CS 333: Natural Language Processing

Fall 2025

Prof. Carolyn Anderson
Wellesley College

Reminders

- Tuesday -> Monday cycle for final two assignments:
  - HW 7 is due on Monday, 11/17
  - HW 8 will be released on Tuesday, 11/18 and due on Monday, 11/24
- My next help hours: Thursday 4-5
- Two CS Colloquia upcoming!

HW 7: Evaluating LLMs

In this assignment, you will practice evaluating LLMs. You'll work with Llama 1B, and test its capabilities on two tasks: Pig Latin decoding, and common-sense reasoning.

The assignment also asks you to do some setup work for HW 8, and to describe your final project topic.

WELLESLEY CS COLLOQUIUM

Professor Lindsey D. Cameron

Wharton School, University of Pennsylvania

Resocializing the Platform: Patchwork Embeddedness and How Workers ReConstitute Digital Platforms

14 NOV 2025 | 3:30 PM | SCI-H105

Snacks will be provided!

WELLESLEY
COMPUTER SCIENCE

?sb129@wellesley.edu
Accessibility and Disability:
accessibility@wellesley.edu

Two ML talks for the price of one:

Co-designing Tools to Measure Student Learning with Machine Learning and Science Education Research

Dr. Kaitlin Gili

Subgroup Validity in Machine Learning for Echocardiogram Data

Cynthia Feeney

Thursday Nov. 20 at 12:45-1pm in H-105

HW 6 Curiosity Points

- Model experimentation:
  - Comparison of non-linear activation functions
  - Experiments with model architecture changes
- Analysis:
  - Dimensionality reduction and visualization of learned embeddings
  - Visualization of model performance
- Literature exploration:
  - Read related work on non-linear activation functions, music understanding, neural network models at Google, class imbalance, and BERT models
- New tasks:
  - Scraping lyrics from new artists
  - Lyric guessing game
  - Song year prediction task

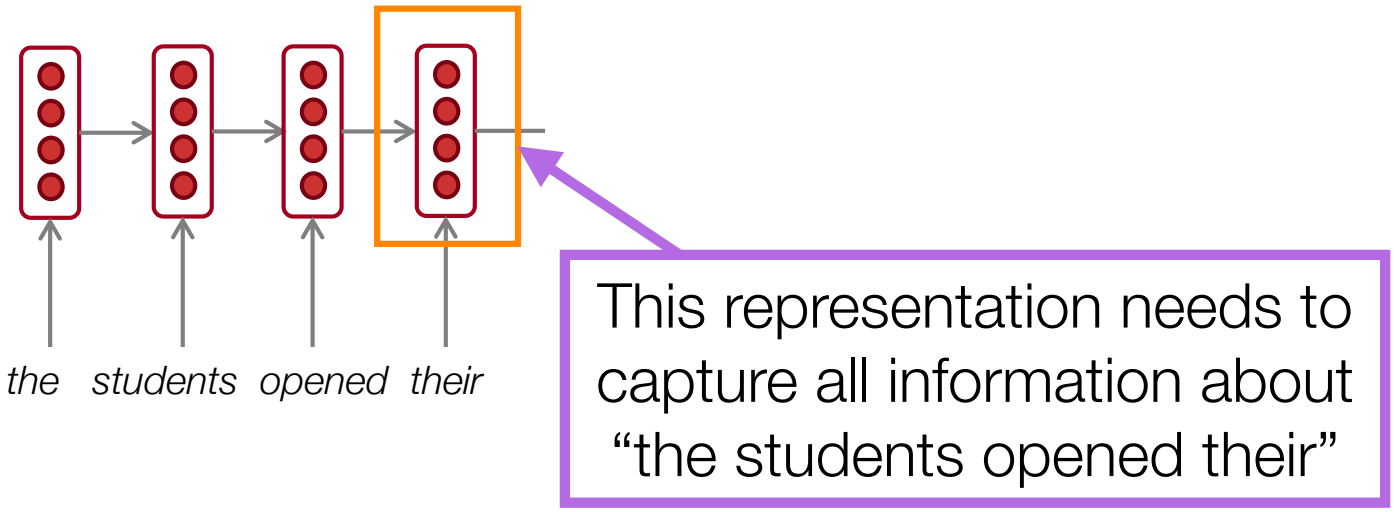
Recap

“you can’t cram the meaning of a whole %&@#&ing sentence into a single $*(&@ing vector!”

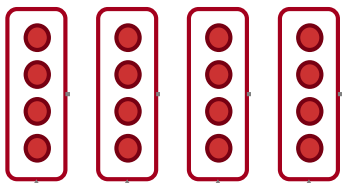
— Ray Mooney (NLP professor at UT Austin)

Slides adapted from Mohit Iyyer

idea: what if we use multiple vectors?



Instead of this, let's try:

the students opened their =  **(all 4 hidden states!)**

Slides adapted from Mohit Iyyer

The solution: **attention**

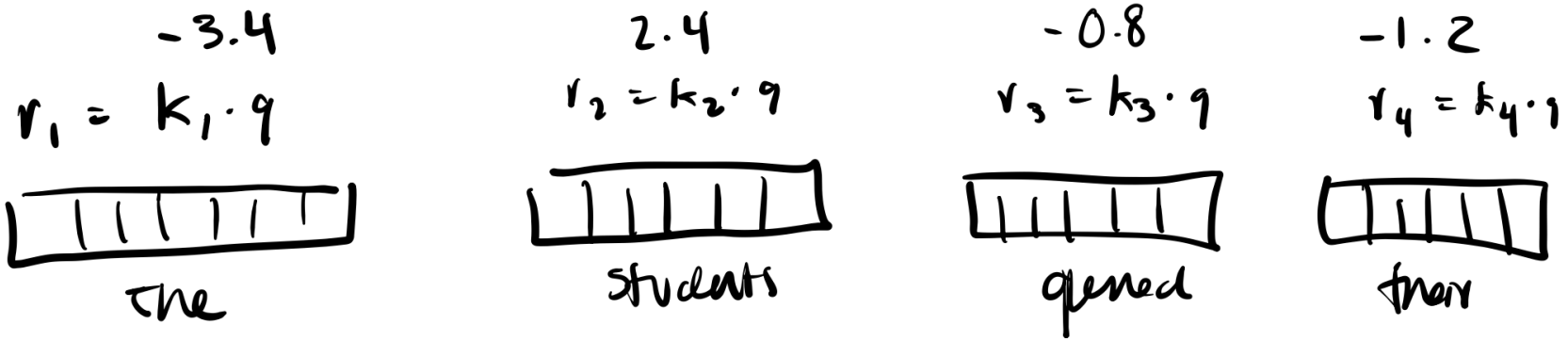
- **Attention mechanisms** (Bahdanau et al., 2015) allow language models to focus on a particular part of the observed context at each time step
  - Originally developed for machine translation, and intuitively similar to *word alignments* between different languages

Slides adapted from Mohit Iyyer

What Is Attention?

Query: a task-specific vector that is "searching" for important things from the past context.

Step 1) Measure the **similarity** (dot product) between query & key

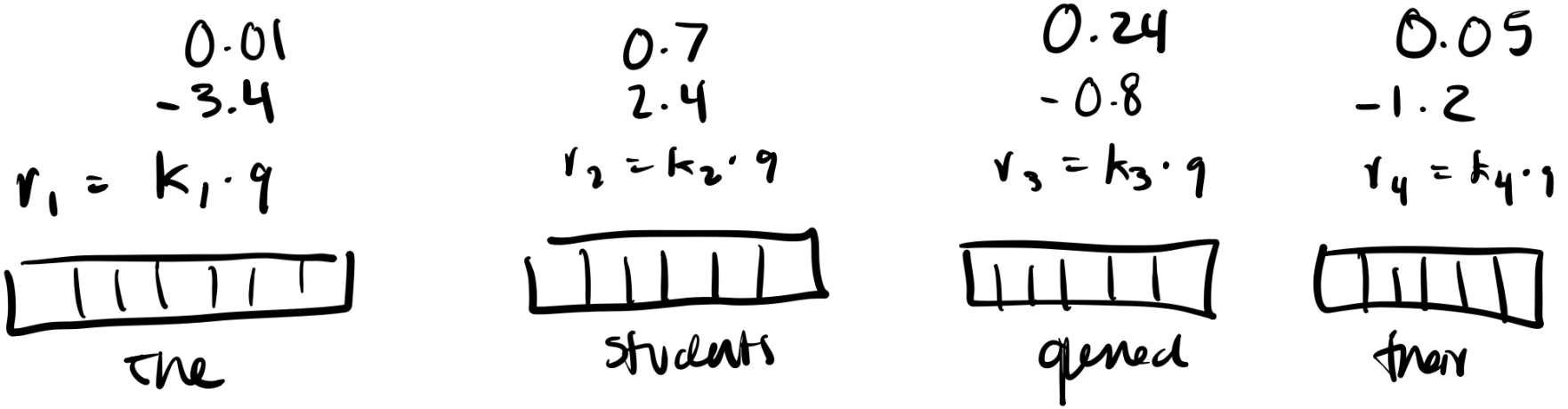


Keys: representations of past context

What Is Attention?

Query: a task-specific vector that is "searching" for important things from the past context.

Step 2: softmax Scale scores to between 0-1



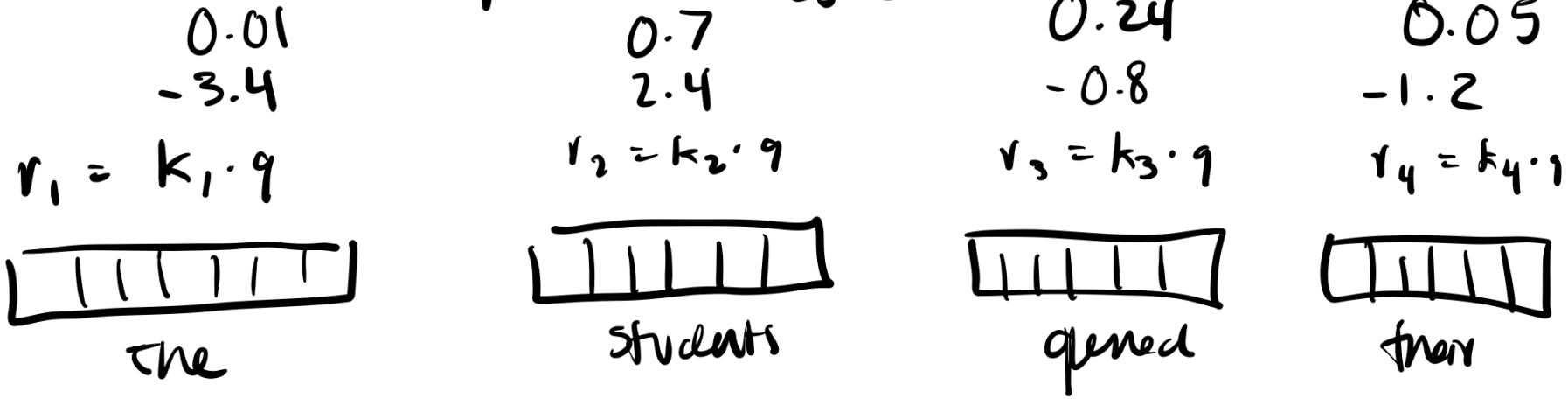
Keys: representations of past context

What Is Attention?

Query: a task-specific vector that is "searching" for important things from the past context.

Step 3: Compute a weighted average of embeddings

Output: a vector of the same dimension as one of the word embeddings.



Keys: representations of past context


Vicki
@vboykis
They don’t tell you this in the paper (well they do but you have to read it like 15 times)
Multiplying
a lot of vectors
a lot of times
with scaled softmax
Attention
imgflip.com
6:20 PM · Feb 22, 2023 · 88.1K Views

Why dot product?

- Dot product provides a measure of similarity between keys and queries.
- But you might be wondering: *why do we want to pay attention to words that are similar to the current word*?

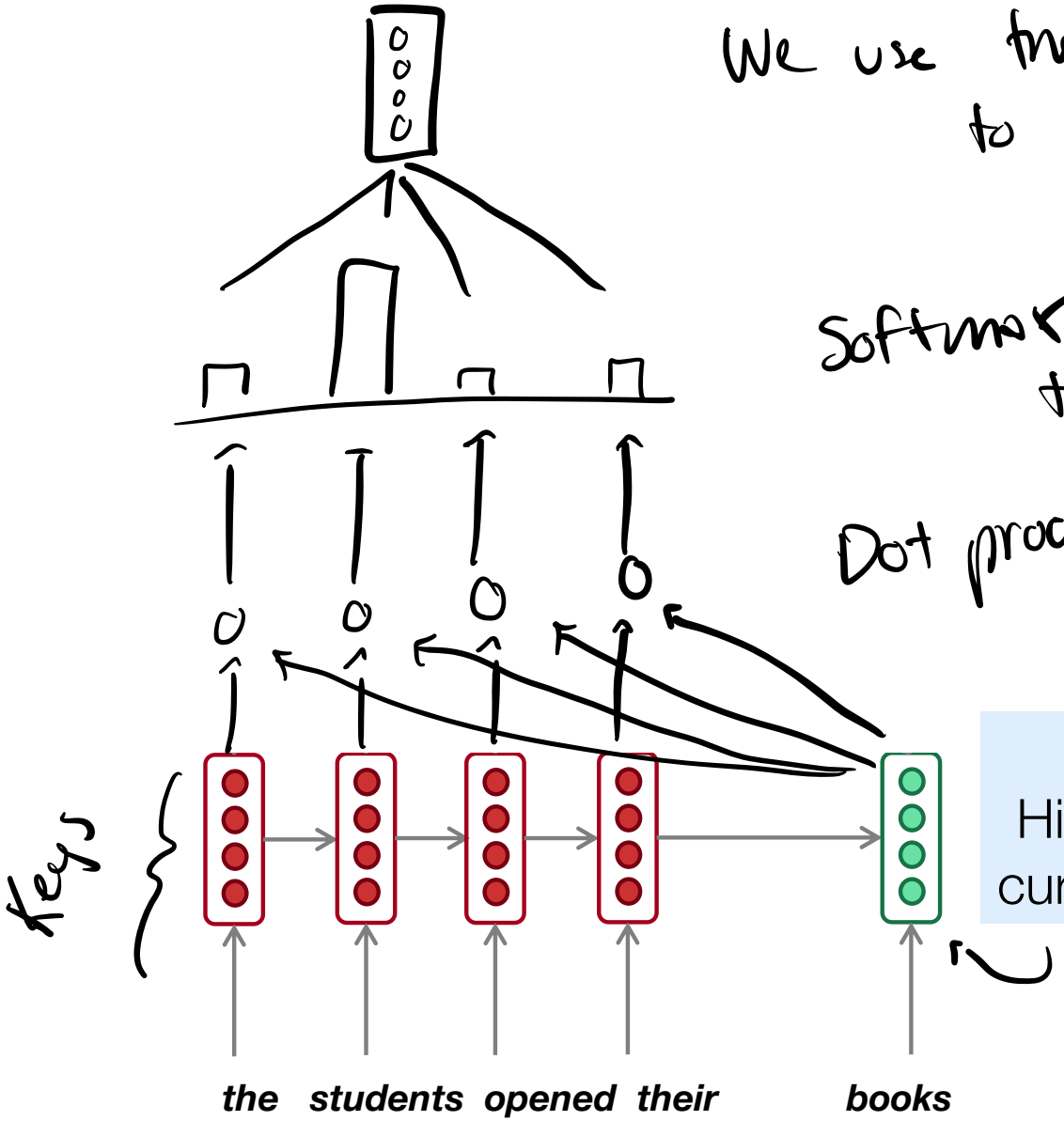
Why dot product?

- Dot product provides a measure of similarity between keys and queries.
- But you might be wondering: *why do we want to pay attention to words that are similar to the current word*?

Consider:

My brother, a chemist, was late yesterday because he missed the bus. When he arrived, he was surprised to find that his lab _____

Attention mechanisms in neural language models



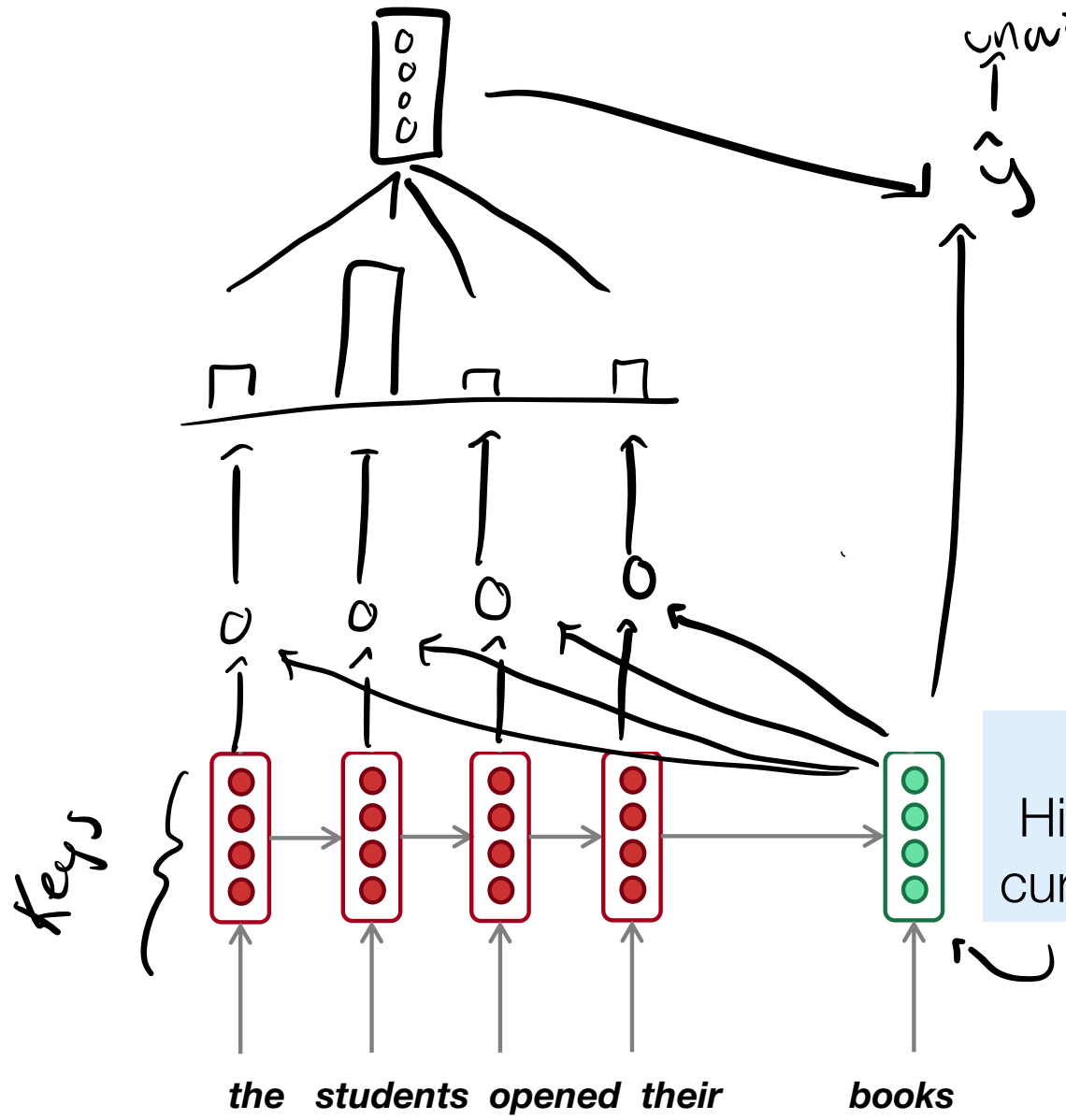
We use the attention distribution to compute a weighted average of the keys

Softmax over the dot products to turn into a probability dist.

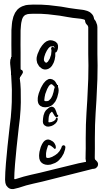
Dot product between $q \cdot k$

Query 1:
Hidden state at current time step

Attention mechanisms in neural language models

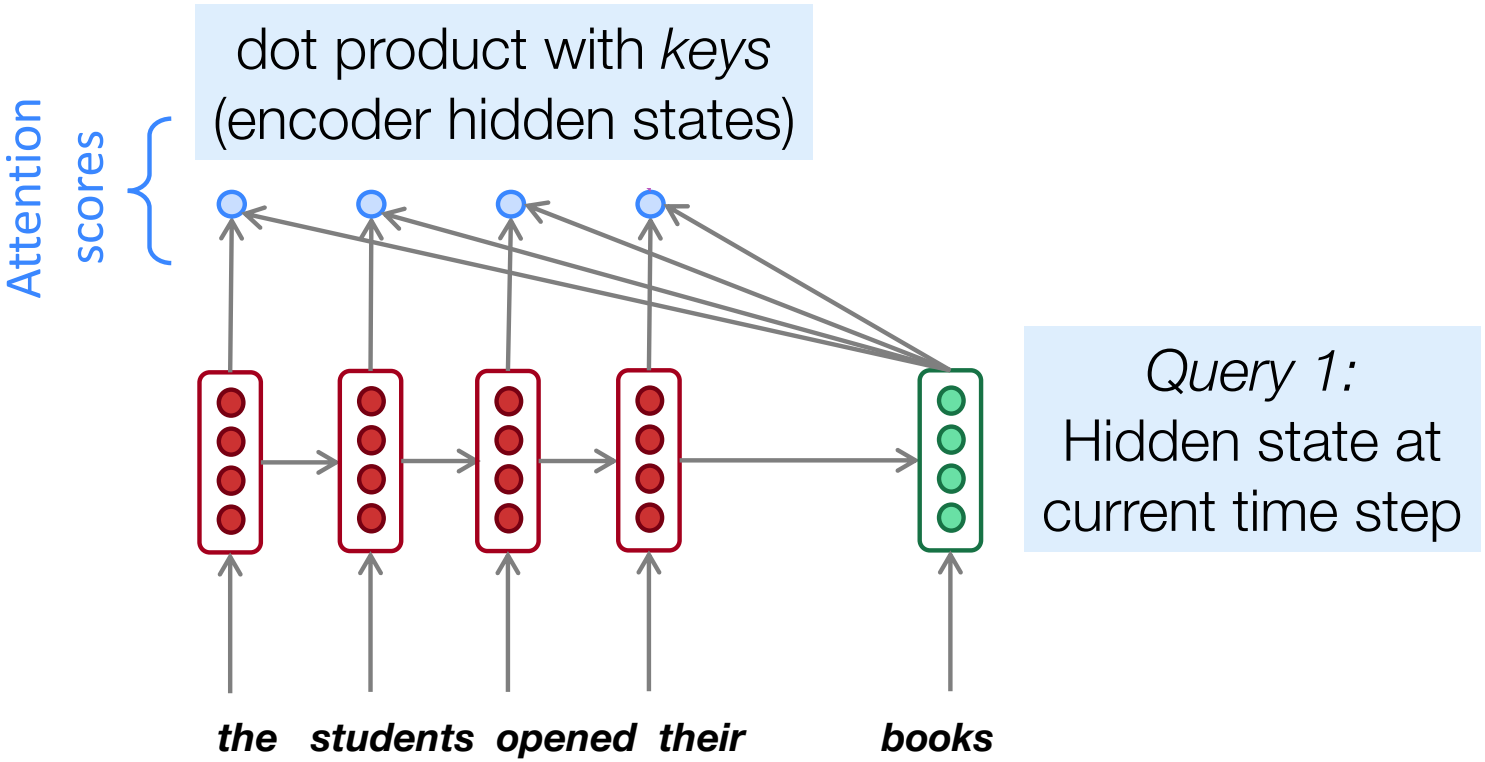


Concatenate (or otherwise compose) the attention output w/ the current word embedding (query), then pass through softmax to predict a next word.

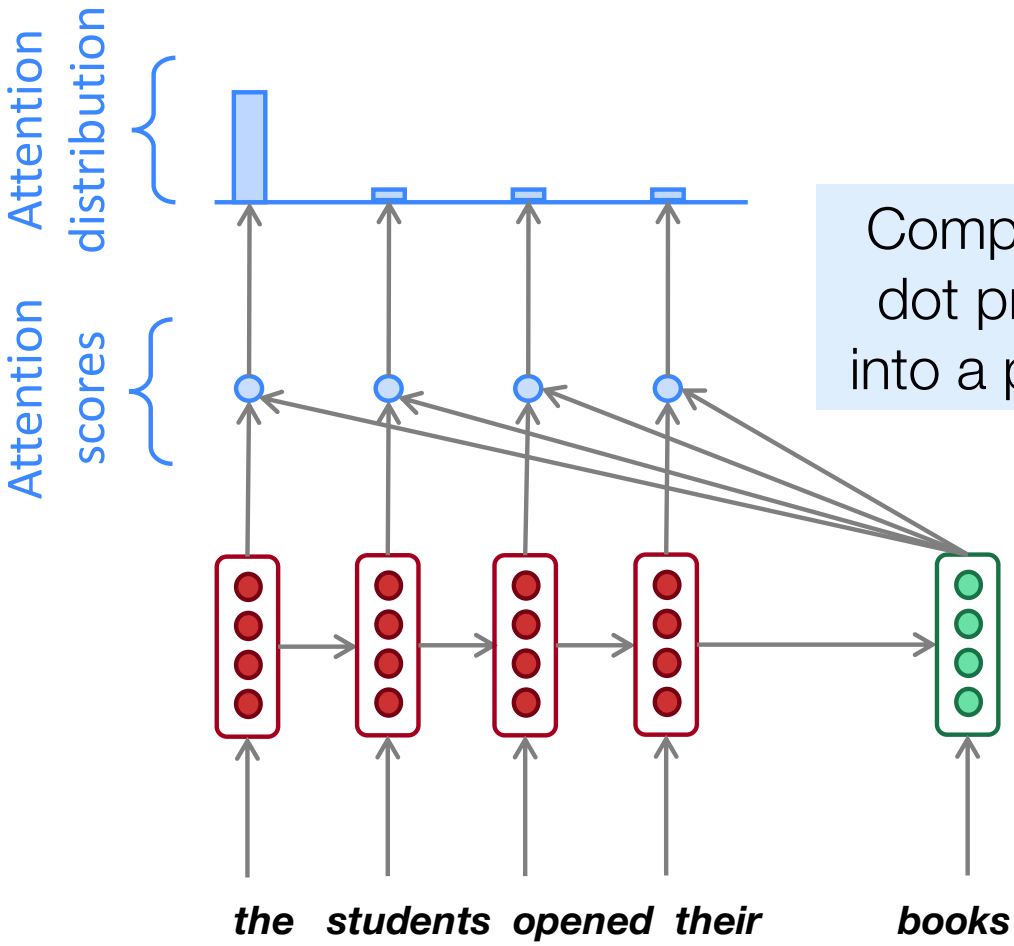


unwillingly is now the query

Attention mechanisms in neural language models

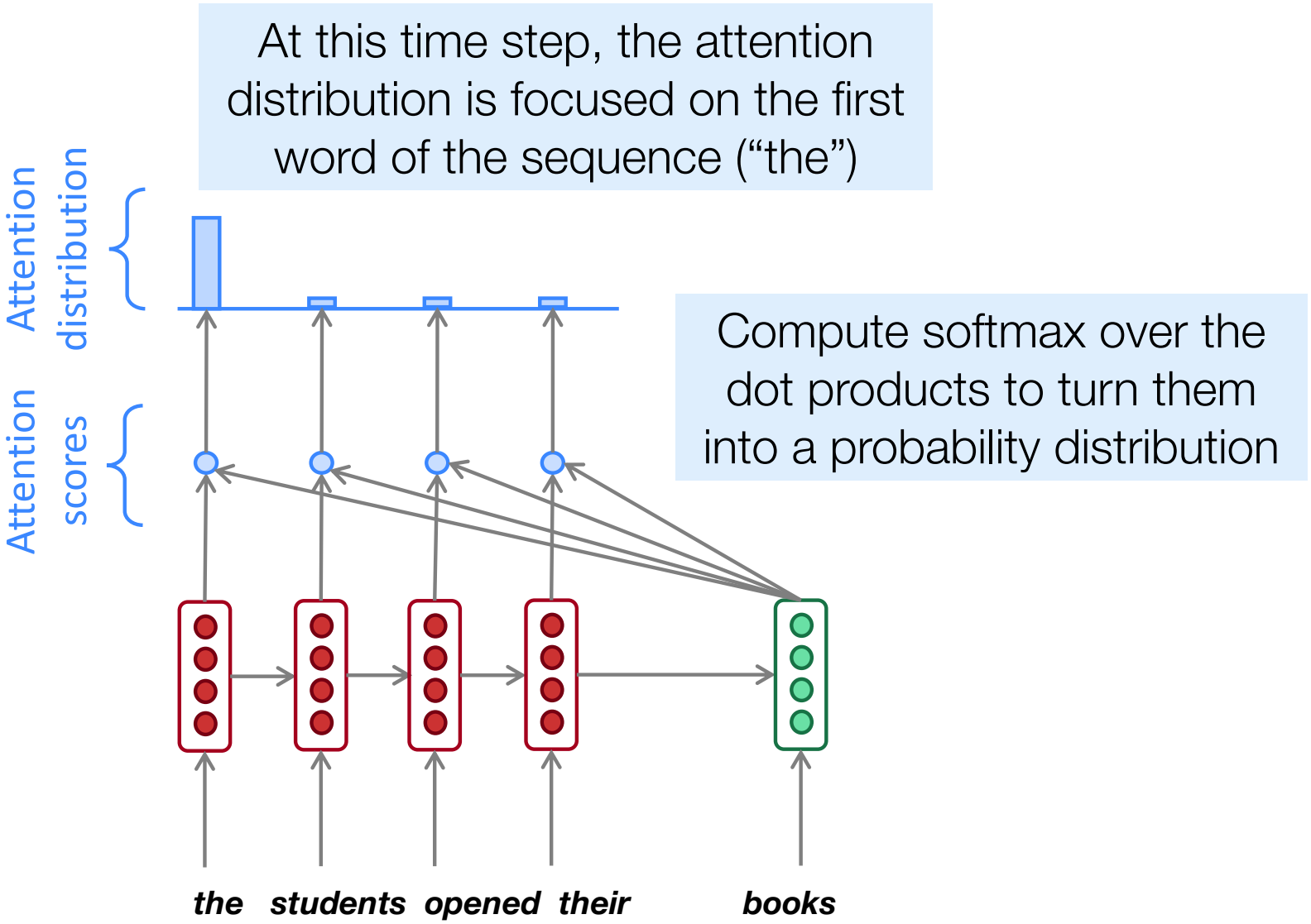


Attention mechanisms in neural language models

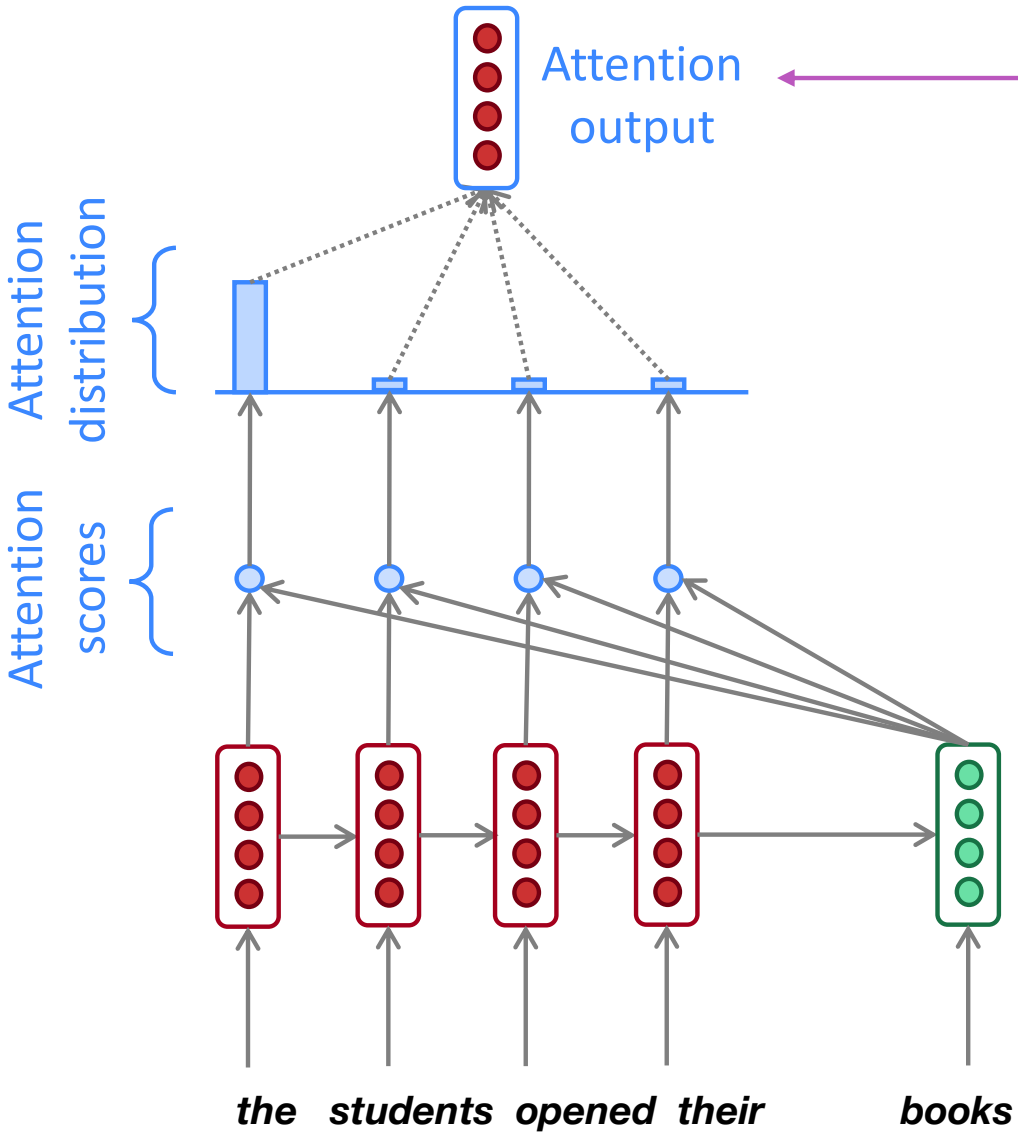


Compute softmax over the dot products to turn them into a probability distribution

Attention mechanisms in neural language models



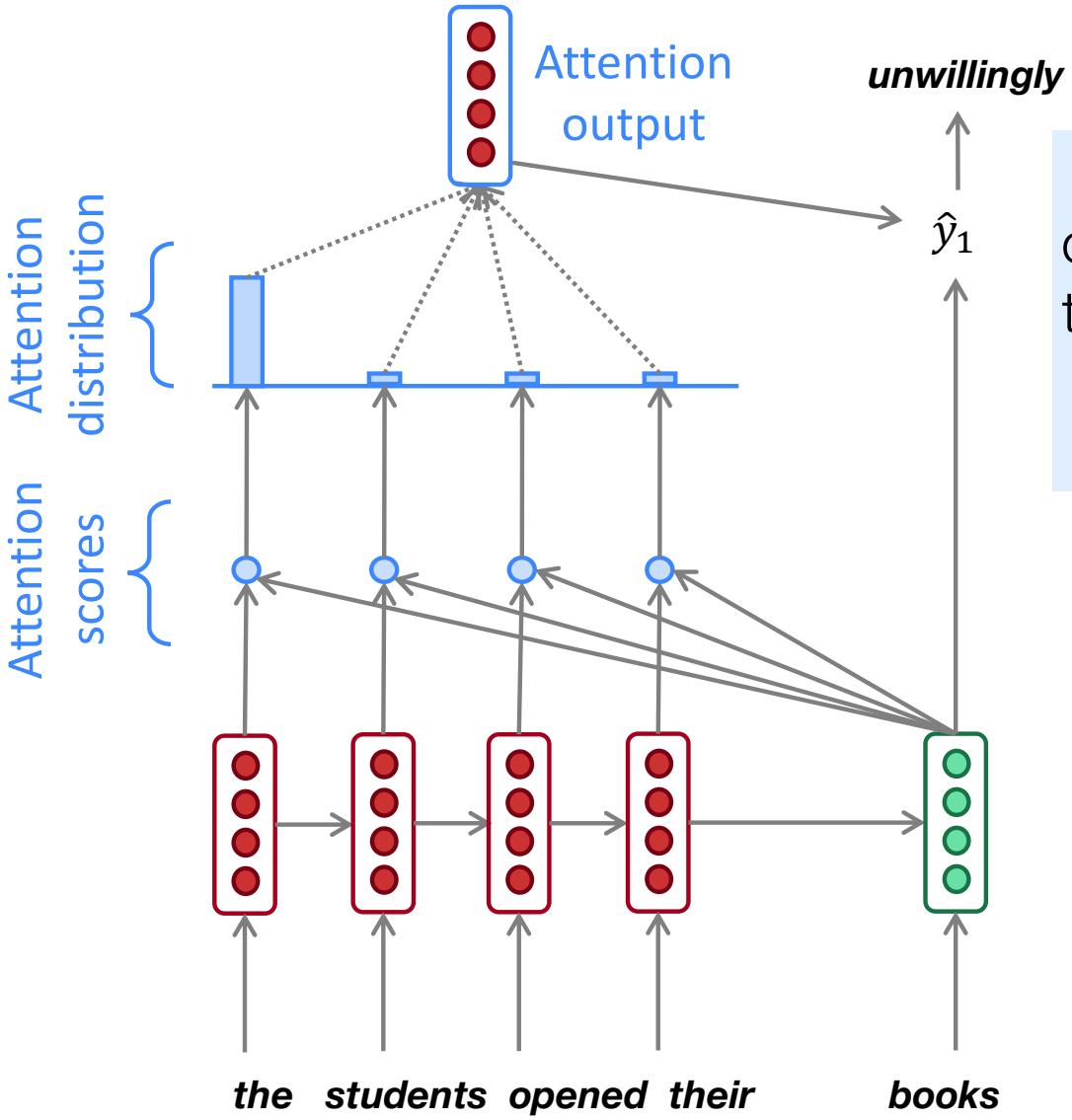
Attention mechanisms in neural language models



We use the attention distribution to compute a weighted average of the hidden states.

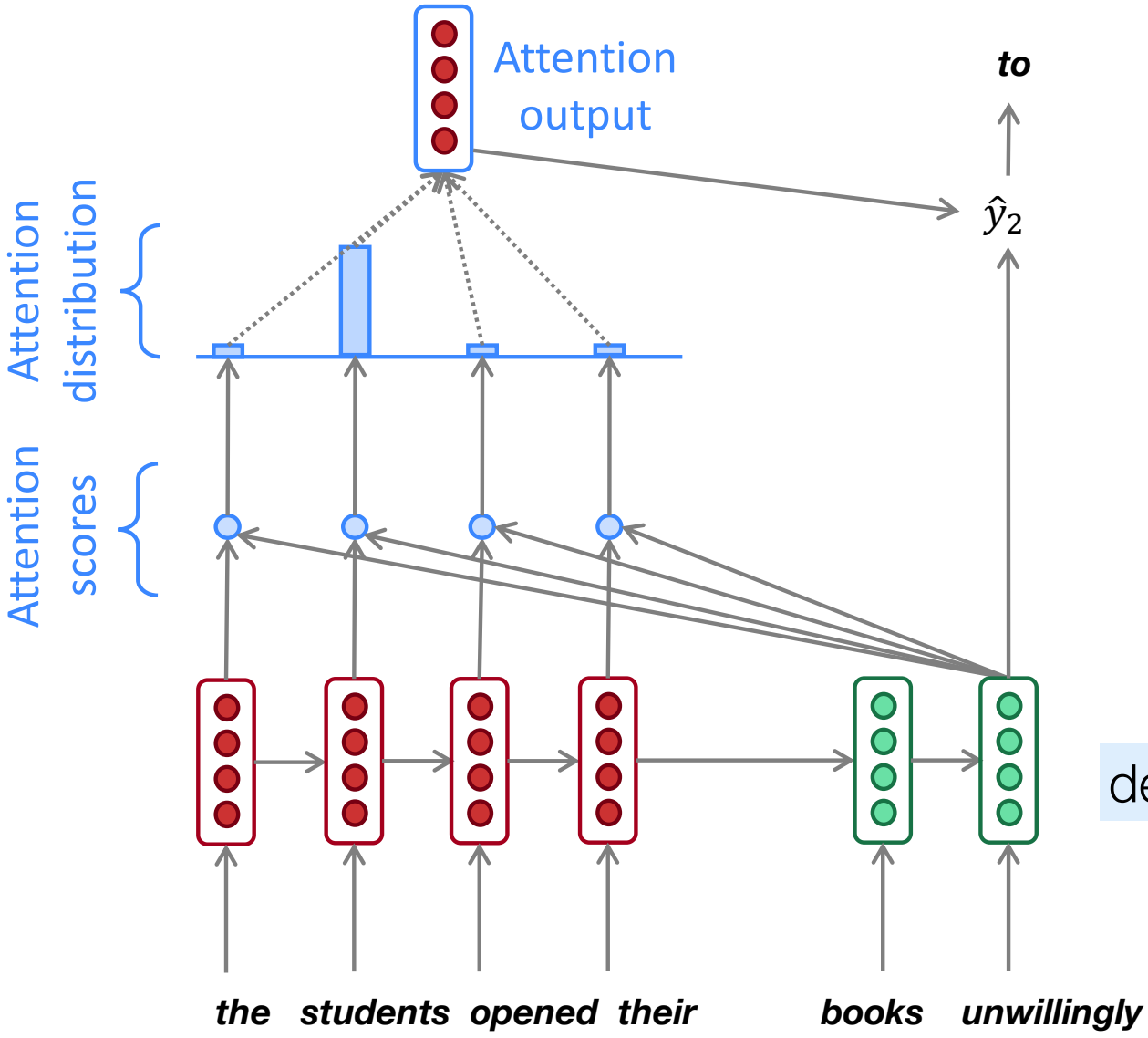
Intuitively, the resulting attention output contains information from hidden states that received high attention scores

Sequence-to-sequence with attention



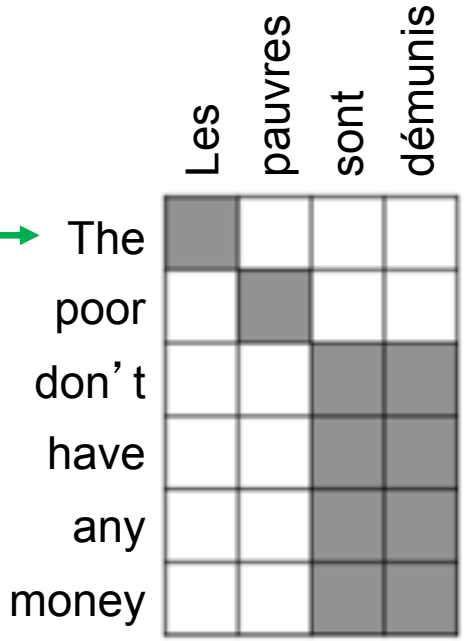
Concatenate (or otherwise compose) the attention output with the current hidden state, then pass through a softmax layer to predict the next word

Sequence-to-sequence with attention



decoder, second time step

- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck
- Attention helps with vanishing gradient problem
  - Provides shortcut to faraway states
- Attention provides some interpretability
  - By inspecting attention distribution, we can see what the decoder was focusing on
  - We get alignment for free!
  - This is cool because we never explicitly trained an alignment system
  - The network just learned alignment by itself



Variants of Attention

- Original formulation: $a(\mathbf{q}, \mathbf{k}) = w_2^T \tanh(W_1[\mathbf{q}; \mathbf{k}])$
- Bilinear product: $a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^T W \mathbf{k}$ Luong et al., 2015
- Dot product: $a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^T \mathbf{k}$ Luong et al., 2015
- Scaled dot product: $a(\mathbf{q}, \mathbf{k}) = \frac{\mathbf{q}^T \mathbf{k}}{\sqrt{|\mathbf{k}|}}$ Vaswani et al., 2017

Self-Attention

Parallelizing Self-Attention