N-gram Model [underflow problem, (smoothing) Zero-probability problem] N-gram Evaluation

An n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus.

condition Probability :

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

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

More Variables :

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

Chain Rule :-

$$P(x_1, x_2 \ldots x_n) = P(x_1)\ P(x_2|x_1)\ P(x_3|x_1, x_2) \ldots P(x_n|x_1, \ldots x_{n-1})$$

Types :-

- Unigram (1-gram): No history is used
- Bigram (2-gram): one word history is used
- Trigram (3-gram): Two words " " "
- four-gram (4-gram): Three " " " "
- five-gram (5-gram): Four " " " "

Adv :-

- Easy to understands, implement
- can be easily convert to any gram

Disadv :-

→ Underflow due to multiplication of probabilities

solution : use log. Add prob

→ Zero probability problem

solution : use laplace smoothing

Ex :-

Consider the corpus. and find the most probable stmt using 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>

Ans :-
first statement is more probable.

1. <S> I like college </S>

$= P(I|<S>) \times P(like|I) \times P(college|like) \times P(</S>|college)$

$= 3/7 \times 3/6 \times 3/5 \times 3/3 = 9/70 = 0.13$

$= \log(3/7) + \log(3/6) + \log(3/5) + \log(3/3)$

$= -2.0513.$

2. <S> Do I like Henry </S>

$= P(do|<S>) \times P(I|do) \times P(like|I) \times P(Henry|like) \times P(</S>|Henry)$

$= 3/7 \times 2/4 \times 3/6 \times 2/5 \times 3/5 = 9/350 = 0.025$

$= \log(3/7) + \log(2/4) + \log(3/6) + \log(2/5) + \log(3/5) =$

-3.6607

Find the probability for the following statements

1. <s> like college </s>
2. <s> do I like Henry </s>

1. <s> like college </s>

$= P(\text{like} \mid \langle s \rangle) \times P(\text{college} \mid \text{like}) \times P(\langle /s \rangle \mid \text{college})$

$= 0/7 \times 3/5 \times 3/3$

$= 0.$

2. <s> do I like Henry </s>

$= P(\text{do} \mid \langle s \rangle) \times P(\text{I} \mid \text{do}) \times P(\text{like} \mid \text{I}) \times$

$P(\text{Henry} \mid \text{like}) \times P(\langle /s \rangle \mid \text{Henry})$

$= 3/7 \times 2/4 \times 3/6 \times 2/5 \times 3/5$

$= 0.0257.$ (second statement is probable)

In the first stmt, the other like words have more frequency but it doesn't appear after start so it becomes zero.

To overcome zero probability problem, we can apply Laplace smoothing

Laplace Smoothing:-

Find the unique words in the corpus.

<s></s> I, Henry, do, like, am, college.

= 8 unique words

But we exclude <s> as it never comes in bi-gram bcoz

Total unique words = 7.

Give the following bi-gram prob estimated by laplace model.

add 1 at numerator & add 7 in denominator.

① <s> like college </s>

$= p(\text{like} \mid \langle s \rangle) \times p(\text{college} \mid \text{like}) \times p(\langle /s \rangle \mid \text{college})$

$= (0+1)/(7+7) * \frac{(3+1)}{5+7} * \frac{3+1}{3+7}$

$= 1/14 \times 4/12 \times 4/10 = 0.0095$

② <s> do I like Henry </s>

$= p(\text{do} \mid \langle s \rangle) * p(\text{I} \mid \text{do}) * p(\text{like} \mid \text{I}) * p(\text{Henry} \mid \text{like}) * p(\langle /s \rangle \mid \text{Henry})$

$= (8+1)/(7+7) * \frac{(2+1)}{(4+7)} * \frac{(3+1)}{(6+7)} * \frac{(2+1)}{(5+7} * \frac{(3+1)}{(5+7)}$

$= 4/14 * 3/11 \times 4/13 \times 3/12 \times 4/12 = 0.0020$

first statement is more probable.

Language Model Evaluation :- N-gram Evaluation

Language model is better if it is assigning a high probability to the real, frequently observed and grammatical sentence over false, rarely observed and ungrammatical sentences.

Two different criteria for evaluation:

1) Extrinsic
2) Intrinsic.

Extrinsic Evaluation :-

It evaluates the language model when solving a specific task.

for e.g. speech recognition accuracy, Machine translation accuracy, spelling correction accuracy - compare 2 (or more) models, and check which works best.

Disadvantage :-

↳ Expensive
↳ Time consuming.

Intrinsic Evaluation :-

The language model is best when it predicts an unseen test set.

It uses perplexity:

Definition of perplexity :-

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

→ Lower the value of perplexity : Better model

→ More value of perplexity : Confused for prediction

$$PP(W) = P(W_1, W_2 \dots W_N)^{\frac{1}{N}}$$

$$PP(W) = \left[\prod_i \frac{1}{P(W_i | W_1, W_2 \dots W_{i-1})} \right]^{1/N}$$

WSJ corpus :-

Contains 38 million words : Training
1.5 million words : Test

N-gram order	unigram	Bigram	Trigram
Perplexity Measure	962	170	109.

Perplexity of tri-gram is very less than this it is the best model for predicting the next word.

Ex:-

Perplexity for Bi-gram <s> I like college </s>

$= P(I|<s>) \times P(like|I) \times P(college|like) \times P(</s>|college)$

So N=4. $= 3/7 \times 3/6 \times 3/5 \times 3/3 = 0.13.$

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

Note :- This is a small corpus, so we can't take decision. Since both values are same. for large corpus - the values vary.

Perplexity of Trigram <s> I like college </s>

$P(W) = P(like|<s> I) \times P(college|I\ like) * P(</s>|like\ college)$

So N=3. $= 1/3 \times 2/3 \times 3/3 = 0.22$

$PP(W) = (1/0.22)^{1/3} = 1.66$