Attention mechanisms

CS 585, Fall 2019

Introduction to Natural Language Processing

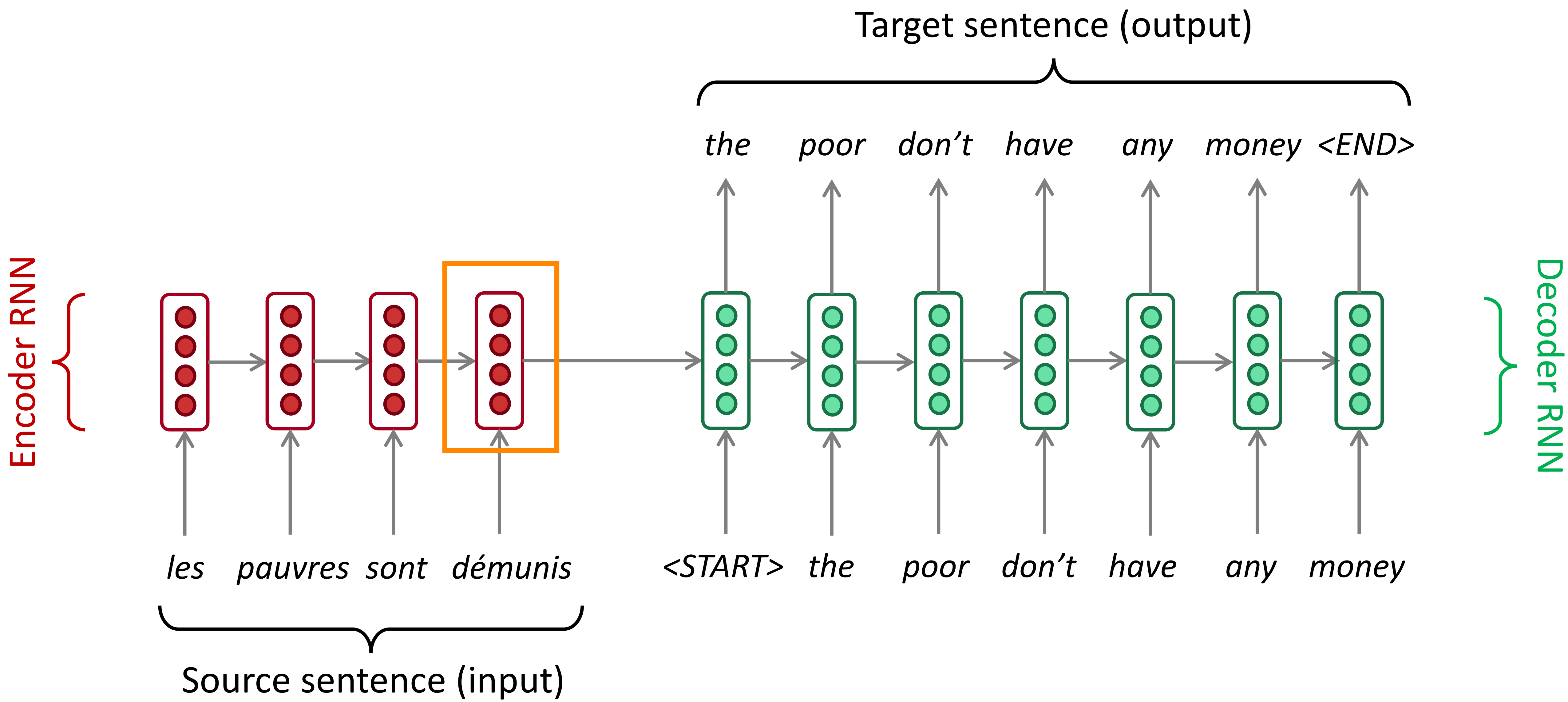
Mohit Iyyer

College of Information and Computer Sciences

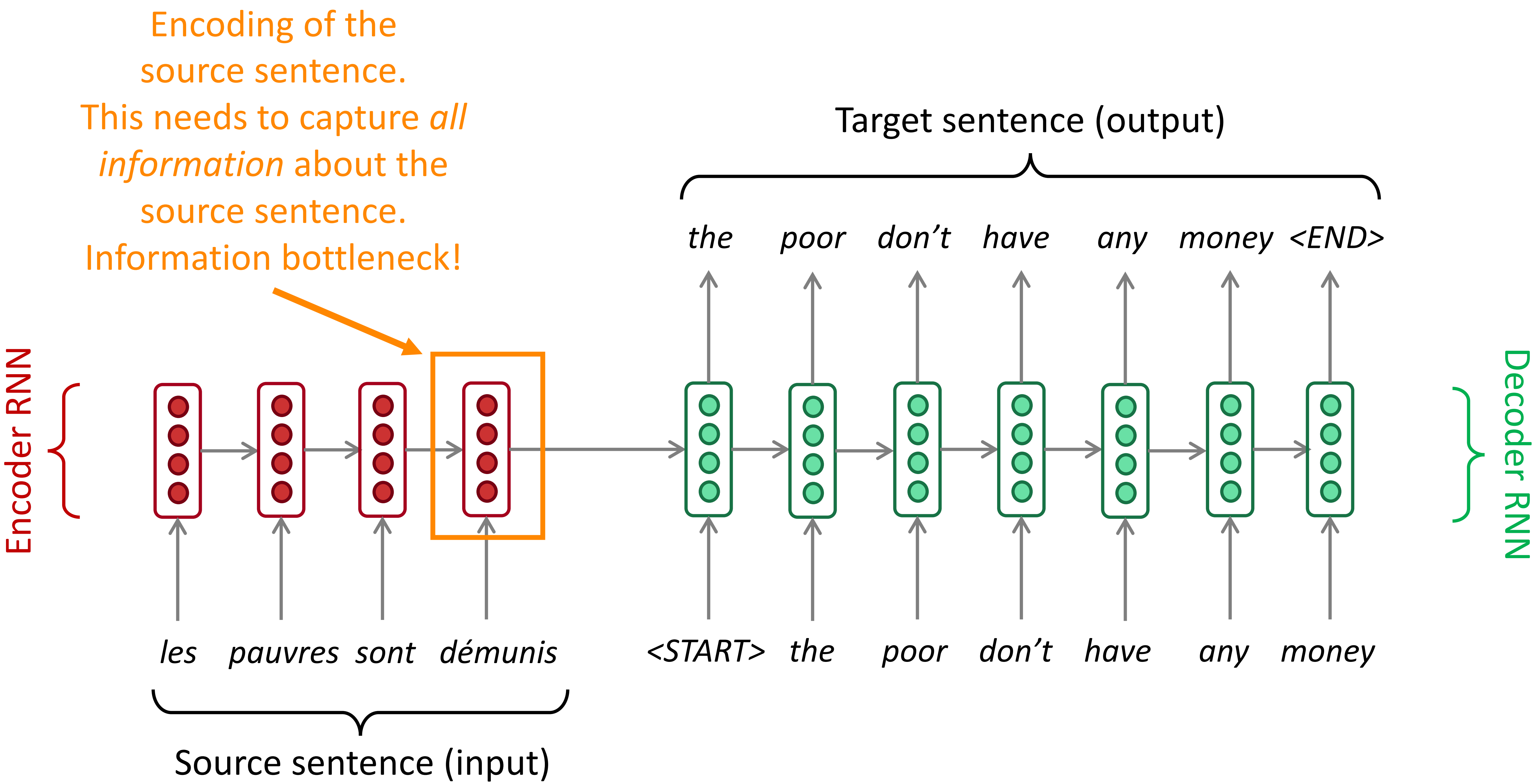
University of Massachusetts Amherst

some slides from Richard Socher

Sequence-to-sequence: the bottleneck problem



Sequence-to-sequence: the bottleneck problem

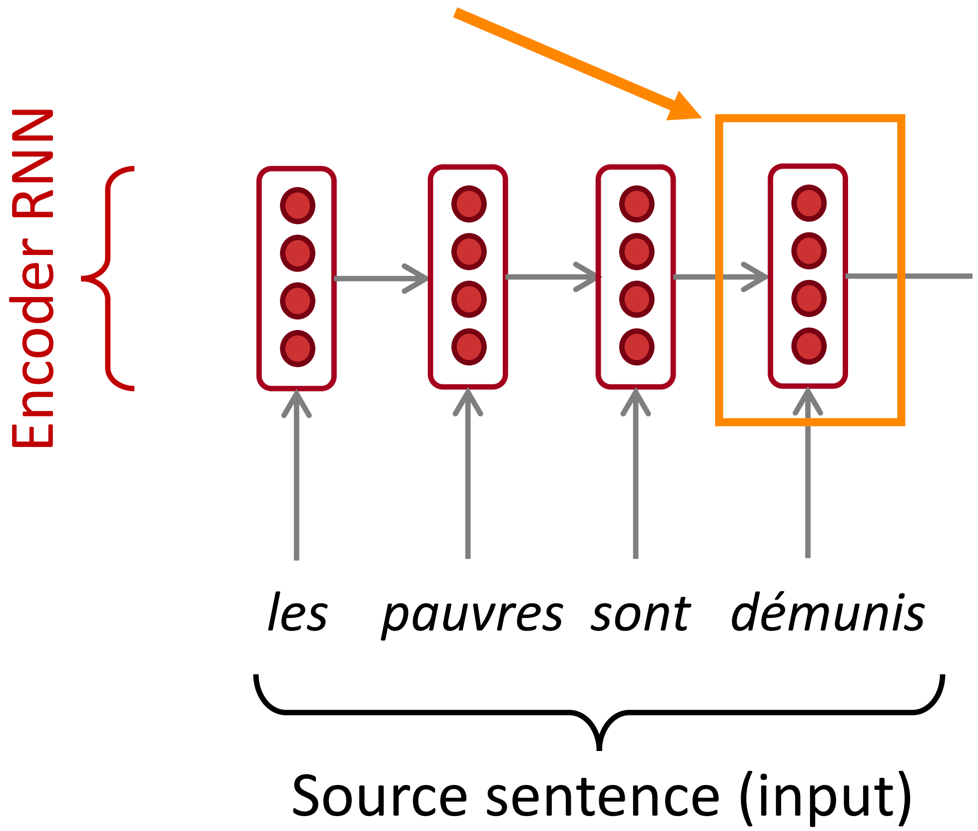


“you can’t cram the meaning of a whole %&@#&ing sentence into a single $*(&@ing vector!”

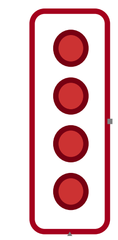
— Ray Mooney (NLP prof at UT Austin)

idea: what if we use multiple vectors?

Encoding of the source sentence. This needs to capture *all information* about the source sentence. Information bottleneck!

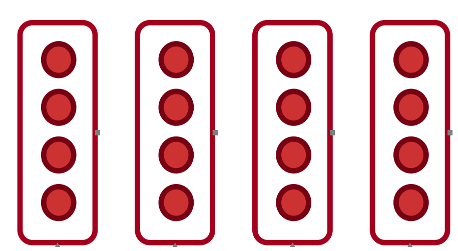


Instead of:

les pauvres sont démunis = 

Let's try:

les pauvres sont démunis =

 (all 4 hidden states!)

The solution: **attention**

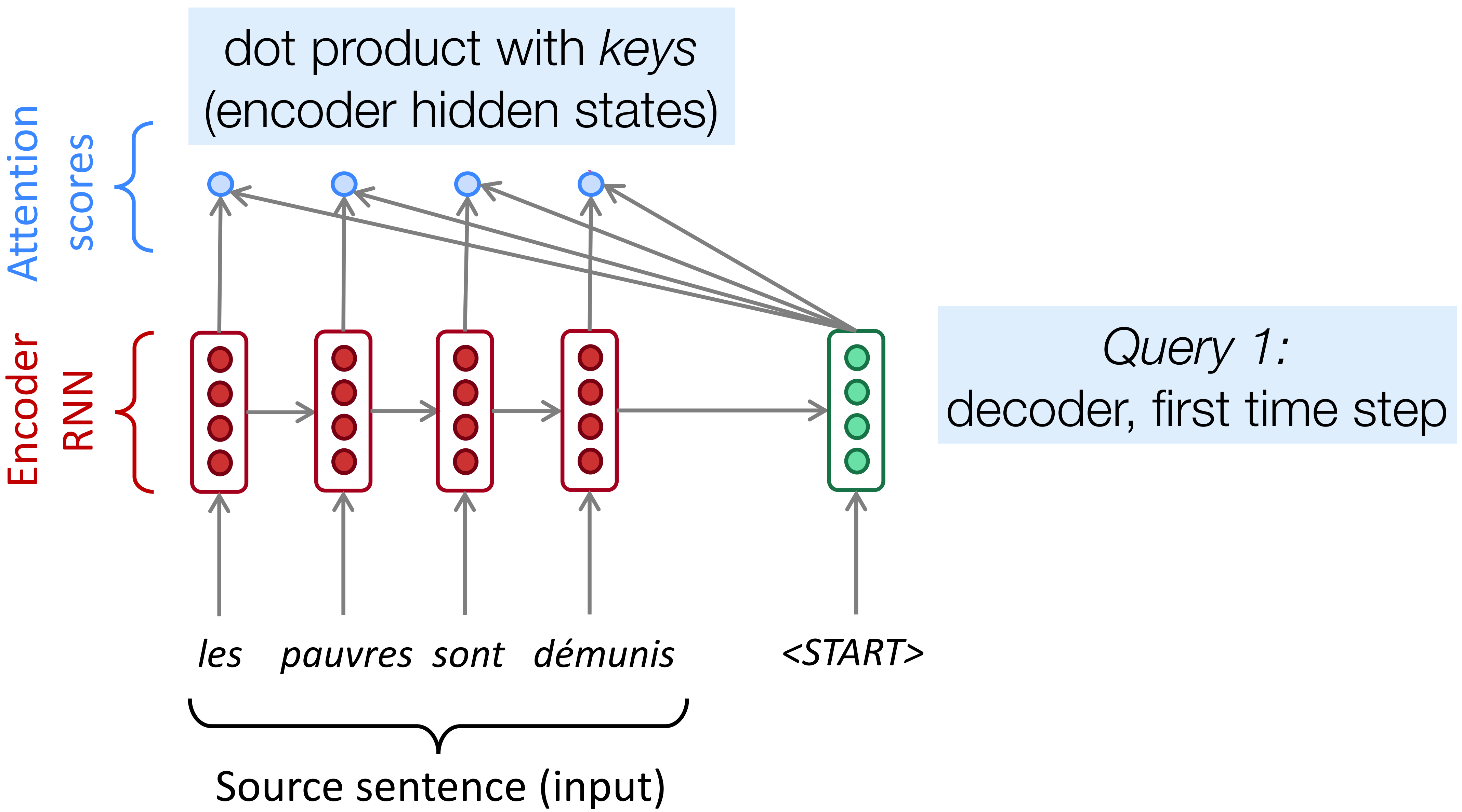
- **Attention mechanisms** (Bahdanau et al., 2015) allow the decoder to focus on a particular part of the source sequence at each time step
  - Conceptually similar to *word alignments*

How does it work?

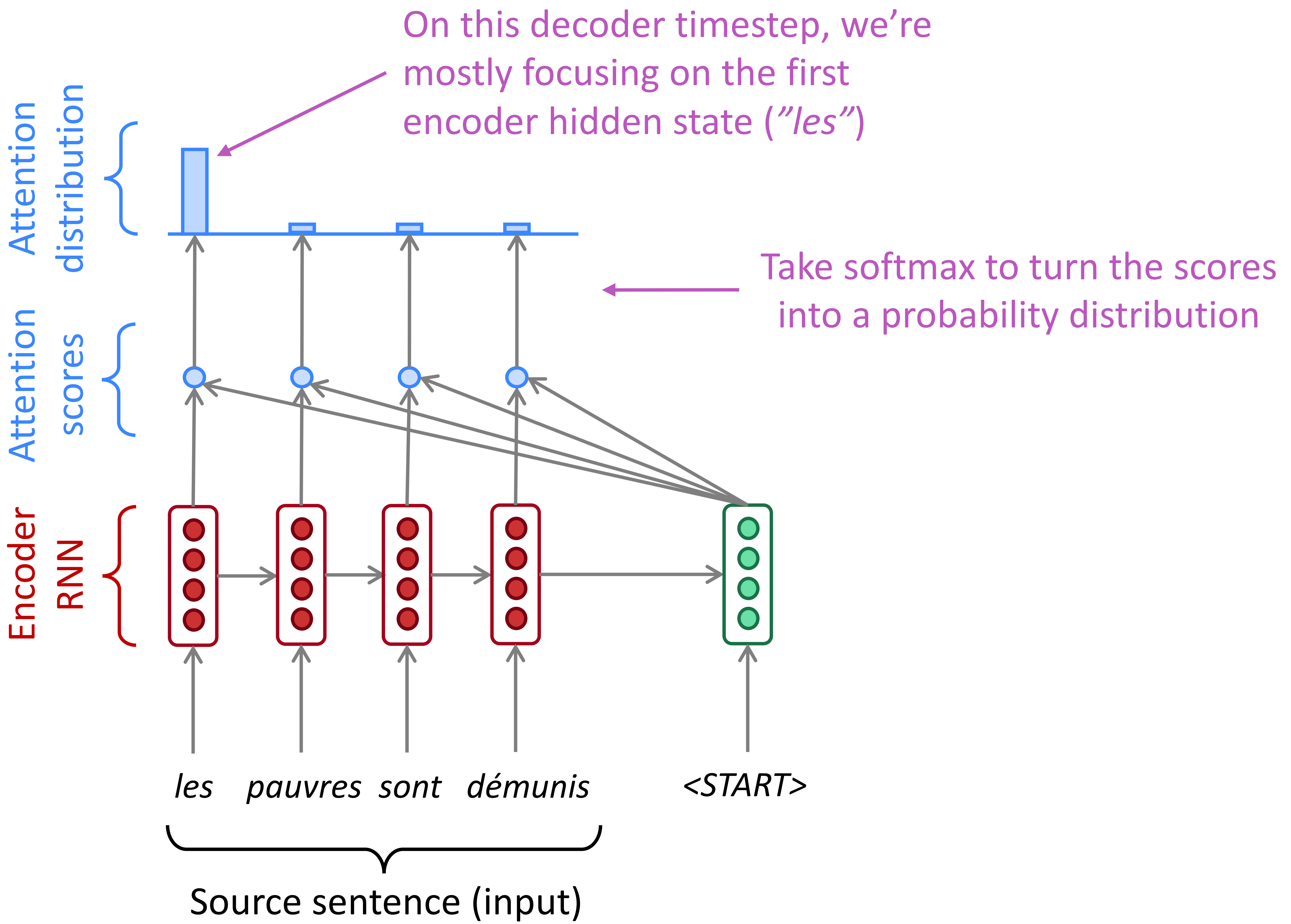
- in general, we have a single *query* vector and multiple *key* vectors. We want to score each query-key pair

in machine translation, what are the queries and keys?

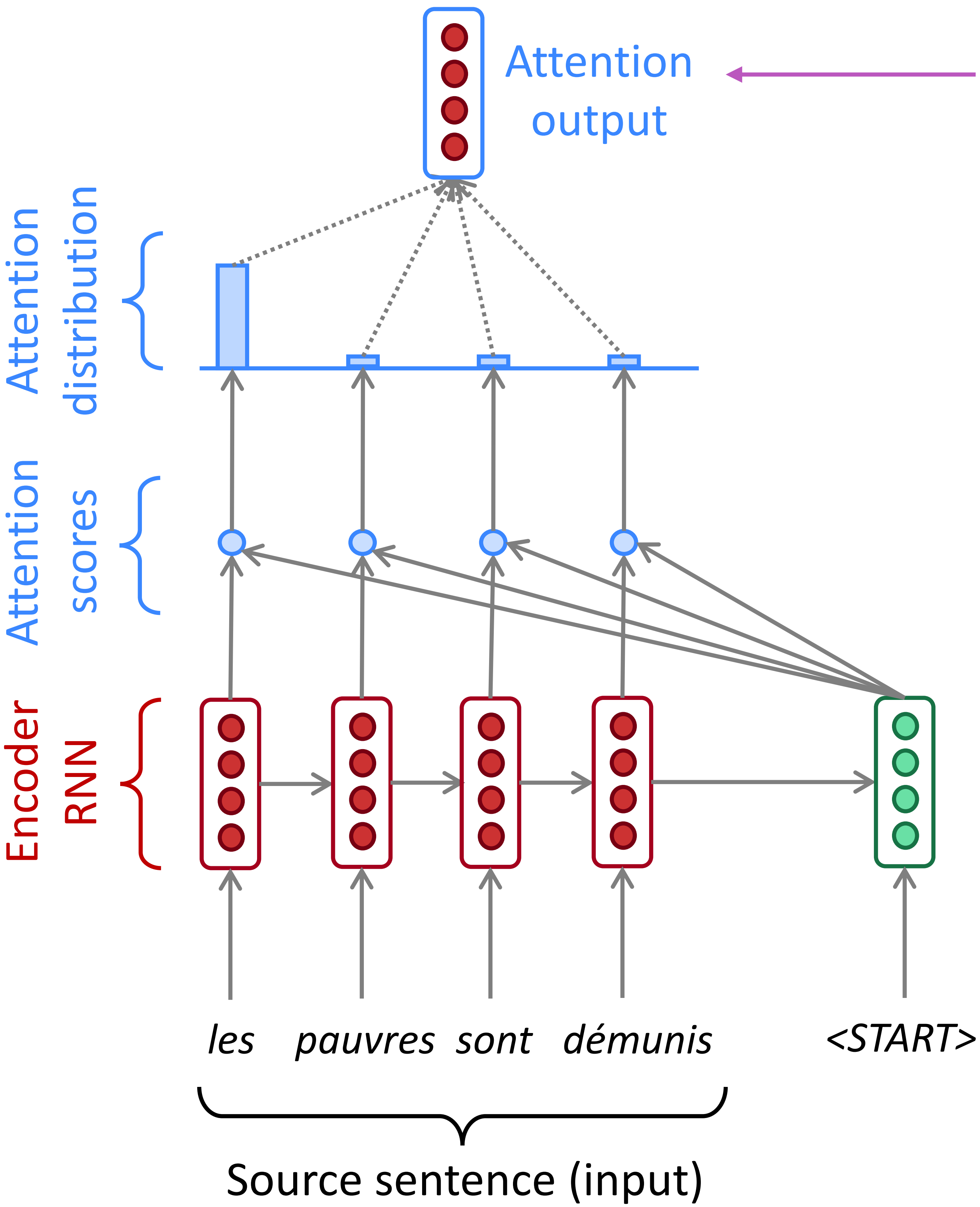
Sequence-to-sequence with attention



Sequence-to-sequence with attention



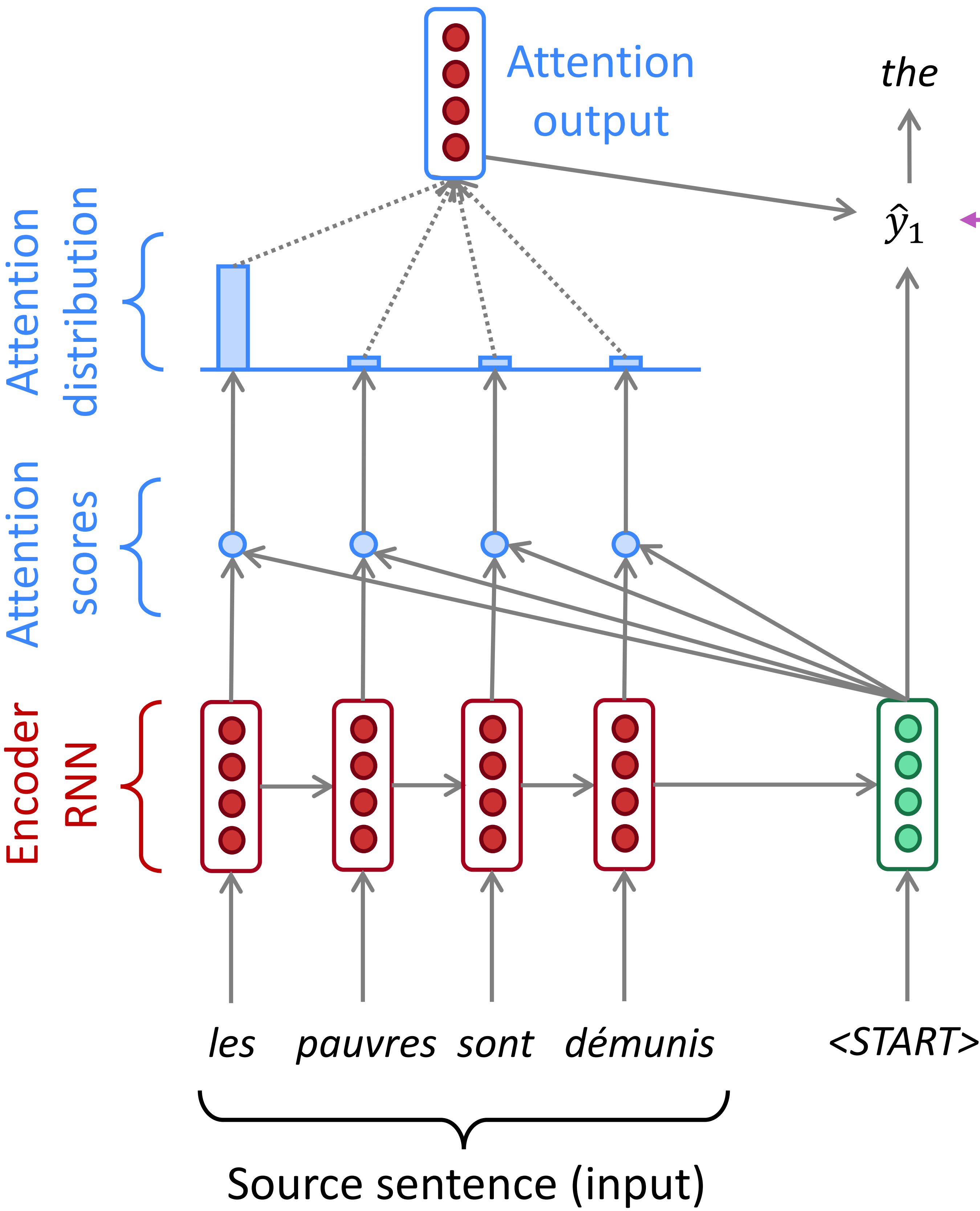
Sequence-to-sequence with attention



Use the attention distribution to take a **weighted sum** of the encoder hidden states.

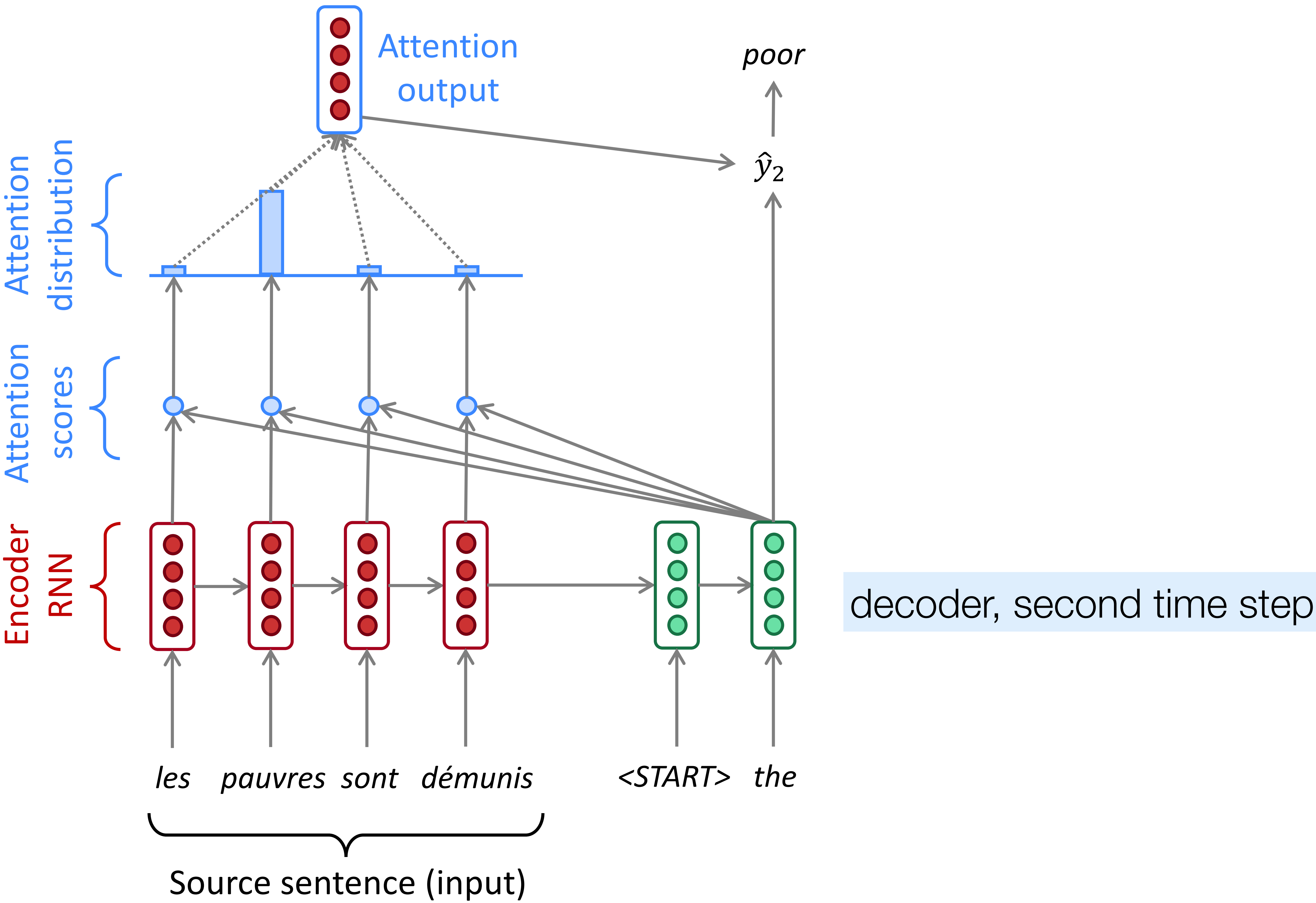
The attention output mostly contains information the hidden states that received high attention.

Sequence-to-sequence with attention



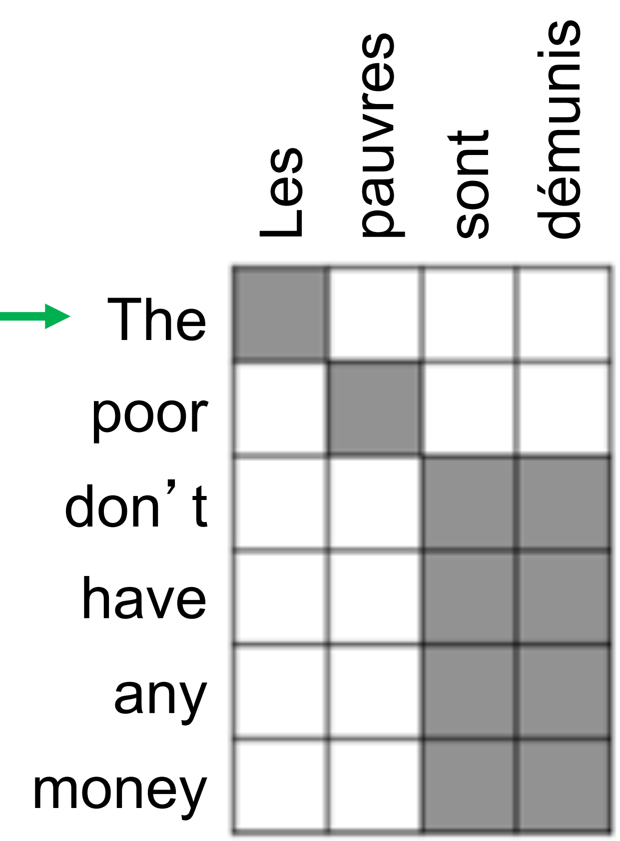
Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

Sequence-to-sequence with attention



Attention is great

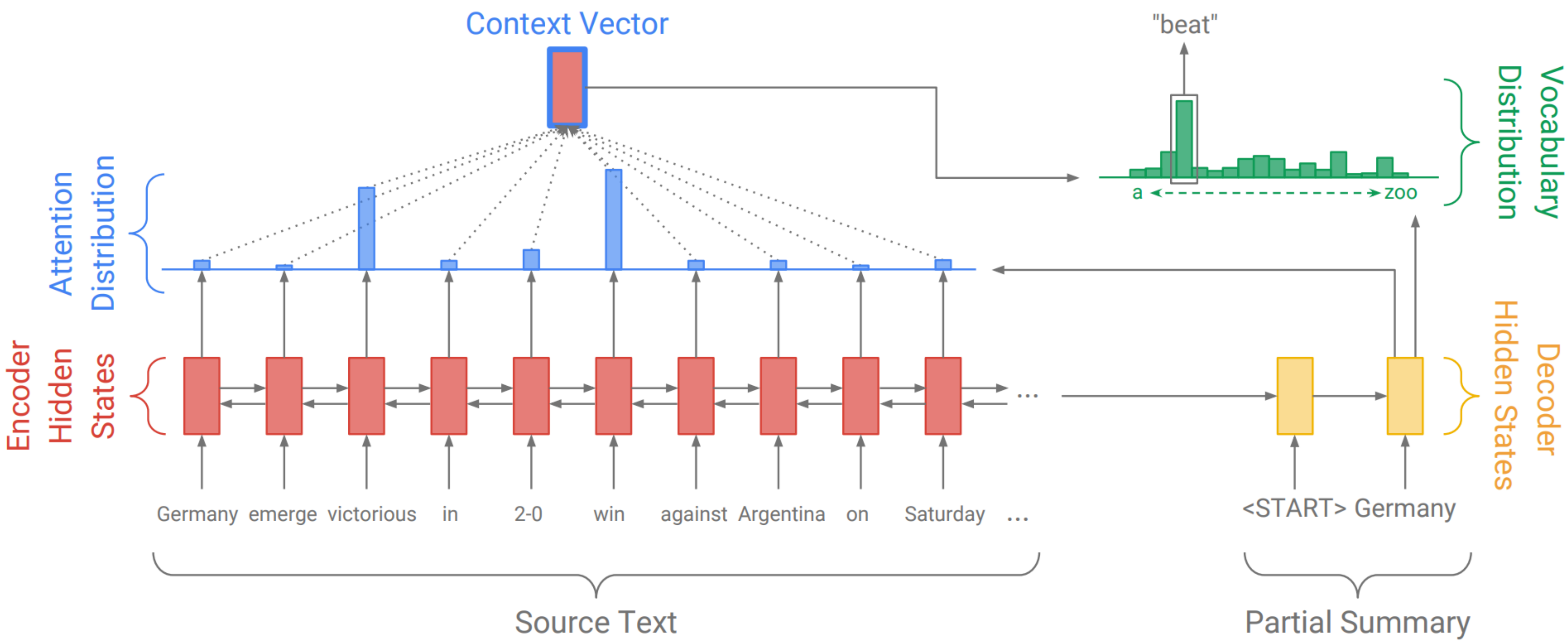
- Attention significantly improves NMT performance
  - It's very useful to allow decoder to focus on certain parts of the source
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck
- Attention helps with vanishing gradient problem
  - Provides 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



Many variants of attention

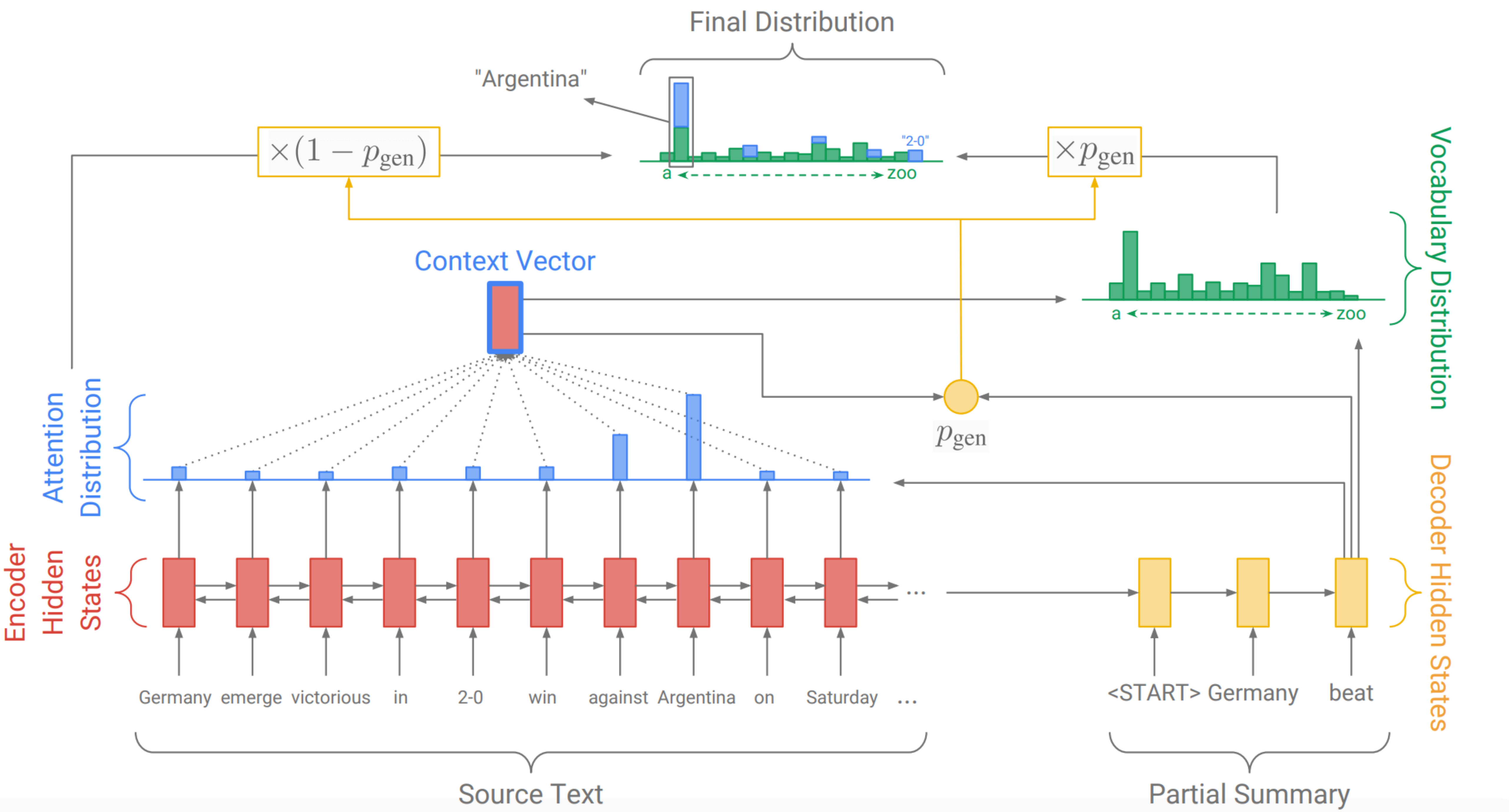
- Original formulation: $a(\mathbf{q}, \mathbf{k}) = w_2^T \tanh(W_1[\mathbf{q}; \mathbf{k}])$
- Bilinear product: $a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^T W \mathbf{k}$ Luong et al., 2015
- Dot product: $a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^T \mathbf{k}$ Luong et al., 2015
- Scaled dot product: $a(\mathbf{q},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

Attention is not just for MT!



Here we have a standard seq2seq model for summarization

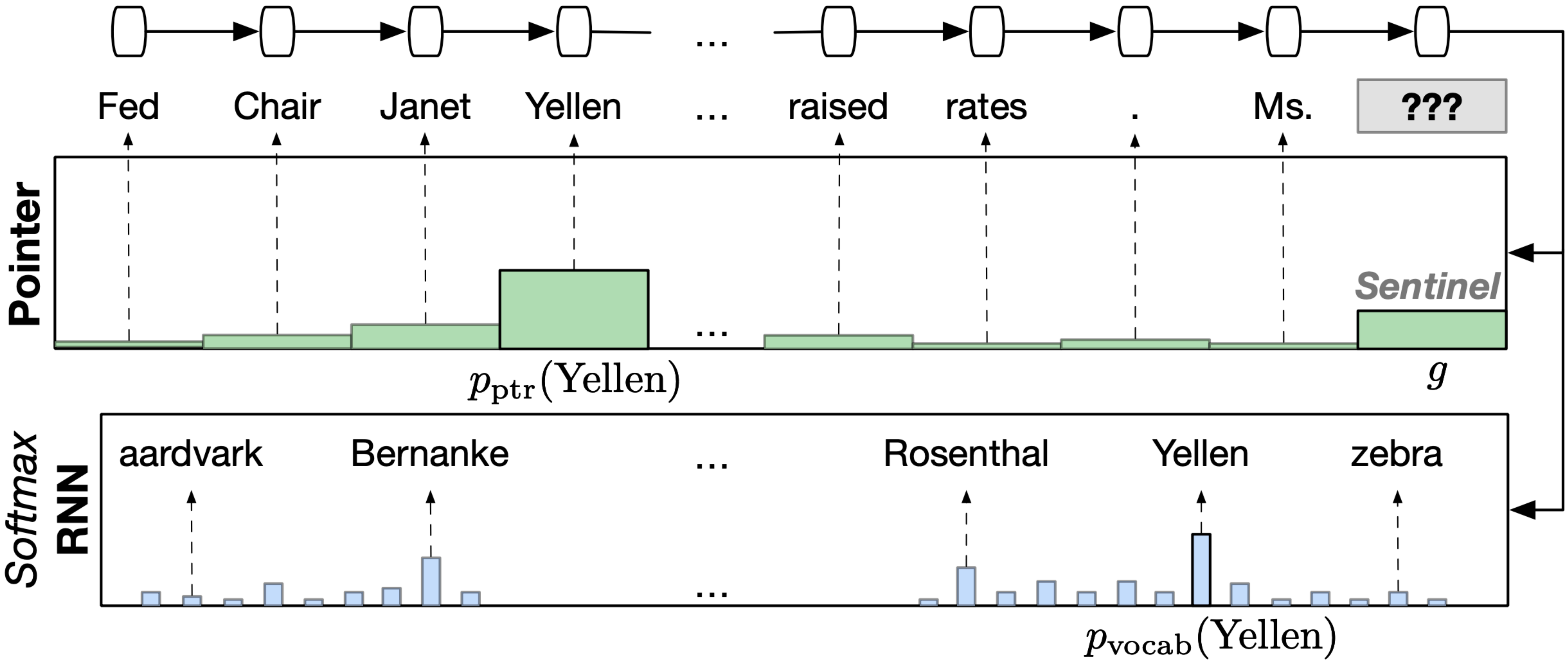
See et al., 2017



Here we have a seq2seq model with a **copy mechanism** for summarization

See et al., 2017

Target-side attention (in LMs or more complex MT models)



$$p(\mathrm{Yellen}) = g\ p_{\mathrm{vocab}}(\mathrm{Yellen}) + (1-g)\ p_{\mathrm{ptr}}(\mathrm{Yellen})$$

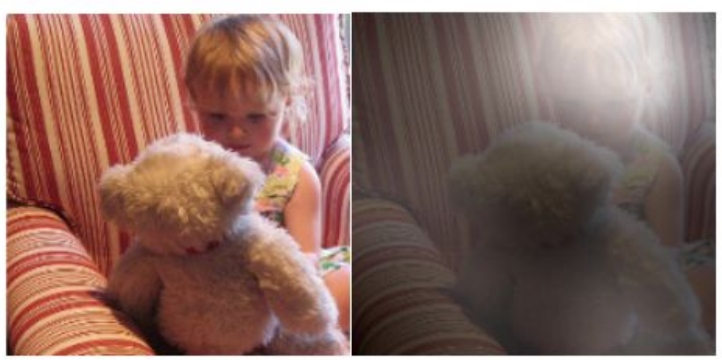
Merity et al., 2016

Image Captioning with Attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Xu et al., 2015

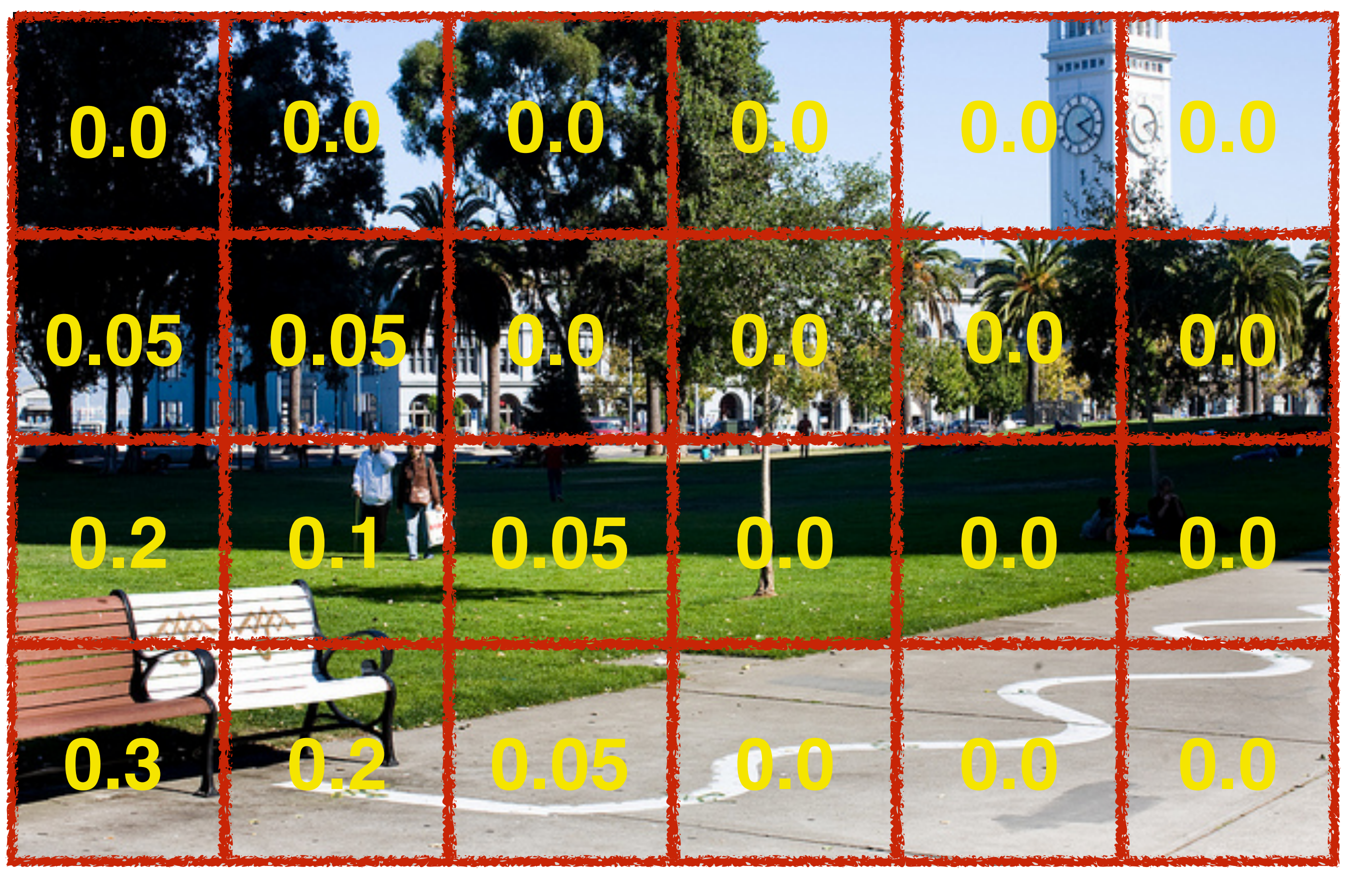
visual attention

- Use the question representation *q* to determine where in the image to look

How many benches are shown?

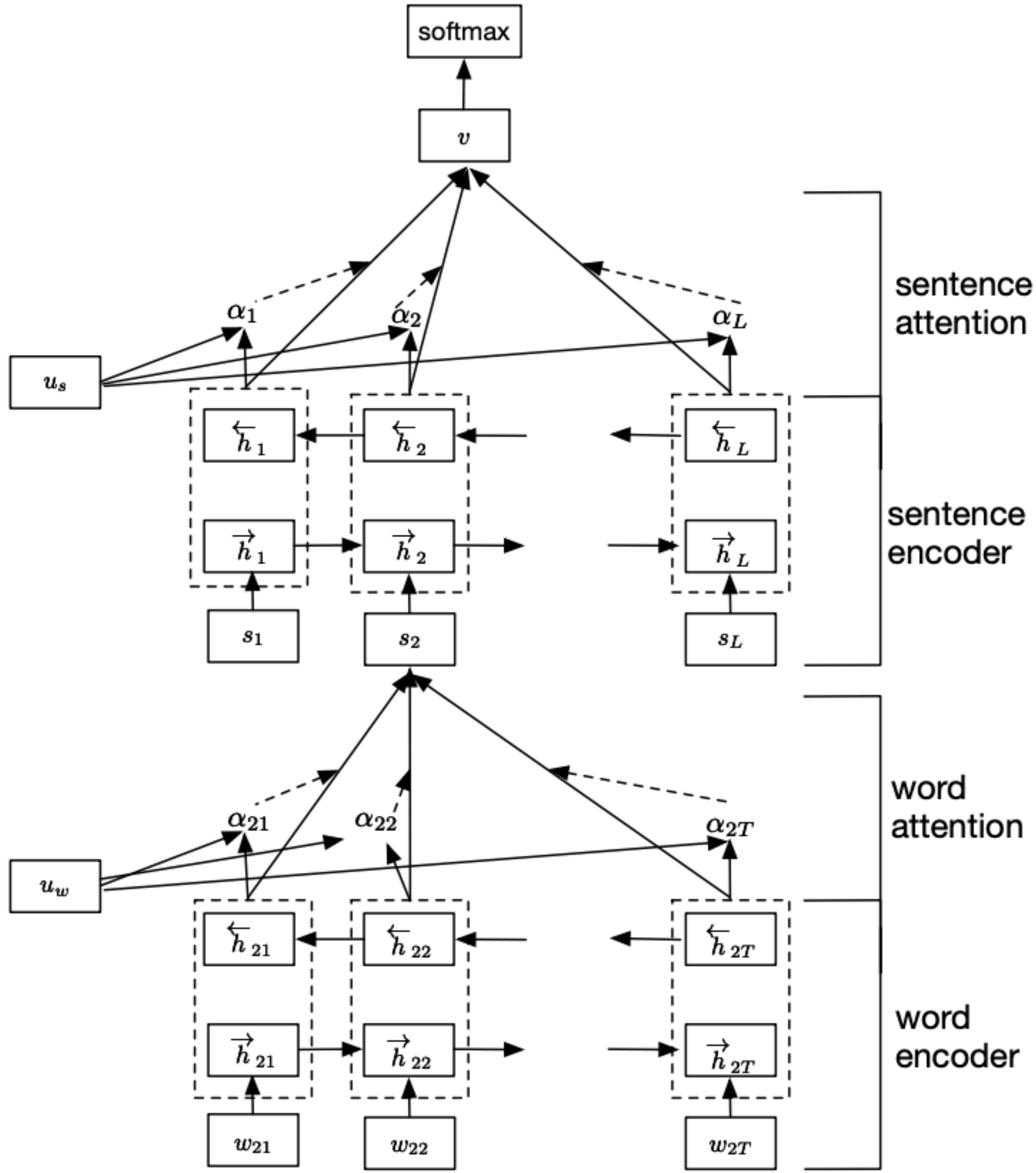
softmax:
predict answer

attention over final convolutional layer in network: 196 boxes, captures color and positional information



How many benches are shown?

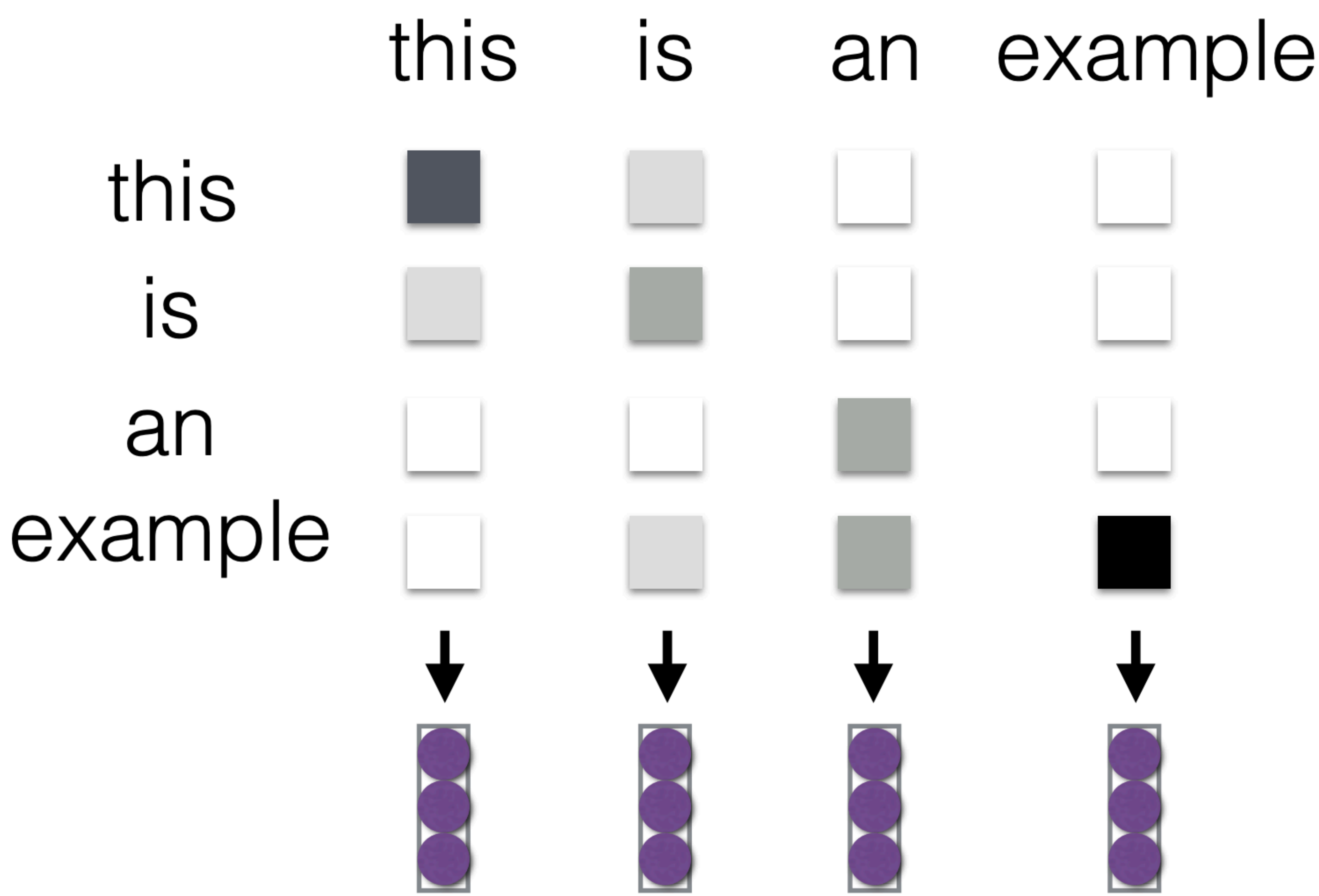
Hierarchical attention



Yang et al., 2016

Self-attention as an encoder!

(core component of Transformer)

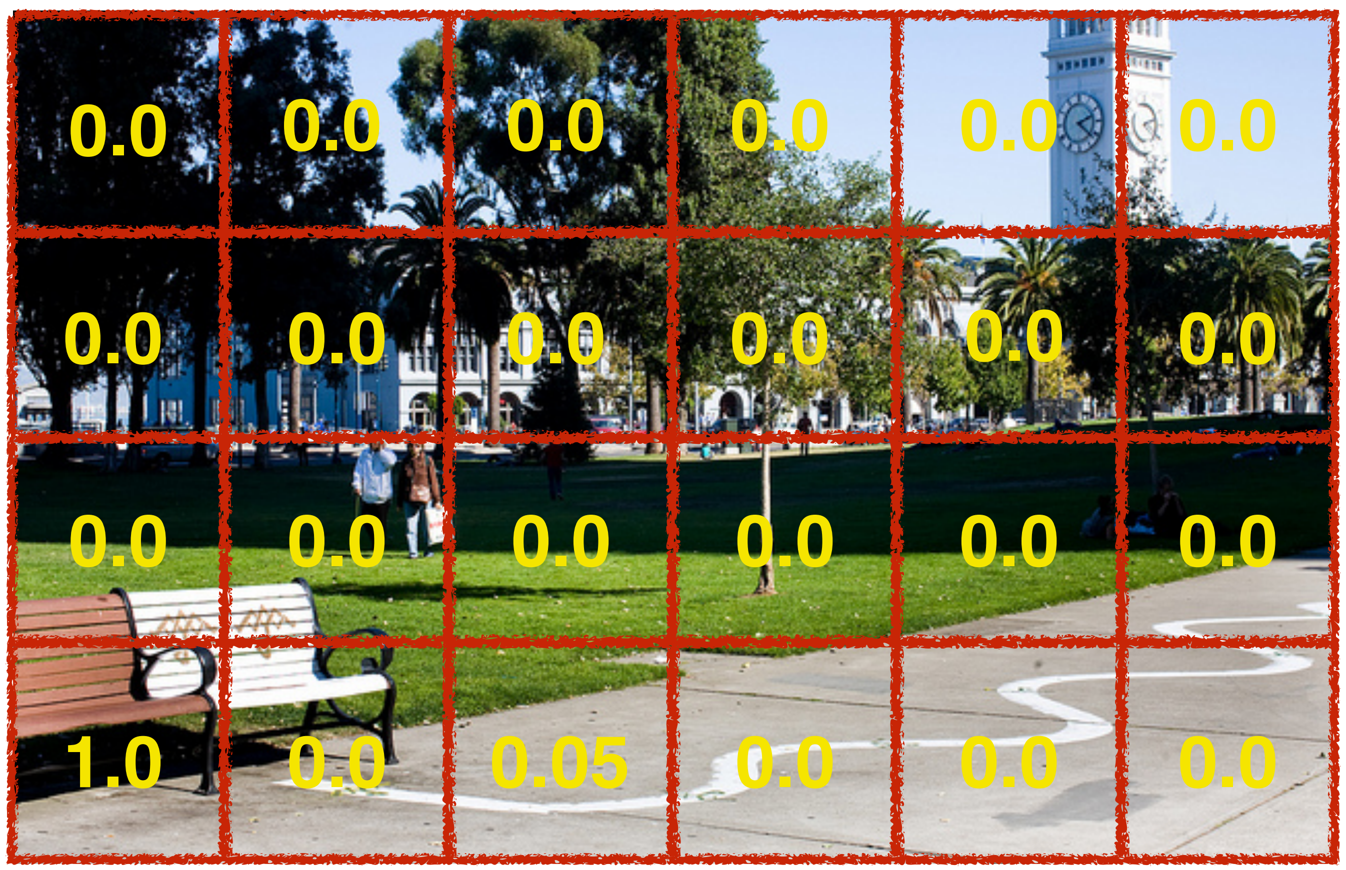


Attention variants

hard attention

attention over final convolutional layer in network: 196 boxes, captures color and positional information

softmax: predict answer



we can use *reinforcement learning* to focus on just one box

How many benches are shown?

Xu et al., 2015

Multi-headed attention

- Preview of next class!
- Intuition: *k* different attentions, each of which is computed independently and focuses on different parts of the sentence
- Transformers = stacked layers of multi-headed self-attention