Statistical Language Model (SLM)

A SLM is a model that assigns probabilities of sequences of words in a language.

Instead of relying on predefined grammar rules,

↳ slm learns these probabilities from large dataset (corpora) & use statistical method to estimate.

→ These models are based on the idea that the probability of a word depends on the words that came before it, & they calculate this probability by analyzing patterns in a text.

eg:- "The cat chased,"
the model should predict the most probable next word (eg, "the", "a", "mouse")

Types of SLM:-

* n-grams
* Exponential models
* skip-gram models

1. N-grams :-

An n-gram is a sequence of n items from a sample of text or speech, such as syllables, letters, words or base pair.

N-gram model use the frequency of these sequences in a training corpus to predict the likelihood of word sequences.

Ex :- types :-
Uni-gram
bi-gram
tri-gram
& so. on.

The n-grams typically are collected from a text or speech corpus.

Conditional probability :

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

$$P(A,B) = P(B|A) \cdot P(A)$$

more variables :

$$P(A,B,C,D) = P(A)\, P(B|A)\, P(C|A,B)\, P(D|A,B,C)$$

Ex:

P("about five minutes from") =

P(about) * p(five | about) *

P(minutes | about five) *

P(from | about five minutes)

∴ probability of words in sentences:

$$P(w_1, w_2, \ldots w_n) = \Pi [P[W_i | W_1, W_2 \ldots W_{i-1})$$

Types :-

Unigram (1-gram) : No history is used

bigram (2-gram) : One previous word history

Trigram (3-gram) : two previous words history

fourgram (4-gram) : Four Previous Three words history

fivegram (5-gram) : four previous words history.

Ex :- I

Unigram (1-gram)

"about five minutes from ____"

Assume in corpus dinner word is present with highest probability.

Unigram doesn't take into account probabilities with previous words like from, minutes

so unigram will predicts dinner

"about five minutes from dinner"

[Note :-

Corpus : is a large collection of text data. consists of texts, spoken language transcription written or even structured document].

Ex II :- Estimating Bi-gram probabilities

What is the most probable next word predicted by the model for the following sequence?

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>

Ex 1) <s> Do ?

Soln :-

Word	Frequency
<s>	7
</s>	7
I	6
am	2
Henry	5
like	5
college	3
do	4

Next word prediction probability w_{i-1} = do -

do appear 4 times.

Next Word	Prob Next word = $\frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$
P(</s> | do) (it means how many time end sentence </s> appear after do)	0/4
P(<I> | do)	2/4 (prob of I appear after do)
P(<am> | do)	0/4
P(<Henry> | do)	1/4
P(<like> | do)	1/4
P(<college | do)	0/4
P(do | do)	0/4

I is more probable so

Do I (is the answer)

Ex (2) <s> I like Henry ? {use bi gram}

Next Word	prob. Next word = $\frac{N}{D} = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$
P(</s> | Henry)	3/5
P(<I> | Henry)	1/5
P(<am> | Henry)	0
P(<Henry | Henry>	0
P(<like | Henry>	1/5

P(college | Henry)	0
P(do | Henry)	0

<1s> is more probable

<s> I like Henry </s>

Ex 3) <s> Do I like ? use Tri-gram

P <I like> = 3

Tri-gram use 2 prev. words to predict next word.

Next word prediction probability

if W_{i-2} = I & W_{i-1} = like

$$\text{Prob. Next word} = \frac{\text{count}(W_{i-2}, W_{i-1}, W_i)}{\text{count}(W_{i-2}, W_{i-1})}$$

Next word	Prob. Next word
P(</s> | I like)	0/3 (prob. of </s> appear after I like)
P(<I> | I like	0/3 (I appear after I like)
P(<am> | I like)	0/3 (prob. of am appear after I like)
P(<Henry> | I like)	1/3 (prob. of Henry appear after I like)
P(<like> | I like)	0/3 (prob. of like appear after I like)

P(<college> | I like)	2/3
P(do | I like)	0/3

College is probable

∴ Do I like college

Ex :- 4) :- <s> Do I like college ? use 4-gram.

4-gram uses previous three words

P(I like college) = 2.

Next word prediction probability :-

$W_{i-3} = I$, $W_{i-2} = like$, $W_{i-1} = college$

Next word	Prob. Next word = $\frac{count(W_{i-3}, W_{i-2}, W_{i-1}, W_i)}{count(W_{i-3}, W_{i-2}, W_{i-1})}$
P(</s> | I like college)	2/2
P(<I> | I like college)	0/2
P(<am> | I like college)	0/2
P(<Henry> | I like college)	0/2
P(<like> | I like college)	0/2
P(<college> | I like college)	0/2
P(do | I like college)	0/2

</s> is more probable

∴ <s> Do I like college </s>

Ex 5 :- Which of the following sentence is better.

i.e Gets a higher probability in this model (use Bi-gram)

<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>

Soln :-

Lets :- take any two sentence & find prob. <s> I like college </s> &
<s> Do I like Henry </s>.

Word	freq
<s>	7
</s>	7
< I >	6
am	2
Henry	5
like	5
college	3
do	4.

1.) <s> I like college </s>

$= P(I \mid <s>) \times P(like \mid I) \times$

$P(college \mid like) \times P(</s> \mid college)$

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

$= 9/70$

$= 0.13$

2.) <s> Do I like Henry </s>

$= P(do \mid <s>) \times P(I \mid do) * P(like \mid I)$

$\times P(Henry \mid like) \times P(</s> \mid Henry)$

$= 3/7 * 2/4 * 3/6 * 2/5 * 3/5$

$= 9/350$

$= 0.0257$

First statement is more probable.