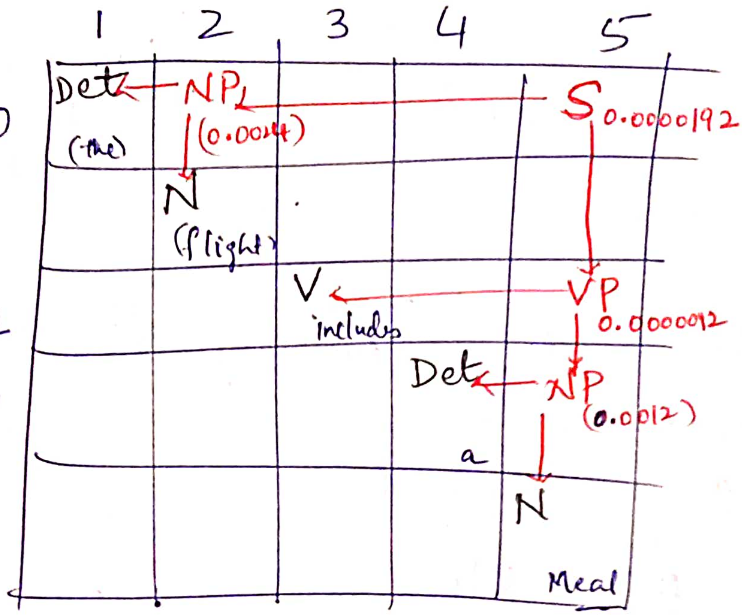
Det → the [0.4]

Det → a [0.4]

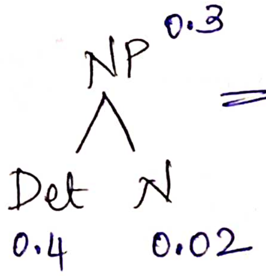
N → Meal [0.01]

N → Flight [0.02]

0 The 1 fight 2 includes 3 a 4 Meal 5

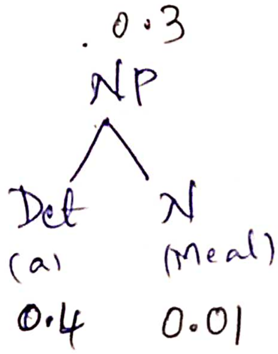


Probabilistic CKY



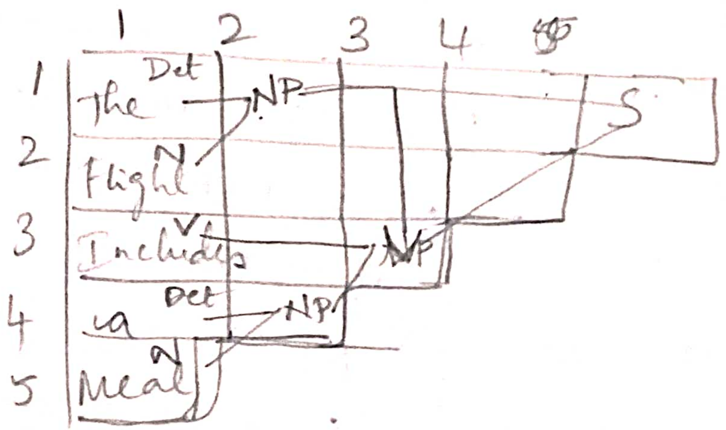
⇒ $0.3 \times 0.4 \times 0.02$

$= 0.0024$



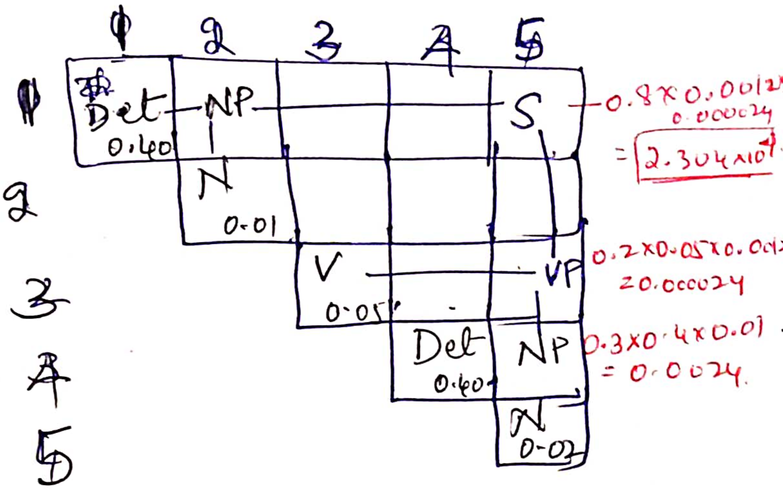
⇒ $0.3 \times 0.4 \times 0.01$

$= 0.0012$



Ex: 0 The 1 Price 2 includes 3 a 4 facemask 5

Production rules	Probability
S → NP VP	0.80
NP → Det N	0.30
VP → V NP	0.20
V → includes	0.05
Det → the	0.40
Det → a	0.40
N → Price	0.01
N → Facemask	0.02



Ex: The Girl wrote an essay

S → NP VP

VP − Verb NP

NP → Det Noun

Det → an | the

Verb → ~~Ato~~ wrote

~~Noun~~ − Girl.

Noun − essay

i = col.
j = row.