NGram Language Model
Log of Probabilities
Laplace Smoothing
Perplexity

N-gram Model

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.

Conditional Probability: $P(B \mid A) = \dfrac{P(A,B)}{P(A)}$ $\qquad P(A,B) = P(A)P(B \mid A)$

More variables: $P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$

Chain Rule:

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

P("about five minutes from")= P(about) × P(five |about) × P(minutes | about five) × P(from | about five minutes)

Probability of words in sentences:

$$P(w_1, w_2, \ldots w_n) = \prod_i P(w_i \mid w_1, w_2, w_3 \ldots, w_{i-1})$$

Unigram(1-gram): **No history is used.**

Bi-gram(2-gram): **One word history**

Tri-gram(3-gram): **Two words history**

Four-gram(4-gram): **Three words history**

Five-gram(5-gram): **Four words history**

Generally in practical applications, Bi-gram(previous one word), Tri-gram(previous two word, Four-gram (previous three word) are used.

Unigram(1-gram): No history is used.

"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.

Unigram will predict dinner.

"about five minutes from dinner"

Bi-gram(2-gram): **One word history**

$$P(w_1, w_2) = \prod_{i=2} P(w_2 \mid w_1)$$

$$P(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

"about five minutes from….."

Assumption: Next word may be college, class

$$P(\text{college} \mid \text{about five minutes from}) = \frac{\text{count(about five minutes from college)}}{\text{count(about five minutes from)}}$$

$$P(\text{class} \mid \text{about five minutes from}) = \frac{\text{count(about five minutes from class)}}{\text{count(about five minutes from)}}$$

"about five minutes from....."

$$\text{Count(about five minutes from)} = P(\text{about} \mid \langle S \rangle) \times P(\text{five} \mid \text{about}) \times P(\text{minutes} \mid \text{five}) \times P(\text{from} \mid \text{minutes})$$

$$\text{Count(about five minutes from } \textbf{college}) = P(\text{about} \mid \langle S \rangle) \times P(\text{five} \mid \text{about}) \times P(\text{minutes} \mid \text{five}) \times P(\text{from} \mid \text{minutes}) \times P(\text{college} \mid \text{from})$$

$$\text{Count(about five minutes from } \textbf{class}) = P(\text{about} \mid \langle S \rangle) \times P(\text{five} \mid \text{about}) \times P(\text{minutes} \mid \text{five}) \times P(\text{from} \mid \text{minutes}) \times P(\text{class} \mid \text{from})$$

$$P(\text{college} \mid \text{about five minutes from}) = \frac{\text{count(about five minutes from college)}}{\text{count(about five minutes from)}}$$

$$= P(\text{college} \mid \text{from})$$

$$P(\text{class} \mid \text{about five minutes from}) = \frac{\text{count(about five minutes from class)}}{\text{count(about five minutes from)}}$$

$$= P(\text{class} \mid \text{from})$$

Tri-gram(2-gram): **Two words history**

$$P(w_1, w_2, w_3) = \prod_{i=3} P(w_3 \mid w_1, w_2) \qquad P(w_i \mid w_{i-1}, w_{i-2}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})}$$

Count(about five minutes from)= P(five | <S>, about) × P(minutes | about, five) × P(from | five , minutes)

Count(about five minutes from **college**)= P(five | <S>, about) × P(minutes | about, five) × P(from | five , minutes) × P(college | minutes from)

Count(about five minutes from **class**)= P(five | <S>, about) × P(minutes | about, five) × P(from | five , minutes) × P(class | minutes from)

$$P(\text{college} \mid \text{about five minutes from}) = \frac{count(\text{about five minutes from college})}{count(\text{about five minutes from})}$$

=P(college | minutes from)

$$P(\text{class} \mid \text{about five minutes from}) = \frac{count(\text{about five minutes from class})}{count(\text{about five minutes from})}$$

=P(class | minutes from)

$$P(\text{college} \mid \text{about five minutes from}) = \frac{\text{count(about five minutes from college)}}{\text{count(about five minutes from)}}$$

=P(college | five minutes from)

$$P(\text{class} \mid \text{about five minutes from}) = \frac{\text{count(about five minutes from class)}}{\text{count(about five minutes from)}}$$

=P(college | five minutes from)

As no. of previous state (history) increases, it is very difficult to match that set of words in corpus.

Probabilities of **larger collection of word is minimum.** To overcome this problem, Bi-gram model is used

Exercise 1: Estimating Bi-gram probabilities

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

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

1) <S> Do ?

<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>

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

Next word	Probability Next Word= $\frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$
P(</S> |do)	0/4
P(<I> | do)	**2/4**
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

2) <S> I like Henry ?

<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>

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}=Henry

Next word	Probability 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

</S> is more probable

3) <S> Do I like ?

Use Tri-gram

P<I like>=3

<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>

Next word prediction probability

W_{i-2}=I and W_{i-1}=like

Next word	Probability Next Word= $\frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})}$
P(</S> |I like)	0/3
P(<I> | I like)	0/3
P(<am>| I like)	0/3
P(<Henry>| I like)	1/3
P(<like | I like)	0/3
P(<college | I like)	2/3
P(do | I like)	0/3

College is probable

4) <S> Do I like college ?

Use Four-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>

Next word prediction probability

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

Next word	Probability Next Word= $\frac{\text{count}(w_{i-3}, w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-3}, w_{i-2}, w_{i-1})}$
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

Which of the following sentence is better. i.e. Gets a higher probability with 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>

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

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

<S> like college </S>=?

=P(I | <S>) × P(like | I) × P(college | like) × P(</S> | 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 | <S>) × P(I | do) × P(like | I) × P(Henry | like) × P(</S> | Henry)

$= 3/7 \times 2/4 \times 3/6 \times 2/5 \times 3/5 = 9/350 = 0.0257$

ANS: First statement is more probable

Which of the following sentence is better. i.e. Gets a higher probability with Bi-gram model.

<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>

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

First statement is more probable

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

=P(I| <S>) × P(like | I) × P(college| like) × P(</S> | college)

=3/7 × 3/6 × 3/5 ×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>) × P(I | do) × P(like | I) × P(Henry | like) × P(</S> | Henry)

=3/7 × 2/4 × 3/6 ×2/5 ×3/5 = 9/350=**0.0257**

=log(3/7)+log(2/4) +log(3/6) + log(2/5) + log(3/5)= **-3.6607**

<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>

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

Second statement is more probable

1. <S> like college </S>

=P(like | <S>) × P(college| like) × P(</S> | college)

=0/7 × 3/5 ×3/3 =**0**

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

=P(do | <S>) × P(I | do) × P(like | I) × P(Henry | like) × P(</S> | Henry)

=3/7 × 2/4 × 3/6 ×2/5 ×3/5 = 9/350=**0.0257**

Laplace Smoothing

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>

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

Features : <S>, </S>, I, am, Henry, like, college, Do

Total number of unique words : 8

But we exclude <S> as it is not used in bigram.

So, total unique word is 7.

Give the following bi-gram probabilities estimated by Laplace model.

Add one Smoothing

1. <S> like college </S>

=P(like | <S>) × P(college | like) × P(</S> | college)

=(0+1)/(7+7) × (3+1)/(5+7) ×(3+1)/(3+7)

=1/14 × 4/12 × 4/10

=**0.0095**

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

=P(do | <S>) × P(I | do) × P(like | I) × P(Henry | like) × P(</S> | Henry)

=(3+1)/(7+7) × (2+1)/(4+7) × (3+1)/(6+7) ×(2+1)/(5+7) ×(3+1)/(5+7)

= 4/14 × 3/11 × 4/13 × 3/12 × 4/12

=**0.0020**

We are given the following corpus, modified from the one in the chapter:

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>
```

Using a bigram language model with add-one smoothing, what is P(Sam | am)? Include <s> and </s> in your counts just like any other token.

Word	Frequency
I	4
am	3
do	1
not	1
like	1
green	1
eggs	1
Sam	4
and	1

$$P(Sam|am) = \frac{2}{3} \text{ (Bigram Model)}$$

$$P(Sam|am) = \frac{3}{14} \text{ (Bigram Model with add-one smoothing)}$$

Language Model Evaluation

- LM 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
  - ✓Extrinsic
  - ✓Intrinsic
- Extrinsic Evaluation
- It evaluates the language model when solving a specific task.
- Ex: Speech recognition accuracy, Machine translation accuracy, spelling correction accuracy
- Compare 2 or more models and check which one is better.
- Disadvantages
  - ✓Expensive and Time consuming

Language Model Evaluation

- Intrinsic Evaluation:
- The language model is best when it predicts an unseen test set.
- 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**

Perplexity (Intrinsic Evaluation model)

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

Definition of Perplexity:

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

$$PP(w) = P(w_1,w_2,w_3,w_4,\ldots,\ldots,wN)^{\frac{1}{N}}$$

$$PP(w) = \left(\prod_i \frac{1}{P(w_i \mid w_1, w_2, \ldots\ldots, w_{i-1})}\right)^{\frac{1}{N}} \qquad PP(w) = \left(\prod_i \frac{1}{P(w_i \mid w_{i-1})}\right)^{\frac{1}{N}}$$

- Lower the value of perplexity : **Better model**
- More value of perplexity : **Confused for prediction**

Perplexity for Bigram <S> I like college </S>

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

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

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

Perplexity for Trigram <S> I like college </S>

$P(w)=P(like \mid \langle S\rangle\, I) \times P(college \mid I\ like) \times P(\langle /S\rangle \mid like\ college)$

$P(w) = 1/3 \times 2/3 \times 3/3 = 2/9 =$ **0.22**

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

Advantages:

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

Disadvantages:

- Underflow due to multiplication of probabilities
- **Solution:** Use log. Add probabilities.
- Zero probability problem
- **Solution:** Use Laplace smoothing

References:

Daniel Jurafsky, James H. Martin —Speech and Language Processing‖ Second Edition, Prentice Hall, 2008.

Christopher D.Manning and Hinrich Schutze, — Foundations of Statistical Natural Language Processing , MIT Press, 1999.

THANK YOU