KI-Campus
Die Lernplattform
für Künstliche Intelligenz

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Language Models

Salar Mohtaj | DFKI

Language models

- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

Language models

- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

What is language modeling?

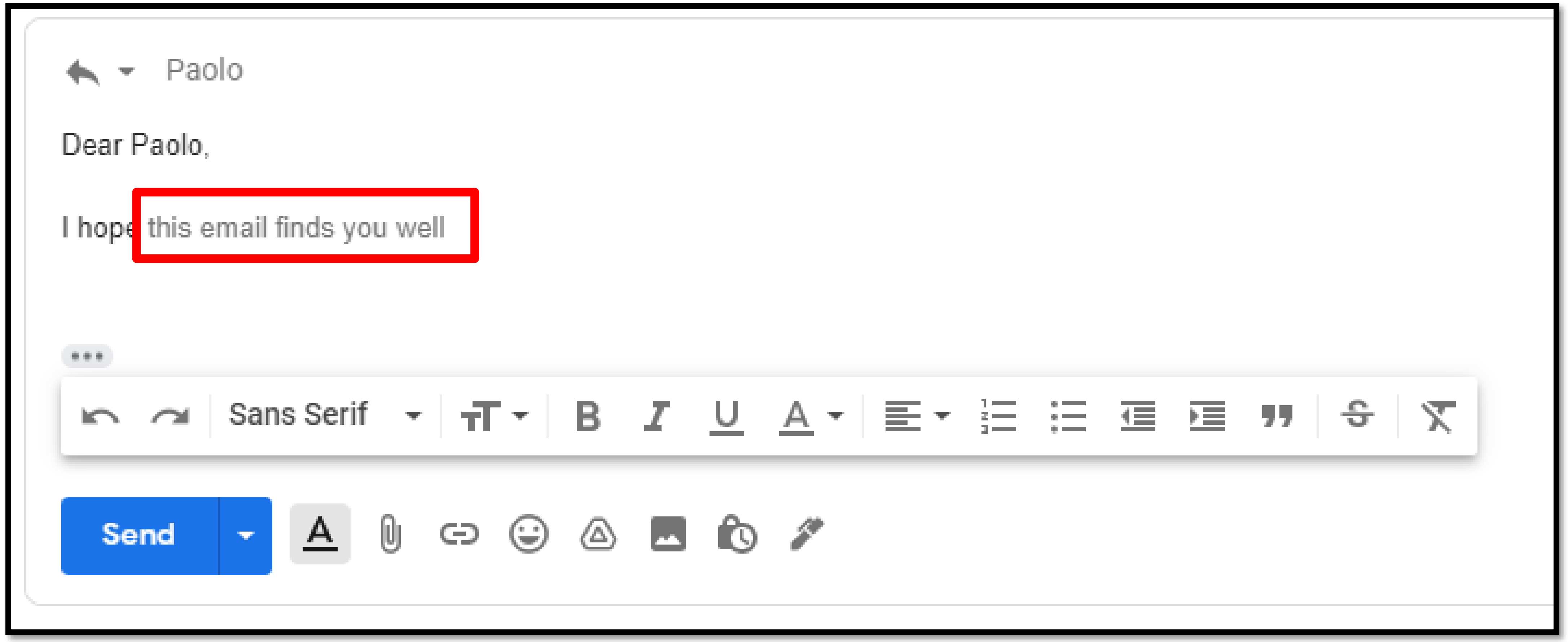
- Models that assign probabilities to sequences of words are called ***language models*** or ***LMs***
  - A language model learns to predict the probability of a sequence of words
  - It is a statistical tool to predict words
- ***Language models*** try to find patterns in the human language
  - They are used to predict the next word in a sentence

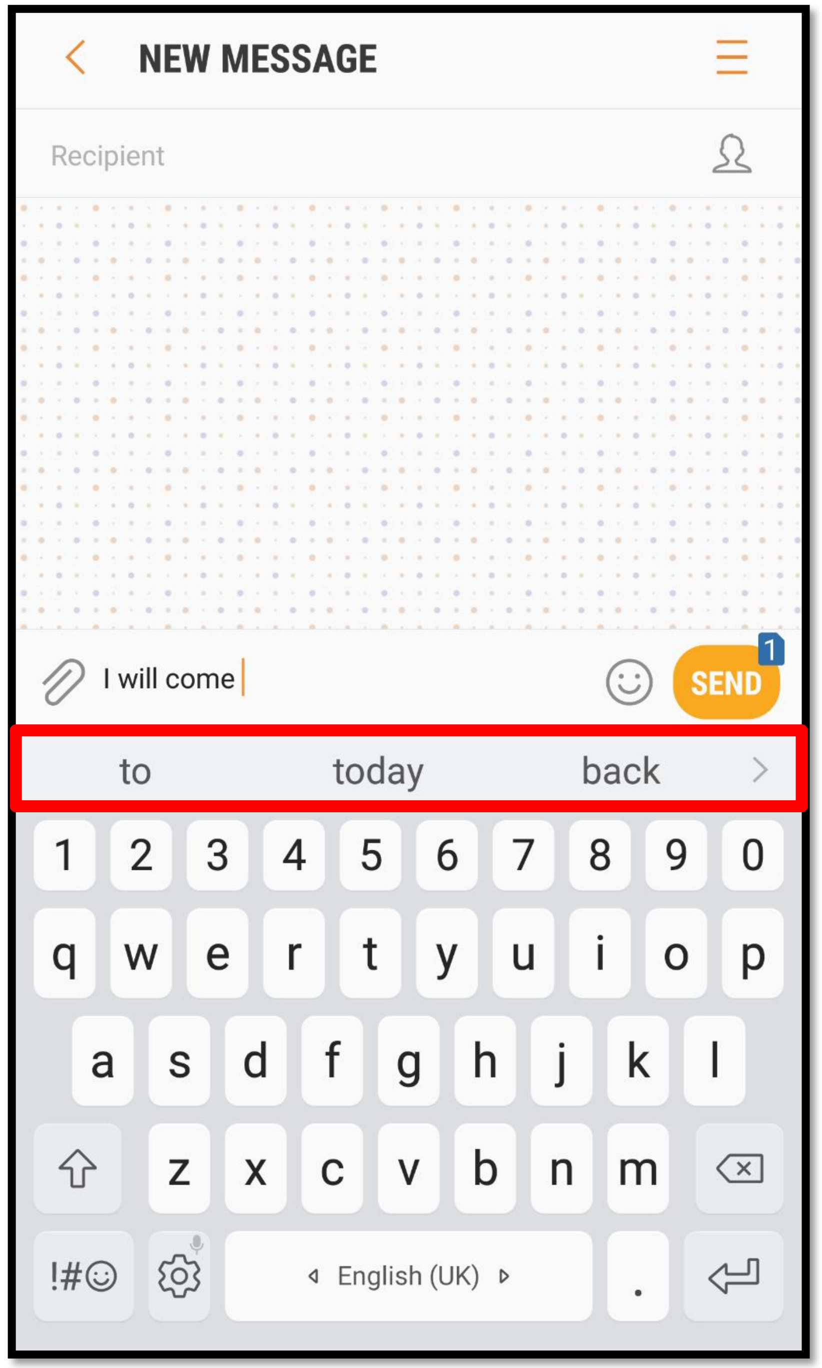
Can you please come ~~time~~?

Can you please come here?

- Language models are a crucial component in the NLP journey

What is language modeling?





Language models

- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

Why language modeling is critical in NLP?

- The overall performance of different NLP tasks can be improved by ***language models***
- Especially in cases where the machine has to generate human language
  - Machine translation
  - Text summarization
  - Image captioning
  - …

It is a cat

Katze es ist eine

es ist eine Katze

Es eine ist Katze

Why language modeling is critical in NLP?

- Text summarization

Why language modeling is critical in NLP?

- Image captioning

man in black shirt is playing guitar

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

two young girls are playing with lego toy

Language models

- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

Statistical language modeling

- Statistical language modeling is the development of ***probabilistic models*** that are able to predict the next word in the sequence given the words that precede it
- The objective is to compute the probability of a word ***w*** given some history ***h***

$\boxed{\text{Can you please come}}$ ______?

$$P(w|h)$$

$$P(w|w_1, \ldots., w_{n-1})$$

Statistical language modeling

Can you please come ______?

Can you please come time?

$$P(time|can\ you\ please\ come)$$

$$= \frac{c(can\ you\ please\ come\ time)}{c(can\ you\ please\ come)}$$

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

Statistical language modeling

Can you please come ______?

Can you please come <u>time</u>?

Can you please come <u>here</u>?

$P(w|h)$

$P(time|can\ you\ please\ come)$

$P(here|can\ you\ please\ come)$

$$= \frac{c(can\ you\ please\ come\ time)}{c(can\ you\ please\ come)}$$

$$= \frac{c(can\ you\ please\ come\ here)}{c(can\ you\ please\ come)}$$

Statistical language modeling

- Joint probability of an entire sequence

Can you please come here?

- $P(Can\ you\ please\ come\ here)$
- Decompose this probability using the ***chain rule of probability***

$$P(w_1, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2) \dots P(w_n|w_{1:n-1})$$

- We could estimate the joint probability of an entire sequence of words by multiplying together a number of conditional probabilities.

Statistical language modeling

$$P(w_1, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \dots P(w_n|w_{1:n-1})$$

Can you please come here?

$$P(Can\ you\ please\ come\ here) = P(can)P(you|can)P(please|can\ you)$$
$$P(come|can\ you\ please)P(here|can\ you\ please\ come)$$

- For 100 words (|*v*|=*100*) and average sentence length of 10:
  - 100^{10} possible sequences

Statistical language modeling

$$P(w_1, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2) \dots P(w_n|w_{1:n-1})$$

- ***Markov assumption***
  - The probability of a word depends only on the ***k*** previous words
  - Markov models are the class of probabilistic models Markov that assume we can predict the probability of some future unit without looking too far into the past

$$P(w_n|w_1, \dots, w_{n-1}) \approx P(w_n|w_{i-k}, \dots, w_{n-1})$$

$\boxed{\text{Can you please come}}$ here?

please come here?

Statistical language modeling

- Instead of computing the probability of a word given its ***entire history***, we can ***approximate*** the history by just the last few words
  - N-gram model

2-gram language model

come here?

$P(here|can\ you\ please\ come)$ → $P(here|come)$

Statistical language modeling

- With a bigram language model, we are approximating:

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$$

$$P(here|can\ you\ please\ come) \approx P(here|come)$$

Statistical language modeling

2-gram

come ____?

Asian ____?

the ____?

3-gram

please come ____?

an Asian ____?

before the ____?

Students should register before the ____?

Statistical language modeling

- With a bigram language model, we are approximating:

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$$

$$P(w_n|w_{n-1}) = \frac{c(w_{n-1}w_n)}{c(w_{n-1})}$$

$$P(here|can\ you\ please\ come) \approx P(here|come)$$

$$P(here|come) = \frac{c(come\ here)}{c(come)}$$

Statistical language modeling

Can you please come here?

$Counting\ number\ of\ times\ that\ the\ sequence\ is\ reapeaded\ in\ the\ corpus$

$P(Can\ you\ please\ come\ here) = P(can)P(you|can)P(please|can\ you)$
$P(come|can\ you\ please)P(here|can\ you\ please\ come)$

$$P(w_n|w_{1:n-1}) \ \approx \ P(w_n|w_{n-1})$$

$$P(Can\ you\ please\ come\ here) = P(can)P(you|can)P(please|\ you)P(come|please)P(here|come)$$

$$P(w_n|w_{n-1}) = \frac{c(w_{n-1}w_n)}{c(w_{n-1})}$$

$$P(here|come) = \frac{c(come\ here)}{c(come)}$$

Statistical language modeling

2-gram

- D_1 = <s> the book is written </s>
- D_2 = <s> the paint is drawn </s>
- D_3 = <s> the texture of the paint is white </s>

$$\boxed{P(the|<s>) = {}^{3}/_{3}}$$

$$P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}$$

$$P(is|book) = {}^{1}/_{1}$$

$$P(written|is) = {}^{1}/_{3}$$

$$P(</s>|written) = {}^{1}/_{1}$$

$$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$$

$$P(is|paint) = {}^{2}/_{2}$$

$$P(drawn|is) = {}^{1}/_{3}$$

$$P(</s>|drawn) = {}^{1}/_{1}$$

$$P(texture|the) = {}^{1}/_{4}$$

$$P(of|texture) = {}^{1}/_{1}$$

$$P(the|of) = {}^{1}/_{1}$$

$$P(white|is) = {}^{1}/_{3}$$

$$P(</s>|white) = {}^{1}/_{1}$$

Statistical language modeling

2-gram

- D_1 = <s> the book is written </s>
- D_2 = <s> the paint is drawn </s>
- D_3 = <s> the texture of the paint is white </s>

$P(the| < s >) = {}^{3}/_{3}$

$$\boxed{P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}}$$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}$

$P(</s > |written) = {}^{1}/_{1}$

$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$

$P(is|paint) = {}^{2}/_{2}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s > |drawn) = {}^{1}/_{1}$

$P(texture|the) = {}^{1}/_{4}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(white|is) = {}^{1}/_{3}$

$P(</s > |white) = {}^{1}/_{1}$

Statistical language modeling

2-gram

- D_1 = <s> the book is written </s>
- D_2 = <s> the paint is drawn </s>
- D_3 = <s> the texture of the paint is white </s>

$P(the| <s>) = {}^{3}/_{3}$

$$\boxed{P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}}$$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}$

$P(</s> |written) = {}^{1}/_{1}$

$$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$$

$P(is|paint) = {}^{2}/_{2}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s> |drawn) = {}^{1}/_{1}$

$P(texture|the) = {}^{1}/_{4}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(white|is) = {}^{1}/_{3}$

$P(</s> |white) = {}^{1}/_{1}$

Statistical language modeling

$P(the| < s >) = {}^{3}/_{3}$

$\boxed{P(written|is) = {}^{1}/_{3}}$

$P(is|paint) = {}^{2}/_{2}$

$P(texture|the) = {}^{1}/_{4}$

$\boxed{P(white|is) = {}^{1}/_{3}}$

$P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}$

$P(</s > |written) = {}^{1}/_{1}$

$\boxed{P(drawn|is) = {}^{1}/_{3}}$

$P(of|texture) = {}^{1}/_{1}$

$P(</s > |white) = {}^{1}/_{1}$

$P(is|book) = {}^{1}/_{1}$

$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$

$P(</s > |drawn) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

- Predict the next word in text:

My friend is written.

Statistical language modeling

$P(the| <s>) = {}^{3}/_{3}$

$P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}$

$P(</s> |written) = {}^{1}/_{1}$

$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$

$P(is|paint) = {}^{2}/_{2}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s> |drawn) = {}^{1}/_{1}$

$P(texture|the) = {}^{1}/_{4}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(white|is) = {}^{1}/_{3}$

$P(</s> |white) = {}^{1}/_{1}$

- Compute the probability of sentences:

D = <s> the book is white </s>

$P(D) = P(the| <s>)P(book|the)$
$P(is|book)P(white|is)\ P(</s> |is)$

$P(D) = {}^{3}/_{3} \times {}^{1}/_{4} \times {}^{1}/_{1} \times {}^{1}/_{3} \times {}^{1}/_{1} = 0.083$

Statistical language modeling

$$\boxed{P(the| < s >) = {}^{3}\!/_{3}} \qquad P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}\!/_{4} \qquad P(is|book) = {}^{1}\!/_{1}$$

$$P(written|is) = {}^{1}\!/_{3} \qquad P(</s > |written) = {}^{1}\!/_{1} \qquad P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}\!/_{4}$$

$$P(is|paint) = {}^{2}\!/_{2} \qquad P(drawn|is) = {}^{1}\!/_{3} \qquad P(</s > |drawn) = {}^{1}\!/_{1}$$

$$P(texture|the) = {}^{1}\!/_{4} \qquad P(of|texture) = {}^{1}\!/_{1} \qquad P(the|of) = {}^{1}\!/_{1}$$

$$P(white|is) = {}^{1}\!/_{3} \qquad P(</s > |white) = {}^{1}\!/_{1}$$

- Generate a sample text (the Shannon visualization method):
  - Choose a random bigram (<s>, w) according to its probability
  - Now choose a random bigram (w, x) according to its probability
  - And so on until we choose </s>

<s>

Statistical language modeling

$P(the|<s>) = {}^{3}/_{3}$

$\boxed{P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}}$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}$

$P(</s>|written) = {}^{1}/_{1}$

$\boxed{P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}}$

$P(is|paint) = {}^{2}/_{2}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s>|drawn) = {}^{1}/_{1}$

$\boxed{P(texture|the) = {}^{1}/_{4}}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(white|is) = {}^{1}/_{3}$

$P(</s>|white) = {}^{1}/_{1}$

- Generate a sample text (the Shannon visualization method):
  - Choose a random bigram (<s>, w) according to its probability
  - Now choose a random bigram (w, x) according to its probability
  - And so on until we choose </s>

<s> the

Statistical language modeling

$P(the| < s >) = {}^{3}/_{3}$

$P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}$

$P(</s > |written) = {}^{1}/_{1}$

$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$

$\boxed{P(is|paint) = {}^{2}/_{2}}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s > |drawn) = {}^{1}/_{1}$

$P(texture|the) = {}^{1}/_{4}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(white|is) = {}^{1}/_{3}$

$P(</s > |white) = {}^{1}/_{1}$

- Generate a sample text (the Shannon visualization method):
  - Choose a random bigram (<s>, w) according to its probability
  - Now choose a random bigram (w, x) according to its probability
  - And so on until we choose </s>

<s> the paint

Statistical language modeling

$P(the| < s >) = {}^{3}/_{3}$

$P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}$

$P(</s > |written) = {}^{1}/_{1}$

$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$

$P(is|paint) = {}^{2}/_{2}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s > |drawn) = {}^{1}/_{1}$

$P(texture|the) = {}^{1}/_{4}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(white|is) = {}^{1}/_{3}$

$P(</s > |white) = {}^{1}/_{1}$

- Generate a sample text (the Shannon visualization method):
  - Choose a random bigram (<s>, w) according to its probability
  - Now choose a random bigram (w, x) according to its probability
  - And so on until we choose </s>

<s> the paint is written </s>

Language models

- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

Challenges of statistical language modeling

- Out of Vocabulary (OOV) words
- Zero probabilities

Out of Vocabulary (OOV) words

- When some words/terms in test set have never seen before
- Two common solutions
  - Make the vocabulary as closed (no new words in test set)
    1. Choose a vocabulary (word list) that is fixed in advance
    2. Convert the other tokens into <UNK>
    3. Estimate the probabilities for <UNK>

Can you please come here?

Can you please <unk> here?

Vocabulary (word list)
can
you
please
here

Out of Vocabulary (OOV) words

- When some words/terms in test set have never seen before
- Two common solutions
  - Make the vocabulary as closed (no new words in test set)
    1. Choose a vocabulary (word list) that is fixed in advance
    2. Convert the other tokens into <UNK>
    3. Estimate the probabilities for <UNK>
  - Replacing words in the training data by <UNK> based on their frequency

Zero probabilities

- A word appear after a word they never appeared after in training
  - It’s not unknown words

Can you please come here?

$P(you|come) = 0$

- Laplace smoothing
  - To add one to all the bigram counts, before we normalize them into probabilities

Statistical language modeling

2-gram

- D_1 = <s> the book is written </s>
- D_2 = <s> the paint is drawn </s>
- D_3 = <s> the texture of the paint is white </s>

$P(the| < s >) = {}^{3}/_{3}$

$P(book|the) = \frac{c(the\ book)}{c(the)} = {}^{1}/_{4}$

$P(is|book) = {}^{1}/_{1}$

$P(written|is) = {}^{1}/_{3}\ {}^{2}/_{4}$

$P(</s > |written) = {}^{1}/_{1}$

$P(paint|the) = \frac{c(the\ paint)}{c(the)} = {}^{2}/_{4}$

$P(is|paint) = {}^{2}/_{2}$

$P(drawn|is) = {}^{1}/_{3}$

$P(</s > |drawn) = {}^{1}/_{1}$

$P(texture|the) = {}^{1}/_{4}\ {}^{2}/_{5}$

$P(of|texture) = {}^{1}/_{1}$

$P(the|of) = {}^{1}/_{1}$

$P(whilte|is) = {}^{1}/_{3}$

$P(</s > |white) = {}^{1}/_{1}$

$\boxed{P(paint|texture) = {}^{0}/_{1}\ {}^{1}/_{2}}$

Language models

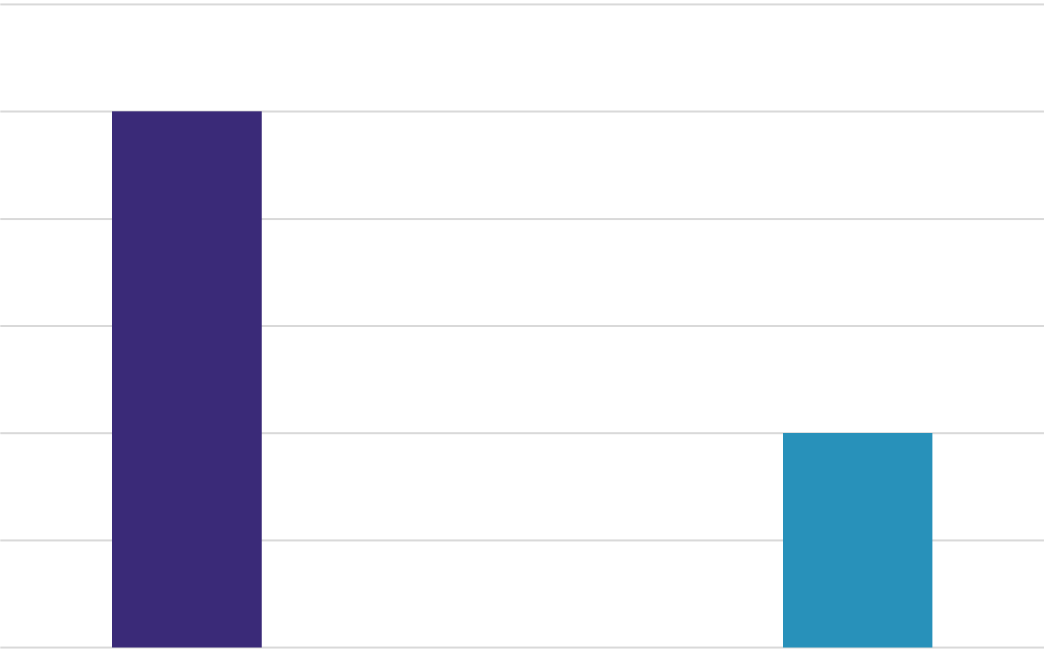
- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

Evaluation of language models

- There are two main approaches for evaluating language models:
  - Extrinsic evaluation
  - Intrinsic evaluation

Language model A

Language model B

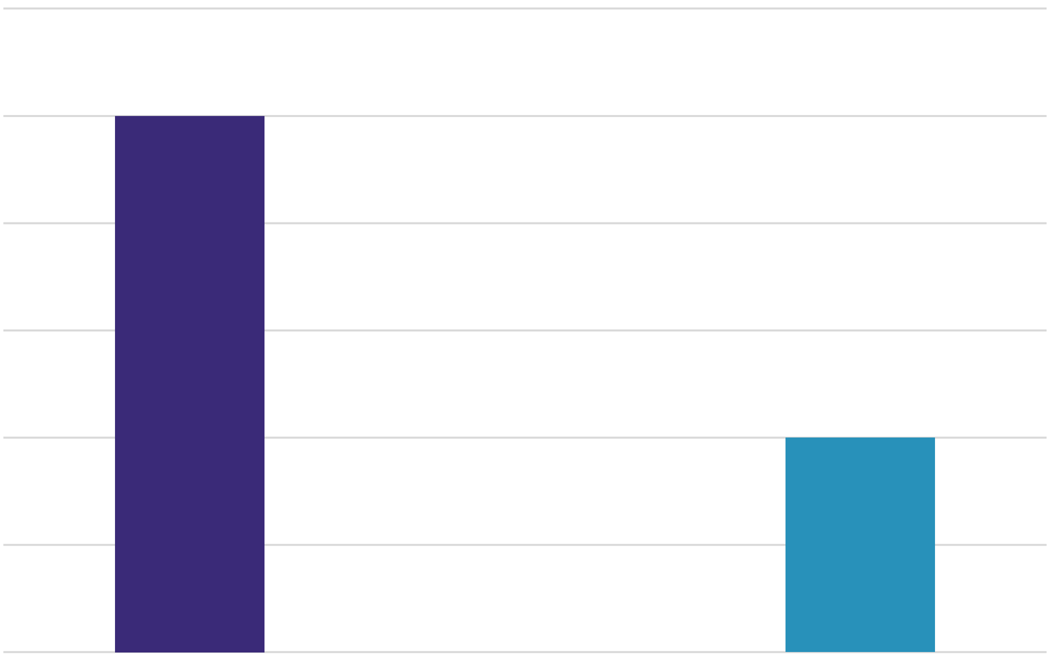


Evaluation of language models

- Intrinsic evaluation
  - Whichever model assigns a higher probability to the test set (meaning it more accurately predicts the test set) is a better model
  - Given two probabilistic models, the better model is the one that has a tighter fit to the test data or that better predicts the details of the test data, and hence will assign a higher probability to the test data

Language model A

Language model B



Evaluation of language models

- Perplexity
  - Measurement of how well a probability distribution or probability model predicts a sample
  - A ***low perplexity*** indicates the probability distribution is good at predicting the sample

$$Perplexity(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

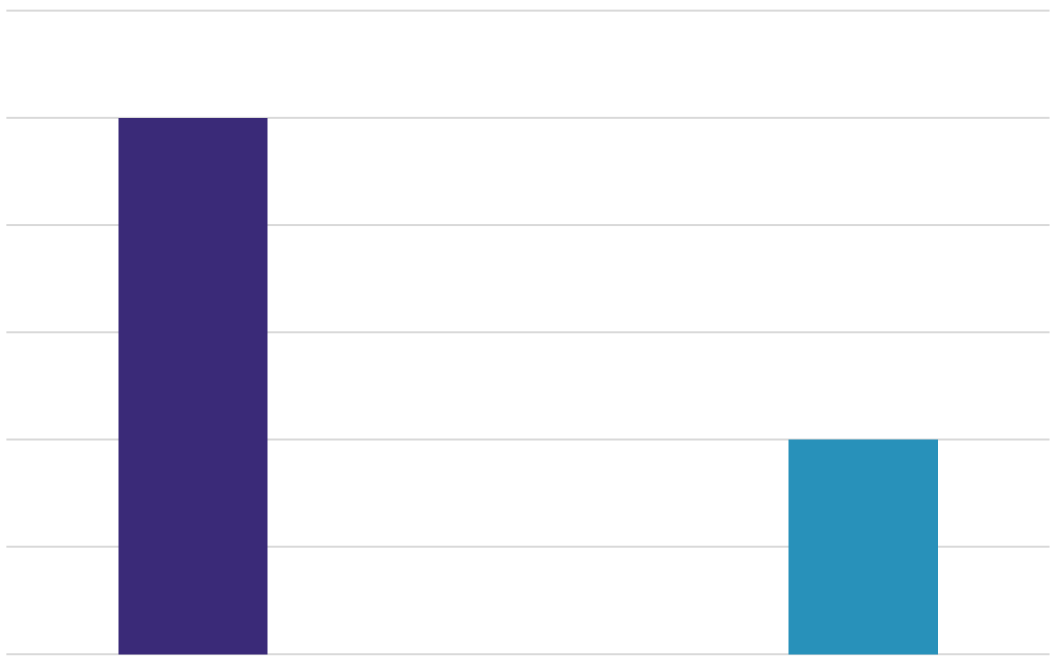
$$W = w_1 w_2 \ldots w_N$$

Evaluation of language models

$$Perplexity(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

$W = w_1 w_2 \ldots w_N$

W = Can you please come here?



Language model A	$P(you\|can) = 0.5$	$P(please\|you) = 0.7$
Language model B	$P(you\|can) = 0.02$	$P(please\|you) = 0.26$

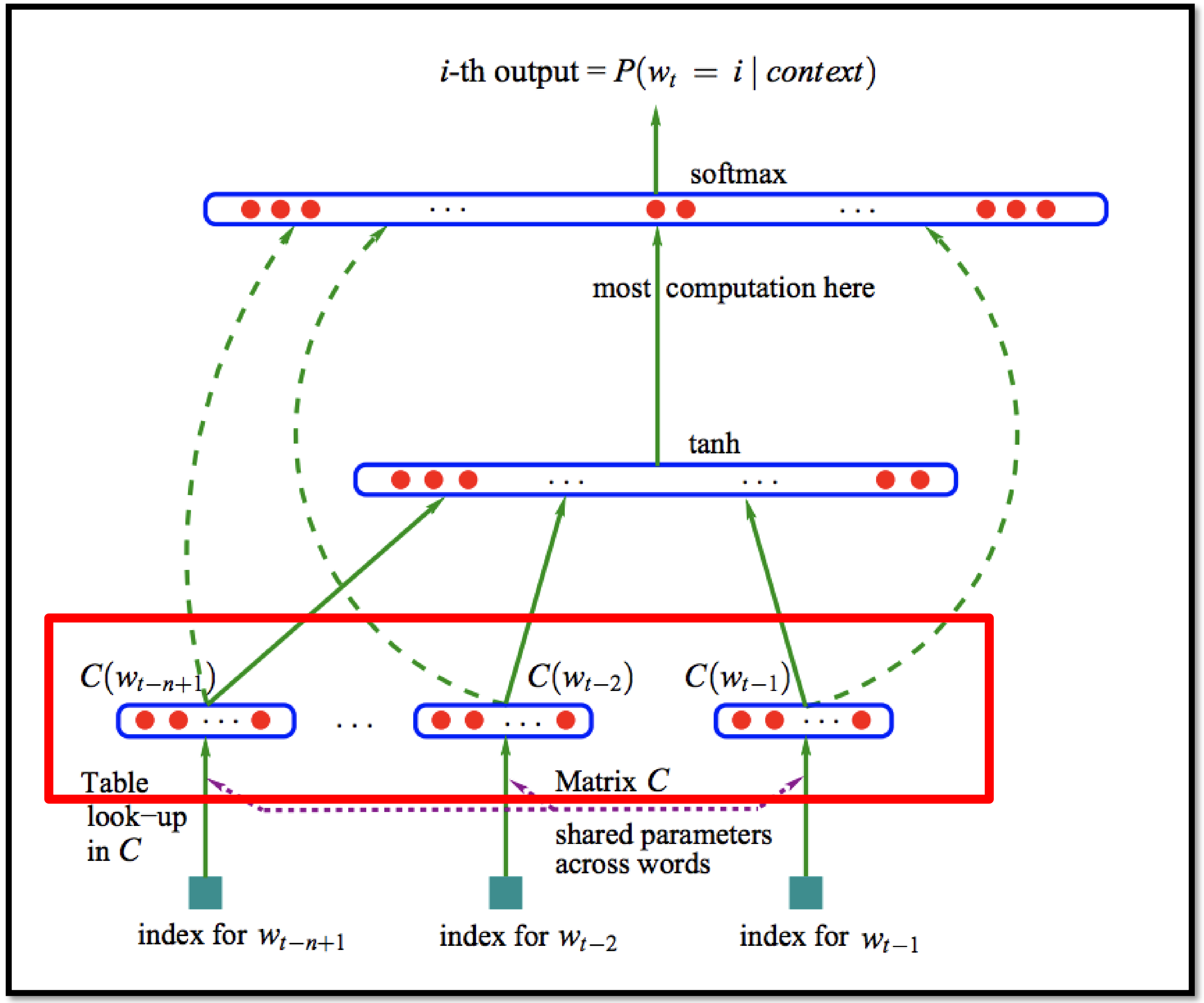
Language models

- What is language modeling?
- Why language modeling is critical in NLP?
- Statistical language modeling
- Challenges of statistical language modeling
- Evaluation of language models
- Neural language models

Neural language models

- Language models based on neural networks
- Neural language models advantages:
  - Can handle much ***longer histories***
  - Can ***generalize*** over contexts of similar words
  - Has much ***higher predictive accuracy*** than an n-gram language model
  - Don’t need ***smoothing***
- Neural language models are ***too slower*** than traditional language models to train

Neural language models



Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. The journal of machine learning research, 3, 1137-1155.

Can you please come here?

you	–	come	please	was	here	she	can	time	just
0.17	0.04	0.06	0.10	0.09	0.26	0.08	0.05	0.04	0.11

you	–	come	please	was	here	she	can	time	just
1.7	0.3	0.7	1.1	1	2.1	0.98	0.43	0.23	1.21

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

can		
0.1	-0.2	-0.4

you		
0.3	0.1	0.9

please		
-0.7	0.2	-0.7

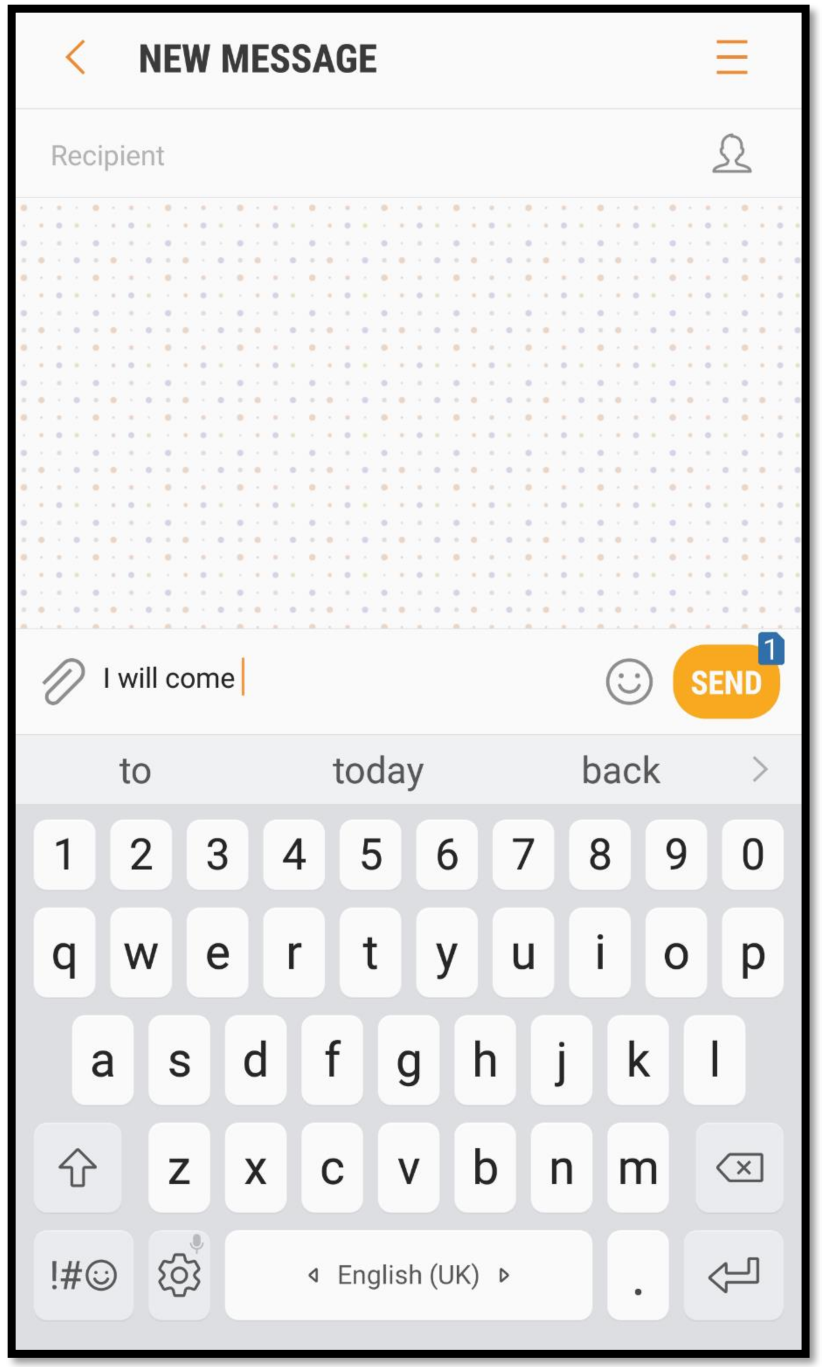
Neural language models

- Character based language models

C

- Advantages
  - The model is smaller (97 English-language characters in common, includes all punctuation marks)
  - Flexibility in handling any words
- Disadvantages
  - Lack of semantic content of the input (characters are meaningless)
  - Longer sequences increase computational expense

Summary



Can you please come here?

Summary

2-gram language model

come here?

$P(here|can\ you\ please\ come) \rightarrow P(here|come)$

Can you please <unk> here?

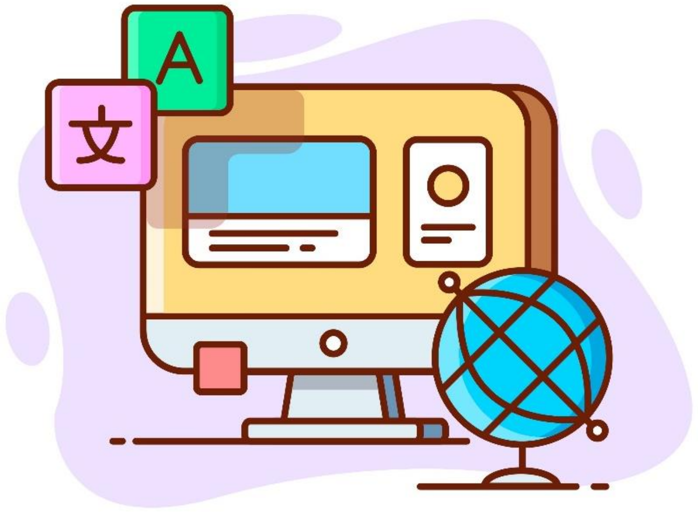
Can you please come here?

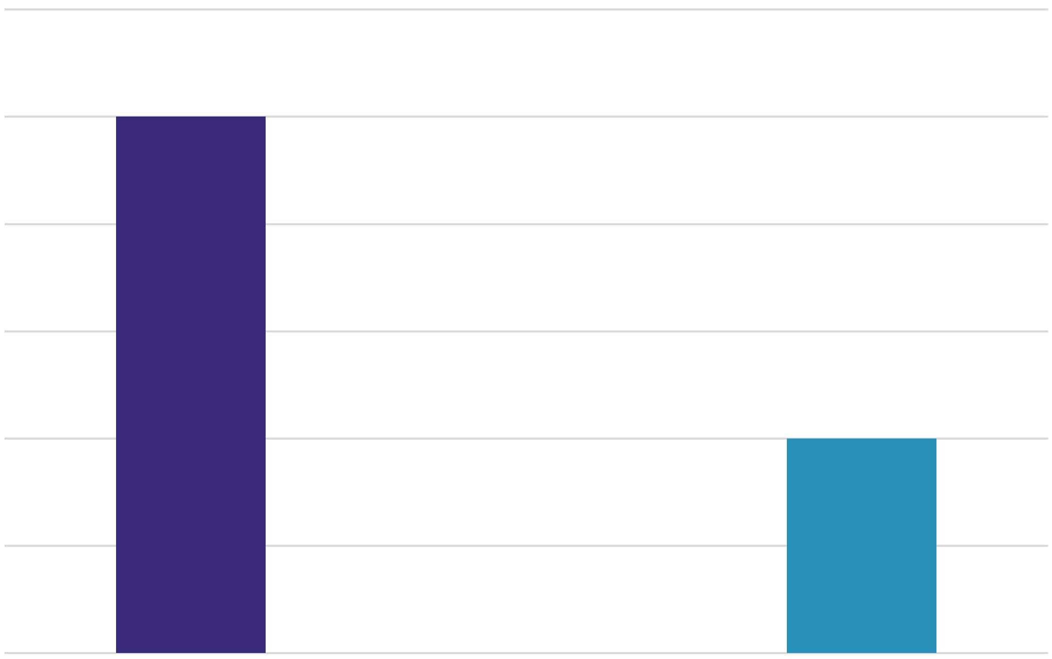
$P(you|come) = 0$

Summary

Language model A

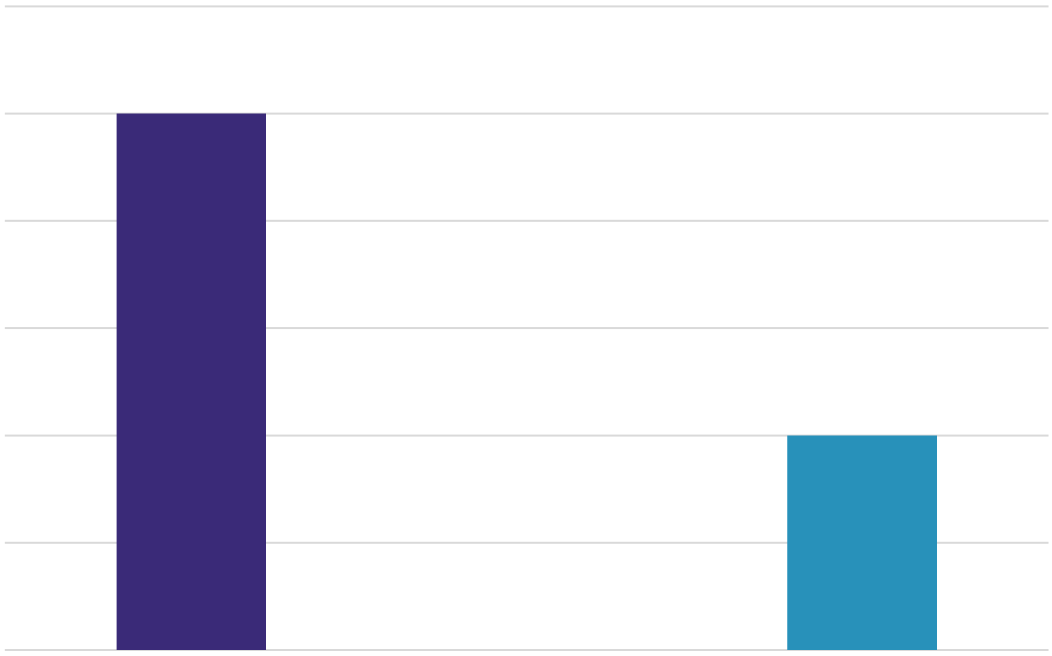
Language model B



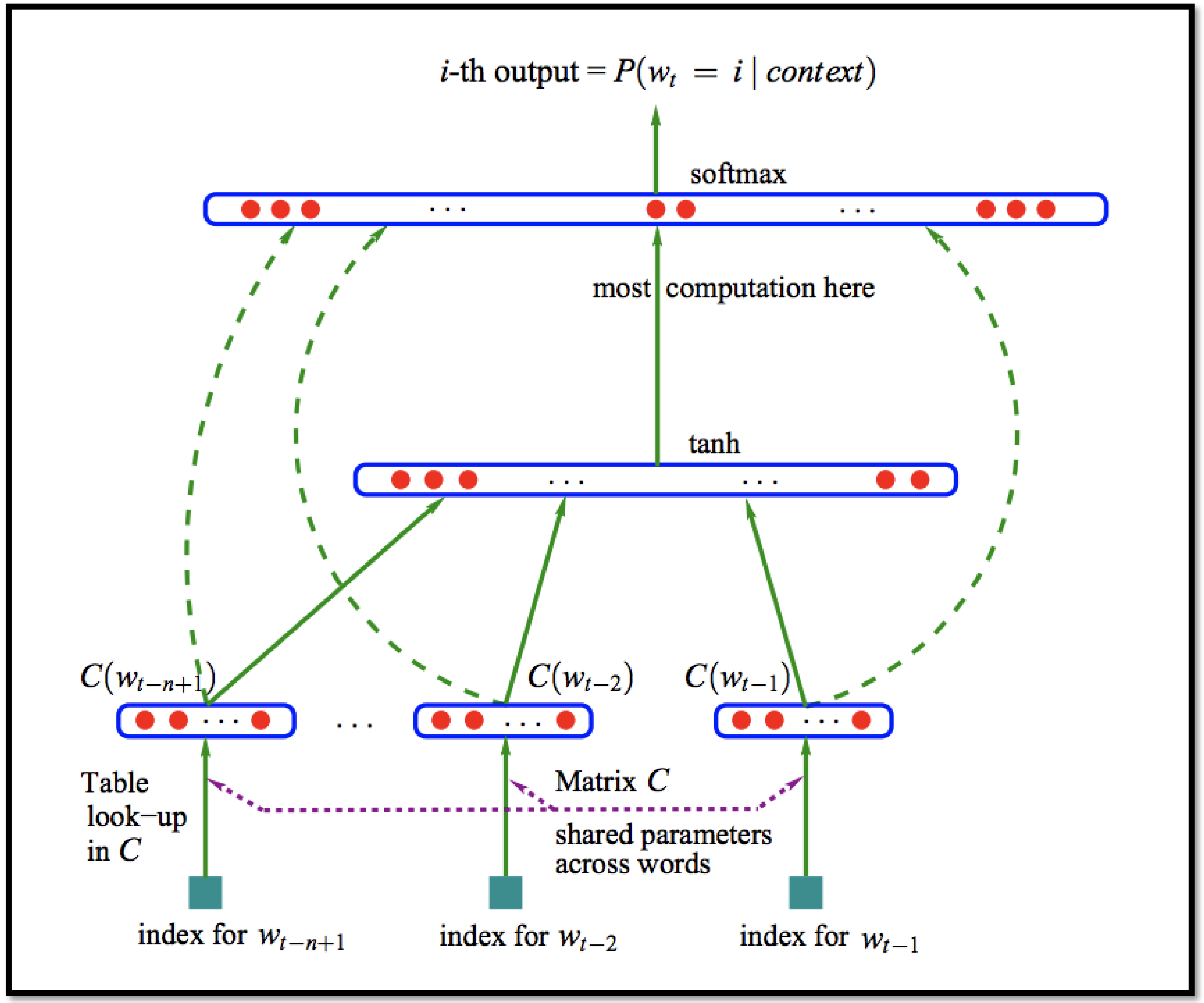


Language model A

Language model B



Summary



you	I	come	please	was	here	she	can	time	just
0.17	0.04	0.06	0.10	0.09	0.26	0.08	0.05	0.04	0.11

can		
0.1	-0.2	-0.4

you		
0.3	0.1	0.9

please		
-0.7	0.2	-0.7