

# Copyright Notice

These slides are distributed under the Creative Commons License.

[DeepLearning.AI](#) makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite [DeepLearning.AI](#) as the source of the slides.

For the rest of the details of the license, see <https://creativecommons.org/licenses/by-sa/2.0/legalcode>

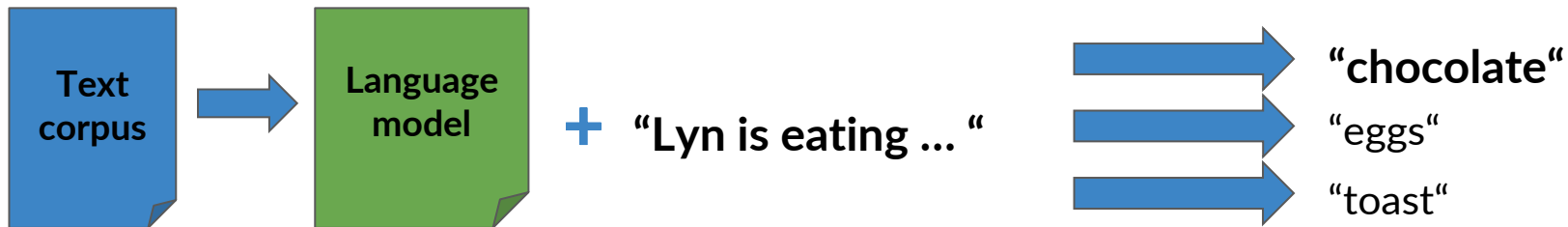


deeplearning.ai

# N-Grams: Overview

# What you'll be able to do!

- Create **language model (LM)** from text corpus to
  - Estimate probability of word sequences
  - Estimate probability of a word following a sequence of words
- Apply this concept to **autocomplete a sentence** with most likely suggestions



# Other Applications

## Speech recognition



$P(\text{I saw a van}) > P(\text{eyes awe of an})$

## Spelling correction



“He entered the **ship** to buy some groceries” - “ship” a dictionary word

- $P(\text{entered the shop to buy}) > P(\text{entered the ship to buy})$

## Augmentative communication

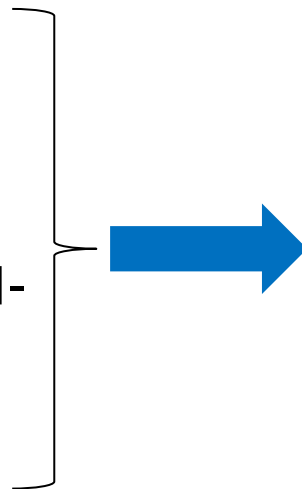


Predict most likely word from menu for people unable to physically talk or sign.

(Newell et al., 1998)

# Learning objectives

- Process text corpus to N-gram language model
- Out of vocabulary words
- Smoothing for previously unseen N-grams
- Language model evaluation



Sentence  
auto-complete



deeplearning.ai

# N-grams and Probabilities

# Outline

- What are N-grams?
- N-grams and conditional probability from corpus

# N-gram

An N-gram is a sequence of N words

Corpus: I am happy because I am learning

Unigrams: { I , am , happy , because , learning }

Bigrams: { I am , am happy , happy because ... }

✗ I happy

Trigrams: { I am happy , am happy because, ... }



# Sequence notation

Corpus: This is great ... teacher drinks tea.  $m = 500$

$w_1 \ w_2 \ w_3$   $w_{498} \ w_{499} \ w_{500}$

$$w_1^m = w_1 \ w_2 \ \dots \ w_m$$

$$w_1^3 = w_1 \ w_2 \ w_3$$

$$w_{m-2}^m = w_{m-2} \ w_{m-1} \ w_m$$

# Unigram probability

Corpus: I am happy because I am learning

Size of corpus  $m = 7$

$$P(I) = \frac{2}{7}$$

$$P(happy) = \frac{1}{7}$$

Probability of unigram:

$$P(w) = \frac{C(w)}{m}$$

# Bigram probability

Corpus: I am happy because I am learning

$$P(am|I) = \frac{C(I \text{ } am)}{C(I)} = \frac{2}{2} = 1$$

$$P(happy|I) = \frac{C(I \text{ } happy)}{C(I)} = \frac{0}{2} = 0 \quad \times \text{ I happy}$$

$$P(learning|am) = \frac{C(am \text{ } learning)}{C(am)} = \frac{1}{2}$$

Probability of a bigram: 
$$P(y|x) = \frac{C(x \text{ } y)}{\sum_w C(x \text{ } w)} = \frac{C(x \text{ } y)}{C(x)}$$

# Trigram Probability

Corpus: I am happy because I am learning

$$P(happy|I\ am) = \frac{C(I\ am\ happy)}{C(I\ am)} = \frac{1}{2}$$

Probability of a trigram:  $P(w_3|w_1^2) = \frac{C(w_1^2\ w_3)}{C(w_1^2)}$

$$C(w_1^2\ w_3) = C(w_1\ w_2\ w_3) = C(w_1^3)$$

# N-gram probability

Probability of N-gram:  $P(w_N | w_1^{N-1}) = \frac{C(w_1^{N-1} w_N)}{C(w_1^{N-1})}$

$$C(w_1^{N-1} w_N) = C(w_1^N)$$

## Quiz

**Objective:** Apply n-gram probability calculation on sample corpus and 3-gram.

### Question:

**Corpus:** “In every place of great resort the monster was the fashion. They sang of it in the cafes, ridiculed it in the papers, and represented it on the stage. ” (Jules Verne, Twenty Thousand Leagues under the Sea)

In the context of our corpus, what is the probability of word “papers” following the phrase “it in the”.

**Type:** Multiple Choice, single answer

### Options and solution:

1.  $P(\text{papers}|\text{it in the}) = 0$

2.  $P(\text{papers}|\text{it in the}) = 1$

3.  $P(\text{papers}|\text{it in the}) = 2/3$

4.  $P(\text{papers}|\text{it in the}) = \frac{1}{2}$

$= C(\text{it in the papers})/C(\text{it in the})$



deeplearning.ai

# Sequence Probabilities

---

# Outline

- Sequence probability
- Sequence probability shortcomings
- Approximation by N-gram probabilities



# Probability of a sequence

- Given a sentence, what is its probability?

$$P(\textit{the teacher drinks tea}) = ?$$

- Conditional probability and chain rule reminder

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

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

# Probability of a sequence

$$P(\textit{the teacher drinks tea}) =$$

$$P(\textit{the})P(\textit{teacher}|\textit{the})P(\textit{drinks}|\textit{the teacher}) \\ P(\textit{tea}|\textit{the teacher drinks})$$

# Sentence not in corpus

- Problem: Corpus almost never contains the exact sentence we're interested in or even its longer subsequences!

Input: the teacher drinks tea

$$P(\text{the teacher drinks tea}) = P(\text{the})P(\text{teacher}|\text{the})P(\text{drinks}|\text{the teacher})P(\text{tea}|\text{the teacher drinks})$$

$$P(\text{tea}|\text{the teacher drinks}) = \frac{C(\text{the teacher drinks tea})}{C(\text{the teacher drinks})}$$

← Both  
← likely  
0

# Approximation of sequence probability

the teacher drinks tea

$$P(\text{tea}|\text{the teacher drinks}) \approx P(\text{tea}|\text{drinks})$$

$$\begin{aligned} &P(\text{teacher}|\text{the}) \\ &P(\text{drinks}|\text{teacher}) \\ &P(\text{tea}|\text{drinks}) \end{aligned}$$

$$P(\text{the teacher drinks tea}) = P(\text{the})P(\text{teacher}|\text{the})P(\text{drinks}|\text{the teacher})P(\text{tea}|\text{the teacher drinks})$$



$$P(\text{the})P(\text{teacher}|\text{the})P(\text{drinks}|\text{teacher})P(\text{tea}|\text{drinks})$$

# Approximation of sequence probability

- Markov assumption: only last N words matter

- Bigram  $P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-1})$

- N-gram  $P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-N+1}^{n-1})$

- Entire sentence modeled with bigram  $P(w_1^n) \approx \prod_{i=1}^n P(w_i | w_{i-1})$

$$P(w_1^n) \approx \boxed{P(w_1)} P(w_2 | w_1) \dots P(w_n | w_{n-1})$$

## Quiz

**Objective:** Apply sequence probability approximation with bigrams.

### Question:

Given these conditional probabilities

$P(\text{Mary})=0.1$ ;  $P(\text{likes})=0.2$ ;  $P(\text{cats})=0.3$   
 $P(\text{Mary}|\text{likes})=0.2$ ;  $P(\text{likes}|\text{Mary})=0.3$ ;  $P(\text{cats}|\text{likes})=0.1$ ;  $P(\text{likes}|\text{cats})=0.4$

Approximate the probability of the following sentence with bigrams: "Mary likes cats"

**Type:** Multiple Choice, single answer

### Options and solution:

- |    |                                     |    |                                     |
|----|-------------------------------------|----|-------------------------------------|
| 1. | $P(\text{Mary likes cats}) = 0$     | 2. | $P(\text{Mary likes cats}) = 1$     |
| 3. | $P(\text{Mary likes cats}) = 0.003$ | 4. | $P(\text{Mary likes cats}) = 0.008$ |



deeplearning.ai

# Starting and Ending Sentences

# Outline

- Start of sentence symbols `<s>`
- End of sentence symbol `</s>`



# Start of sentence token <s>

the teacher drinks tea

$$P(\text{the teacher drinks tea}) \approx P(\text{the})P(\text{teacher}|\text{the})P(\text{drinks}|\text{teacher})P(\text{tea}|\text{drinks})$$



<s> the teacher drinks tea

$$P(\text{<s> the teacher drinks tea}) \approx P(\text{the}|\text{<s>})P(\text{teacher}|\text{the})P(\text{drinks}|\text{teacher})P(\text{tea}|\text{drinks})$$

# Start of sentence token <s> for N-grams

- Trigram:

$$P(\text{the teacher drinks tea}) \approx$$

$$P(\text{the})P(\text{teacher}|\text{the})P(\text{drinks}|\text{the teacher})P(\text{tea}|\text{teacher drinks})$$

the teacher drinks tea => <s> <s> the teacher drinks tea

$$P(w_1^n) \approx P(w_1|\text{<s> <s>})P(w_2|\text{<s> } w_1) \dots P(w_n|w_{n-2} w_{n-1})$$

- N-gram model: add N-1 start tokens <s>

## End of sentence token </s> - motivation

$$P(y|x) = \frac{C(x \ y)}{\sum_w C(x \ w)} = \frac{C(x \ y)}{C(x)}$$

Corpus:

<s> Lyn drinks chocolate  
<s> John drinks

$$\sum_w C(drinks \ w) = 1$$

$$C(drinks) = 2$$

# End of sentence token </s> - motivation

Corpus

<s> yes no

<s> yes yes

<s> no no

Sentences of length 2:

<s> yes yes

<s> yes no

<s> no no

<s> no yes

$$P(< s > \text{ yes yes}) =$$

$$P(\text{yes} \mid < s >) \times P(\text{yes} \mid \text{yes}) =$$

$$\frac{C(< s > \text{ yes})}{\sum_w C(< s > w)} \times \frac{C(\text{yes yes})}{\sum_w C(\text{yes } w)} =$$

$$\frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$$

# End of sentence token $\langle /s \rangle$ - motivation

Corpus

$\langle s \rangle$  yes no

$\langle s \rangle$  yes yes

$\langle s \rangle$  no no

Sentences of length 2:

$\langle s \rangle$  yes yes

$\langle s \rangle$  yes no

$\langle s \rangle$  no no

$\langle s \rangle$  no yes

$$P(\langle s \rangle \text{ yes yes}) = \frac{1}{3}$$

$$P(\langle s \rangle \text{ yes no}) = \frac{1}{3}$$

$$P(\langle s \rangle \text{ no no}) = \frac{1}{3}$$

$$P(\langle s \rangle \text{ no yes}) = 0$$

$$\sum_{2 \text{ word}} P(\dots) = 1$$

# End of sentence token $\langle /s \rangle$ - motivation

Corpus

$\langle s \rangle$  yes no

$\langle s \rangle$  yes yes

$\langle s \rangle$  no no

Sentences of length 3:

$\langle s \rangle$  yes yes yes

$\langle s \rangle$  yes yes no

...

$\langle s \rangle$  no no no

$$P(\langle s \rangle \text{ yes yes yes}) = \dots$$

$$P(\langle s \rangle \text{ yes yes no}) = \dots$$

$$\dots = \dots$$

$$P(\langle s \rangle \text{ no no no}) = \dots$$

$$\sum_{\text{3 word}} P(\dots) = 1$$

# End of sentence token </s> - motivation

## Corpus

<s> yes no

<s> yes yes

<s> no no

$$\sum_{\text{2 word}} P(\dots) + \sum_{\text{3 word}} P(\dots) + \dots = 1$$

# End of sentence token $\text{</s>}$ - solution

- Bigram

$\text{<s> the teacher drinks tea} \Rightarrow \text{<s> the teacher drinks tea </s>}$

$$P(\text{the}|\text{<s>})P(\text{teacher}|\text{the})P(\text{drinks}|\text{teacher})P(\text{tea}|\text{drinks})P(\text{</s>}|\text{tea})$$

Corpus:

$\text{<s> Lyn drinks chocolate </s>}$

$\text{<s> John drinks </s>}$

$$\sum_w C(\text{drinks } w) = 2$$
$$C(\text{drinks}) = 2$$



# End of sentence token </s> for N-grams

- N-gram => just one </s>

E.g. Trigram:

the teacher drinks tea => <s> <s> the teacher drinks tea </s>

# Example - bigram

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

$$P(\text{John}|\text{<s>}) = \frac{1}{3}$$

$$P(\text{chocolate}|\text{eats}) = \frac{1}{2}$$

$$P(\text{sentence}) = \boxed{\frac{2}{3}} * \boxed{\frac{1}{2}} * \boxed{\frac{1}{2}} * \boxed{\frac{2}{2}} = \frac{1}{6}$$

$$P(\text{</s>}|\text{tea}) = \frac{1}{1}$$

$$P(\text{Lyn}|\text{<s>}) = ? = \frac{2}{3}$$

## Quiz

**Objective:** Apply sequence probability approximation with bigrams after adding start and end word.

### Question:

Given these conditional probabilities

$P(\text{Mary})=0.1$ ;       $P(\text{likes})=0.2$ ;       $P(\text{cats})=0.3$

$P(\text{Mary}|\text{<s>})=0.2$ ;       $P(\text{</s>}|\text{cats})=0.6$

$P(\text{likes}|\text{Mary})=0.3$ ;       $P(\text{cats}|\text{likes})=0.1$

Approximate the probability of the following sentence with bigrams: "<s> Mary likes cats </s>"

**Type:** Multiple Choice, single answer

### Options and solution:

1.       $P(\text{<s> Mary likes cats </s>}) = 0$

2.       $P(\text{<s> Mary likes cats </s>}) = 0.0036$

3.       $P(\text{<s> Mary likes cats </s>}) = 0.003$

4.       $P(\text{<s> Mary likes cats </s>}) = 1$



deeplearning.ai

# The N-gram Language Model

---

# Outline

- Count matrix
- Probability matrix
- Language model
- Log probability to avoid underflow
- Generative language model

# Count matrix

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

- Rows: unique corpus (N-1)-grams
- Columns: unique corpus words

Corpus: <s>I study I learn</s>

- Bigram count matrix

“study I” bigram

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	1	1
study	0	0	1	0	0
learn	0	1	0	0	0

# Probability matrix

- Divide each cell by its row sum

Corpus: <s>I study I learn</s>

Count matrix (bigram)

	<s>	</s>	I	study	learn	sum
<s>	0	0	1	0	0	1
</s>	0	0	0	0	0	0
I	0	0	0	1	1	2
study	0	0	1	0	0	1
learn	0	1	0	0	0	1



Probability matrix

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	0.5	0.5
study	0	0	1	0	0
learn	0	1	0	0	0

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

$$\text{sum}(\text{row}) = \sum_{w \in V} C(w_{n-N+1}^{n-1}, w) = C(w_{n-N+1}^{n-1})$$

# Language model

- probability matrix => language model
  - Sentence probability
  - Next word prediction

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	0.5	0.5
study	0	0	1	0	0
learn	0	1	0	0	0

Sentence probability:

<s> I learn </s>

$$\begin{aligned} P(\text{sentence}) &= \\ P(I|\text{<s>})P(\text{learn}|I)P(\text{</s>}|\text{learn}) &= \\ 1 \times 0.5 \times 1 &= \\ 0.5 \end{aligned}$$



# Log probability

$$P(w_1^n) \approx \prod_{i=1}^n P(w_i | w_{i-1})$$

- All probabilities in calculation  $\leq 1$  and multiplying them brings risk of underflow

- Logarithm properties reminder  $\log(a * b) = \log(a) + \log(b)$

- Use log of the probabilities in Probability matrix and calculations

$$\log(P(w_1^n)) \approx \sum_{i=1}^n \log(P(w_i | w_{i-1}))$$

- Converts back from log  
 $P(w_1^n) = \exp(\log(P(w_1^n)))$

# Generative Language model

Corpus:

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

1. (<s>, Lyn) or (<s>, John)?
2. (Lyn,eats) or (Lyn,drinks) ?
3. (drinks,tea) or (drinks,chocolate)?
4. (tea,</s>) - always

Algorithm:

1. Choose sentence start
2. Choose next bigram starting with previous word
3. Continue until </s> is picked

## Quiz

**Objective:** Apply sum when calculating log probability instead of product.

### Question:

Given the logarithm of these conditional probabilities:

$$\log(P(\text{Mary}|\langle s \rangle)) = -2; \quad \log(P(\langle /s \rangle | \text{cats})) = -1$$

$$\log(P(\text{likes} | \text{Mary})) = -10; \quad \log(P(\text{cats} | \text{likes})) = -100$$

Approximate the log probability of the following sentence with bigrams : “ $\langle s \rangle$  Mary likes cats  $\langle /s \rangle$ ”

**Type:** Multiple Choice, single answer

### Options and solution:

1.  $\log(P(\langle s \rangle \text{ Mary likes cats } \langle /s \rangle)) = -113$

2.  $\log(P(\langle s \rangle \text{ Mary likes cats } \langle /s \rangle)) = 2000$

3.  $\log(P(\langle s \rangle \text{ Mary likes cats } \langle /s \rangle)) = 113$

4.  $\log(P(\langle s \rangle \text{ Mary likes cats } \langle /s \rangle)) = -112$



deeplearning.ai

# Language Model Evaluation

---

# Outline

- Train/Validation/Test split
- Perplexity

# Test data

- Split corpus to Train/Validation/Test

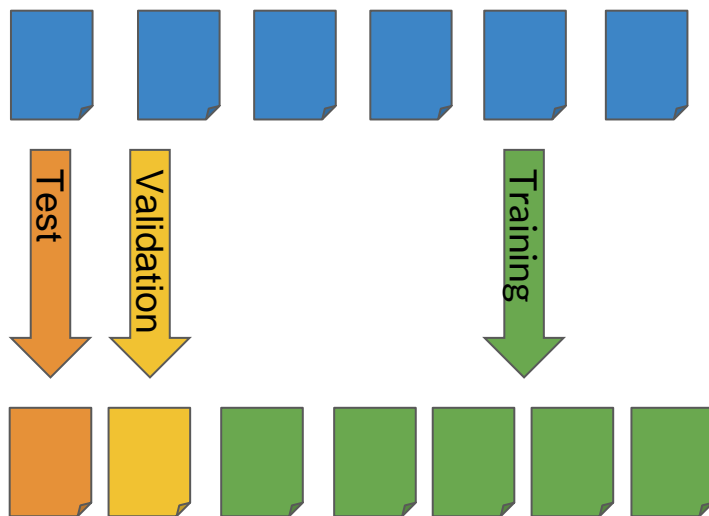
 Evaluate on Training dataset

- For smaller corpora
  - 80% Train
  - 10% Validation
  - 10% Test

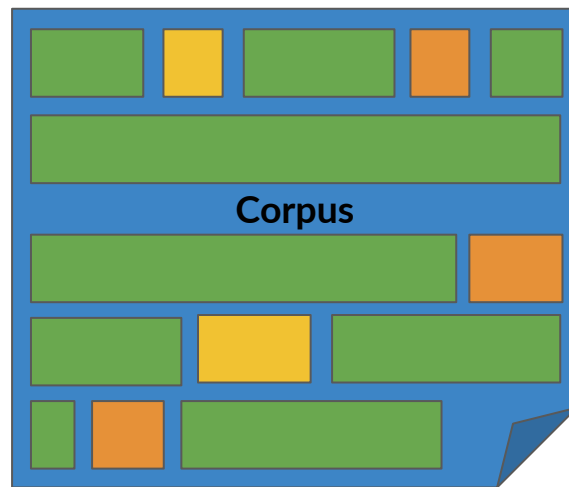
- For large corpora (typical for text)
  - 98% Train
  - 1% Validation
  - 1% Test

# Test data - split method

- Continuous text



- Random short sequences



# Perplexity



$$PP(W) = P(s_1, s_2, \dots, s_m)^{-\frac{1}{m}}$$

$W \rightarrow$  test set containing  $m$  sentences  $s$

$s_i \rightarrow$   $i$ -th sentence in the test set, each ending with  $\langle /s \rangle$

$m \rightarrow$  number of all words in entire test set  $W$  including  $\langle /s \rangle$  but not including  $\langle s \rangle$



# Perplexity

E.g.  $m=100$

$$P(W) = 0.9 \Rightarrow PP(W) = 0.9^{-\frac{1}{100}} = 1.00105416$$

$$P(W) = 10^{-250} \Rightarrow PP(W) = (10^{-250})^{-\frac{1}{100}} \approx 316$$

- Smaller perplexity = better model
- Character level models  $PP <$  word-based models  $PP$

# Perplexity for bigram models

$$PP(W) = \sqrt[m]{\prod_{i=1}^m \prod_{j=1}^{|s_i|} \frac{1}{P(w_j^{(i)} | w_{j-1}^{(i)})}}$$

$w_j^{(i)}$  → j-th word in i-th sentence

- concatenate all sentences in W

$$PP(W) = \sqrt[m]{\prod_{i=1}^m \frac{1}{P(w_i | w_{i-1})}}$$

$w_i$  → i-th word in test set

## Log perplexity

$$PP(W) = \sqrt[m]{\prod_{i=1}^m \frac{1}{P(w_i|w_{i-1})}}$$



$$\log PP(W) = -\frac{1}{m} \sum_{i=1}^m \log_2(P(w_i|w_{i-1}))$$

# Examples

Training 38 million words, test 1.5 million words, WSJ corpus  
Perplexity Unigram: 962 Bigram: 170 Trigram: 109

## Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

## Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

## Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

[Figure from *Speech and Language Processing* by Dan Jurafsky et. al]

## Quiz

**Objective:** Calculate log perplexity from log probabilities using sum and correct normalization coefficient (not including  $\langle s \rangle$ ).

### Question:

Given the logarithm of these conditional probabilities:

$$\log(P(\text{Mary}|\langle s \rangle)) = -2; \quad \log(P(\langle /s \rangle|\text{cats})) = -1$$

$$\log(P(\text{likes}|\text{Mary})) = -10; \quad \log(P(\text{cats}|\text{likes})) = -100$$

Assuming our test set is  $W = \langle s \rangle \text{ Mary likes cats } \langle /s \rangle$ , what is the model's perplexity.

**Type:** Multiple Choice, single answer

### Options and solution:

1.  $\log PP(W) = -113$

2.  $\log PP(W) = (-1/4)*(-113)$

3.  $\log PP(W) = (-1/5)*(-113)$

4.  $\log PP(W) = (-1/5)*113$



deeplearning.ai

# Out of Vocabulary Words

# Outline

- Unknown words
- Update corpus with <UNK>
- Choosing vocabulary

# Out of vocabulary words

- Closed vs. Open vocabularies
- Unknown word = Out of vocabulary word (OOV)
- special tag <UNK> in corpus and in input



# Using <UNK> in corpus

- Create vocabulary V
- Replace any word in corpus and not in V by <UNK>
- Count the probabilities with <UNK> as with any other word

# Example

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>



Corpus

<s> Lyn drinks chocolate </s>

<s> <UNK> drinks <UNK> </s>

<s> Lyn <UNK> chocolate </s>

Min frequency  $f=2$

Vocabulary

Lyn, drinks, chocolate

Input query

<s> Adam drinks chocolate </s>



<s> <UNK> drinks chocolate </s>

# How to create vocabulary $V$

- Criteria:
  - Min word frequency  $f$
  - Max  $|V|$ , include words by frequency
- Use  $\langle \text{UNK} \rangle$  sparingly
- Perplexity - only compare LMs with the same  $V$

## Quiz

**Objective:** Create corpus vocabulary based on minimum frequency.

### Question:

Given the training corpus and minimum word frequency=2, how would the vocabulary for corpus preprocessed with <UNK> look like?

“<s> I am happy I am learning </s> <s> I am happy I can study </s>”

**Type:** Multiple Choice, single answer

### Options and solution:

1. V = (I,am,happy)

2. V = (I,am,happy,learning,can,study)

3. V = (I,am,happy,I,am)

4. V =  
(I,am,happy,learning,can,study,<UNK>)



deeplearning.ai


# Smoothing

---

# Outline

- Missing N-grams in corpus
- Smoothing
- Backoff and interpolation

# Missing N-grams in training corpus

- Problem: N-grams made of known words still might be missing in the training corpus      “John” , “eats” in corpus       “John eats”
- Their counts cannot be used for probability estimation

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

← Can be 0

←

# Smoothing

- Advanced methods:  
Kneser-Ney smoothing  
Good-Turing smoothing

- Add-one smoothing (Laplacian smoothing)

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}, w_n) + 1}{\sum_{w \in V} (C(w_{n-1}, w) + 1)} = \frac{C(w_{n-1}, w_n) + 1}{C(w_{n-1}) + V}$$

- Add-k smoothing

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}, w_n) + k}{\sum_{w \in V} (C(w_{n-1}, w) + k)} = \frac{C(w_{n-1}, w_n) + k}{C(w_{n-1}) + k * V}$$



# Backoff

- If N-gram missing => use (N-1)-gram, ...
  - Probability discounting e.g. Katz backoff
  - “Stupid” backoff

## Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

$$P(\textit{chocolate} | \textit{John drinks}) = ?$$



$$0.4 \times P(\textit{chocolate} | \textit{drinks})$$

# Interpolation

$$\hat{P}(\textit{chocolate} | \textit{John drinks}) = 0.7 \times P(\textit{chocolate} | \textit{John drinks}) \\ + 0.2 \times P(\textit{chocolate} | \textit{drinks}) + 0.1 \times P(\textit{chocolate})$$

$$\hat{P}(w_n | w_{n-2} \ w_{n-1}) = \lambda_1 \times P(w_n | w_{n-2} \ w_{n-1}) \\ + \lambda_2 \times P(w_n | w_{n-1}) + \lambda_3 \times P(w_n) \qquad \sum_i \lambda_i = 1$$

## Quiz

**Objective:** Apply n-gram probability with add-k smoothing for phrase not present in the corpus.

**Question:**

Corpus: "I am happy I am learning"

In the context of our corpus, what is the estimated probability of word "can" following the word "I" using the bigram model and add-k-smoothing where  $k=3$ .

**Type:** Multiple Choice, single answer

**Options and solution:**

1.  $P(\text{can}|\text{I}) = 0$

2.  $P(\text{can}|\text{I}) = 1$

3.  $P(\text{can}|\text{I}) = 3/(2+3*4)$

4.  $P(\text{can}|\text{I}) = 3/(3*4)$



deeplearning.ai

# Week Summary

# Summary

- N-Grams and probabilities
- Approximate sentence probability from N-Grams
- Build language model from corpus
- Fix missing information
  - Out of vocabulary words with <UNK>
  - Missing N-Gram in corpus with smoothing, backoff and interpolation
- Evaluate language model with perplexity
- Coding assignment!