Machine Learning for Language Modelling

Part 1: N-gram models

Marek Rei

UNIVERSITY OF TARTU

UNIVERSITY OF CAMBRIDGE

Course structure

- Lectures:
  1. N-gram models
  2. N-gram smoothing
  3. Neural network models
  4. Neural network optimisation

- Practical
- Homework

Course structure

Participate in the lectures and practical

Complete homework exercises

- Implement N-gram language model
- Complete neural network language model
- Submit code and system output
- More details on the course homepage

Result: pass/fail

Course structure

Course homepage:

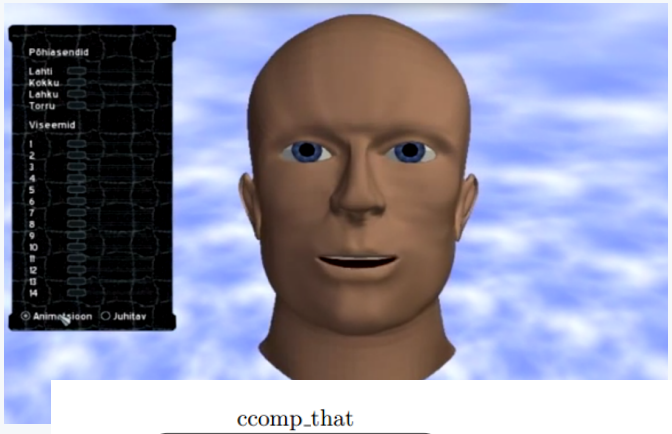
www.marekrei.com/teaching/mllm/

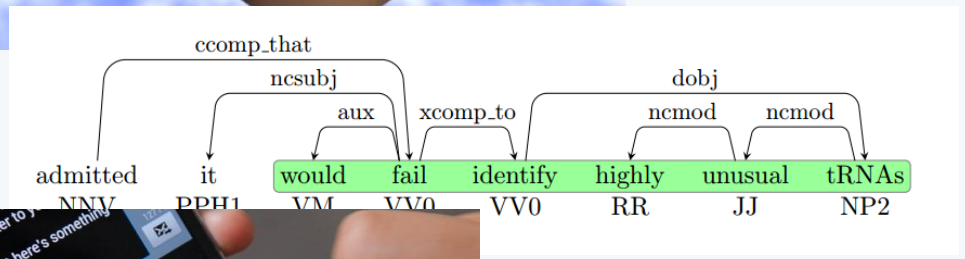
Contains:

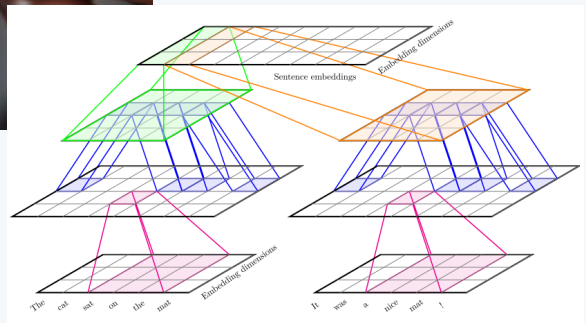
- Lecture slides
- Datasets for language modelling
- References for further information
- An online testing system for homework

About me

- Tallinn University of Technology
- University of Cambridge
- SwiftKey
- University of Cambridge







What is a language model (LM)?

Calculates the probability of a sentence

$$P(today\ is\ a\ windy\ day) = ?$$

Calculates the probability of a word, given previous words

$$P(word \mid context)$$

$$P(day \mid today\ is\ a\ windy) = ?$$

What is a language model (LM)?

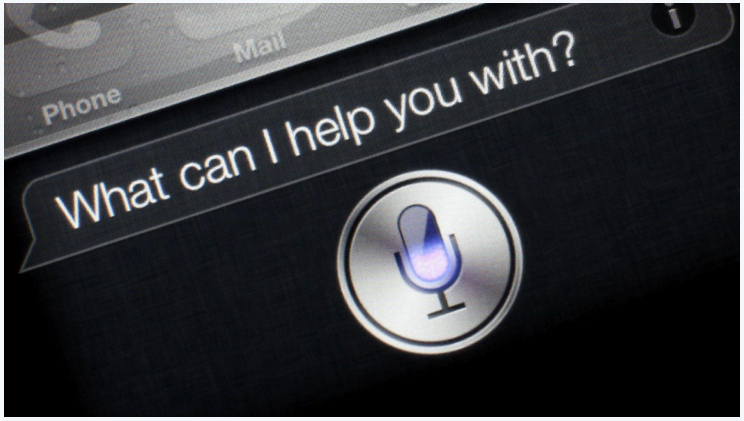
Can rank sentences based on probability

$$P(\textit{today is windy}) = 0.0001$$
$$P(\textit{stochastic gradient descent}) = 0.000001$$
$$P(\text{gradient windy today}) = 0.00000000001$$

Can rank words based on probability

$$P(\textit{windy} \mid \textit{today is}) > P(\textit{yellow} \mid \textit{today is})$$

Applications: Speech recognition



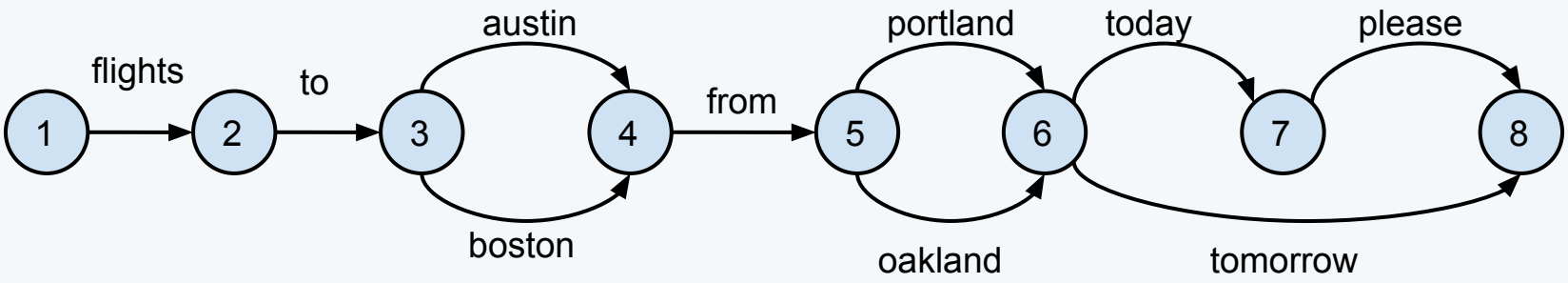
P(*where is the nearest beach*)

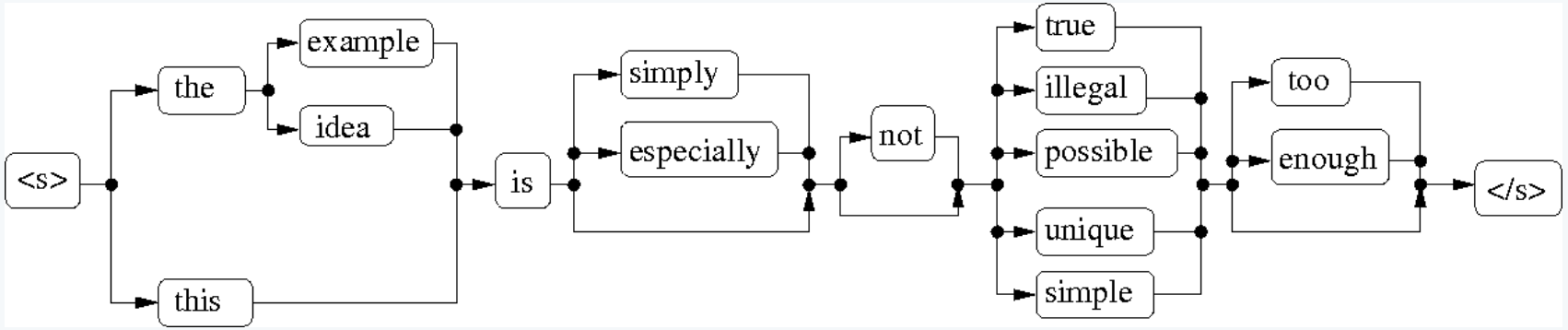
>

P(*where is the nearest breach*)

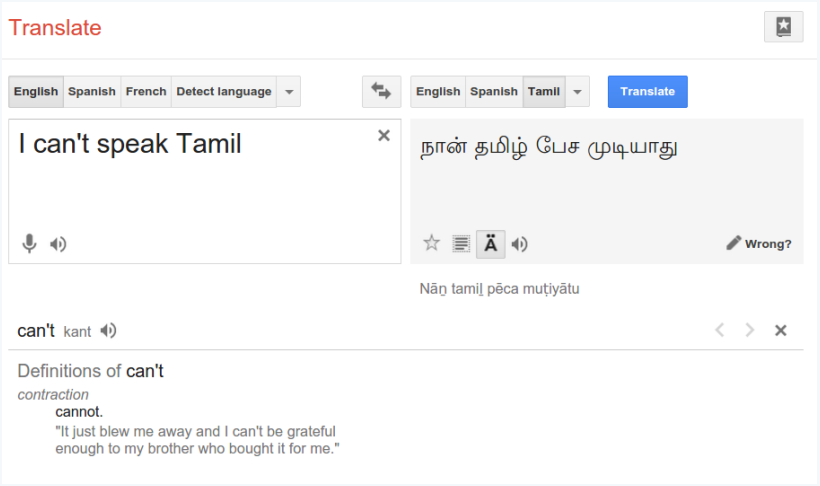
Applications: Speech recognition

Language model helps choose the best path through the speech lattice



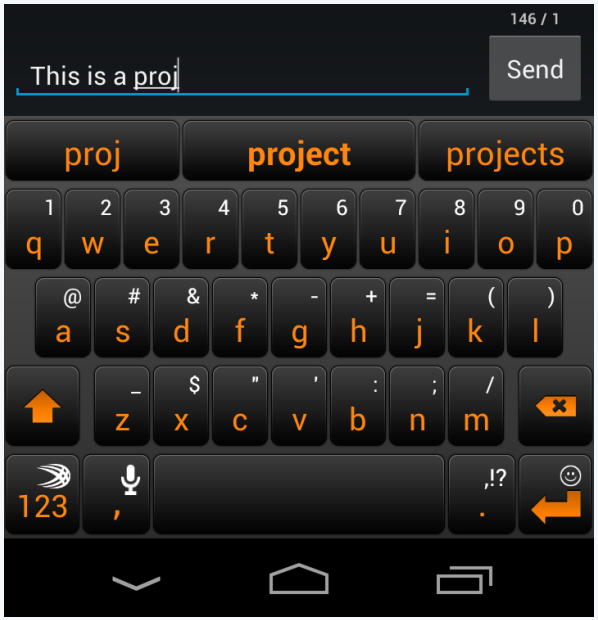


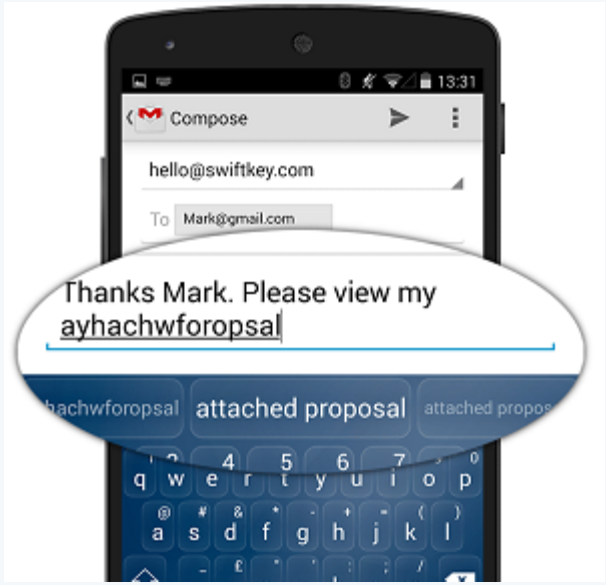
Applications: Machine translation



$$P(\textit{bears are strong}) > P(\text{bears are durable})$$

Applications:
Text prediction/correction





$$P(\textit{proposal} \mid \textit{view my}) > P(\textit{foropsal} \mid \textit{view my})$$

Applications:
Text prediction/correction

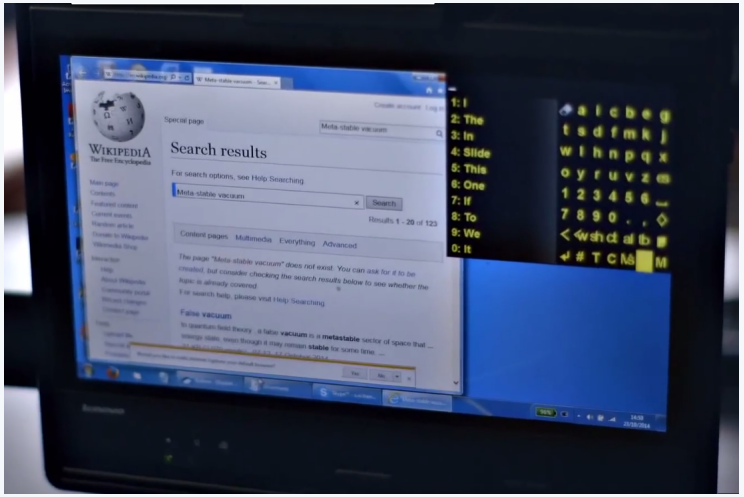
- SwiftKey founded in 2008
- Based on an accurate, fast and adaptive language model
- One of the most popular mobile apps
- 160 employees
- $21.6 Million funding

Applications: Accessibility

Stephen Hawking types by moving his cheek

A cursor goes through all letters and he can stop it at the right one

A language model predicts the word, so he has to type less



Applications: more

- Question answering
- Summarisation
- Text generation
- Information retrieval
- Artificial intelligence
- etc...

Probability of a word

$$P(\text{word}) = \frac{\text{number of times we see this word in the text}}{\text{total number of words in the text}}$$

Probability of a word

The process of machining the fastest wheels in automotive history has begun . The aluminium discs will be fitted to the Bloodhound Supersonic Car , which will endeavour to break the world land speed record (763 mph) later this year . Castle Engineering near Glasgow is leading the industrial consortium that is preparing the wheels . These 90cm discs are a crucial element of the Bloodhound concept , and will have to endure huge loads as they spin at over 170 revolutions per second . Calculations indicate that at peak speed , the wheels will be generating 50,000 radial g at their rim . That 's 50,000 times the pull of gravity . " What does that mean ? It means that a bag of sugar sitting on the wheel when it 's stationary would weigh more than an articulated lorry when the wheel is turning at full speed , " explained Conor La Grue , the components chief on the Bloodhound project . "There are parts of this car where if we have a problem , the driver Andy Green can simply shut them off and bring the vehicle to a stop . But if we have a problem with a wheel , Andy is going to crash . So the design and performance of the discs are absolutely mission-critical , " he told BBC News .

Probability of a word

The process of machining the fastest wheels in automotive history has begun . The aluminium discs will be fitted to the Bloodhound Supersonic Car , which will endeavour to break the world land speed record (763 mph) later this year . Castle Engineering near Glasgow is leading the industrial consortium that is preparing the wheels . These 90cm discs are a crucial element of the Bloodhound concept , and will have to endure huge loads as they spin at over 170 revolutions per second . Calculations indicate that at peak speed , the wheels will be generating 50,000 radial g at their rim . That 's 50,000 times the pull of gravity . " What does that mean ? It means that a bag of sugar sitting on the **wheel** when it 's stationary would weigh more than an articulated lorry when the **wheel** is turning at full speed , " explained Conor La Grue , the components chief on the Bloodhound project . "There are parts of this car where if we have a problem , the driver Andy Green can simply shut them off and bring the vehicle to a stop . But if we have a problem with a **wheel** , Andy is going to crash . So the design and performance of the discs are absolutely mission-critical , " he told BBC News .

$$P(\text{wheel}) = 3/209 = 0.014$$

Probability of a word

The process of machining the fastest wheels in automotive history has begun . The aluminium discs will be fitted to the Bloodhound Supersonic Car , which will endeavour to break the world land speed record (763 mph) later this year . Castle Engineering near Glasgow is leading the industrial consortium that is preparing the wheels . These 90cm discs are a crucial element of the Bloodhound concept , and will have to endure huge loads as they spin at over 170 revolutions per second . Calculations indicate that at peak speed , the wheels will be generating 50,000 radial g at their rim . That 's 50,000 times the pull of gravity . " What does that mean ? It means that a bag of sugar sitting on the wheel when it 's stationary would weigh more than an articulated lorry when the wheel is turning at full speed , " explained Conor La Grue , the components chief on the Bloodhound project . "There are parts of this car where if we have a problem , the driver Andy Green can simply shut them off and bring the vehicle to a stop . But if we have a problem with a wheel , Andy is going to crash . So the design and performance of the discs are absolutely mission-critical , " he told BBC News .

$$P(\text{wheel}) = 3/209 = 0.014$$

$$P(\text{the}) = 16/209 = 0.077$$

Words and frequency

Text: *hi hello hello world !*

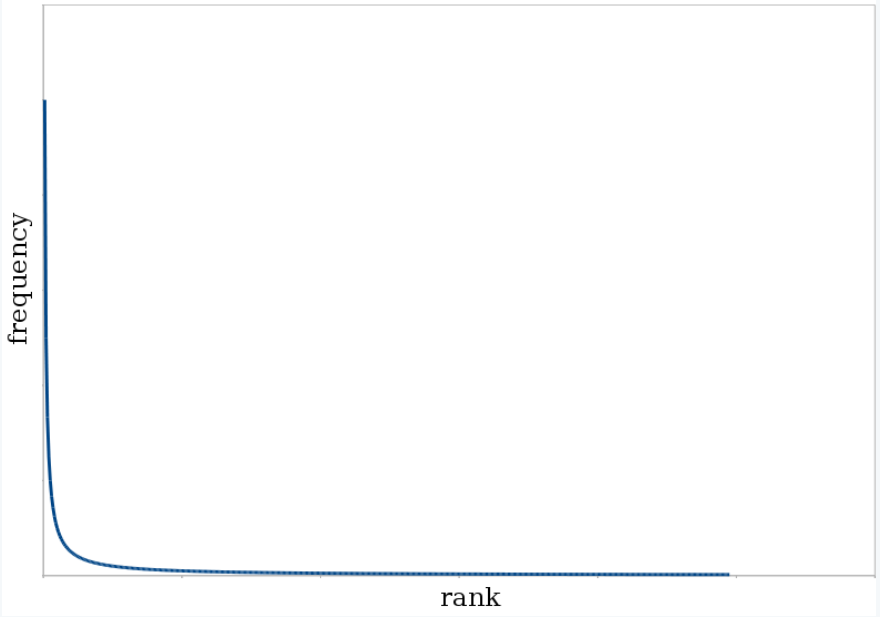
number of tokens

N = 5

number of word types

(vocabulary size)

V = 4



Conditional probability of a word

$$P(\text{word} \mid \text{context}) = \frac{\text{number of times we see context followed by word}}{\text{number of times we see context}}$$

Conditional probability of a word

The process of machining the fastest wheels in automotive history has begun . The aluminium discs will be fitted to the Bloodhound Supersonic Car , which will endeavour to break the world land speed record (763 mph) later this year . Castle Engineering near Glasgow is leading the industrial consortium that is preparing the wheels . These 90cm discs are a crucial element of the Bloodhound concept , and will have to endure huge loads as they spin at over 170 revolutions per second . Calculations indicate that at peak speed , the wheels will be generating 50,000 radial g at their rim . That 's 50,000 times the pull of gravity . " What does that mean ? It means that a bag of sugar sitting on the wheel when it 's stationary would weigh more than an articulated lorry when the wheel is turning at full speed , " explained Conor La Grue , the components chief on the Bloodhound project . "There are parts of this car where if we have a problem , the driver Andy Green can simply shut them off and bring the vehicle to a stop . But if we have a problem with a wheel , Andy is going to crash . So the design and performance of the discs are absolutely mission-critical , " he told BBC News .

$$P(\text{wheel} \mid \text{the}) = 2/16 = 0.125$$

$$P(\text{the} \mid \text{the}) = 0/16 = 0.0$$

Try it

to more extreme weather conditions . Speaking
spring as warm weather arrives in UK
treated to sunny weather . Saturday is
, with warm weather expected for the
for UK including weather warnings , temperature
the recent warm weather has hit sales
often brings sunny weather . Spring can

$$P(\text{weather} \mid \text{warm}) = 3/3 = 1.0$$

$$P(\text{warnings} \mid \text{weather}) = 1/7 = 0.143$$

Useful tips

- Use a much bigger corpus (dataset)
- What counts as a word?

```
at peak speed, the wheels
at peak speed , the wheels
```

Tokenisation helps

- Capitalisation counts

The != *the*

Probability of a sentence

P(the weather is nice) = ?

Use the chain rule in probability theory

$$P(w_1, \ldots, w_N) = \prod_{i=1}^{N} P(w_i | w_1, \ldots, w_{i-1})$$

P(the weather is nice) =

P(the) * P(weather | the) *

P(is | the weather) *

P(nice | the weather is)

Markov assumption

P(begun | The process of machining the fastest wheels in automotive history has) = ?

Let's choose a number **N**, and say only **N-1** previous words affect the probability.

P(begun | history has)

$$P(w_i|w_1 \ldots w_{i-1}) \approx P(w_i|w_{i-2}w_{i-1})$$

Special tokens

We can choose to represent sentence start and end with special tokens

<s> <s> This is a sentence </s>

We can represent rare words with a special token

The roloway monkey is endangered

The <UNK> monkey is endangered

N-gram language model

P(the weather is nice) =

P(the) * P(weather) *

P(is) * P(nice)

N = 1

unigram

P(the weather is nice) =

P(the | <s>) * P(weather | the) *

P(is | weather) * P(nice | is)

N = 2

bigram

P(the weather is nice) =

P(the | <s> <s>) *

P(weather | the <s>) *

P(is | the weather) *

P(nice | weather is)

N = 3

trigram

Markov assumption

Long-range dependencies are not captured

The student who went to the field trip in South Africa has graduated

Question

Using a unigram language model, which sentence has a higher probability?

P(a the it)

P(clouds are moving)

What about using a trigram (3-gram) language model?

Generating text

Given the context, sample the next word from the language model, based on its probability

We could always just pick the most probable word, but

1. It would always generate the same text
2. It wouldn't take into account the word probabilities.

Generating text

Current sentence:

<s> <s>

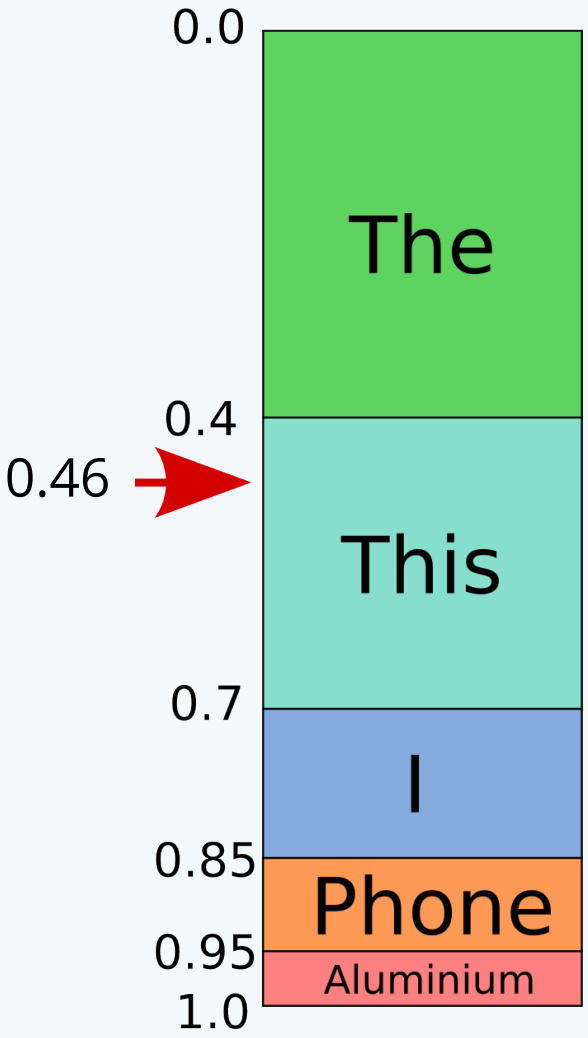
w	P(w|<s> <s>)
The	0.4
This	0.3
I	0.15
Phone	0.1
Aluminum	0.05

Generating text

Current sentence:

<s> <s>

w	P(w|<s> <s>)
The	0.4
This	0.3
I	0.15
Phone	0.1
Aluminum	0.05

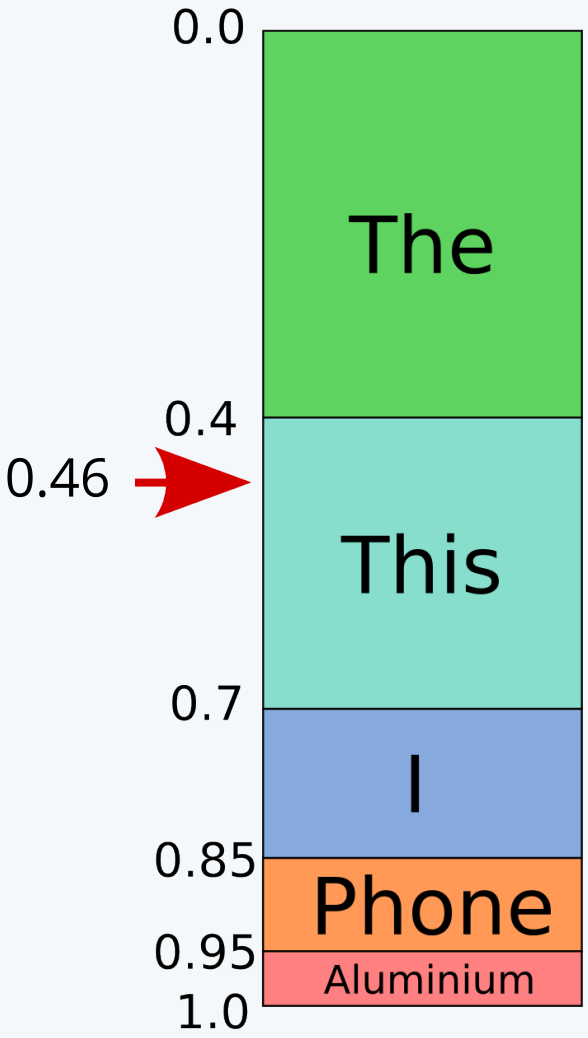


Generating text

Current sentence:

<s> <s> This

w	P(w|<s> <s>)
The	0.4
This	0.3
I	0.15
Phone	0.1
Aluminum	0.05



Generating text

Current sentence:

<s> <s> This

w	P(w| <s> This)
is	0.35
thing	0.3
house	0.2
concert	0.11
hatchback	0.04

Generating text

Current sentence:

<s> <s> This

w	P(w| <s> This)
is	0.35
thing	0.3
house	0.2
concert	0.11
hatchback	0.04

Generating text

Current sentence:

<s> <s> This hatchback

w	P(w| <s> This)
is	0.35
thing	0.3
house	0.2
concert	0.11
hatchback	0.04

Generating text

Current sentence:

<s> <s> This hatchback

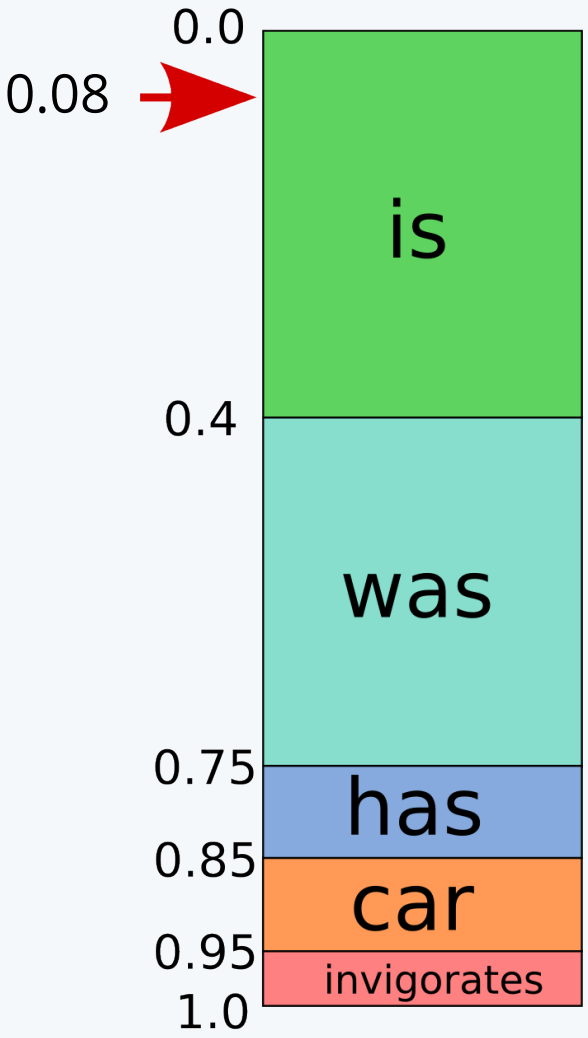
w	P(w| This hatchback)
is	0.4
was	0.35
has	0.1
car	0.1
invigorates	0.05

Generating text

Current sentence:

<s> <s> This hatchback

w	P(w| This hatchback)
is	0.4
was	0.35
has	0.1
car	0.1
invigorates	0.05



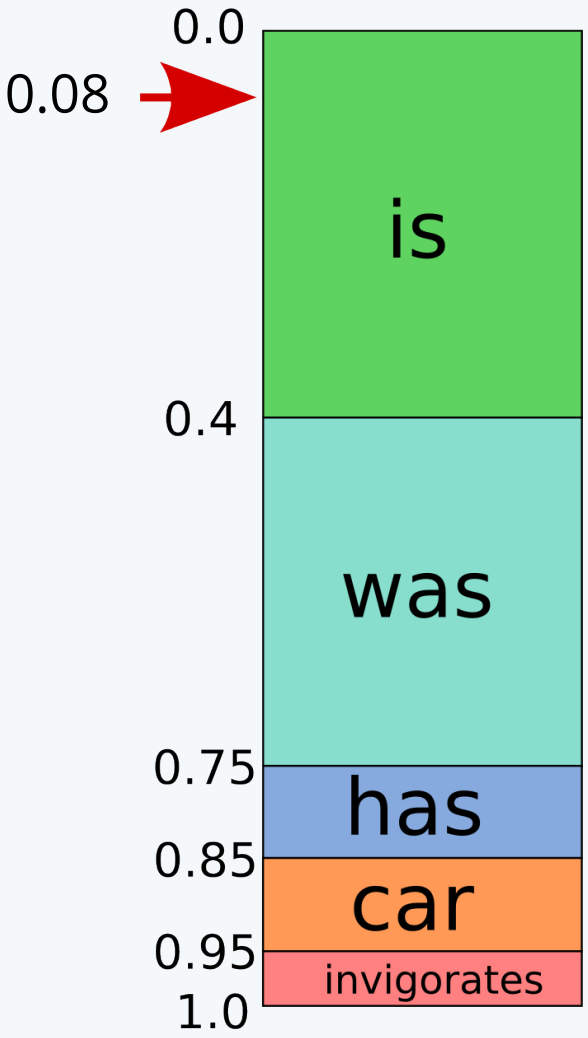
Generating text

Current sentence:

<s> <s> This hatchback

is

w	P(w| This hatchback)
is	0.4
was	0.35
has	0.1
car	0.1
invigorates	0.05

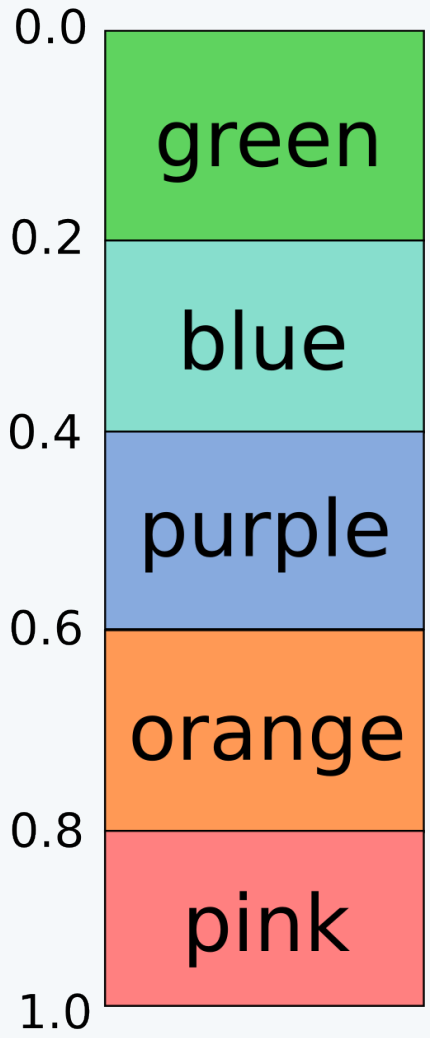


Generating text

Current sentence:

<s> <s> This hatchback is

w	P(w| hatchback is)
green	0.2
blue	0.2
purple	0.2
orange	0.2
pink	0.2

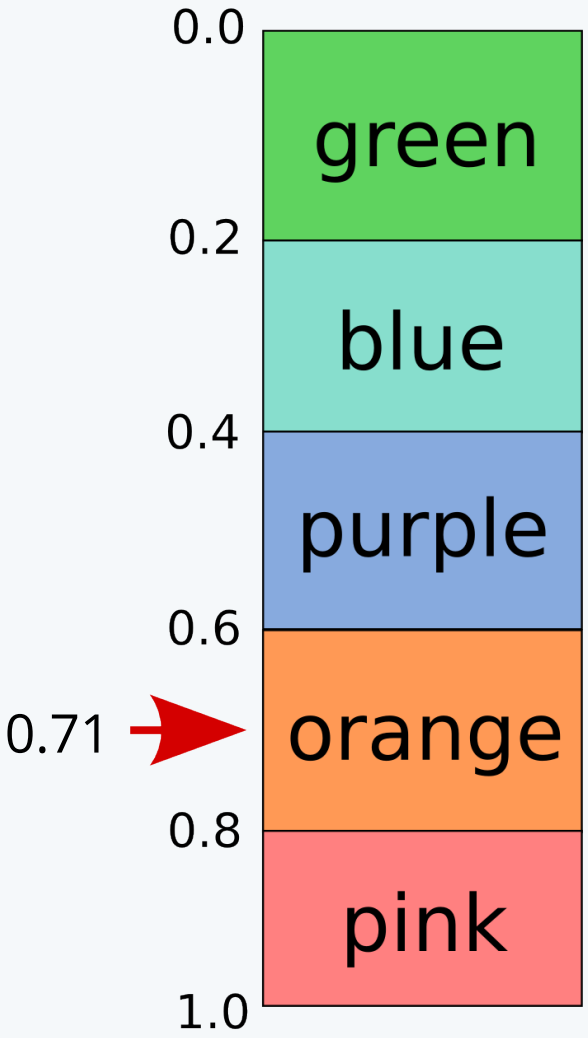


Generating text

Current sentence:

<s> <s> This hatchback is

w	P(w| hatchback is)
green	0.2
blue	0.2
purple	0.2
orange	0.2
pink	0.2

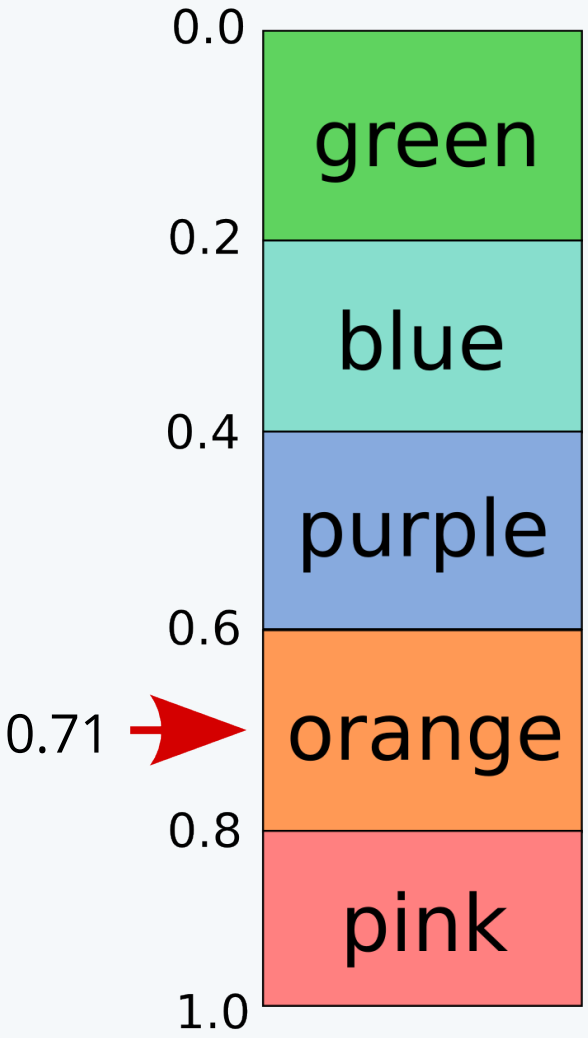


Generating text

Current sentence:

<s> <s> This hatchback is orange

w	P(w| hatchback is)
green	0.2
blue	0.2
purple	0.2
orange	0.2
pink	0.2



Generating text

N = 1

from same post long limited august pogonotrophy in springfield at is some city of in run the building .

N = 2

he became a tower was designed to the district of last night , relocating airfields in 1617 , from the 1996) is required to three sons inheriting equal rights

N = 3

the next season and both emperor of india came to an interest in blues " band mix " cd to include settlement of river street nearest u.s. route 167

N = 4

subsequently he was elected chairman of the tramways committee in 1898 , the forest wood hoopoe and the white-headed wood hoopoe .

N = 5

when concerned , it rears up the anterior portion (usually one-third) of its body when extending the neck , showing the fangs and hissing loudly .

Generating text (N = 3)

Wikipedia

mainland china , based on shared properties (stanford university school of archaeology , cancer research center , the criminal code was close for comfort , rather than the conventional song structures are generated .

WSJ

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

Shakespeare

Sweet prince, Falstaff shall die. Harry of Monmouth’s grave.

This shall forbid it should be branded, if renown made it empty.

Jurafsky & Martin, 2000

Recap

Language models assign probabilities to sentences and words

$$P(sentence) = ?$$

$$P(word \mid context) = ?$$

Used for

- Machine translation
- Speech recognition
- Spelling correction
- Text generation
- and more

Recap

To calculate the probability of a text we use the chain rule

$$P(w_1 \ \ldots \ w_N) = \prod_{i=1}^{N} P(w_i | w_1 \ \ldots \ w_{i-1})$$

and the Markov assumption

$$P(w_1 \ \ldots \ w_N) \approx \prod_{i=1}^{N} P(w_i | w_{i-2} \ w_{i-1})$$

References

Speech and Language Processing
Daniel Jurafsky & James H. Martin (2000)

Evaluating language models. Julia Hockenmaier.
https://courses.engr.illinois.edu/cs498jh/

Language Models. Nitin Madnani, Jimmy Lin. (2010)
http://www.umiacs.umd.edu/~jimmylin/cloud-2010-Spring/

An Empirical Study of Smoothing Techniques for Language Modeling
Stanley F. Chen, Joshua Goodman. (1998)
http://www.speech.sri.com/projects/srilm/manpages/pdfs/chen-goodman-tr-10-98.pdf

Natural Language Processing
Dan Jurafsky & Christopher Manning (2012)
https://www.coursera.org/course/nlp

Extra materials

Try it

```
Alice goes running
  Bob goes running
  Bob goes swimming
  Bob goes running
```

$P(\text{Bob}) = ?$

$P(\text{Alice}) = ?$

$P(\text{goes} \mid \text{Bob}) = ?$

$P(\text{running} \mid \text{goes}) = ?$

$P(\text{running} \mid \text{Bob goes}) = ?$

$P_{\text{unigrams}}(\text{Bob goes running}) = ?$

$P_{\text{bigrams}}(\text{Bob goes running}) = ?$ (using <S>)

Try it

```
Alice goes running
  Bob goes running
  Bob goes swimming
  Bob goes running
```

$P(\text{Bob}) = 3/12 = 0.25$

$P(\text{Alice}) = 1/12 = 0.08$

$P(\text{goes} \mid \text{Bob}) = 3/3 = 1$

$P(\text{running} \mid \text{goes}) = 3/4 = 0.75$

$P(\text{running} \mid \text{Bob goes}) = 2/3 = 0.66$

$P_{\text{unigrams}}(\text{Bob goes running}) = 3/12 * 4/12 * 3/12 = 0.02$

$P_{\text{bigrams}}(\text{Bob goes running}) = 3/4 * 3/3 * 3/4 = 0.56$