

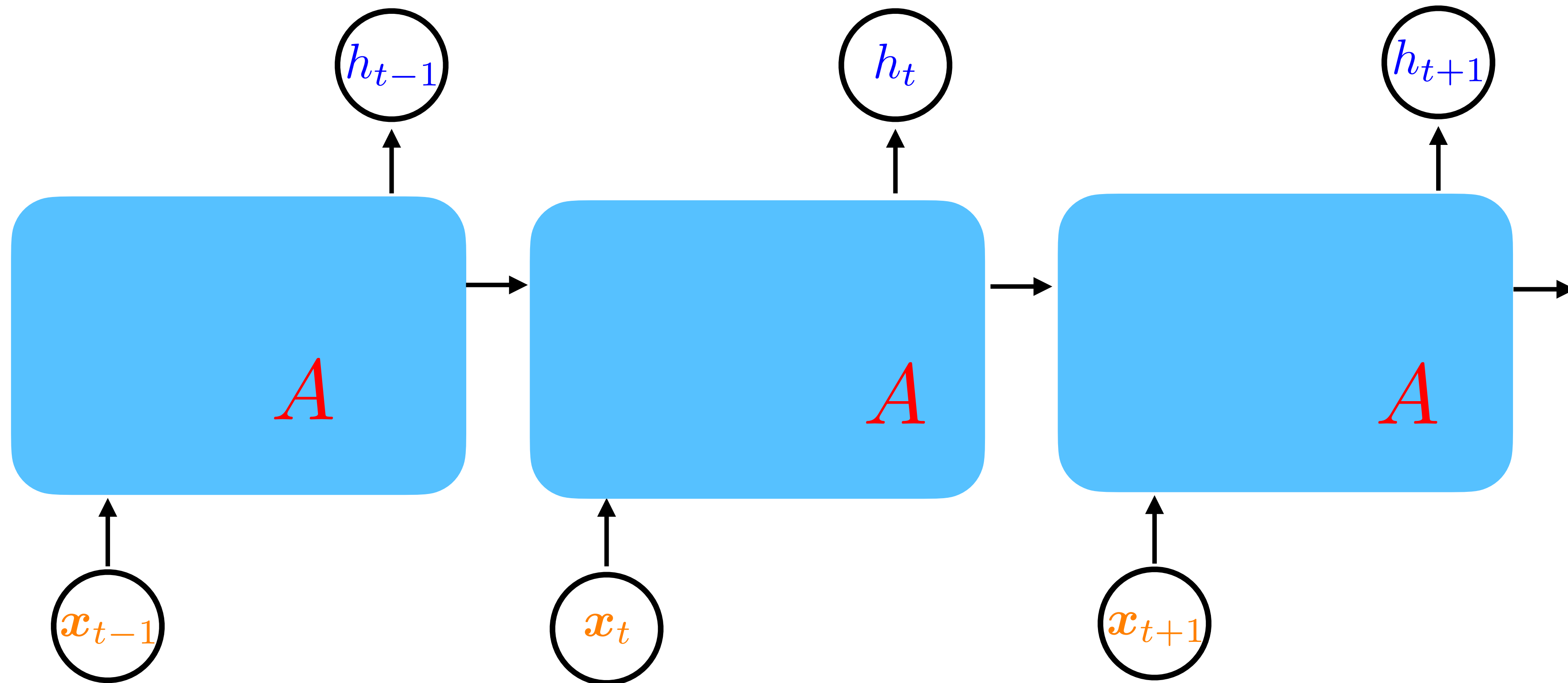
Sequence to Sequence Model and Attention

COMP7607 — Lecture 3

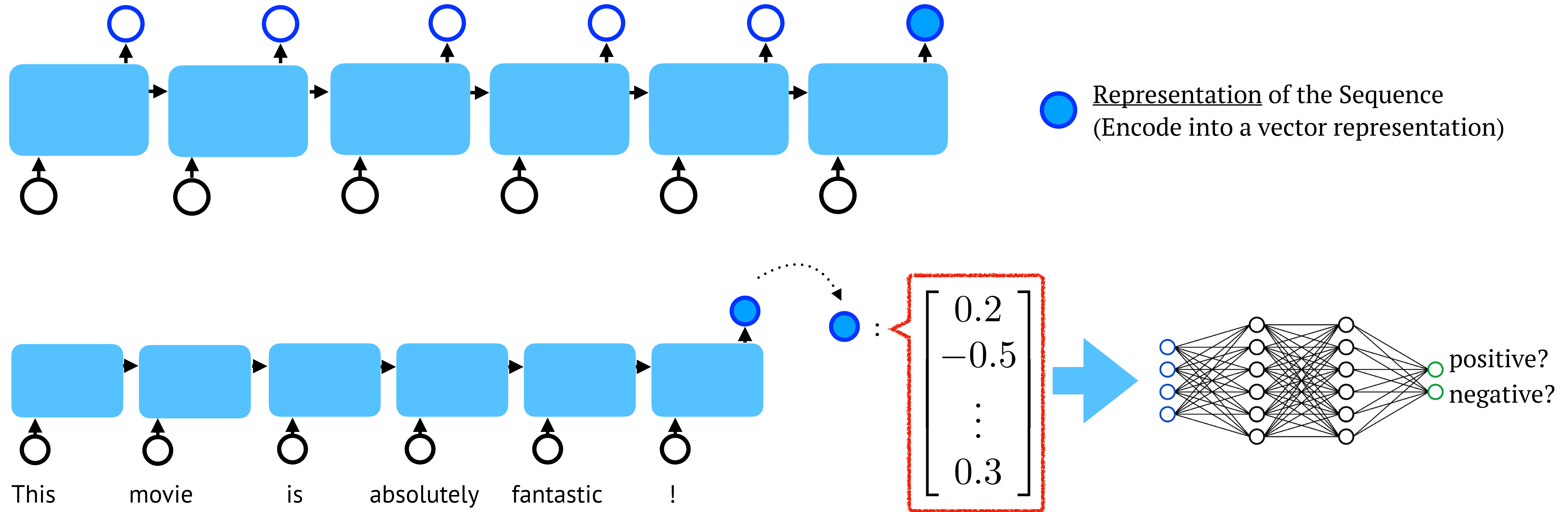
Lingpeng Kong

Department of Computer Science, The University of Hong Kong

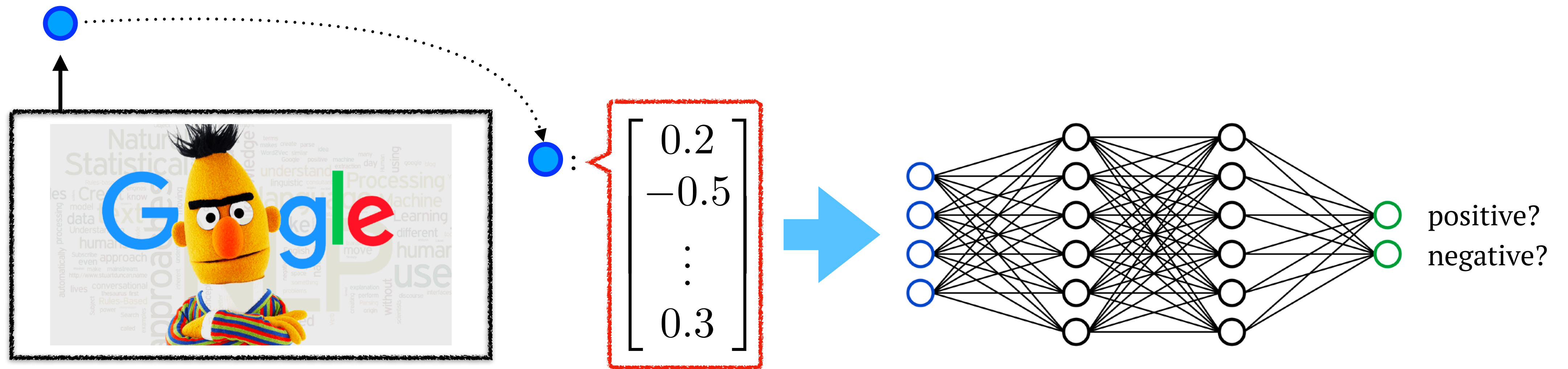
Recurrent Neural Network



RNN as Encoder

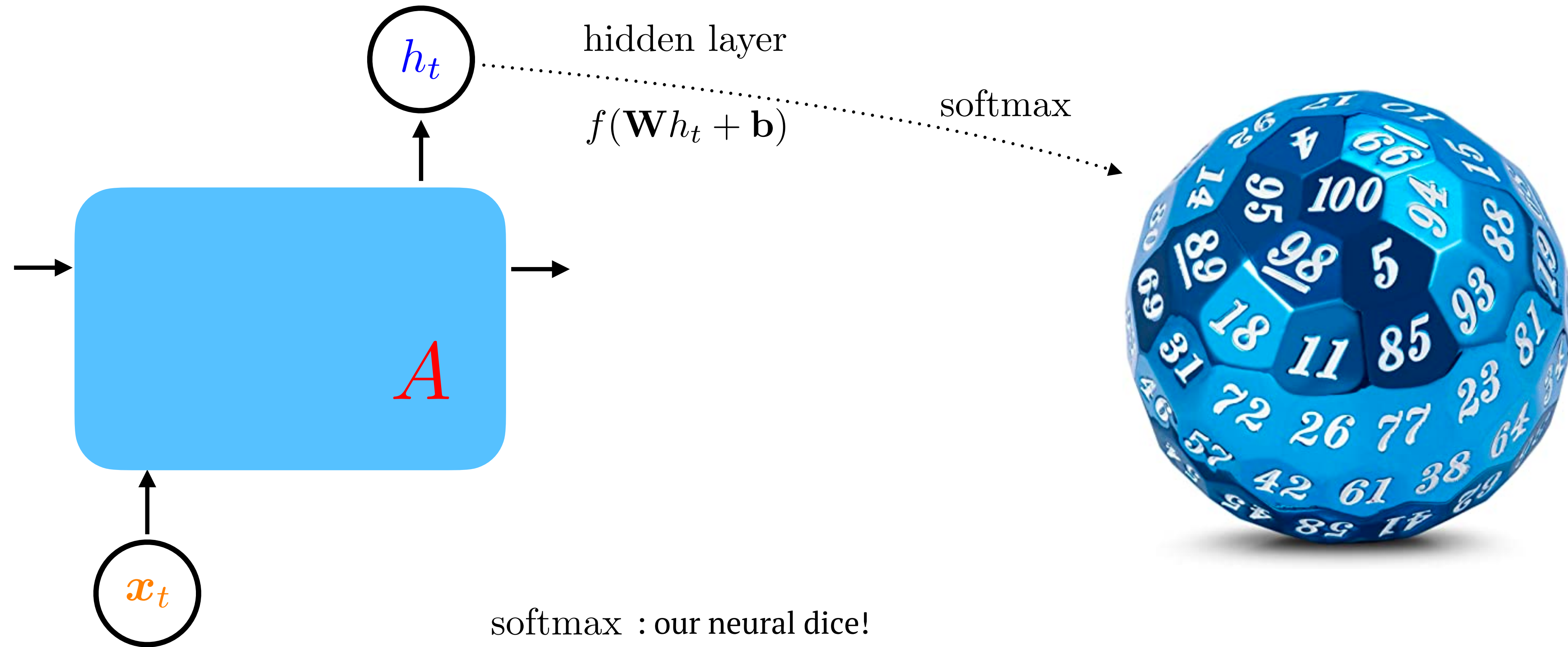


Flashback: Pretraining and Fine-tuning



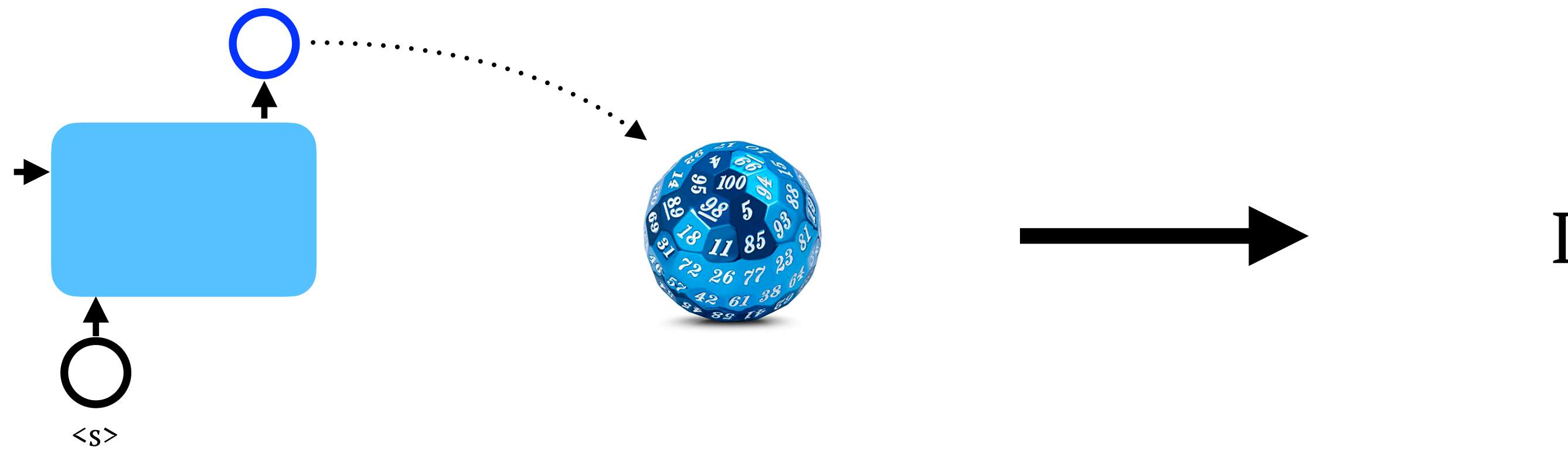
This movie is absolutely fantastic !

Sample a sentence from RNNLMs



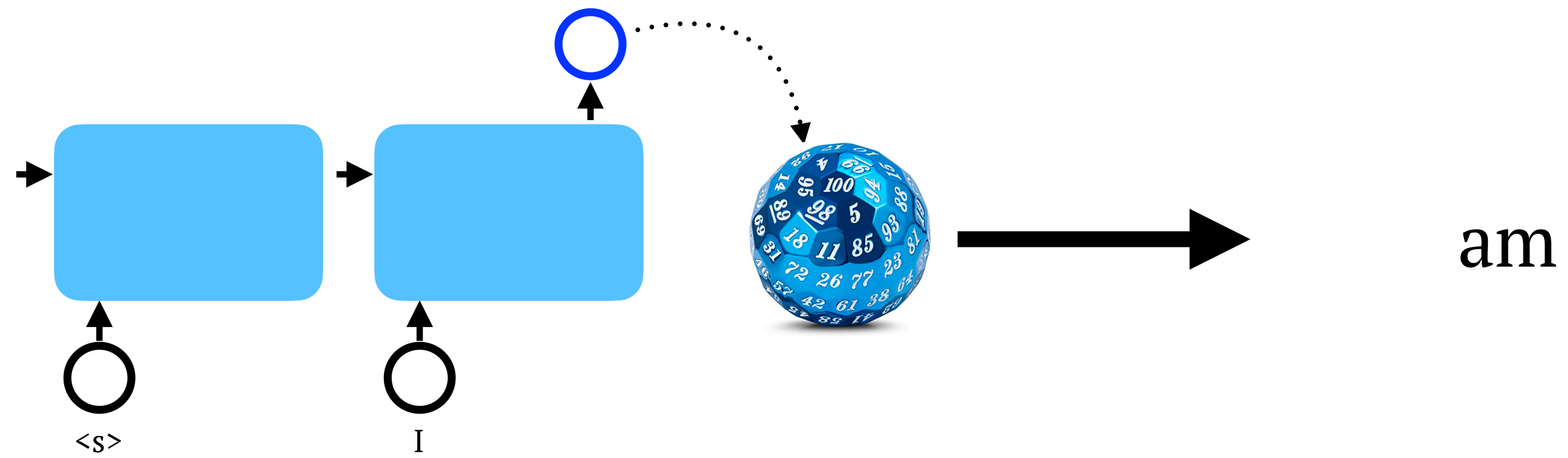
Sample a sentence from RNNLMs

I



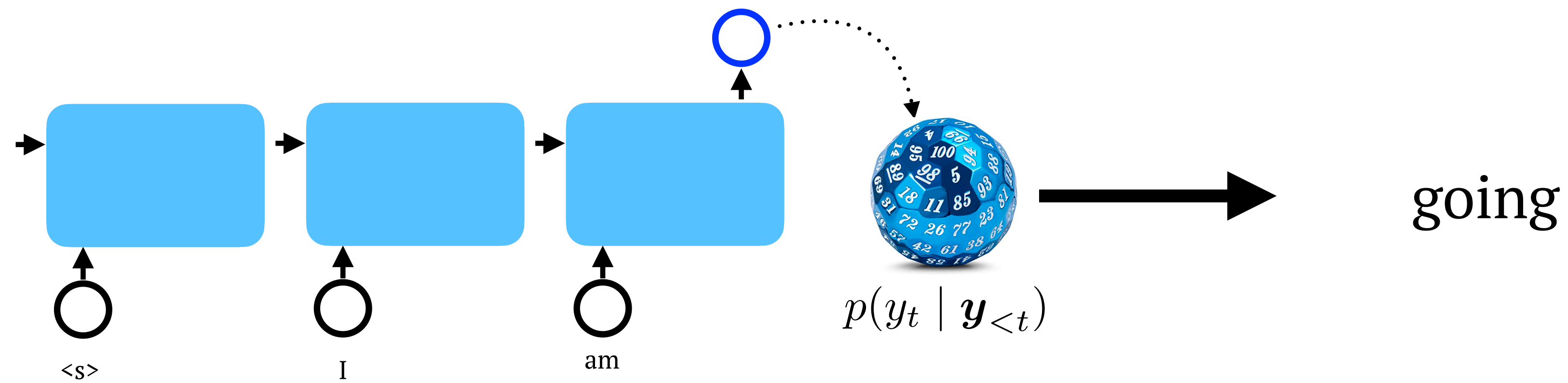
Sample a sentence from RNNLMs

I am



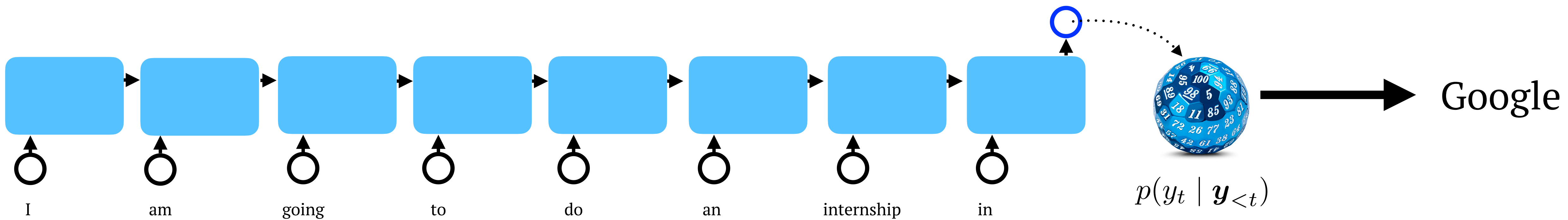
Sample a sentence from RNNLMs

I am going

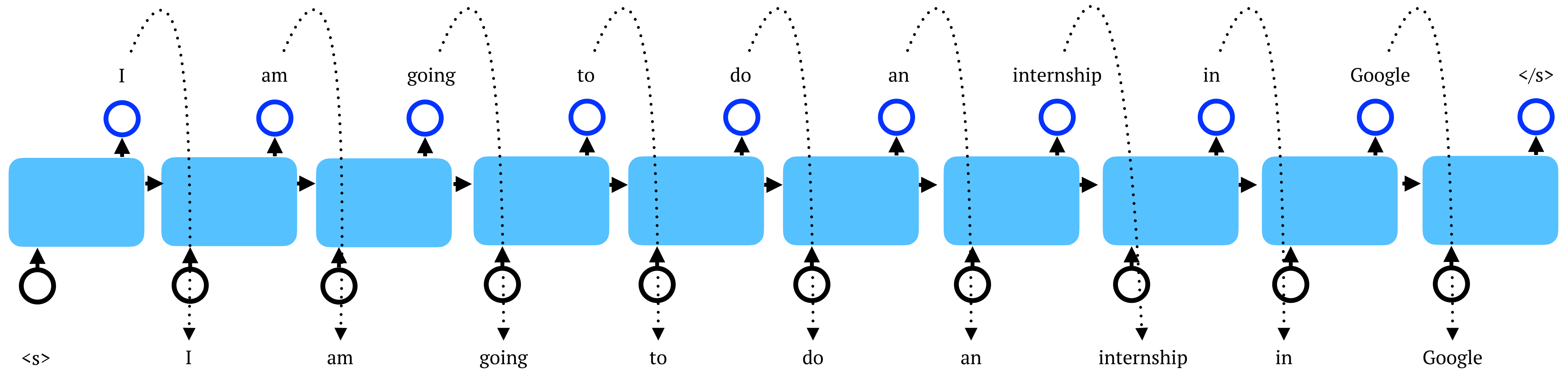


Sample a sentence from RNNLMs

I am going to do an internship in Google



RNN as Decoder (RNNDLM)



$$p(y_t | \mathbf{y}_{<t})$$

Machine Translation

中秋快樂！

x

Happy mid autumn festival !

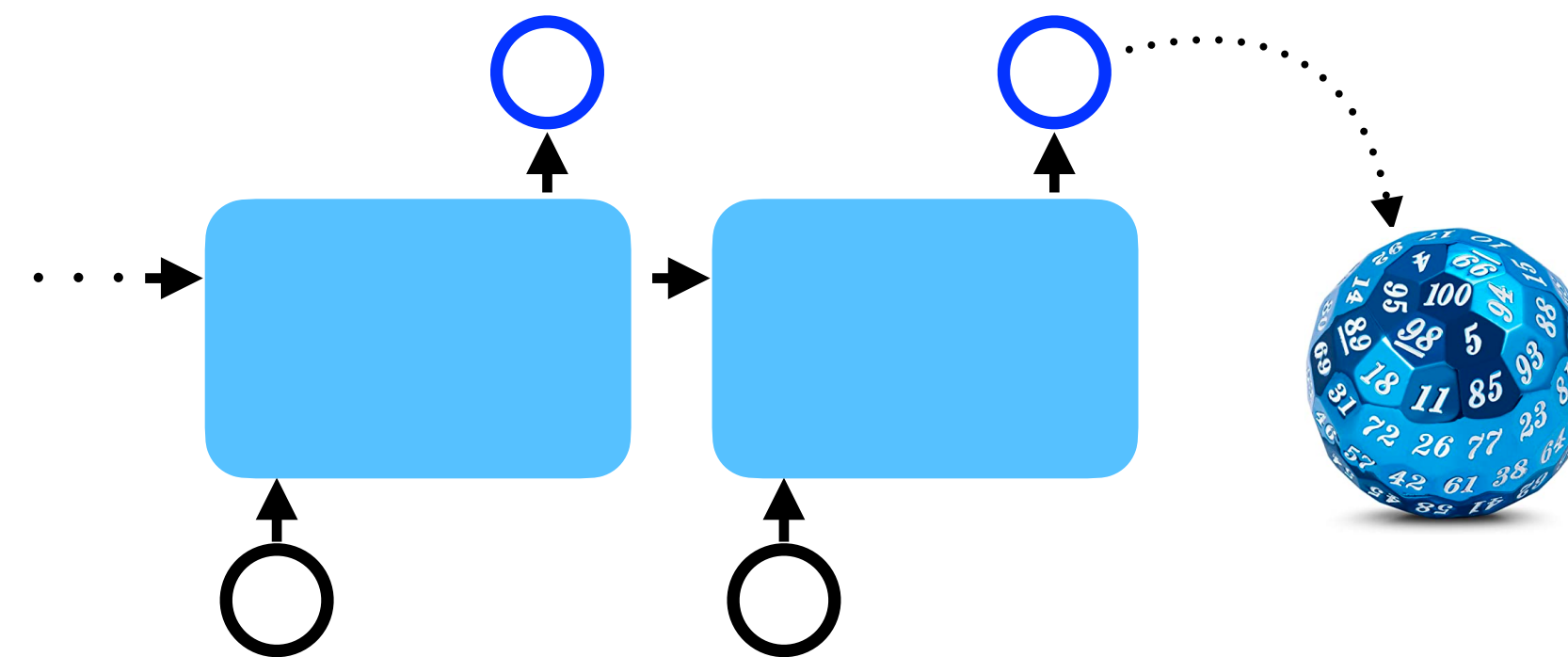
y

Happy mid autumn festival !

$$p(\mathbf{y}) = p(y_1 \dots y_n) = \prod_{t=1}^n p(y_t | \mathbf{y}_{<t})$$



$p(y_t | \mathbf{y}_{<t})$



Machine Translation

中秋快樂！

x

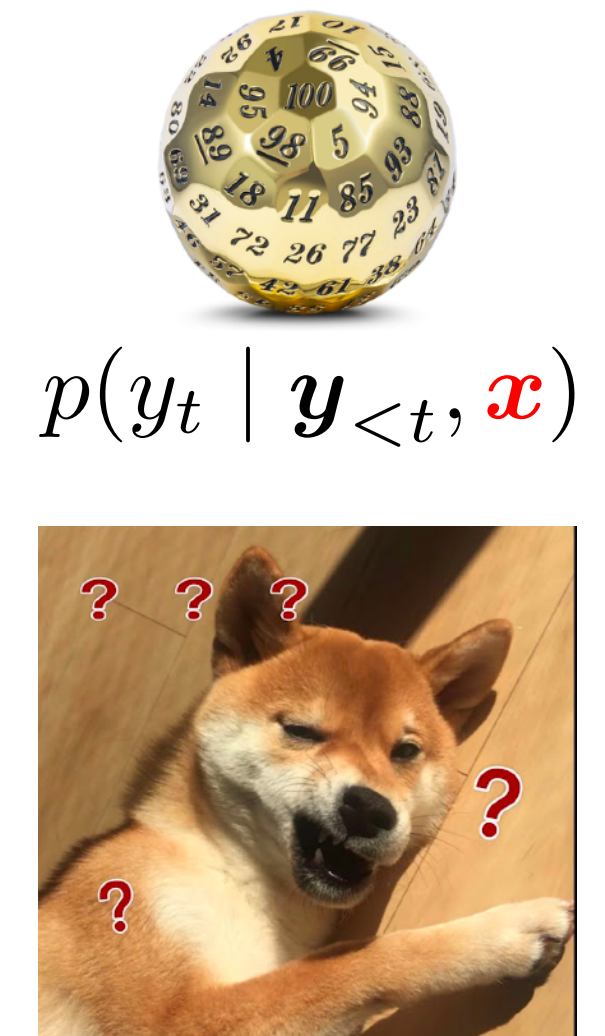
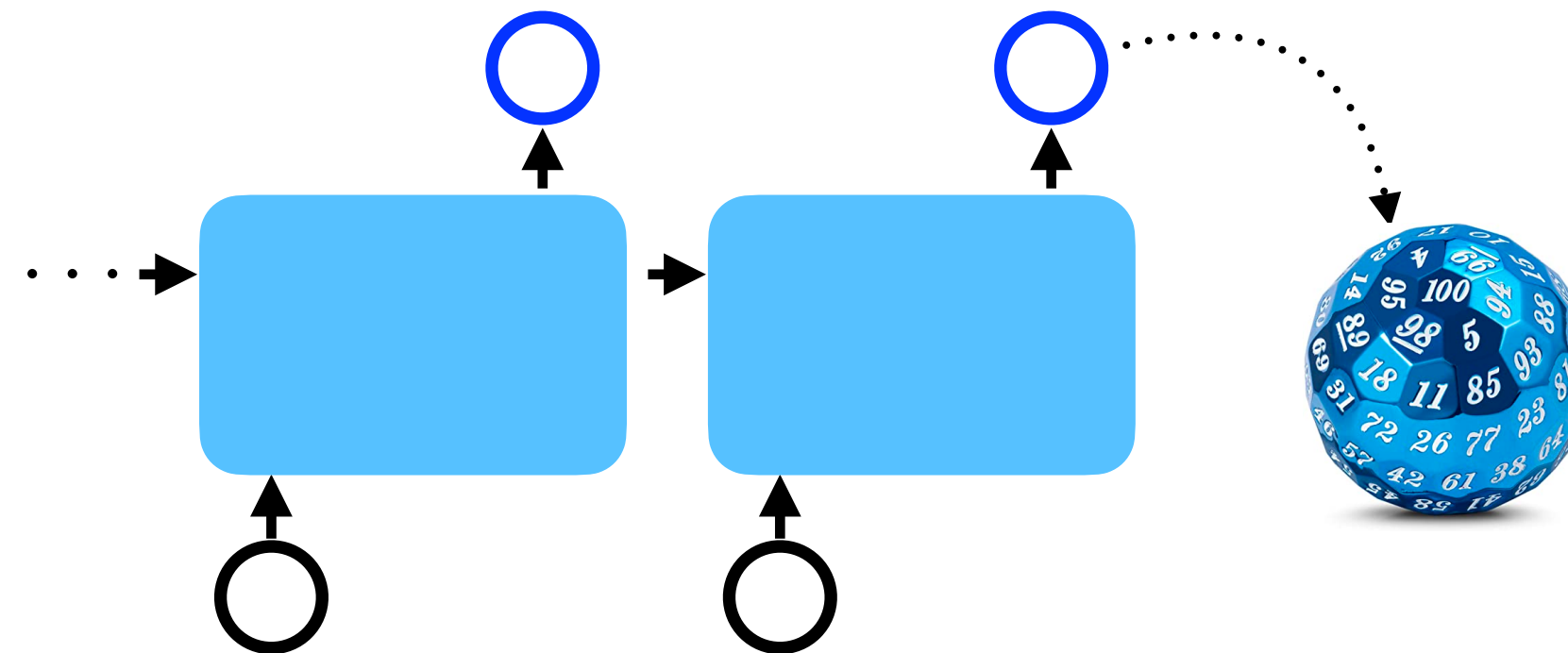
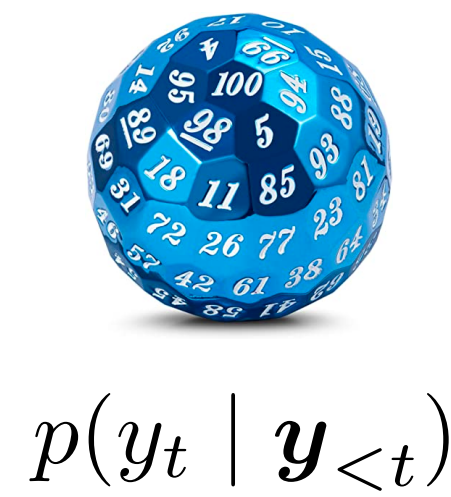
Happy mid autumn festival !

y

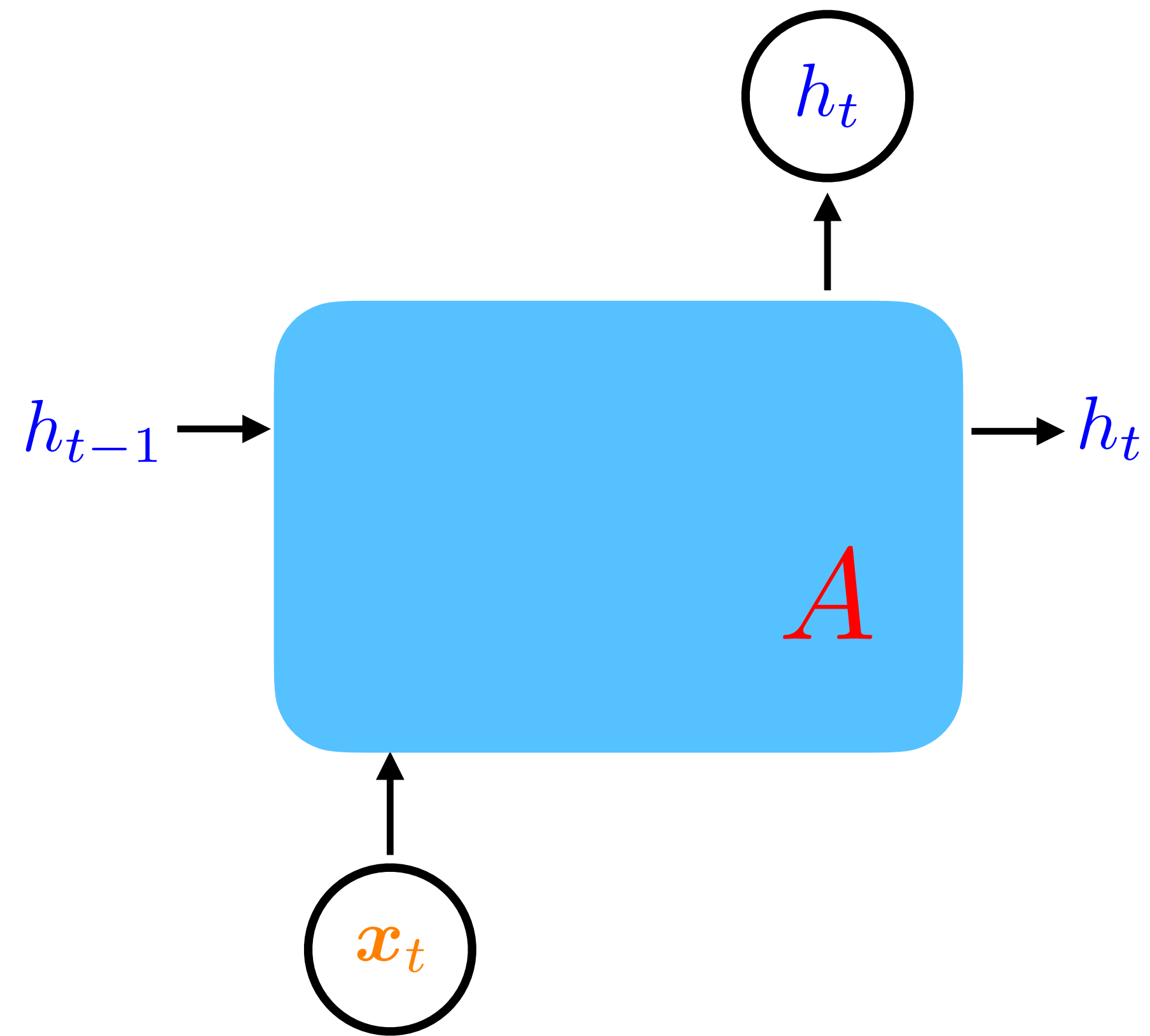
$$p(\mathbf{y} | \mathbf{x}) = p(y_1 \dots y_n | x_1 \dots x_m) = \prod_{t=1}^n p(y_t | \mathbf{y}_{<t}, \mathbf{x})$$

↑ target
↑ source

Conditional Language Model



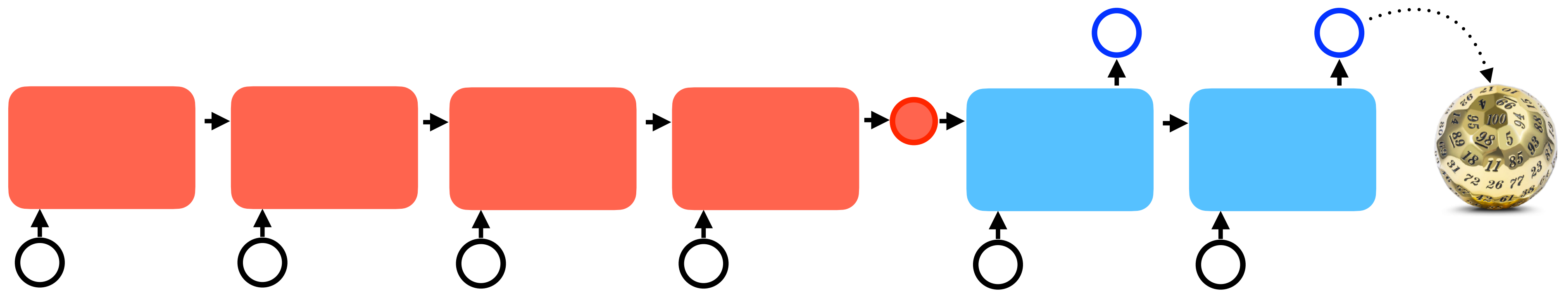
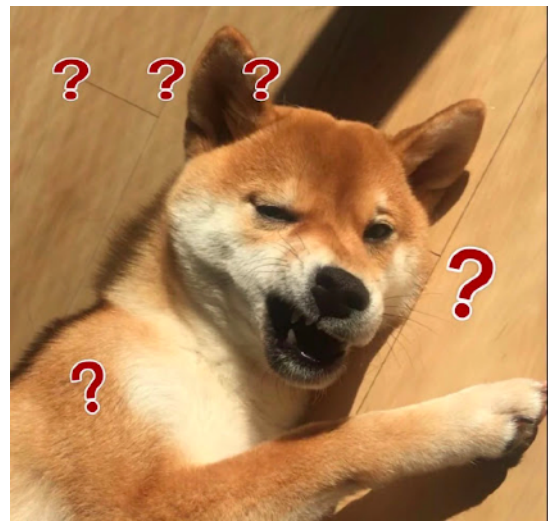
Recurrent Neural Network



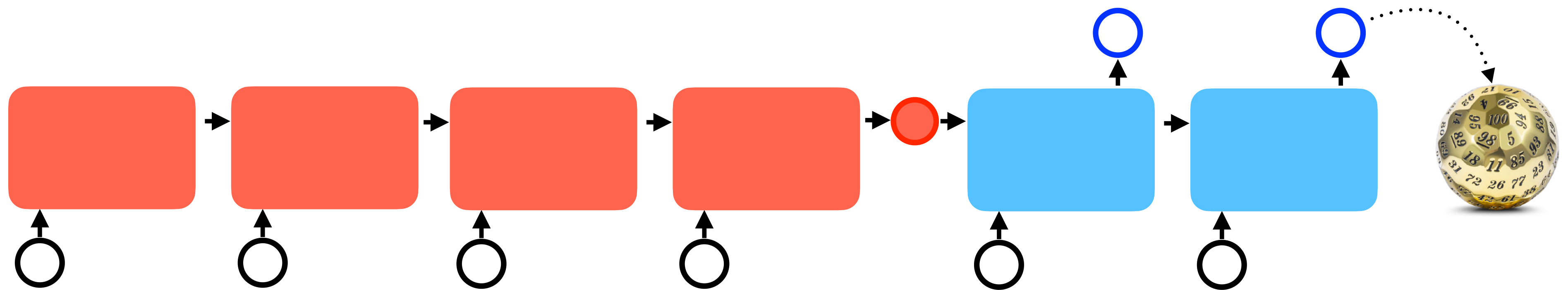
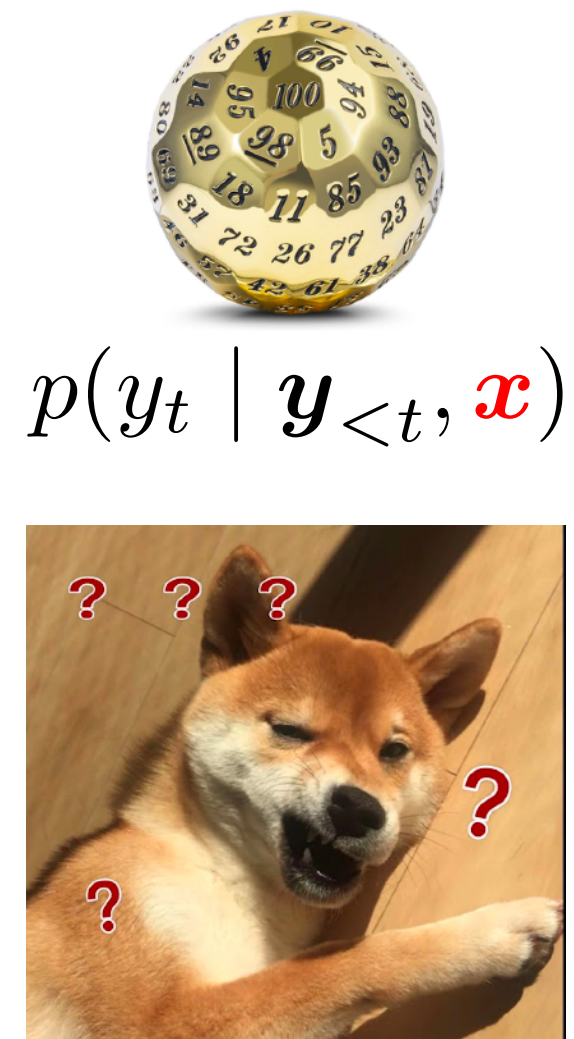
Encoder + Decoder



$$p(y_t | \mathbf{y}_{<t}, \mathbf{x})$$



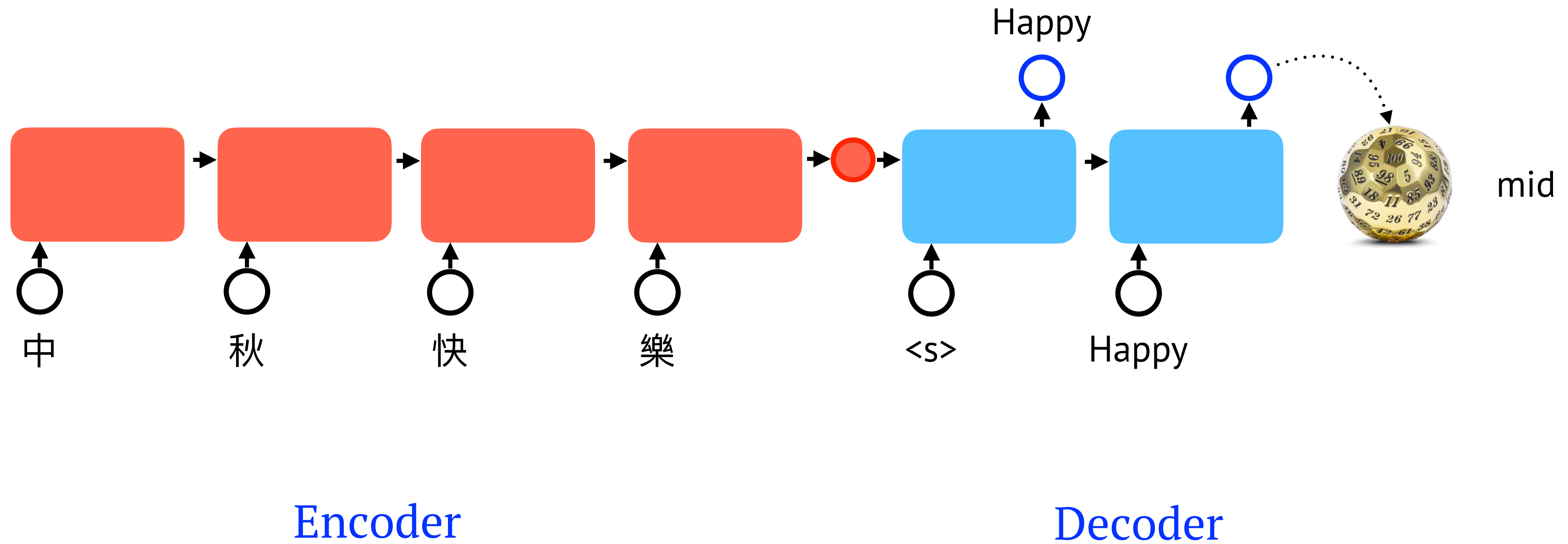
Sequence to Sequence Model



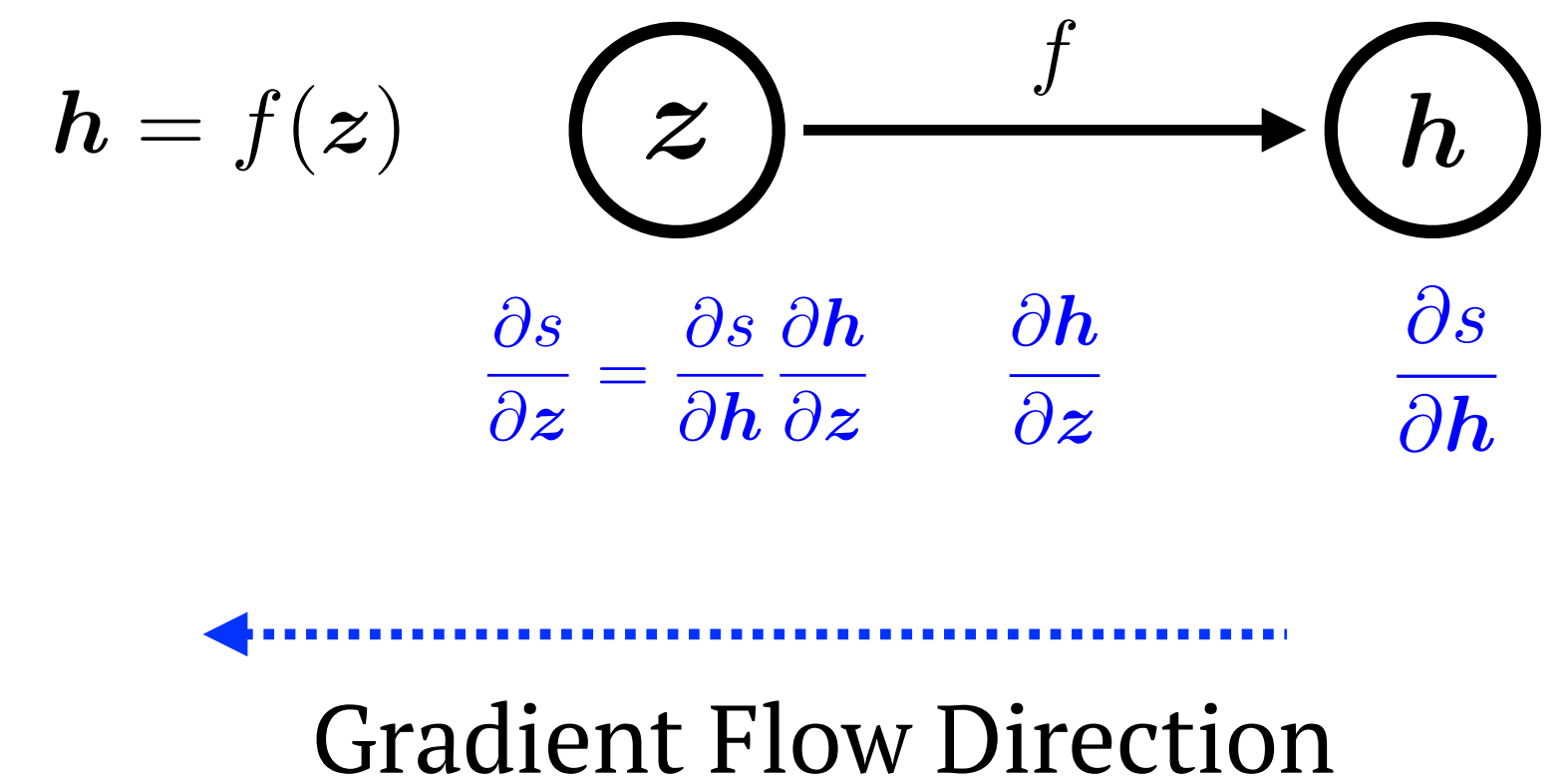
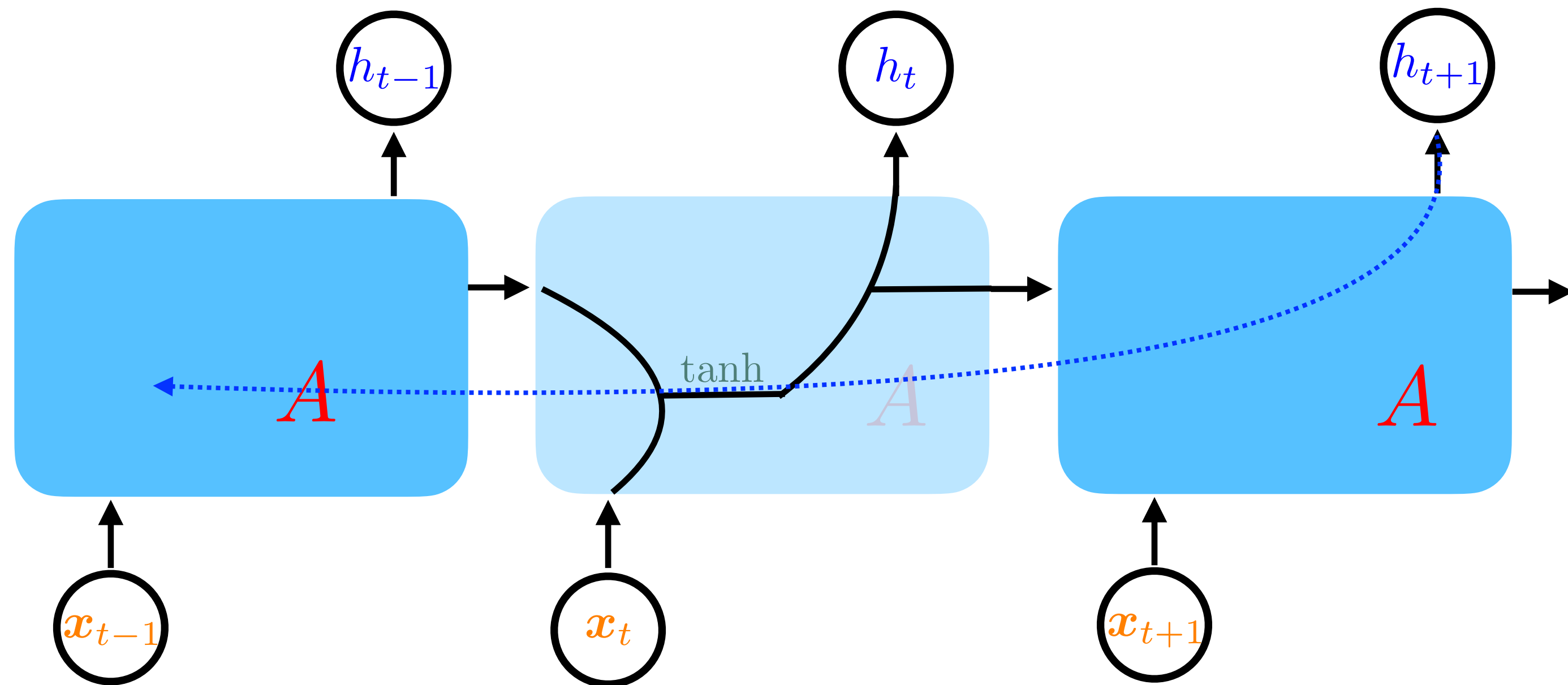
Encoder

Decoder

Sequence to Sequence Model

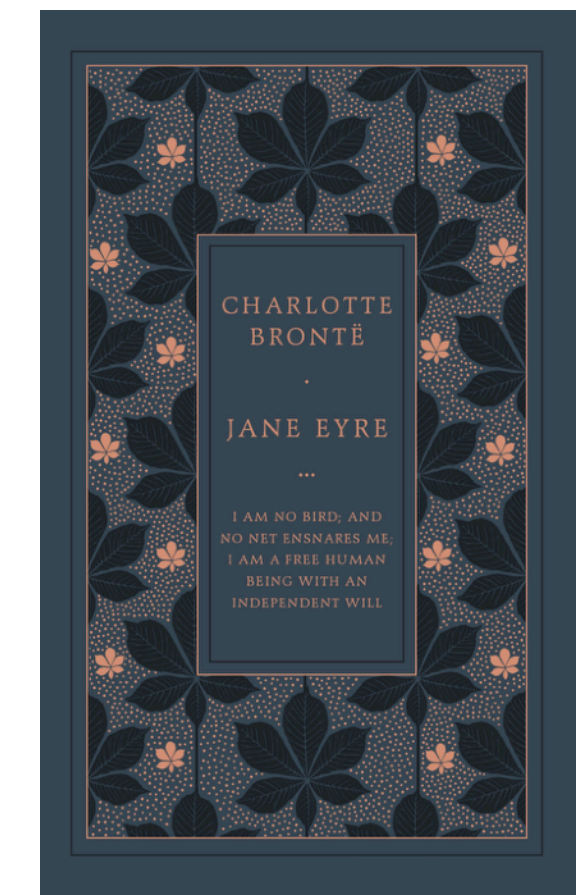
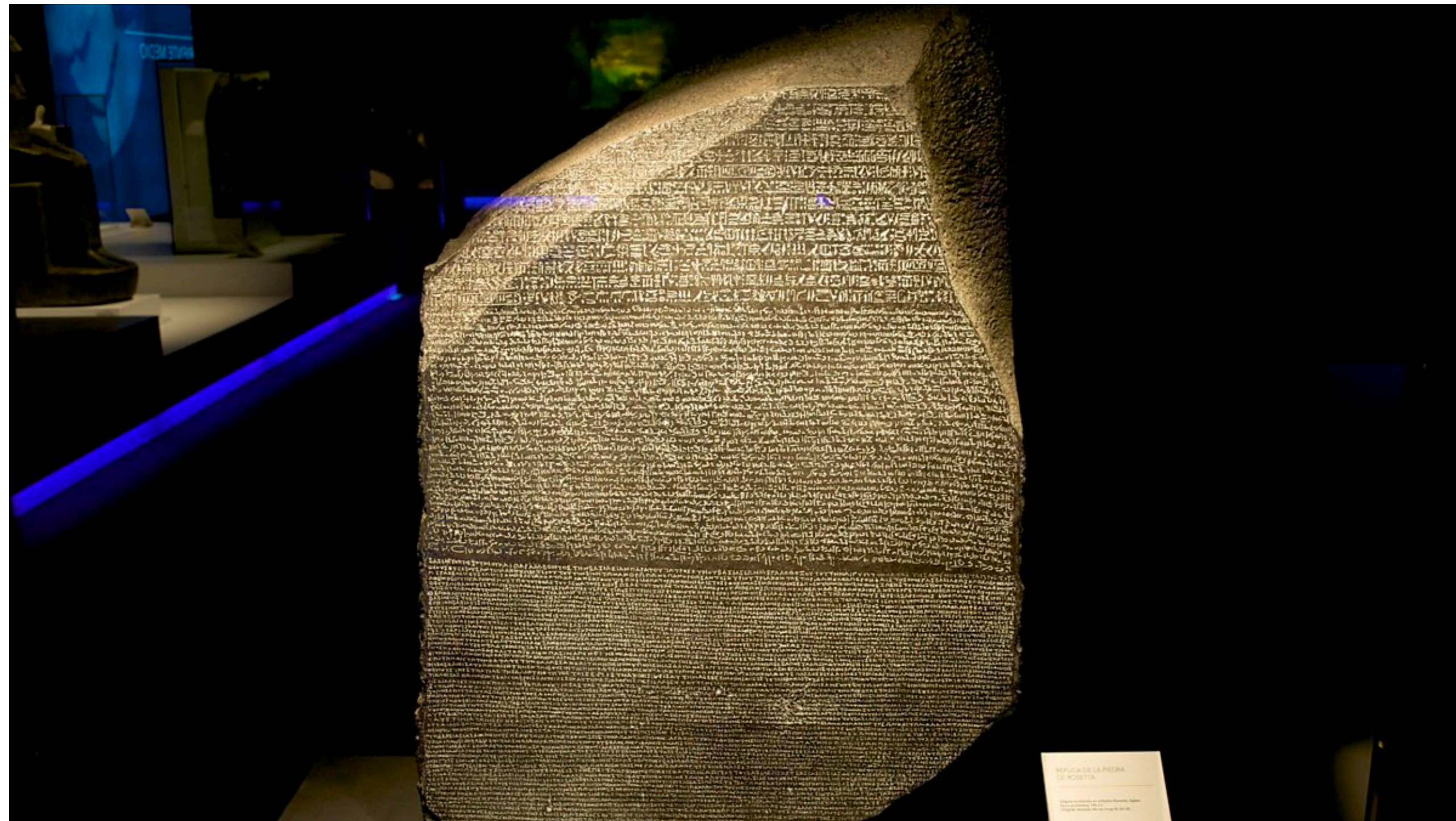


Vanishing Gradient in RNNs



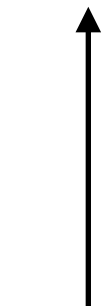
In general, the longer the path, the smaller the gradient signal.

Alignment in Machine Translation



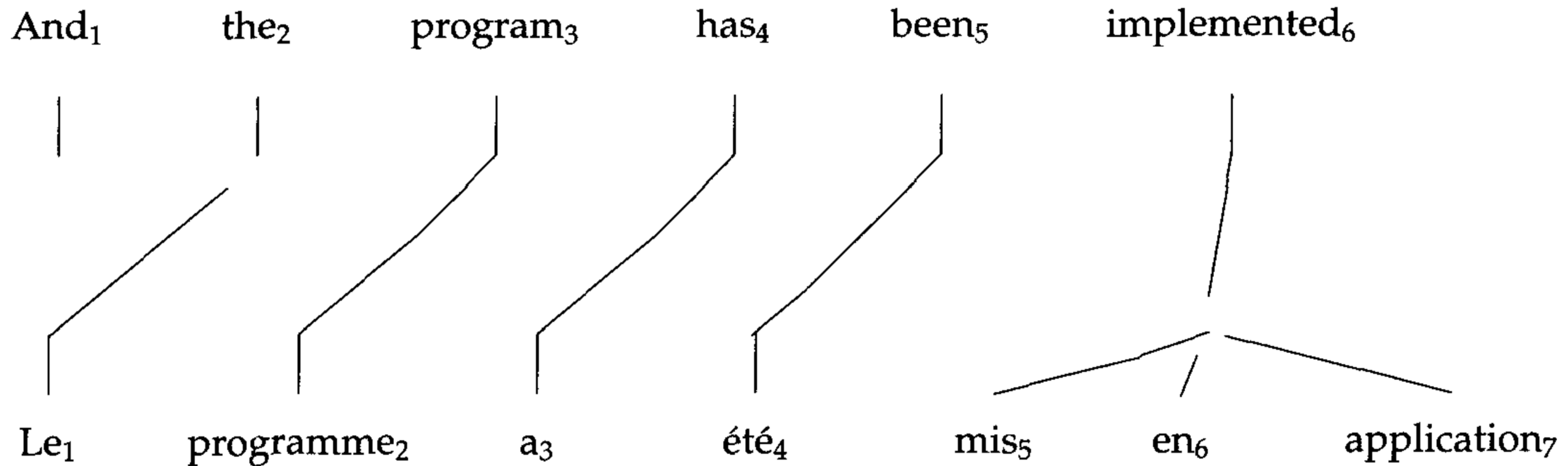
Parallel Corpus

$$p(\mathbf{y} | \mathbf{x})$$



target source

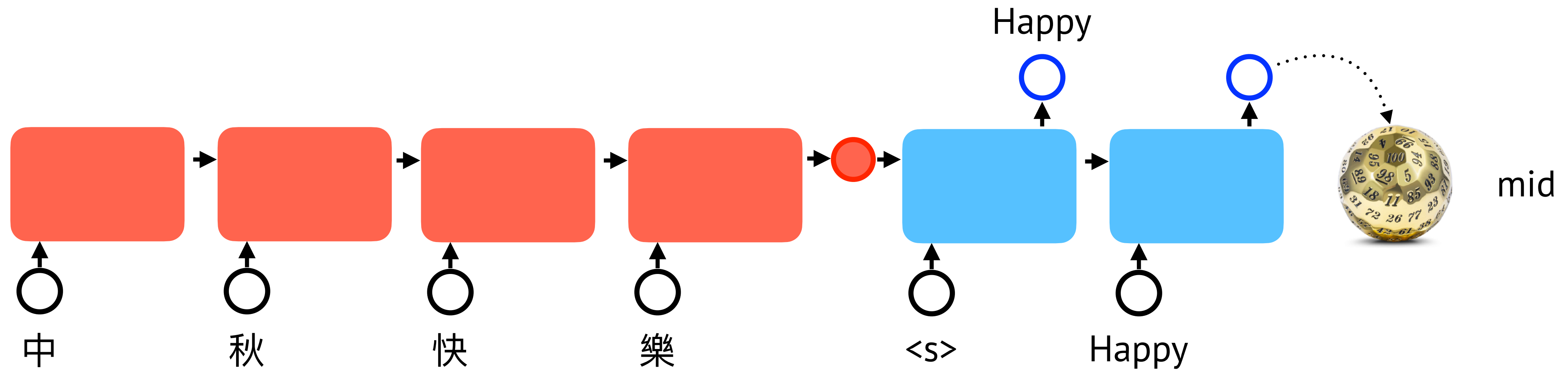
Alignment in Machine Translation




Some words might have no “counter-part”.

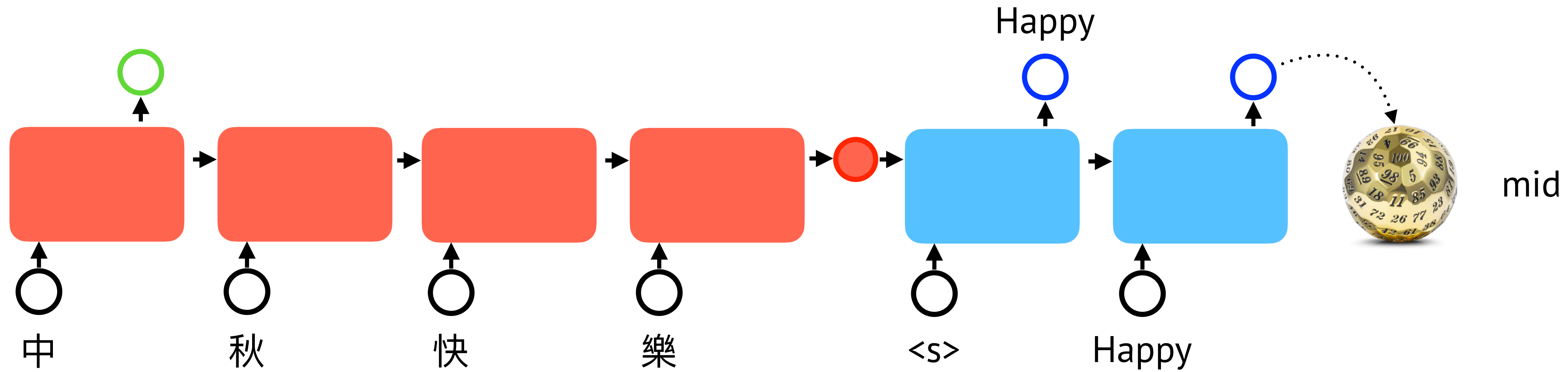
Alignment can be many-to-one (or one-to-many).

Sequence to Sequence Model

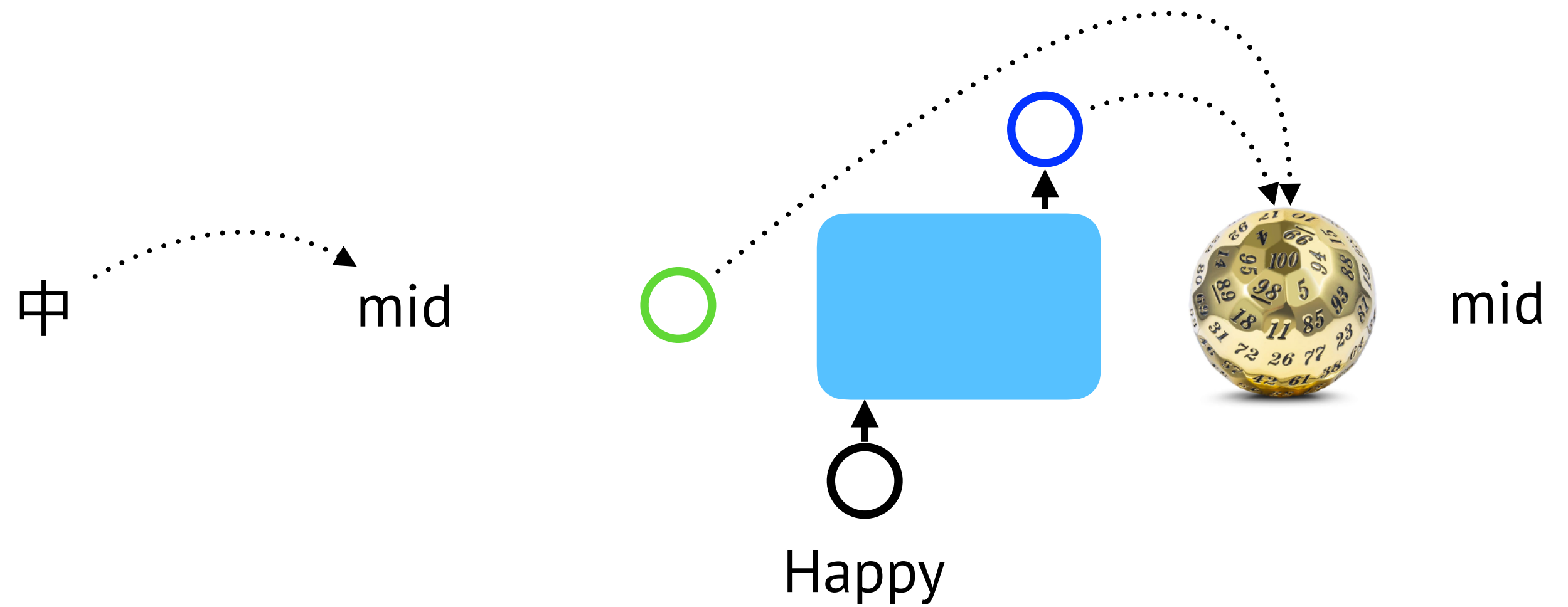
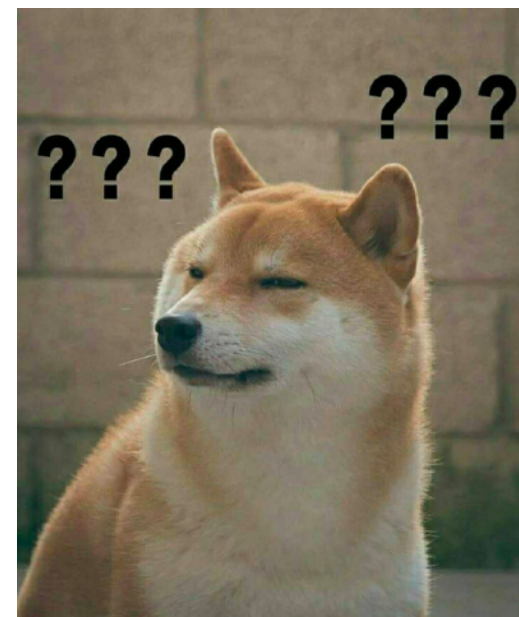



$$p(y_t | \mathbf{y}_{<t}, \mathbf{x})$$

Sequence to Sequence Model

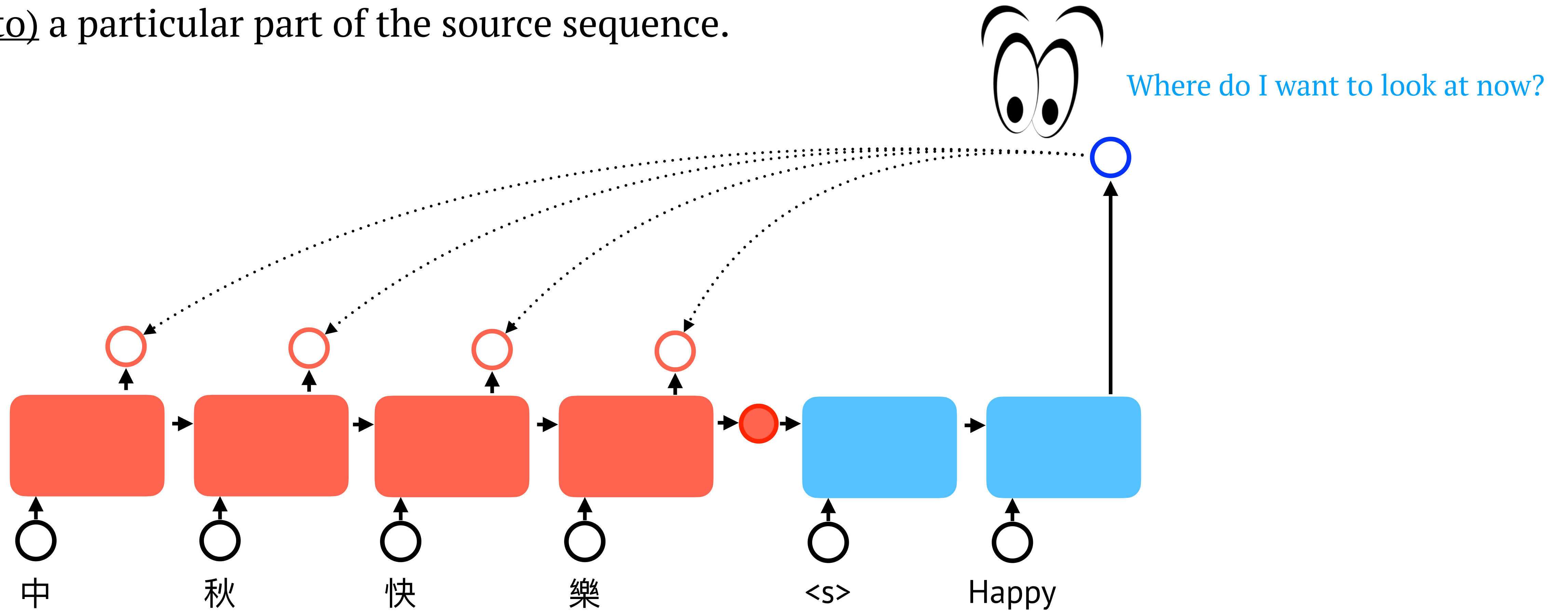


$p(y_t | \mathbf{y}_{<t}, \mathbf{x})$



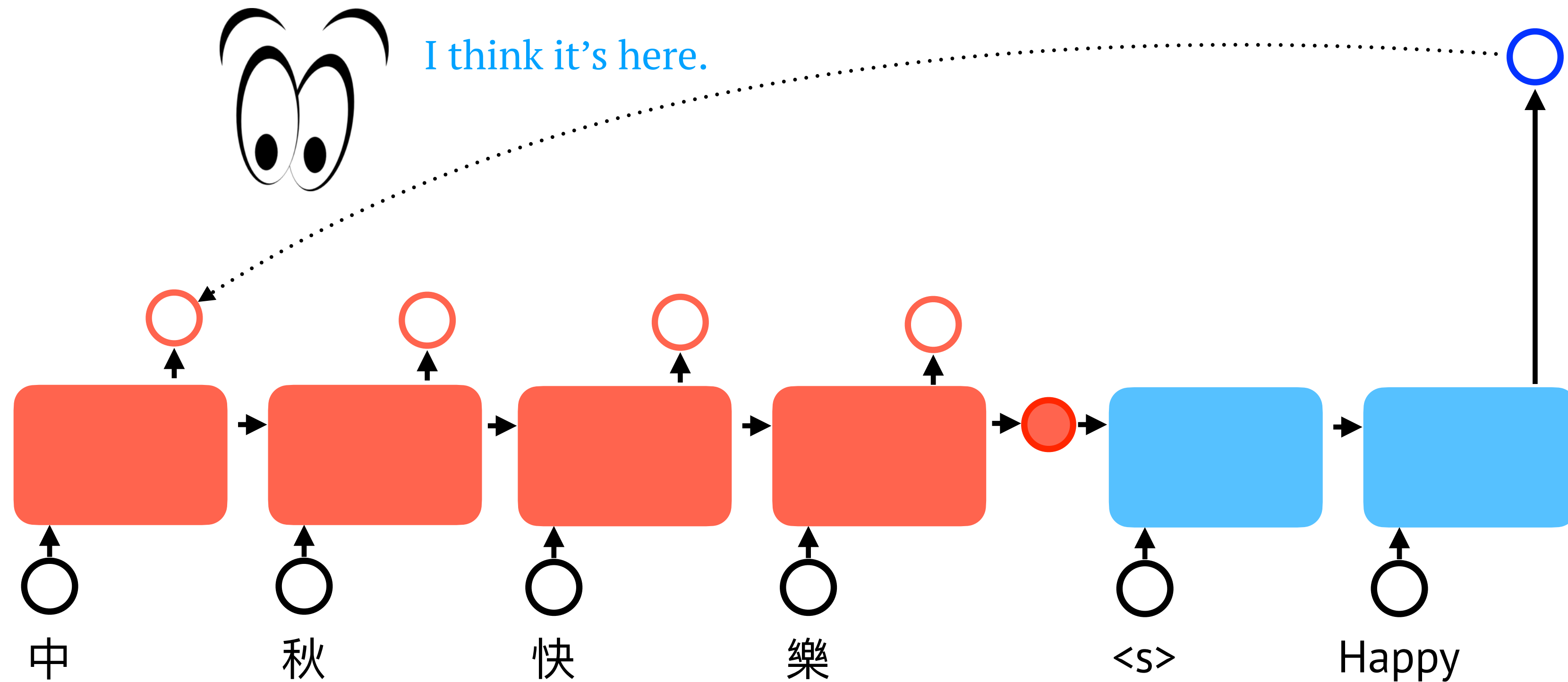
Attention Mechanism

Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.



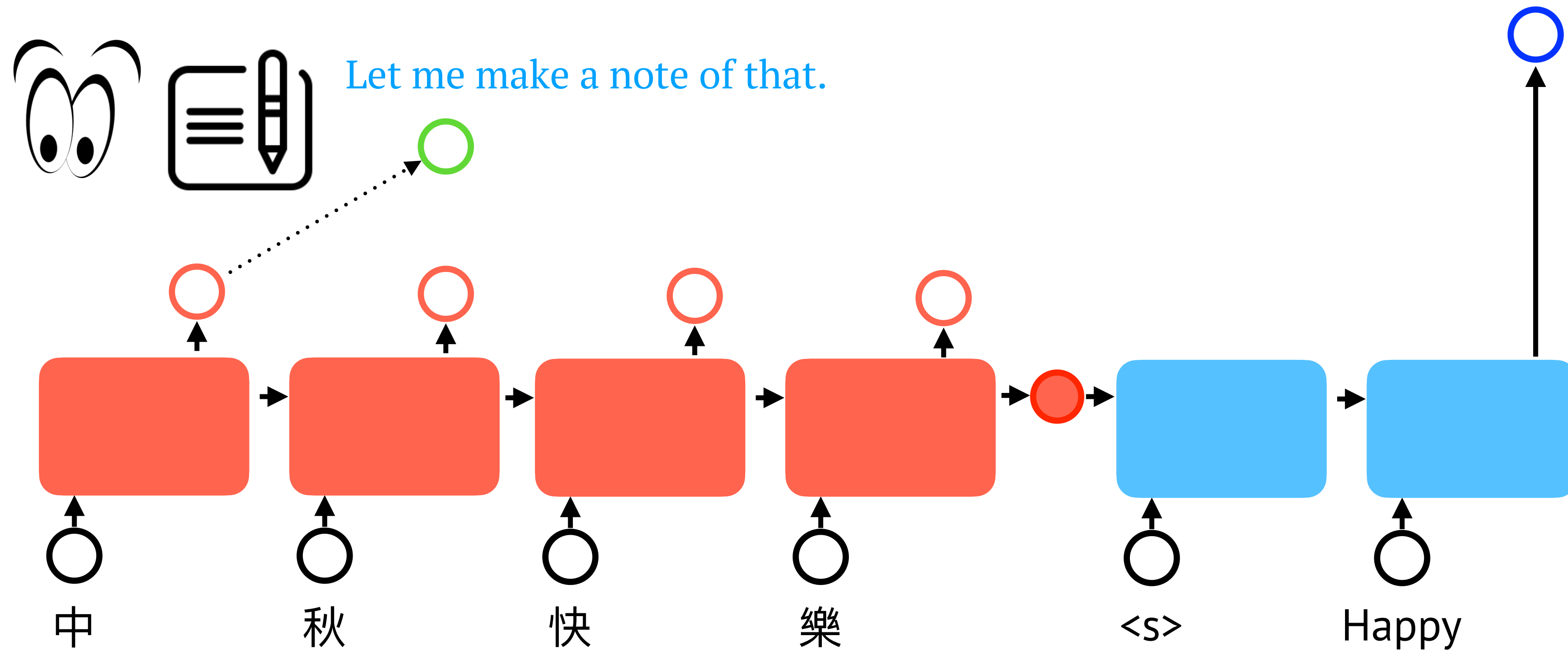
Attention Mechanism

Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.



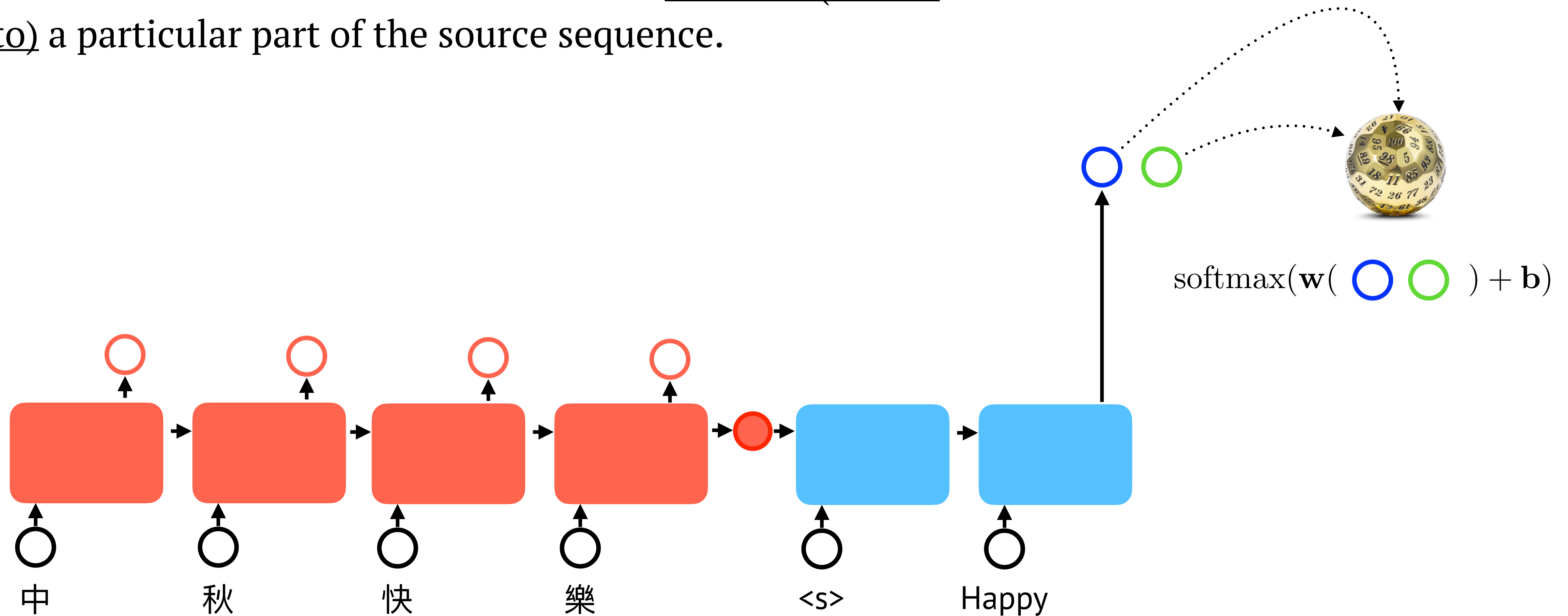
Attention Mechanism

Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.



Attention Mechanism

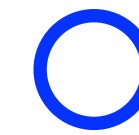
Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.



Memory Abstraction

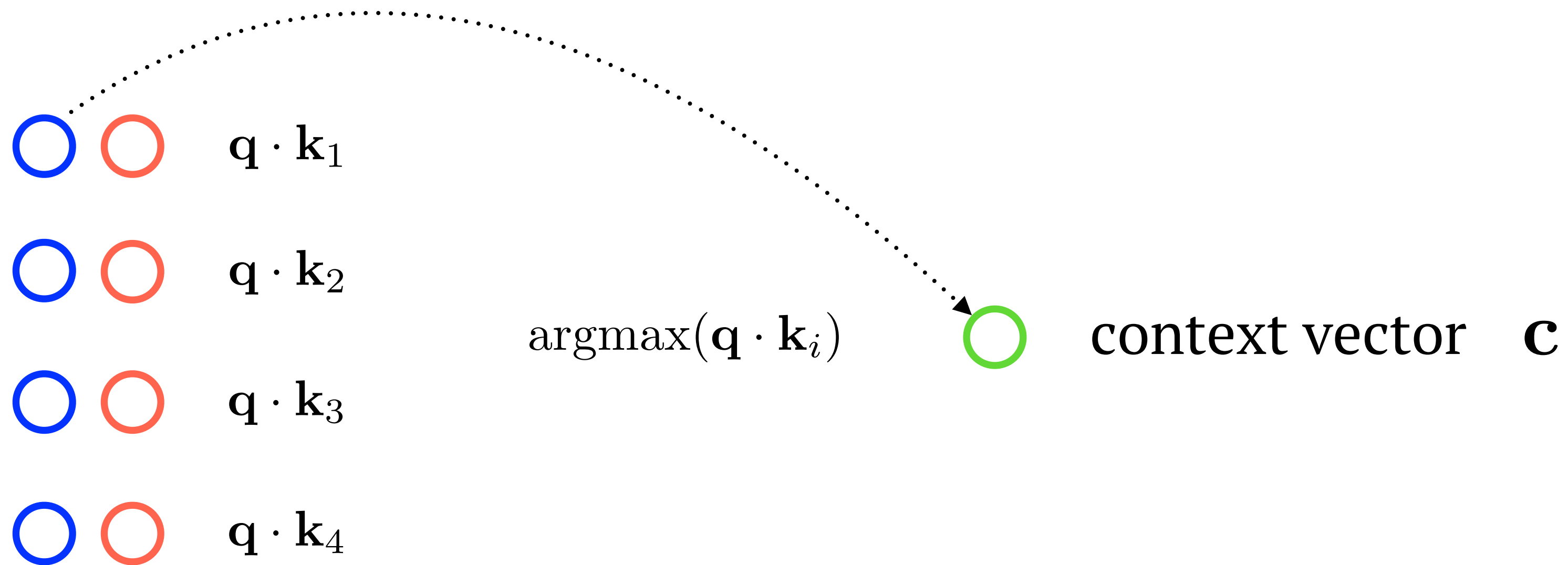


Memory (keys)



Query

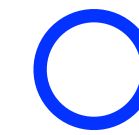
Task: Finding the most “relevant” item in the memory.



Dot-Product-Softmax Attention

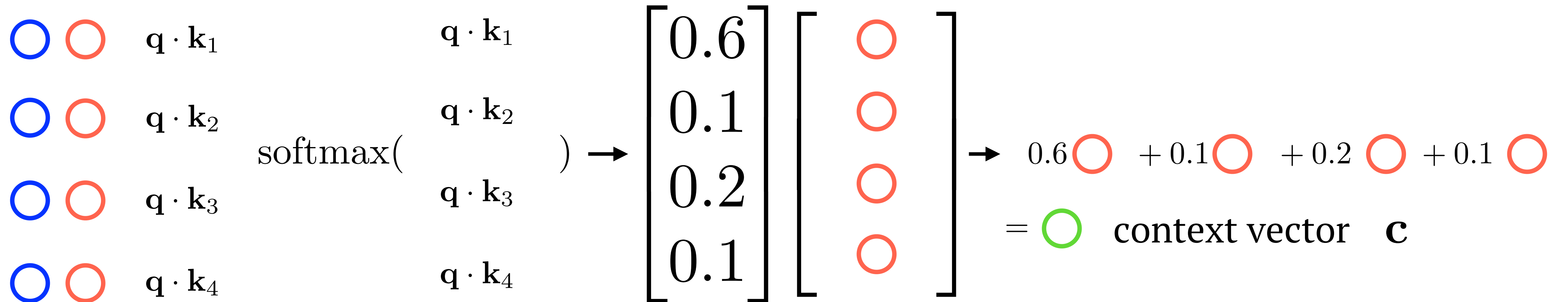


Memory (keys)

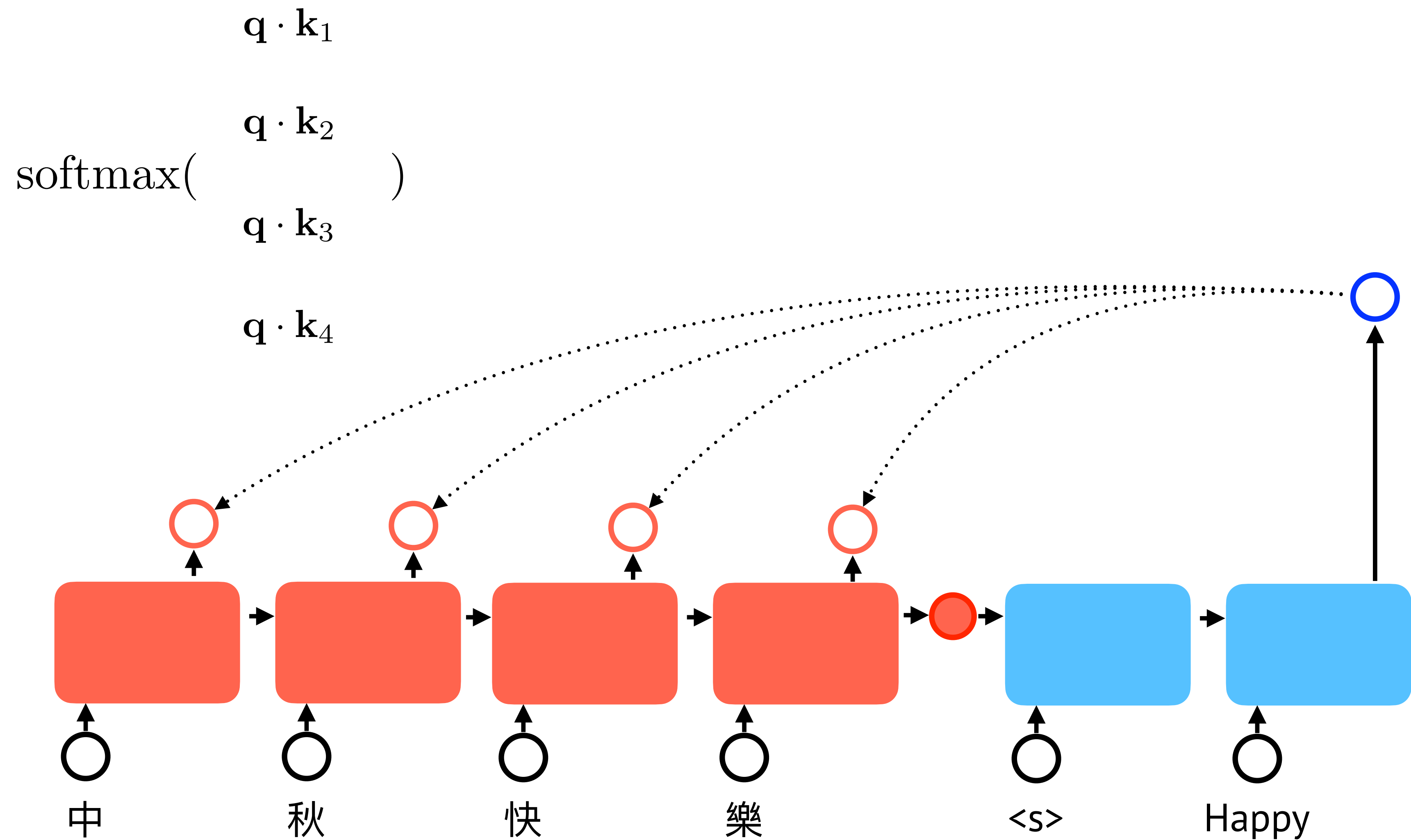


Query

Task: Finding the most “relevant” item in the memory.



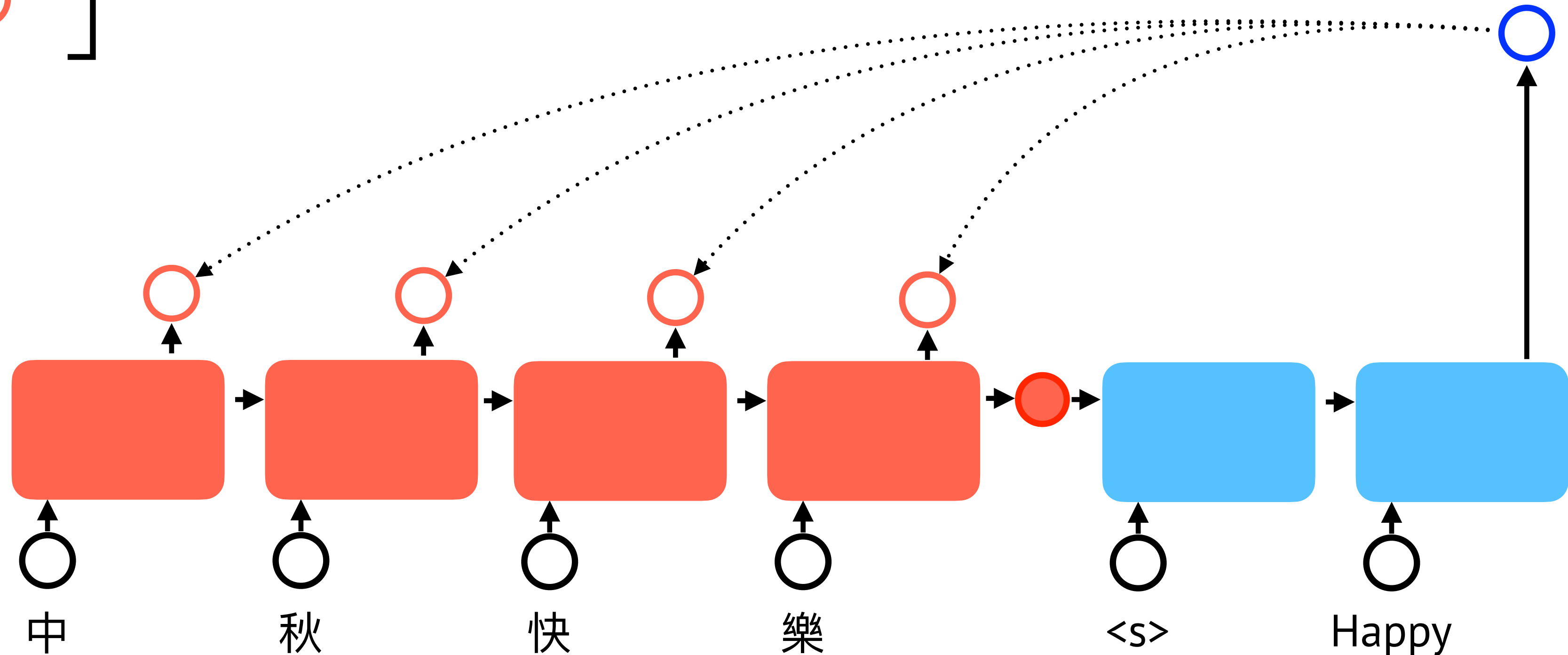
Attention Mechanism



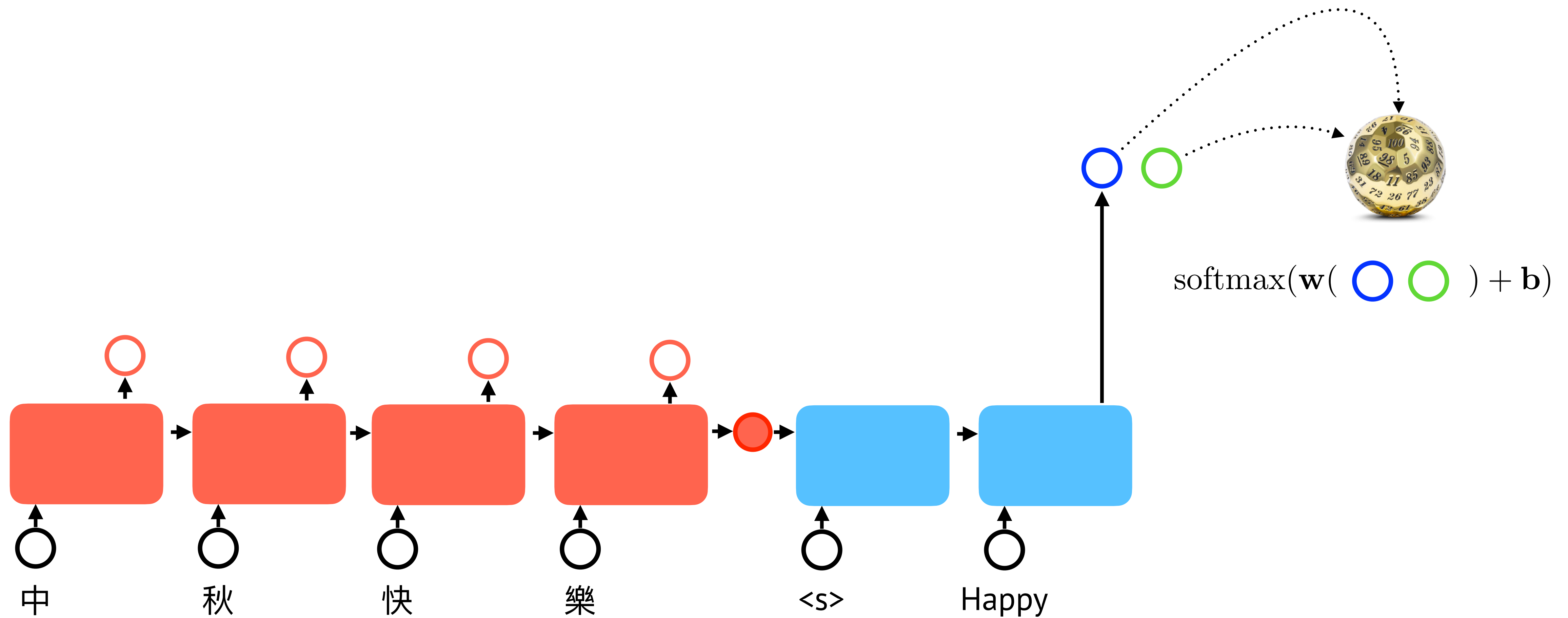
Attention Mechanism

$$\begin{bmatrix} 0.6 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix} \begin{bmatrix} \bigcirc \\ \bigcirc \\ \bigcirc \\ \bigcirc \end{bmatrix}$$

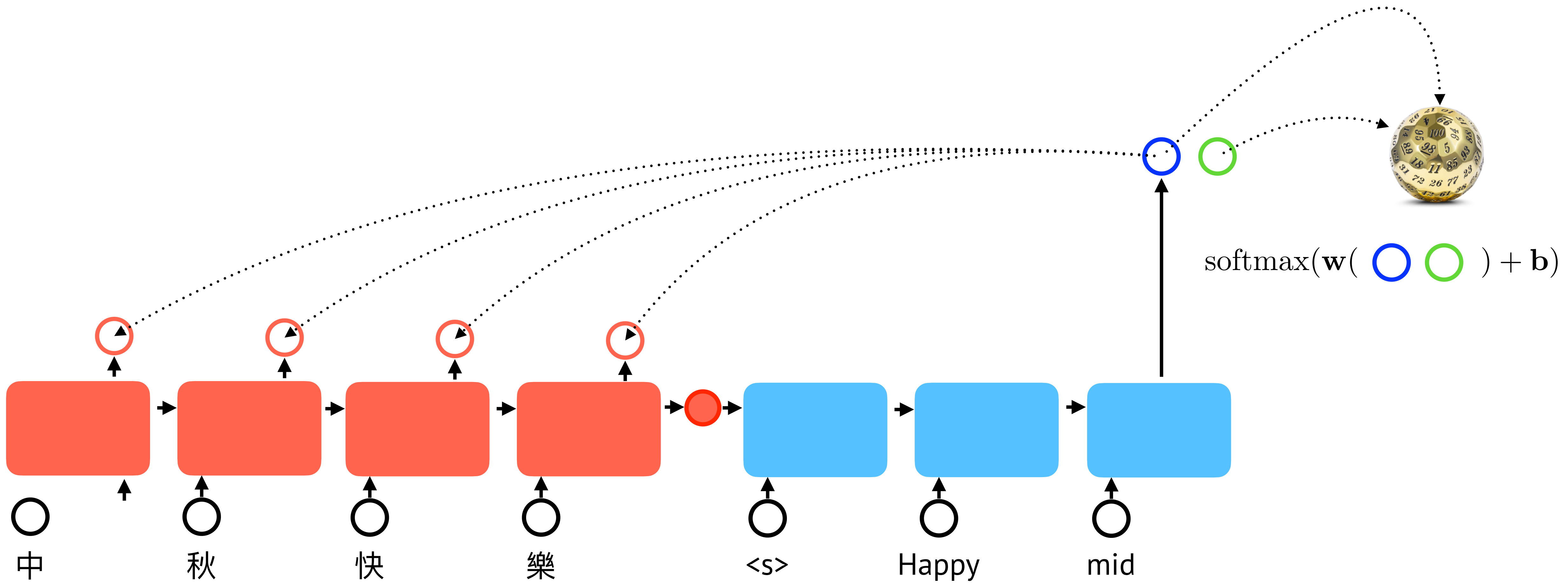
$$\rightarrow 0.6 \bigcirc + 0.1 \bigcirc + 0.2 \bigcirc + 0.1 \bigcirc$$
$$= \bigcirc \text{ context vector } \mathbf{c}$$



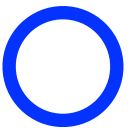
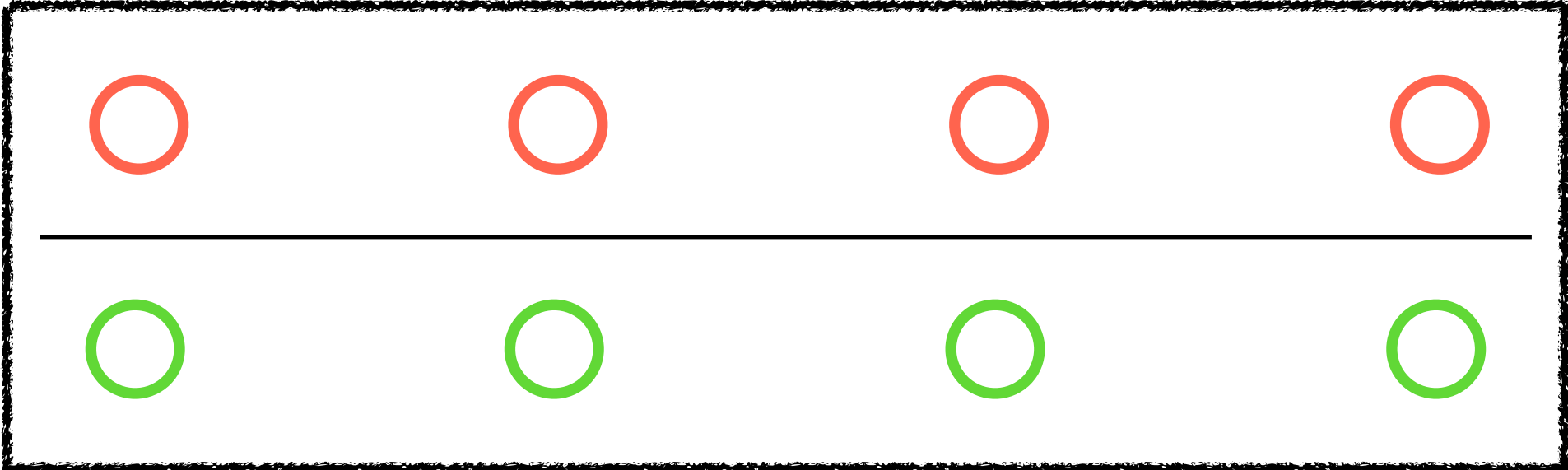
Attention Mechanism



Attention Mechanism



Dot-Product-Softmax Attention



Query

Memory (key-value pairs)

$$\begin{array}{l}
 \text{○} \text{○} \quad \mathbf{q} \cdot \mathbf{k}_1 \quad \mathbf{q} \cdot \mathbf{k}_1 \\
 \text{○} \text{○} \quad \mathbf{q} \cdot \mathbf{k}_2 \quad \mathbf{q} \cdot \mathbf{k}_2 \\
 \text{○} \text{○} \quad \mathbf{q} \cdot \mathbf{k}_3 \quad \mathbf{q} \cdot \mathbf{k}_3 \\
 \text{○} \text{○} \quad \mathbf{q} \cdot \mathbf{k}_4 \quad \mathbf{q} \cdot \mathbf{k}_4
 \end{array}
 \text{softmax}(\quad) \rightarrow
 \begin{bmatrix} 0.6 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix}
 \begin{bmatrix} \text{○} \\ \text{○} \\ \text{○} \\ \text{○} \end{bmatrix}
 \rightarrow
 0.6 \text{○} + 0.1 \text{○} + 0.2 \text{○} + 0.1 \text{○}$$

= ○ context vector \mathbf{c}

Dot-Product-Softmax Attention

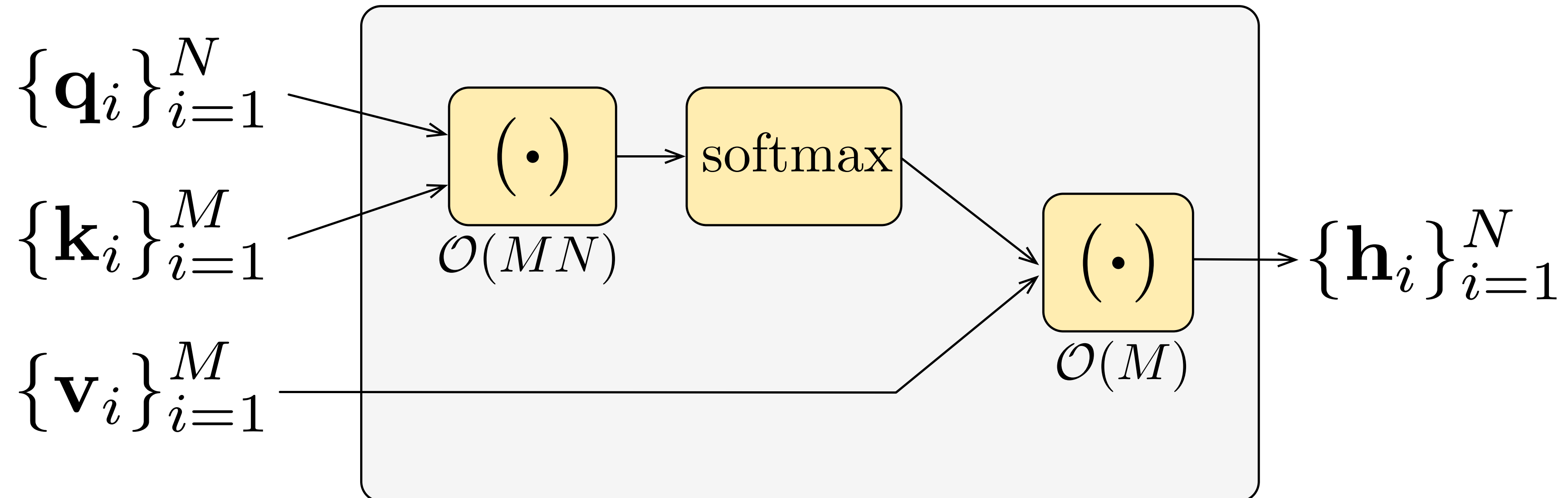
$$\sum_{m=1}^M \frac{\exp(\mathbf{q}_n \mathbf{k}_m)}{\sum_{m'=1}^M \exp(\mathbf{q}_n \mathbf{k}_{m'})} \mathbf{v}_m^\top = \mathbf{V}^\top \text{softmax}(\mathbf{K} \mathbf{q}_n)$$

The diagram highlights the components of the attention mechanism:

- similarity**: $\exp(\mathbf{q}_n \mathbf{k}_m)$ (indicated by a red box)
- normalized similarity**: $\frac{\exp(\mathbf{q}_n \mathbf{k}_m)}{\sum_{m'=1}^M \exp(\mathbf{q}_n \mathbf{k}_{m'})}$ (indicated by a blue box)

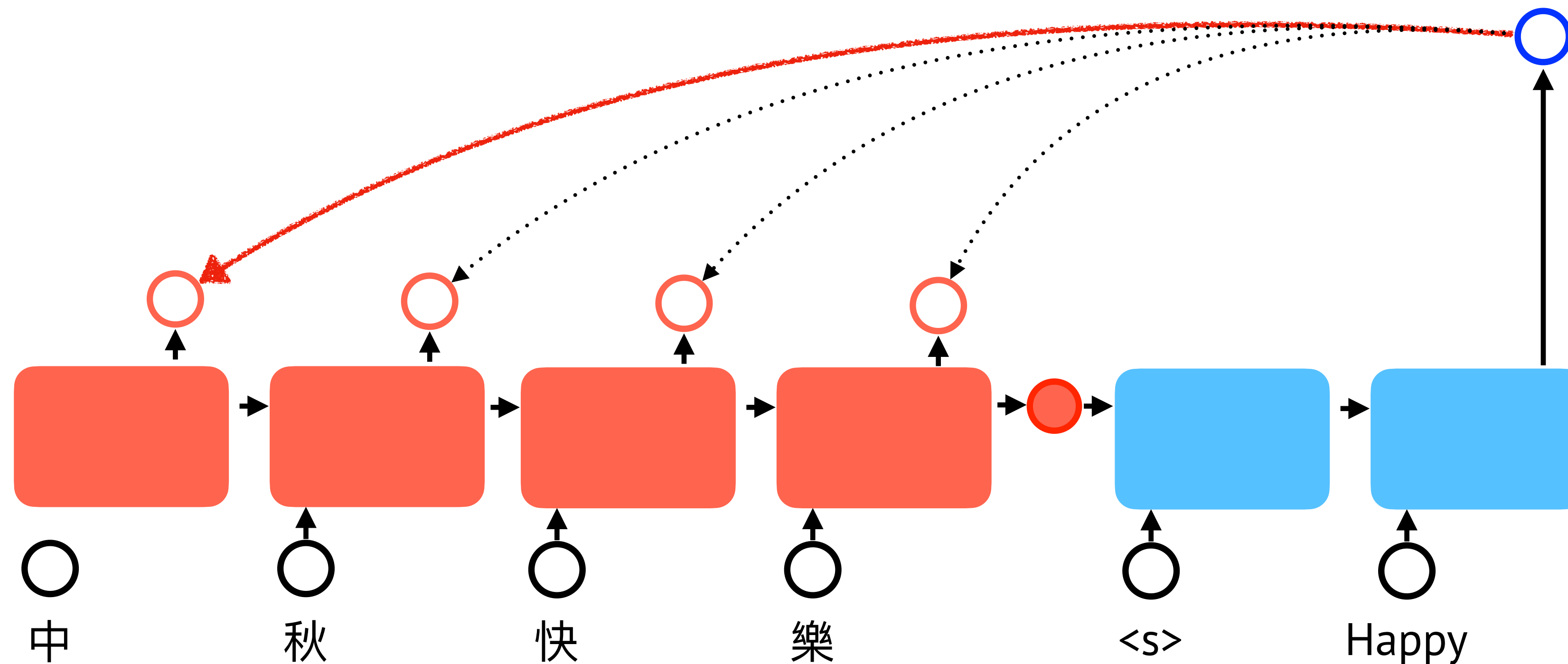
weighted sum

Computational Complexity



Attention Mechanism

It helps with vanishing gradient problem.



Attention Mechanism

It offers some interpretability.

