Count!
N-gram language models

Language modeling

house

rat

This is the …

house ✓

did ✗

we have some intuition for the next word.

did

malt

What's the probability of the next word?

$$p(house \mid this\ is\ the) = ?$$

Toy corpus

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$p(house \mid this\ is\ the) =$

Toy corpus

This is the house *that Jack built.*
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$p(house \mid this\ is\ the) =$

Toy corpus

This is the house *that Jack built.*
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$p(house \mid this\ is\ the) =$

Toy corpus

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$p(house \mid this\ is\ the) = \frac{c(this\ is\ the\ house)}{c(this\ is\ the\ \ldots)} = \frac{1}{4}$$

Toy corpus

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

4-grams

$$p(house \mid this\ is\ the) = \frac{c(this\ is\ the\ house)}{c(this\ is\ the\ \ldots)} = \frac{1}{4}$$

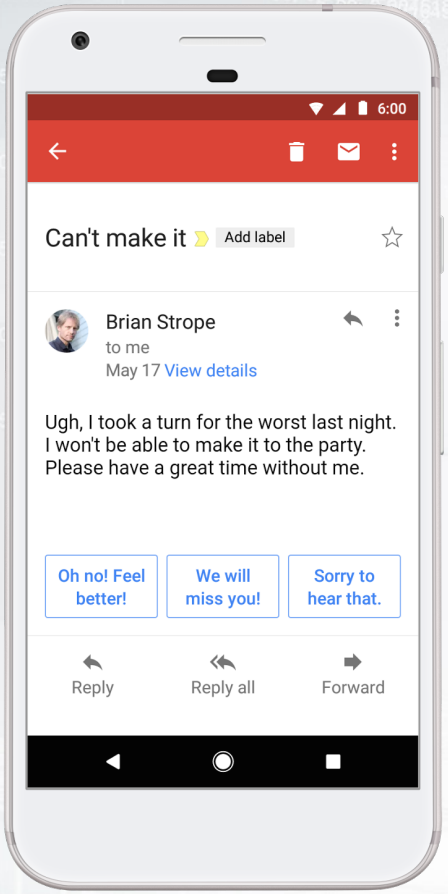
Toy corpus

This is the house ***that Jack*** *built.*
This is the malt
That lay *in the house* ***that Jack*** *built.*
This is the rat,
That ate *the malt*
That lay *in the house* ***that Jack*** *built.*
This is the cat,
That killed *the rat,*
That ate *the malt*
That lay *in the house* ***that Jack*** *built.*

bigrams

$$p(Jack \mid that) = \frac{c(that\ Jack)}{c(that\ldots)} = \frac{4}{10}$$

Where do we need LM?



- Suggestions in messengers
- Spelling correction
- Machine translation
- Speech recognition
- Handwriting recognition
- …

Smart Reply / research.googleblog.com

Language modeling

house

rat

This is the …

did

malt

What's the probability of the whole sequence?

$p(\textit{this is the house}) = ?$

Let's do some math

Predict probability of a sequence of words:

$$\mathbf{w} = (w_1 w_2 w_3 \ldots w_k)$$

Let's do some math

Predict probability of a sequence of words:

$$\mathbf{w} = (w_1 w_2 w_3 \ldots w_k)$$

- **Chain rule:**

$$p(\mathbf{w}) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_1 \ldots w_{k-1})$$

Let's do some math

Predict probability of a sequence of words:

$$\mathbf{w} = (w_1 w_2 w_3 \ldots w_k)$$

- **Chain rule:**

$$p(\mathbf{w}) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_1 \ldots w_{k-1})$$

- **Markov assumption:**

$$p(w_i|w_1 \ldots w_{i-1}) = p(w_i|w_{i-n+1} \ldots w_{i-1})$$

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_{k-1})$$

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_{k-1})$$

Toy corpus:

This is the malt
That lay in the house that Jack built.

p(this is the house) = p(this) p(is| this) p(the| is) p(house| the)

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_{k-1})$$

Toy corpus:

This is the malt
That lay in the house that Jack built.

$$p(\textit{this is the house}) = p(\textit{this})\ p(\textit{is}|\ \textit{this})\ p(\textit{the}|\ \textit{is})\ p(\textit{house}|\ \textit{the})$$

$$= 1/12 \quad 1 \quad 1 \quad 1/2$$

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = \cancel{p(w_1)} p(w_2|w_1) \ldots p(w_k|w_{k-1})$$

$$p(w_1|start)$$

Toy corpus:

This is the malt
That lay in the house that Jack built.

$$\underset{}{p(\text{this is the house})} = \underset{1/2}{p(\text{this})}\ \underset{1}{p(\text{is}|\ \text{this})}\ \underset{1}{p(\text{the}|\ \text{is})}\ \underset{1/2}{p(\text{house}|\ \text{the})}$$

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = \cancel{p(w_1)} p(w_2|w_1) \ldots p(w_k|w_{k-1})$$

$$p(w_1|start)$$

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = \cancel{p(w_1)} p(w_2|w_1) \ldots p(w_k|w_{k-1})$$

$$p(w_1|start)$$

It's normalized separately for each sequence length!

p(this) + p(that) = 1.0

p(this this) + p(this is) + ... + p(built built) = 1.0

...

Bigram language model

So that's what we get for n = 2:

$$p(\mathbf{w}) = \cancel{p(w_1)} p(w_2|w_1) \ldots p(w_k|w_{k-1})$$

$$p(w_1|start)$$

$$p(end|w_k)$$

It's normalized separately for each sequence length!

p(this) + p(that) = 1.0

p(this this) + p(this is) + ... + p(built built) = 1.0

...

Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

p(cat dog cat) =

Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$p(cat\ dog\ cat) = p(cat \mid \_)$

dog	*cat*

Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$p(cat\ dog\ cat) = p(cat \mid \_)$

dog	*cat*

Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$p(cat\ dog\ cat) = p(cat \mid \_)\ p(dog \mid cat)$

dog	*cat tiger*
	cat dog
	cat _

Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$p(cat\ dog\ cat) = p(cat \mid \_)\ p(dog \mid cat)$

dog	*cat tiger*
	cat dog
	cat _

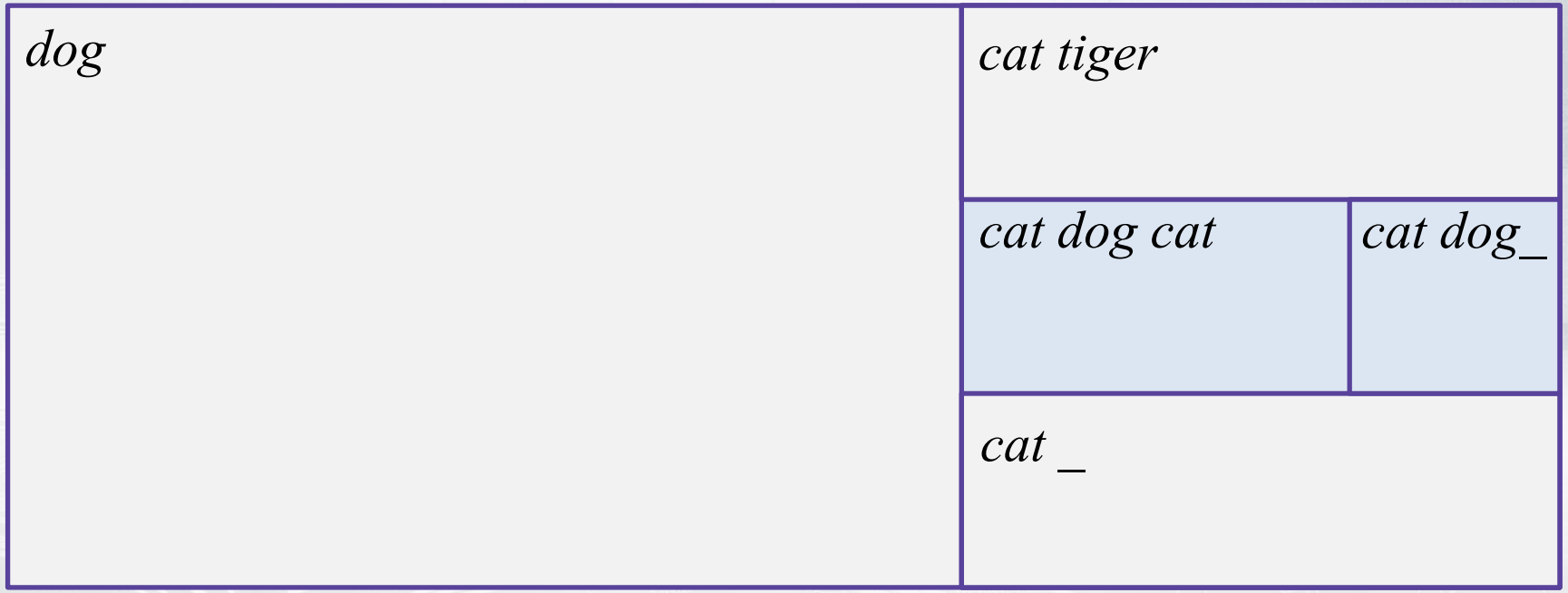
Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$$p(\text{cat dog cat}) = p(\text{cat} \mid \_)\, p(\text{dog} \mid \text{cat})\, p(\text{cat} \mid \text{dog})$$



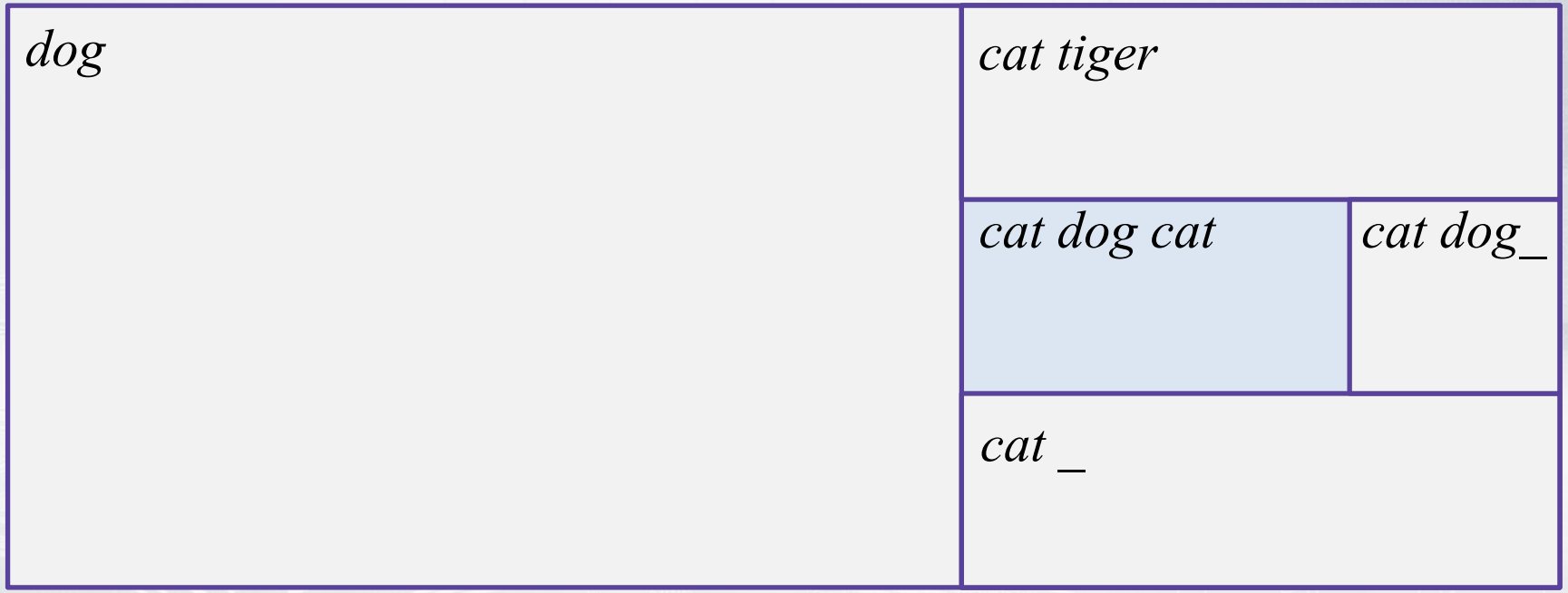
Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$$p(\text{cat dog cat}) = p(\text{cat} \mid \_)\, p(\text{dog} \mid \text{cat})\, p(\text{cat} \mid \text{dog})$$



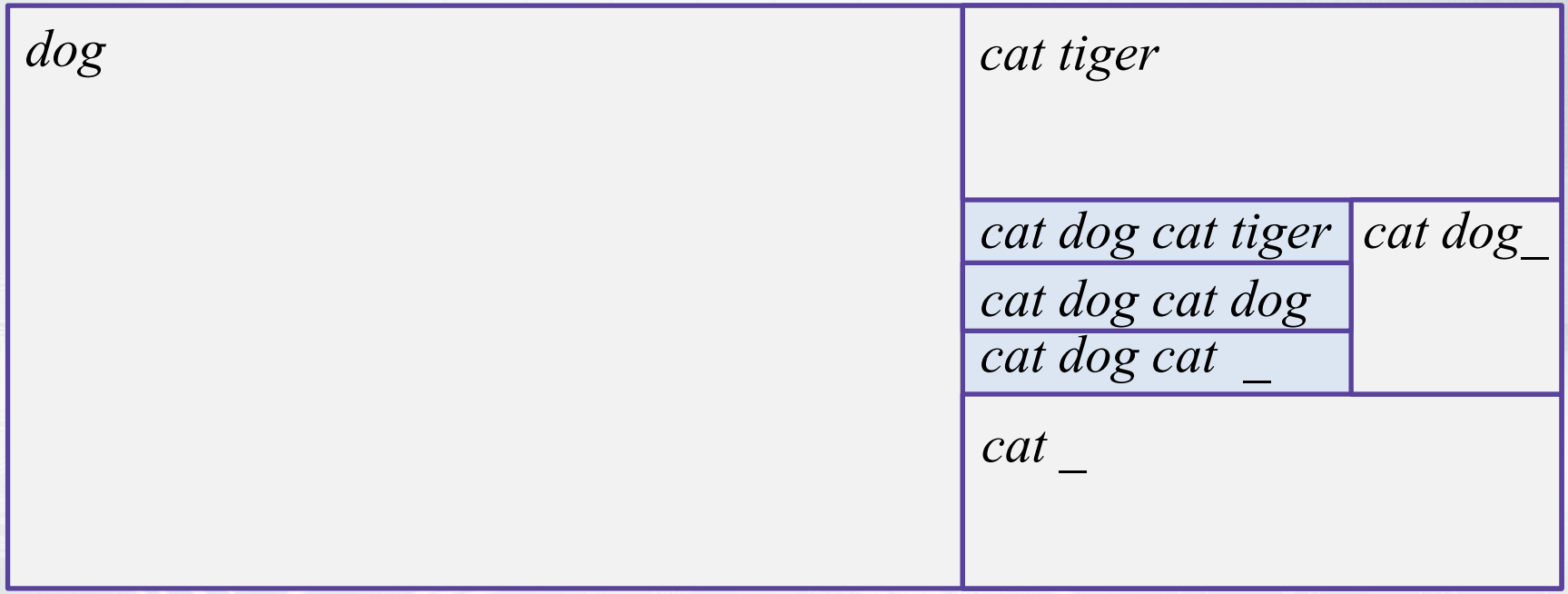
Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$$p(\text{cat dog cat}) = p(\text{cat} \mid \_)\, p(\text{dog} \mid \text{cat})\, p(\text{cat} \mid \text{dog})\, p(\_ \mid \text{cat})$$



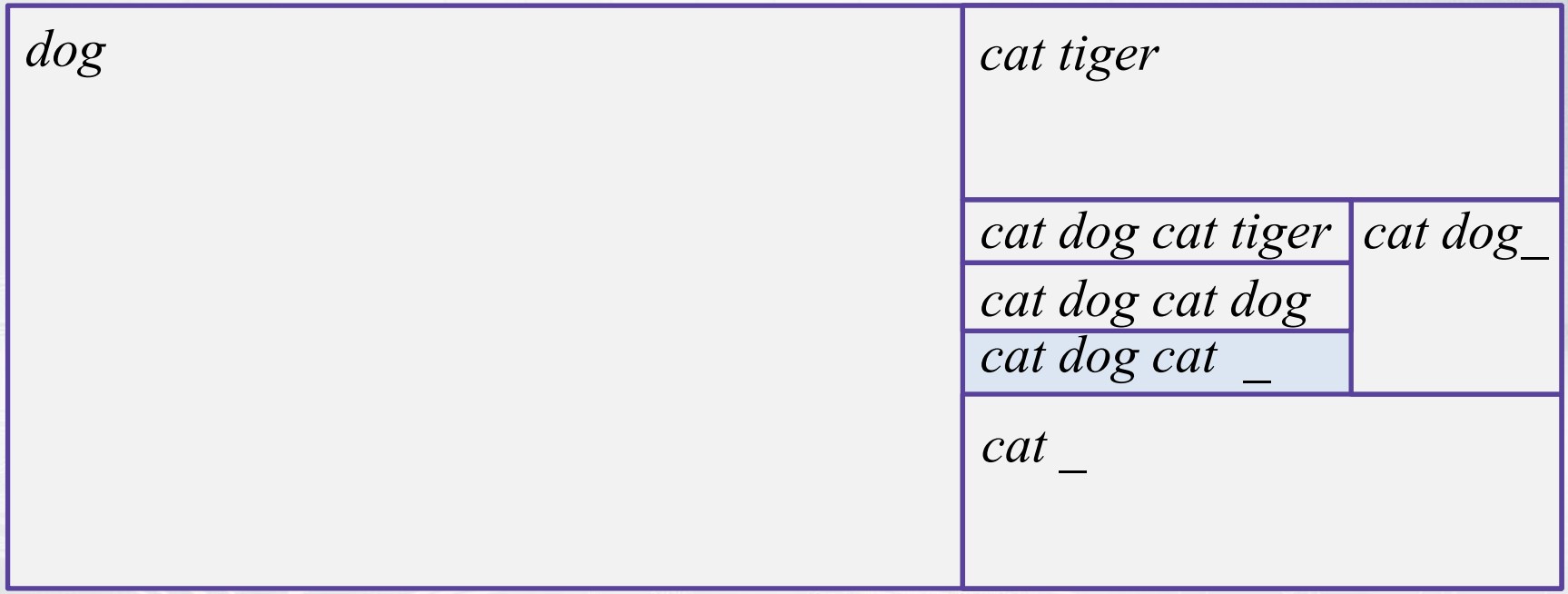
Let's check the model

_ dog _

_ dog cat tiger _

_ cat dog cat _

$$p(\text{cat dog cat}) = p(\text{cat} \mid \_)\, p(\text{dog} \mid \text{cat})\, p(\text{cat} \mid \text{dog})\, p(\_ \mid \text{cat})$$



Resume: bigram language model

Define the model:

$$p(\mathbf{w}) = \prod_{i=1}^{k+1} p(w_i|w_{i-1})$$

Estimate the probabilities:

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)} = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$$

It's all about counting!