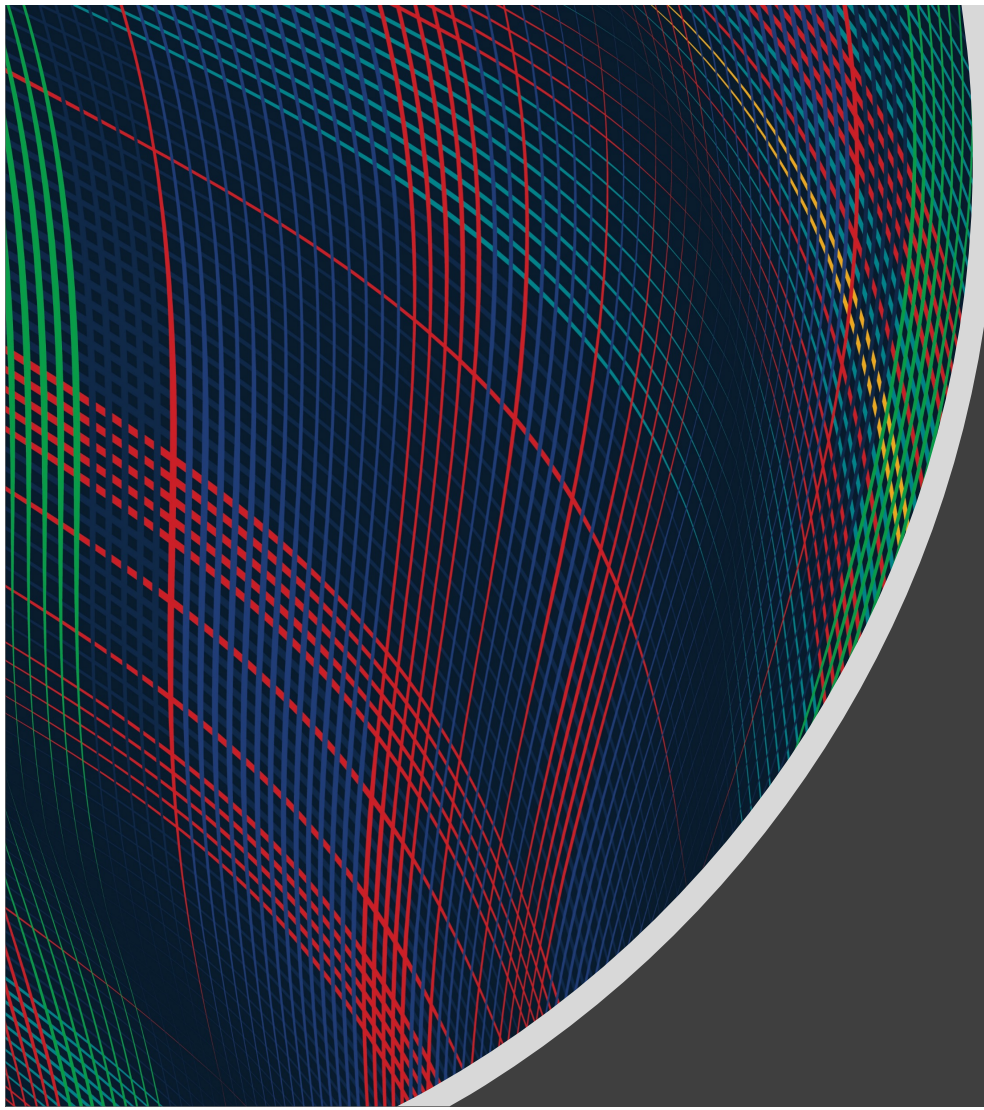


07-280
AI & ML I

Natural Language Processing (NLP): N-gram Language Models

Instructors:
Aviral Kumar& Nihar Shah

Natural Language Processing (NLP)

Quick NLP Intro

Tokenization

N-grams

Language Models (LMs)

LM Sampling

Natural Language Processing (NLP)

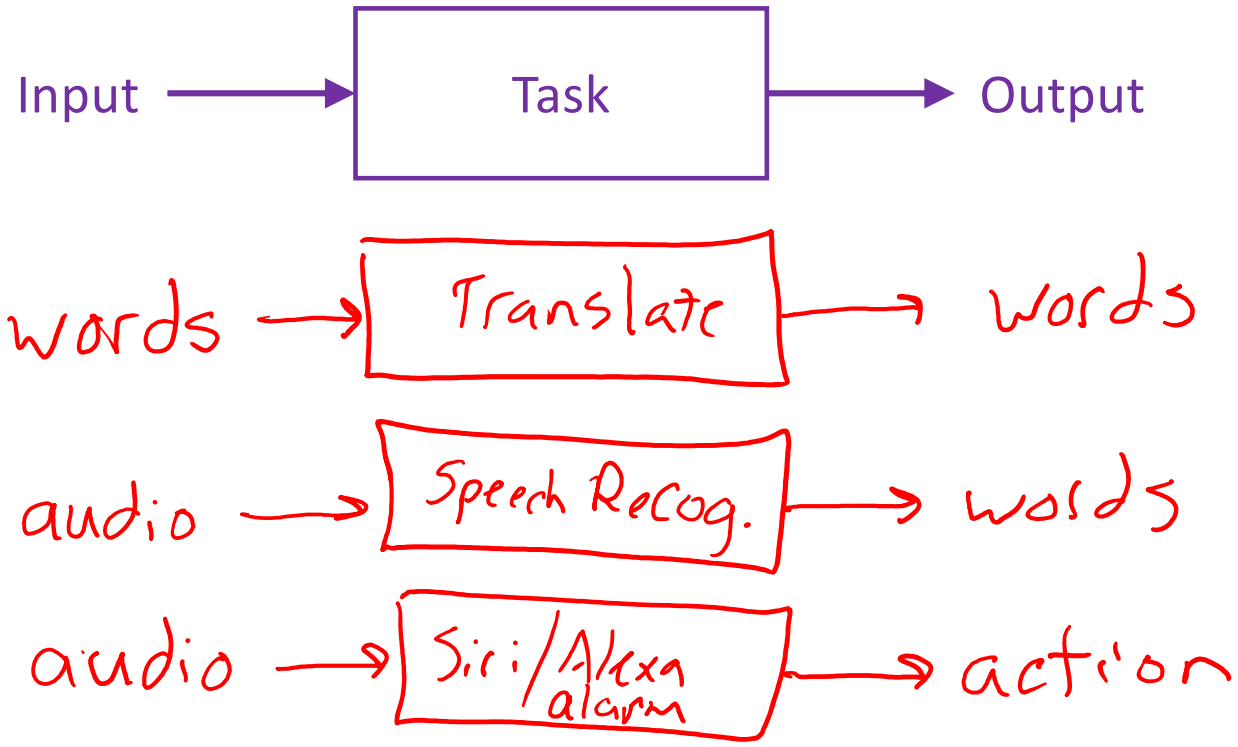
Language is Hard

Emphasis can drastically change meaning

I didn’t eat your dog

NLP Tasks Examples

How many different NLP Input/Output agents can you think of?

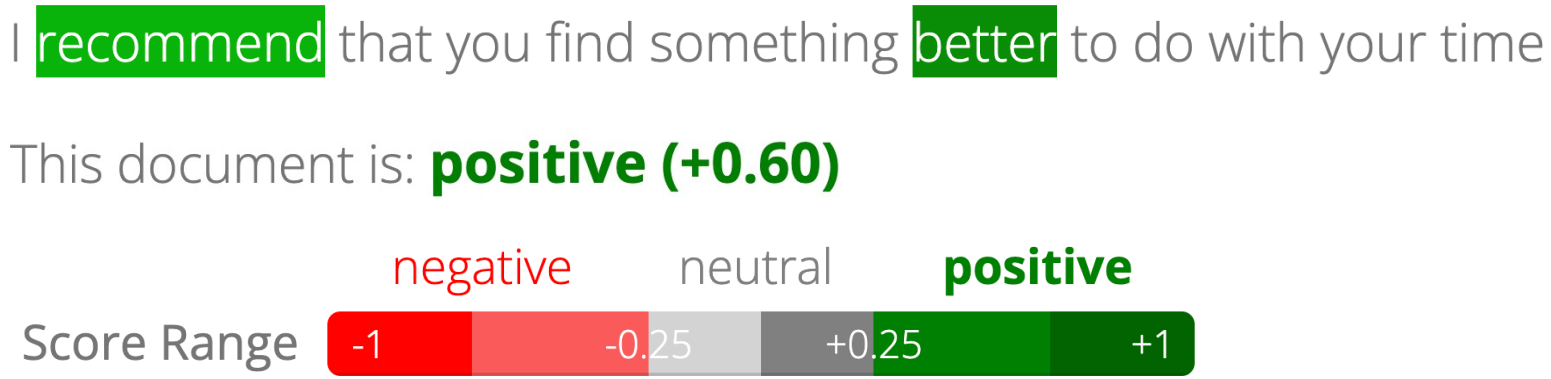


NLP Task: Sentiment Analysis

Sentiment analysis demo

https://text2data.com/Demo

"I recommend that you find something better to do with your time"



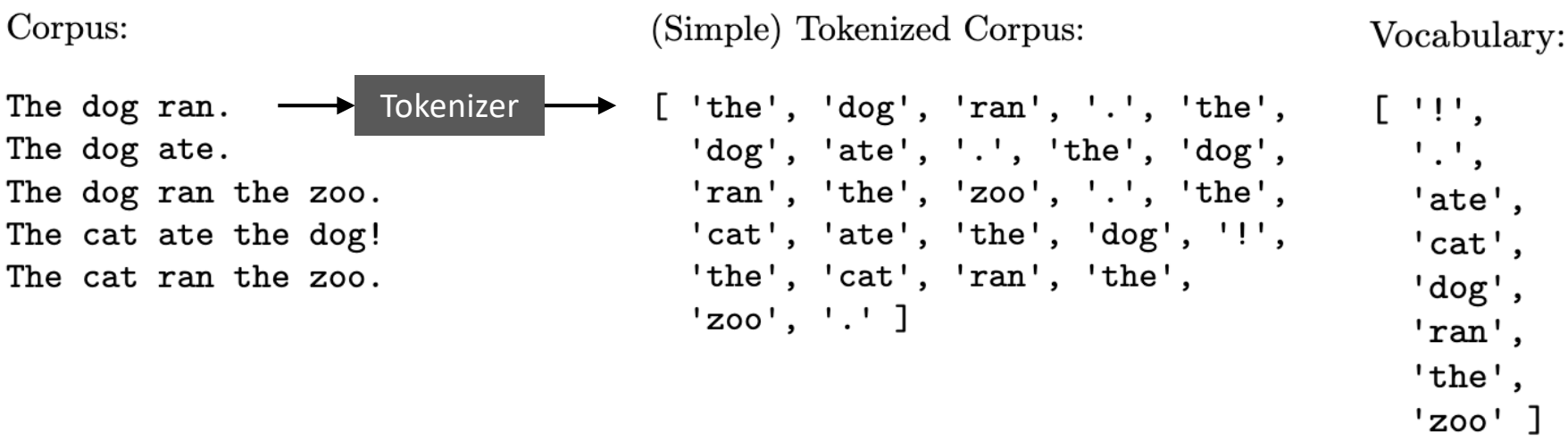
Vocab

Corpus

Tokenization/Tokens

Vocabulary

Context



NLP Vocab

Corpus, Tokenizer

Corpus:

```
The dog ran.
The dog ate.
The dog ran the zoo.
The cat ate the dog!
The cat ran the zoo.
```

→ Tokenizer →

(Simple) Tokenized Corpus:

```
[ 'the', 'dog', 'ran', '.', 'the',
  'dog', 'ate', '.', 'the', 'dog',
  'ran', 'the', 'zoo', '.', 'the',
  'cat', 'ate', 'the', 'dog', '!',
  'the', 'cat', 'ran', 'the',
  'zoo', '.' ]
```

Vocabulary:

```
[ '!',
  '.',
  'ate',
  'cat',
  'dog',
  'ran',
  'the',
  'zoo' ]
```

Tokenization

Tokenization

Text encoding: token → index within vocabulary

What is our "vocabulary", i.e., what are our tokens?

- Characters
- Byte pair encoding
- Words and punctuation

Corpus

```
I am Sam. I am
Sam. Sam I am.
That Sam-I-am.
That Sam-I-am!
```

Tokenization

```
I am Sam. I am Sam. Sam I am.
That Sam-I-am. That Sam-I-am!
```

Text encoding: token → index within vocabulary

What is our "vocabulary", i.e., what are our tokens?

Character token vocabulary

```
 0: !
 1: -
 2: .
 3: I
 4: S
 5: T
 6: a
 7: h
 8: m
 9: t
10: _ (space)
```

BPE token vocabulary

```
 0: !
 1: -
 2: .
 3: I
 4: S
 5: T
 6: a
 7: h
 8: m
 9: t
10: _(space)
11: Sam
12: Th
13: am
14: _Sam
15: _am
```

Word token vocabulary

```
0: !
1: .
2: am
3: I
4: Sam
5: That
6: Sam-I-am
```

Tokenization

```
I am Sam. I am Sam. Sam I am.
That Sam-I-am. That Sam-I-am!
```

Text encoding: token → index within vocabulary

Byte pair encoding (BPE)

Current token vocabulary

(Initialize to characters in corpus)

```
!
-
.
I
S
T
a
h
m
t
_ (space)
```

Pair frequencies

Count frequencies of all pairs of current tokens appearing together in corpus

```
_a: 3
am: 10
_S: 4
Sa: 5
_I: 1
Th: 2
ha: 2
at: 2
```

New tokens

Add most frequent pair to tokens

```
!
-
.
I
S
T
a
h
m
t
_ (space)
am
```

Tokenization

```
I am Sam. I am Sam. Sam I am.
That Sam-I-am. That Sam-I-am!
```

Text encoding: token → index within vocabulary

Byte pair encoding (BPE)

Current token vocabulary

(Initialize to characters in corpus)

```
!
-
.
I
S
T
a
h
m
t
_ (space)
am
```

Pair frequencies

Count frequencies of all pairs of current tokens appearing together in corpus

```
_S:  4
_I:  1
Th:  2
ha:  2
at:  2
_am: 3
Sam: 5
```

New tokens

Add most frequent pair to tokens

```
!
-
.
I
S
T
a
h
m
t
_ (space)
am
Sam
```

Tokenization

```
I am Sam. I am Sam. Sam I am.
That Sam-I-am. That Sam-I-am!
```

Text encoding: token → index within vocabulary

What is our "vocabulary", i.e., what are our tokens?

Character token vocabulary

```
 0: !
 1: -
 2: .
 3: I
 4: S
 5: T
 6: a
 7: h
 8: m
 9: t
10: _ (space)
```

BPE token vocabulary

```
 0: !
 1: -
 2: .
 3: I
 4: S
 5: T
 6: a
 7: h
 8: m
 9: t
10: _(space)
11: Sam
12: Th
13: am
14: _Sam
15: _am
```

Word token vocabulary

```
0: !
1: .
2: am
3: I
4: Sam
5: That
6: Sam-I-am
```

Language Models

Human Language Model



Wheel of Fortune

What is a language model (LM)?

LM Applications: Speech Recognition

Example: Speech Recognition

“artificial ……”

Find most probable next word given “artificial” and the audio for second word.

LM Applications: Speech Recognition

Example: Speech Recognition

"artificial ……."

Find most probable next word given "artificial" and the audio for second word.

Which second word gives the highest probability?

Break down problem

n-gram probability * audio probability

$P(\textbf{limb} \mid \text{artificial}, \text{audio})$

$P(\textbf{limb} \mid \text{artificial}) * P(\text{audio} \mid \textbf{limb})$

$P(\textbf{intelligence} \mid \text{artificial}, \text{audio})$

$P(\textbf{intelligence} \mid \text{artificial}) * P(\text{audio} \mid \textbf{intelligence})$

$P(\textbf{flavoring} \mid \text{artificial}, \text{audio})$

$P(\textbf{flavoring} \mid \text{artificial}) * P(\text{audio} \mid \textbf{flavoring})$

LM Applications: Text Generation

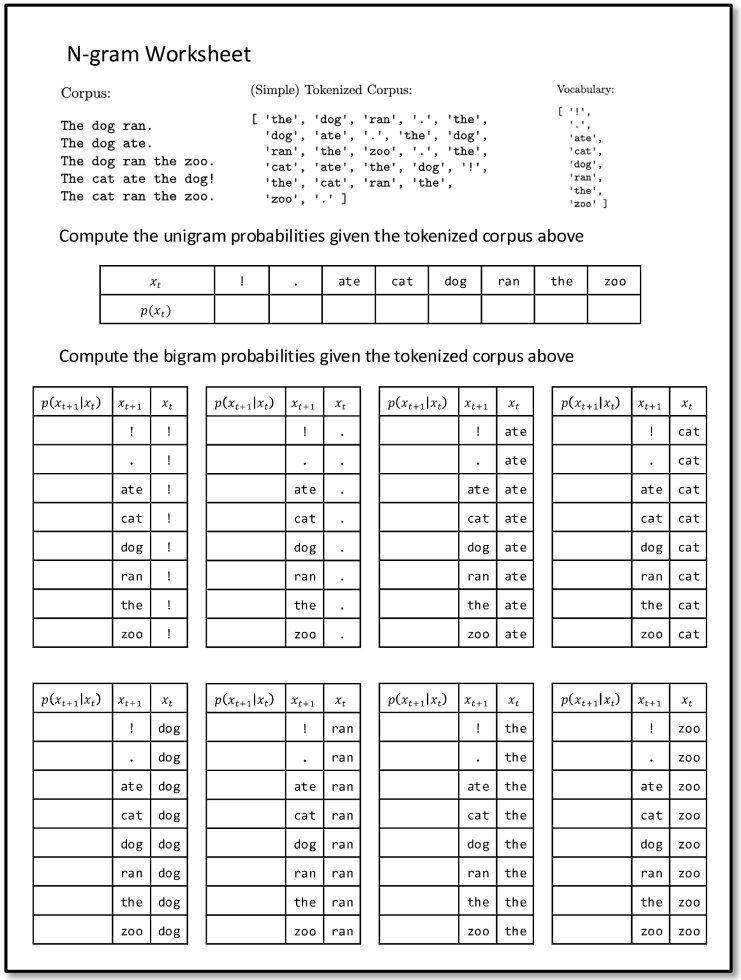
Given a LM that can model the joint distribution any 10-token sequence in the vocabulary, $p(w_1, w_2, \ldots, w_{10})$, how can we get the probability of a 6th token, given 5 previous tokens

Language Models

N-grams

N-gram Worksheet

[Worksheet linked on website]



N-gram Worksheet

Corpus:

```
The dog ran.
The dog ate.
The dog ran the zoo.
The cat ate the dog!
The cat ran the zoo.
```

(Simple) Tokenized Corpus:

```
[ 'the', 'dog', 'ran', '.', 'the',
  'dog', 'ate', '.', 'the', 'dog',
  'ran', 'the', 'zoo', '.', 'the',
  'cat', 'ate', 'the', 'dog', '!',
  'the', 'cat', 'ran', 'the',
  'zoo', '.' ]
```

Vocabulary:

```
[ '!',
  '.',
  'ate',
  'cat',
  'dog',
  'ran',
  'the',
  'zoo' ]
```

Compute the unigram probabilities given the tokenized corpus above

x_t	!	.	ate	cat	dog	ran	the	zoo
$p(x_t)$								

Compute the bigram probabilities given the tokenized corpus above

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	!
	.	!
	ate	!
	cat	!
	dog	!
	ran	!
	the	!
	zoo	!

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	.
	.	.
	ate	.
	cat	.
	dog	.
	ran	.
	the	.
	zoo	.

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	ate
	.	ate
	ate	ate
	cat	ate
	dog	ate
	ran	ate
	the	ate
	zoo	ate

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	cat
	.	cat
	ate	cat
	cat	cat
	dog	cat
	ran	cat
	the	cat
	zoo	cat

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	dog
	.	dog
	ate	dog
	cat	dog
	dog	dog
	ran	dog
	the	dog
	zoo	dog

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	ran
	.	ran
	ate	ran
	cat	ran
	dog	ran
	ran	ran
	the	ran
	zoo	ran

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	the
	.	the
	ate	the
	cat	the
	dog	the
	ran	the
	the	the
	zoo	the

$p(x_{t+1}\|x_t)$	x_{t+1}	x_t
	!	zoo
	.	zoo
	ate	zoo
	cat	zoo
	dog	zoo
	ran	zoo
	the	zoo
	zoo	zoo

N-gram Language Model

If we have unigram and bigram probabilities, do we have a language model??

Unigram: $p(w_t)$ **Bigram:** $p(w_t \mid w_{t-1})$

Example sequence

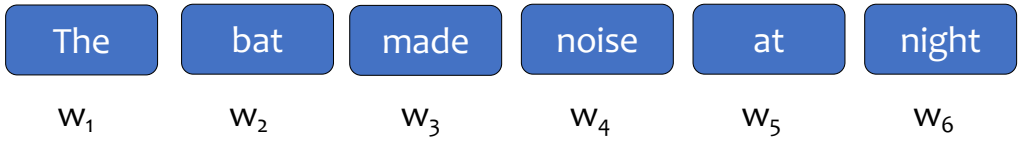
w_1	w_2	w_3	w_4	w_5
the	dog	ate	the	zoo

Joint probability

$$p(w_1, w_2, w_3, w_4, w_5)$$
$$= p(w_1)\, p(w_2 \mid w_1)\, p(w_3 \mid w_2)\, p(w_4 \mid w_3)\, p(w_5 \mid w_4)$$

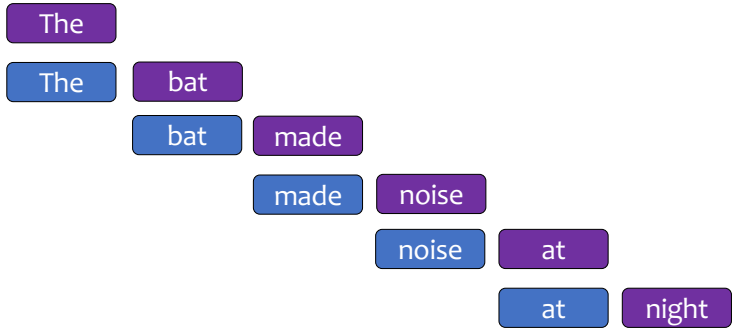
n-Gram Language Model

Question: How can we **define** a probability distribution over a sequence of length T?



n-Gram Model (n=2) $p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_{t-1})$

$p(w_1, w_2, w_3, \ldots, w_6) =$



$p(w_1)$
$p(w_2 \mid w_1)$
$p(w_3 \mid w_2)$
$p(w_4 \mid w_3)$
$p(w_5 \mid w_4)$
$p(w_6 \mid w_5)$

Bigram assumption

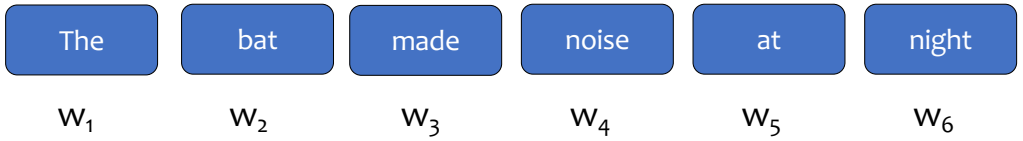
Condition on the previous token only.

$p(w_3 \mid w_1, w_2) \approx p(w_3 \mid w_2)$

$p(w_4 \mid w_1, w_2, w_3) \approx p(w_4 \mid w_3)$

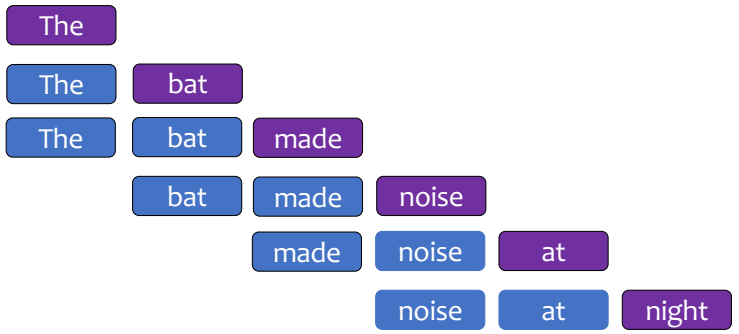
n-Gram Language Model

Question: How can we **define** a probability distribution over a sequence of length T?



n-Gram Model (n=3) $p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_{t-1}, w_{t-2})$

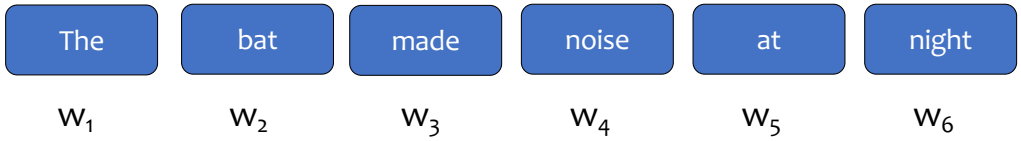
$p(w_1, w_2, w_3, \ldots, w_6) =$



$p(w_1)$
$p(w_2 \mid w_1)$
$p(w_3 \mid w_2, w_1)$
$p(w_4 \mid w_3, w_2)$
$p(w_5 \mid w_4, w_3)$
$p(w_6 \mid w_5, w_4)$

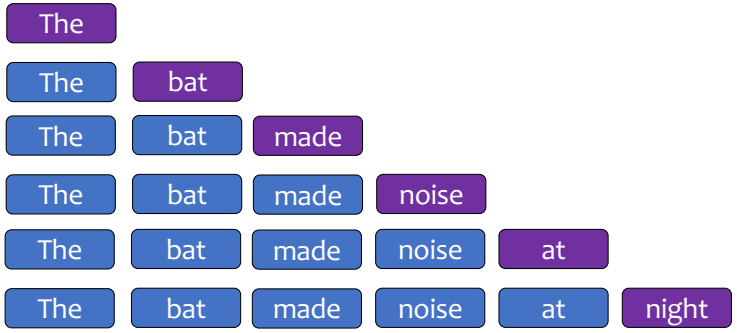
The Chain Rule of Probability

Question: How can we **define** a probability distribution over a sequence of length T?



No independence assumption is required.

$$p(w_1, w_2, w_3, \ldots, w_6) =$$



$$p(w_1)$$
$$p(w_2 \mid w_1)$$
$$p(w_3 \mid w_2, w_1)$$
$$p(w_4 \mid w_3, w_2, w_1)$$
$$p(w_5 \mid w_4, w_3, w_2, w_1)$$
$$p(w_6 \mid w_5, w_4, w_3, w_2, w_1)$$

Markov Chains

N-gram Markov assumption

First-order Markov chain

$$X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_{t-1} \rightarrow X_t \rightarrow X_{t+1} \rightarrow \ldots \rightarrow X_T$$

Conditional independence

$$X_{t-1} \perp\!\!\!\perp X_{t+1} \mid X_t \qquad\qquad X_{t+1} \perp\!\!\!\perp X_1, X_2 \mid X_t$$

$$p(x_{t+1} \mid x_t, x_{1:t-1}) = p(x_{t+1} \mid x_t)$$

Joint distribution

$$p(x_1, \ldots, x_T) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} \mid x_t)$$

LM Training Data

N-gram Training

Where do the n-gram probabilities come from?

Google n-grams demo

N-gram Training Data

Unsupervised learning

Self-supervised learning (auto-regressive)

Example: Jane Austen, *Pride and Prejudice*

Vanity and pride are different things, though the words are often used synonymously. A person may be proud without being vain. Pride relates more to our opinion of ourselves, vanity to what we would have others think of us.

N-gram Training Data

Self-supervised learning (auto-regressive)

Example: Jane Austen, *Pride and Prejudice*

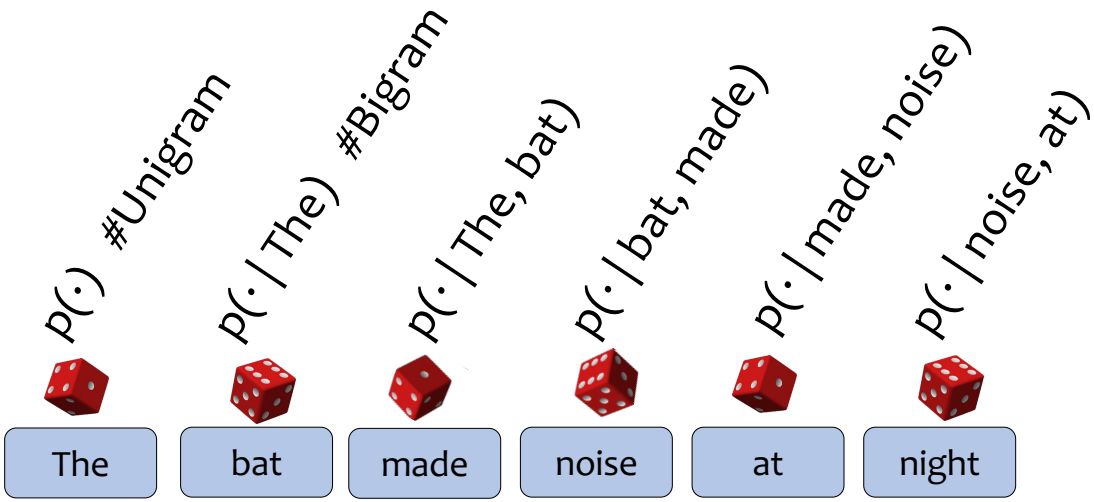
Vanity and pride are different things, though the words are often used synonymously. A person may be proud without being vain. Pride relates more to our opinion of ourselves, vanity to what we would have others think of us.

Sampling from a Language Model

Sampling from a Language Model

How do we sample from a language model?

1. Treat each probability distribution like a (50k-sided) weighted die
2. Pick the die corresponding to $p(w_t \mid w_1, \ldots, w_{t-1})$
3. Roll that die and generate whichever word w_t lands face up
4. Repeat



Slide: CMU, Matt Gormley

N-gram Examples

Random samples from language model trained on Shakespeare:

n=1: "in as , stands gods revenge ! france pitch good in fair hoist an what fair shallow-rooted , . that with wherefore it what a as your . , powers course which thee dalliance all"

n=2: "look you may i have given them to the dank here to the jaws of tune of great difference of ladies . o that did contemn what of ear is shorter time ; yet seems to"

n=3: "believe , they all confess that you withhold his levied host , having brought the fatal bowels of the pope ! ' and that this distemper'd messenger of heaven , since thou deniest the gentle desdemona ,"

n=7: "so express'd : but what of that ? 'twere good you do so much for charity . i cannot find it ; 'tis not in the bond . you , merchant , have you any thing to say ? but little"

This is starting to look a lot like Shakespeare… because it is Shakespeare

Slide: CMU, Zico Kolter

N-gram LM Issues

What are the downsides of N-grams for language models?