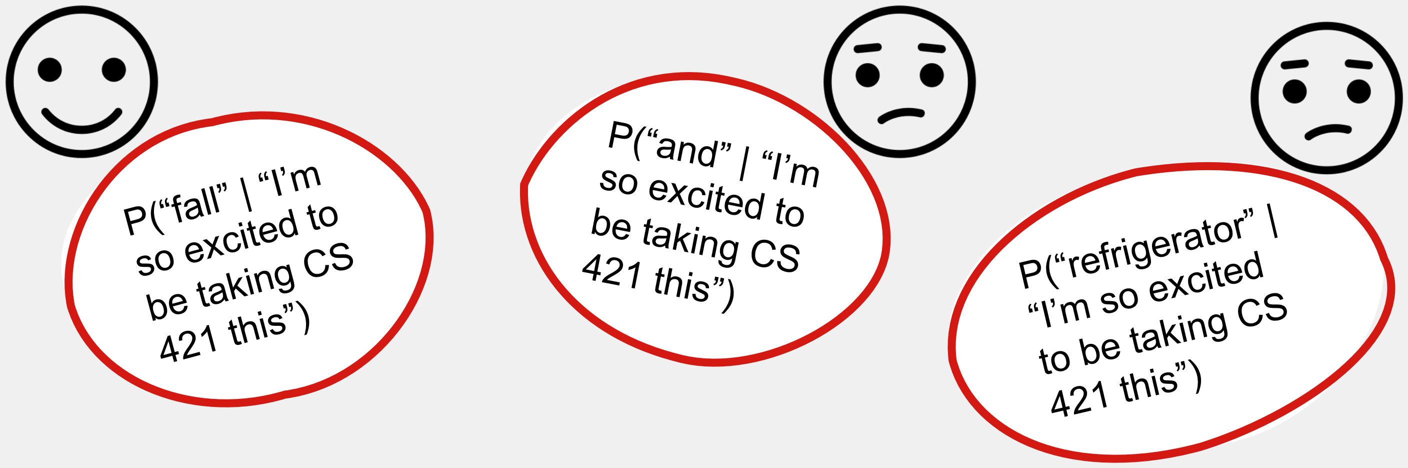
N-Grams and Maximum Likelihood Estimation

Natalie Parde
UIC CS 421

N-Gram Language Models

- Goal: Predict P(word|history)
  - P("spring" | "I'm so excited to be taking CS 421 this")

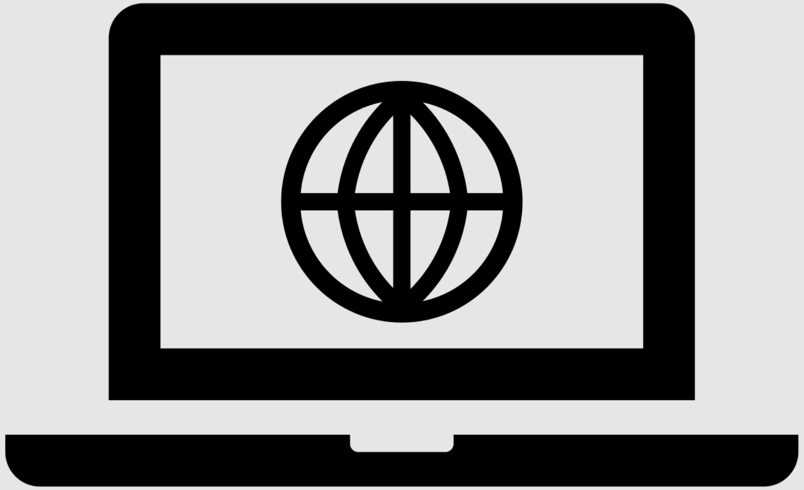


How do we predict these probabilities?

- One method: Estimate it from frequency counts
  - Take a large corpus
  - Count the number of times you see the history
  - Count the number of times the specified word follows the history

P(“spring” | “I’m so excited to be taking CS 521 this”)

= C(“I’m so excited to be taking CS 521 this spring”) / C(“I’m so excited to be taking CS 521 this”)



However, we don’t necessarily want to use our *entire* history.

- What if our history contains uncommon words?
- What if we have limited computing resources?

P(“spring” | “I’m so excited to be taking **Natalie Parde’s** CS 421 this”)

Out of all possible 11-word sequences on the web, how many are “I’m so excited to be taking Natalie Parde’s CS 421 this”?

We need a better way to estimate P(word|history)!

- The solution: Instead of computing the probability of a word given its entire history, **approximate the history using the most recent few words**.
- We do this using fixed-length **n-grams**.

P(“spring” | “taking CS 421 this”)

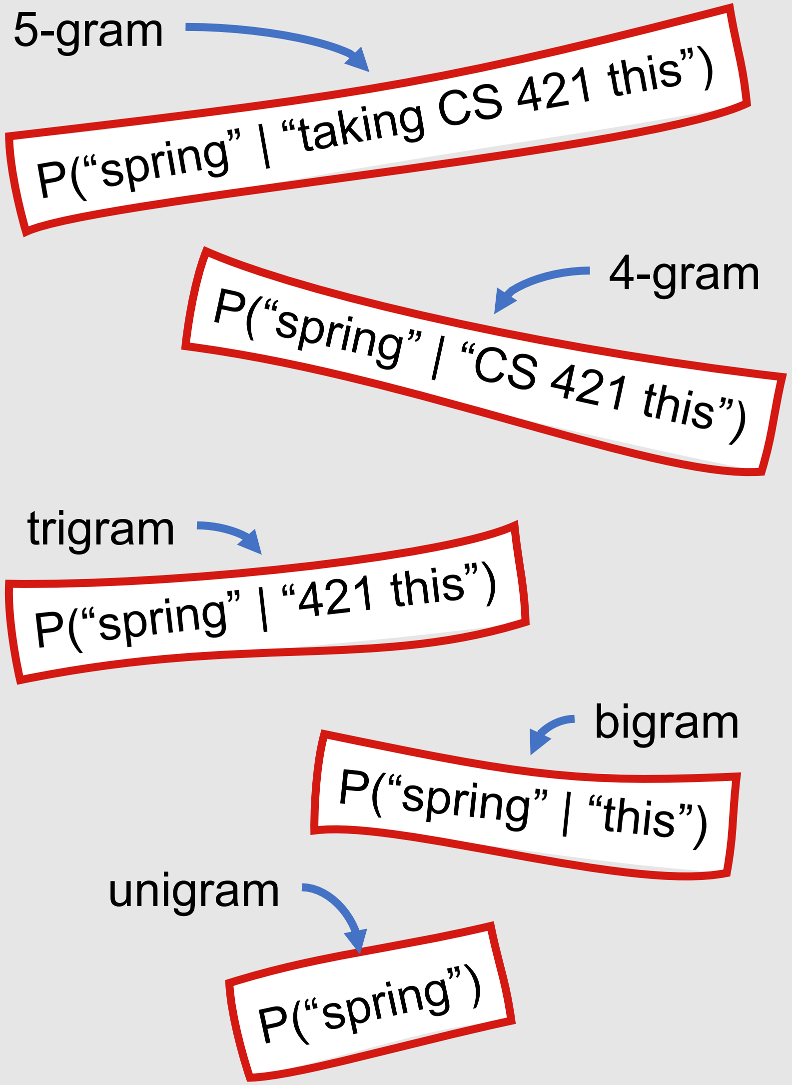
P(“spring” | “CS 421 this”)

P(“spring” | “421 this”)

P(“spring” | “this”)

Special N-Grams

- Most higher-order (n>3) n-grams are simply referred to using the value of *n*
  - 4-gram
  - 5-gram
- However, lower-order n-grams are often referred to using special terms:
  - Unigram (1-gram)
  - Bigram (2-gram)
  - Trigram (3-gram)



N-gram models follow the **Markov assumption.**

- We can predict the probability of some future unit without looking too far into the past
  - **Bigram language model:** Probability of a word depends only on the previous word
  - **Trigram language model:** Probability of a word depends only on the two previous words
  - **N-gram language model:** Probability of a word depends only on the *n*-1 previous words

More formally….

- $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$
- We can then multiply these individual word probabilities together to get the probability of a word sequence
  - $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1})$

P("Summer break is already over?")

P("over?" | "already") * P("already" | "is") * P("is" | "break") * P("break" | "Summer")

To compute n-gram probabilities, maximum likelihood estimation is often used.

- **Maximum Likelihood Estimation (MLE):**
  - Get the requisite n-gram frequency counts from a corpus
  - Normalize them to a 0-1 range
    - $P(w_n \mid w_{n-1})$ = # of occurrences of the bigram $w_{n-1}\ w_n$ / # of occurrences of the unigram w_{n-1}

Example: Maximum Likelihood Estimation

I am cold.

You are cold.

Everyone is cold.

This is Chicago.

Example: Maximum Likelihood Estimation

I am cold. → <s> I am cold. </s>

You are cold. → <s> You are cold. </s>

Everyone is cold. → <s> Everyone is cold. </s>

This is Chicago. → <s> This is Chicago. </s>

Example: Maximum Likelihood Estimation

I am cold. → <s> I am cold. </s>

You are cold. → <s> You are cold. </s>

Everyone is cold. → <s> Everyone is cold. </s>

This is Chicago. → <s> This is Chicago. </s>

Bigram	Frequency
<s> I	1
I am	1
am cold.	1
cold. </s>	3
…	…
is Chicago.	1
Chicago. </s>	1

Example: Maximum Likelihood Estimation

I am cold. → <s> I am cold. </s>

You are cold. → <s> You are cold. </s>

Everyone is cold. → <s> Everyone is cold. </s>

This is Chicago. → <s> This is Chicago. </s>

Bigram	Freq.
<s> I	1
I am	1
am cold.	1
cold. </s>	3
…	…
is Chicago.	1
Chicago. </s>	1

Unigram	Freq.
<s>	4
I	1
am	1
cold.	3
…	…
Chicago.	1
</s>	4

Example: Maximum Likelihood Estimation

I am cold. → <s> I am cold. </s>

You are cold. → <s> You are cold. </s>

Everyone is cold. → <s> Everyone is cold. </s>

This is Chicago. → <s> This is Chicago. </s>

Bigram	Freq.
<s> I	1
I am	1
am cold.	1
cold. </s>	3
…	…
is Chicago.	1
Chicago. </s>	1

Unigram	Freq.
<s>	4
I	1
am	1
cold.	3
…	…
Chicago.	1
</s>	4

P("I" | "<s>") = C("<s> I") / C("<s>") = 1 / 4 = 0.25

Example: Maximum Likelihood Estimation

I am cold. → <s> I am cold. </s>

You are cold. → <s> You are cold. </s>

Everyone is cold. → <s> Everyone is cold. </s>

This is Chicago. → <s> This is Chicago. </s>

Bigram	Freq.
<s> I	1
I am	1
am cold.	1
cold. </s>	3
…	…
is Chicago.	1
Chicago. </s>	1

Unigram	Freq.
<s>	4
I	1
am	1
cold.	3
…	…
Chicago.	1
</s>	4

P("I" | "<s>") = C("<s> I") / C("<s>") = 1 / 4 = 0.25

P("</s>" | "cold.") = C("cold. </s>") / C("cold.") = 3 / 3 = 1.00

Example: Maximum Likelihood Estimation

I am cold. → <s> I am cold. </s>

You are cold. → <s> You are cold. </s>

Everyone is cold. → <s> Everyone is cold. </s>

This is Chicago. → <s> This is Chicago. </s>

Bigram	Freq.
<s> I	1
I am	1
am cold.	1
cold. </s>	3
…	…
is Chicago.	1
Chicago. </s>	1

Unigram	Freq.
<s>	4
I	1
am	1
cold.	3
…	…
Chicago.	1
</s>	4

P("I" | "<s>") = C("<s> I") / C("<s>") = 1 / 4 = 0.25

P("</s>" | "cold.") = C("cold. </s>") / C("cold.") = 3 / 3 = 1.00