Natural Language Processing

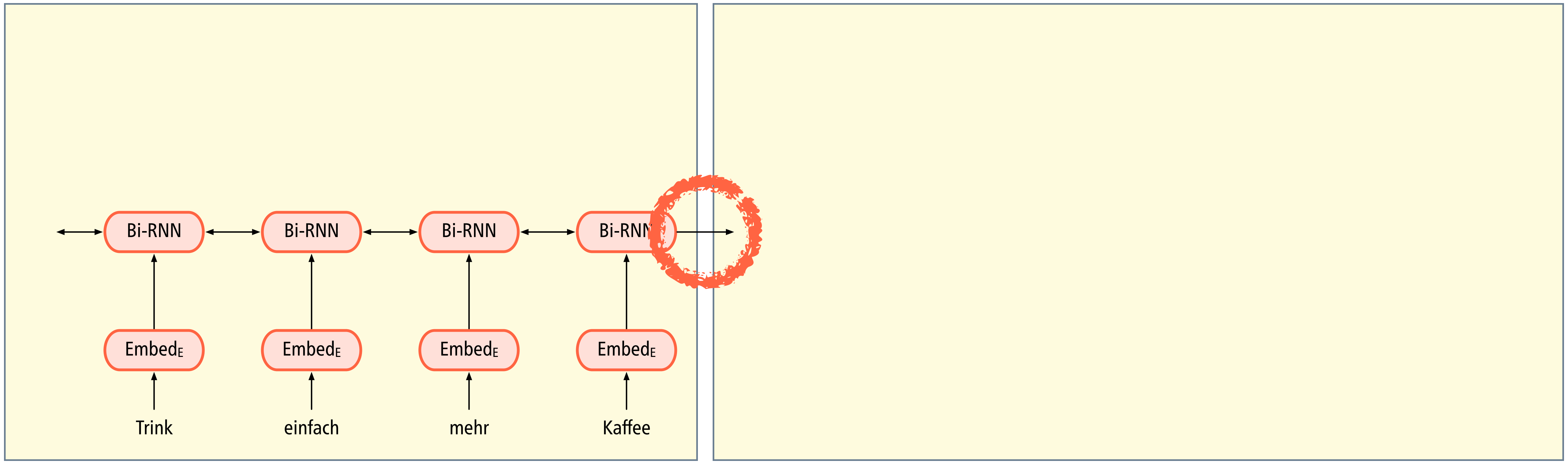
Attention

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING UNIVERSITY

Recency bias in recurrent neural networks

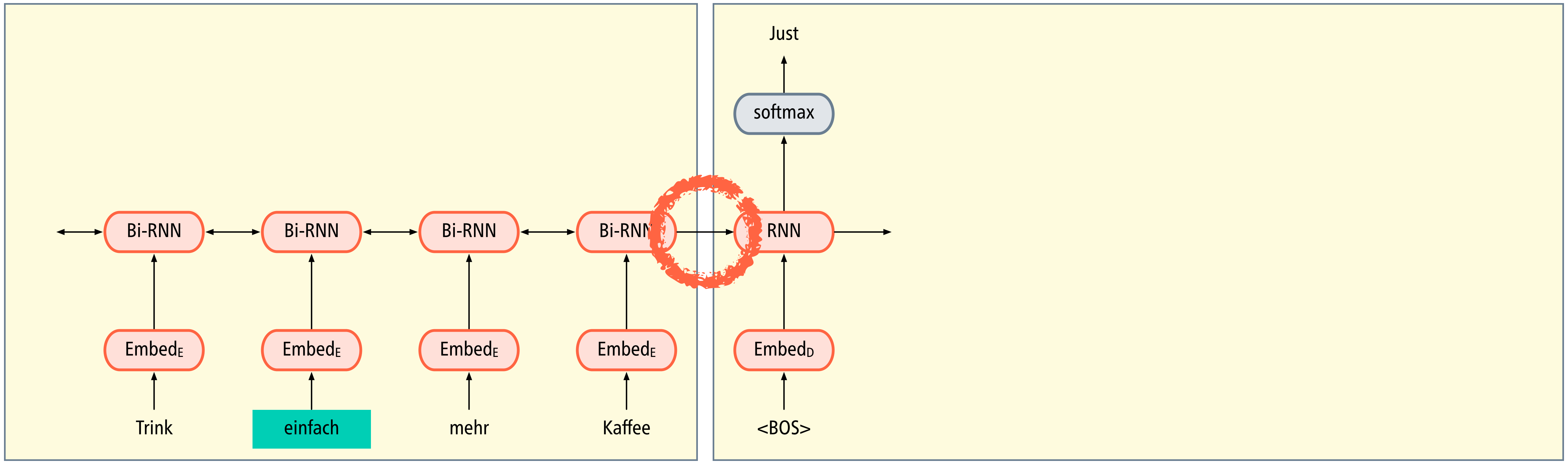


encoder

decoder

Sutskever et al. (2014)

Recency bias in recurrent neural networks

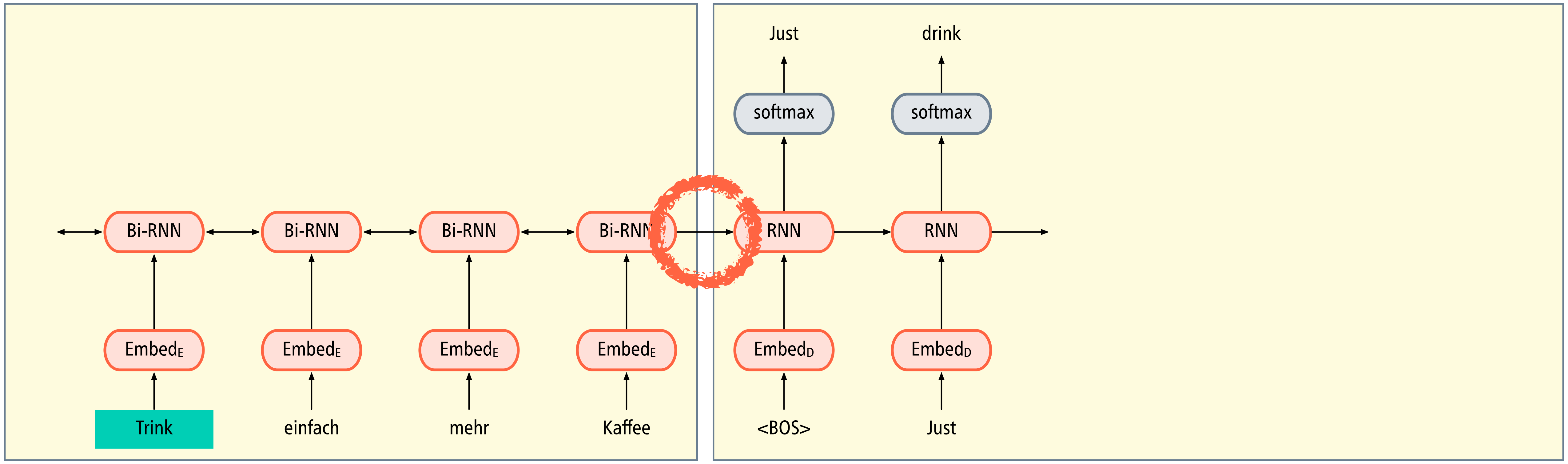


encoder

decoder

Sutskever et al. (2014)

Recency bias in recurrent neural networks



encoder

decoder

Sutskever et al. (2014)

Attention

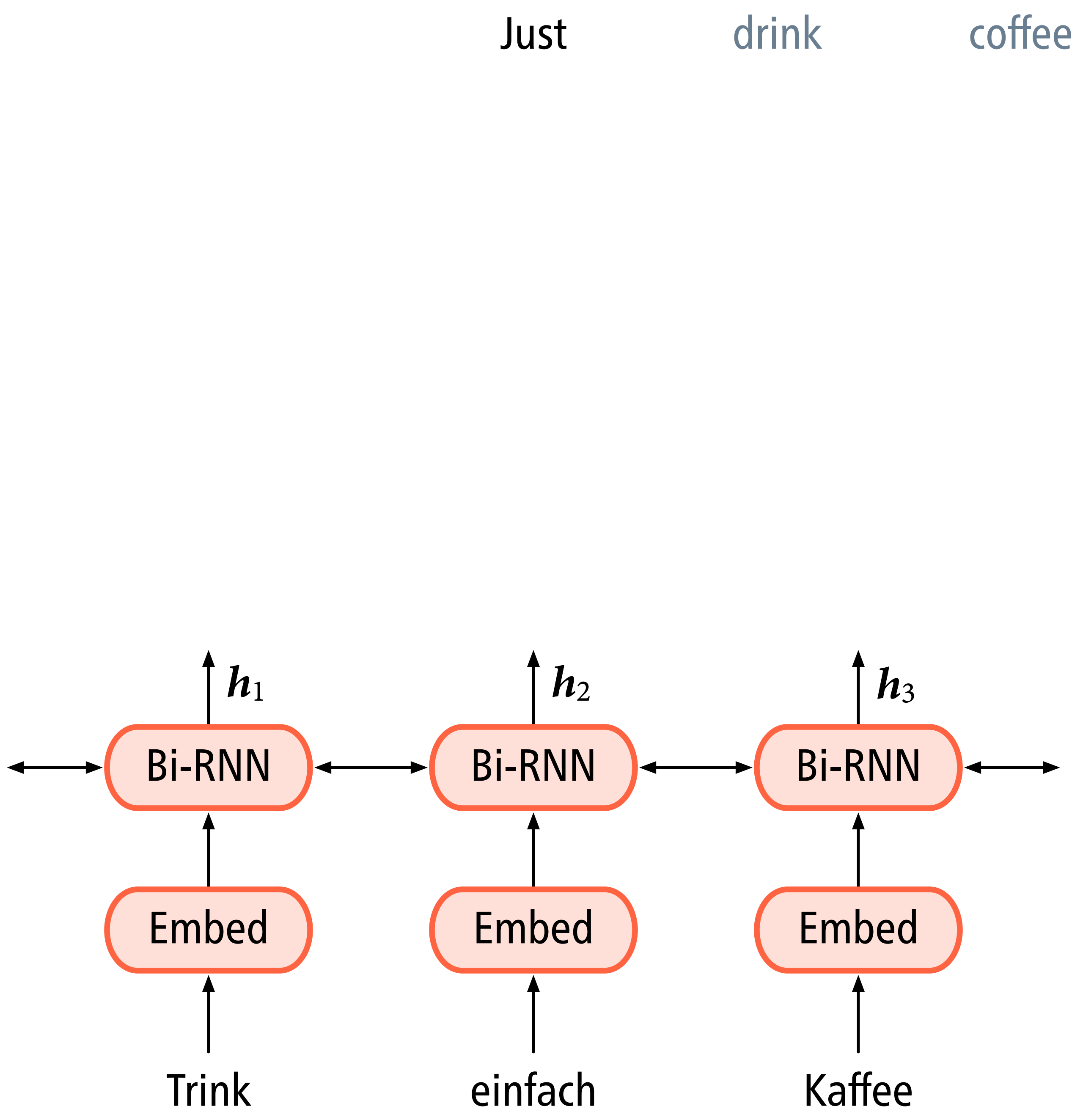
- In the context of machine translation, **attention** enables the model to learn “soft” word alignments.

- Essentially, we compute a set of weights that allow us to score words based on how much the model should “attend to them”.

- Attention was first proposed in the context of the sequence-to-sequence architecture, but is now used in many architectures.

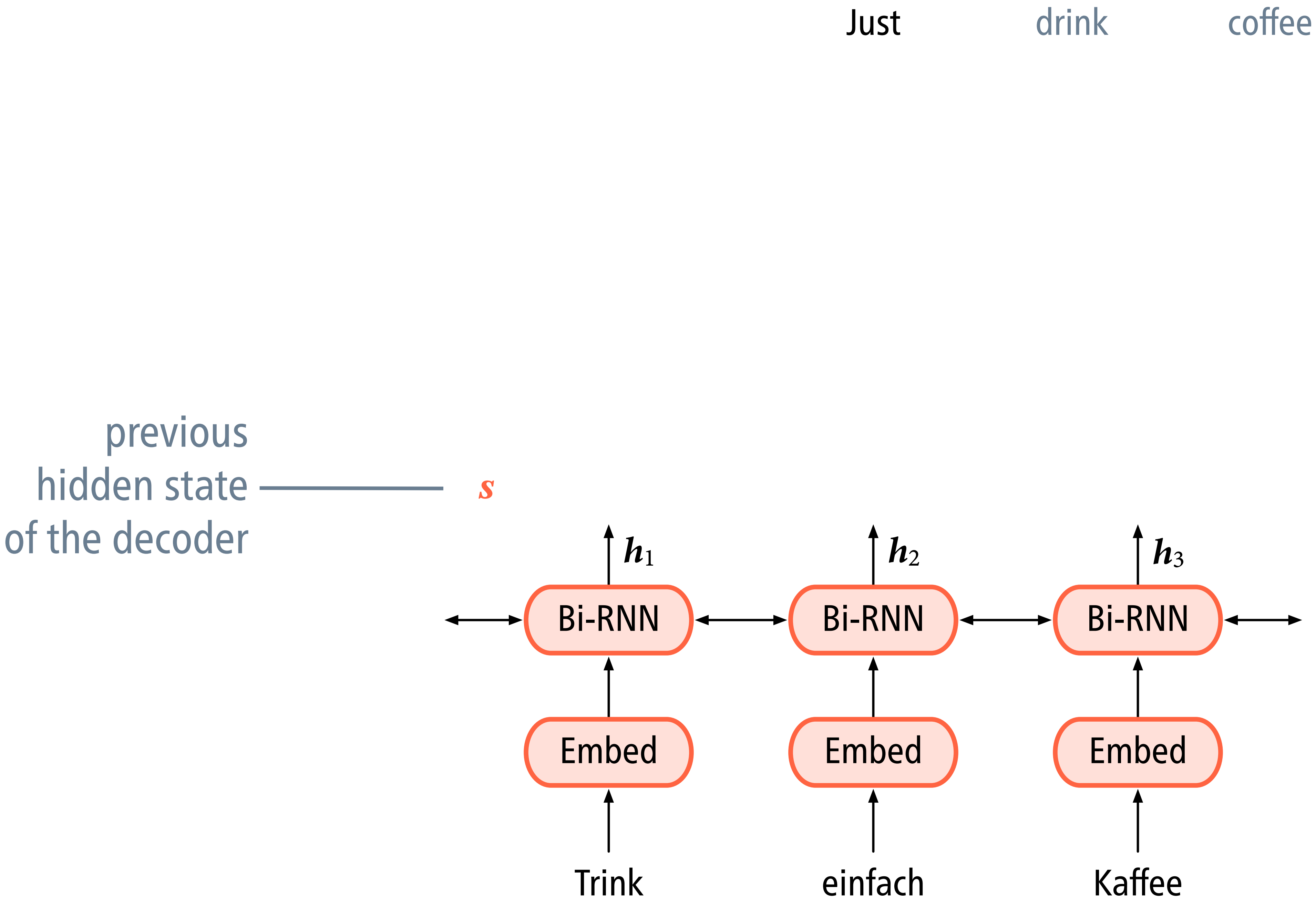
  Bahdanau et al. (2015)

Attention for translation



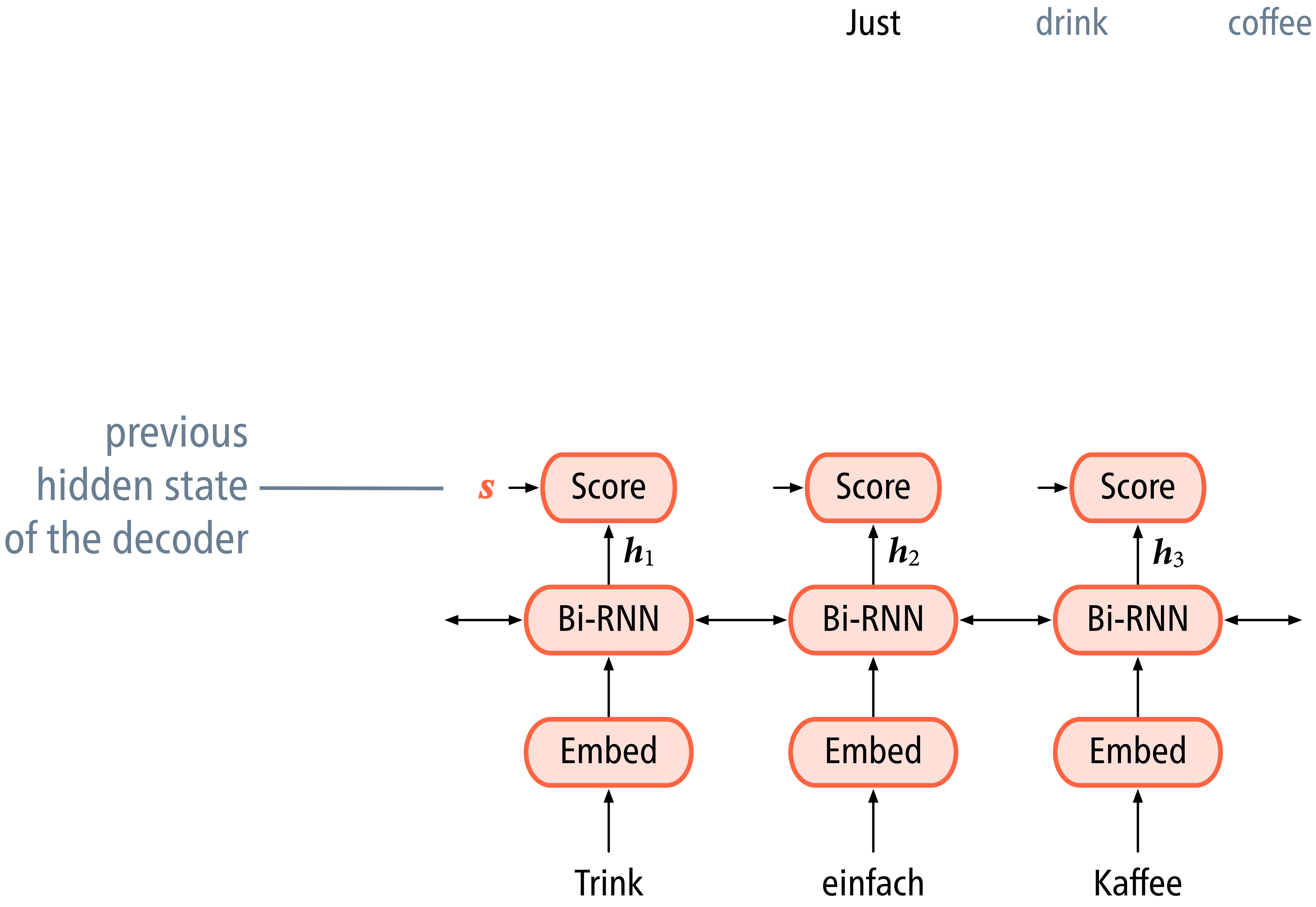
Bahdanau et al. (2015)

Attention for translation



Bahdanau et al. (2015)

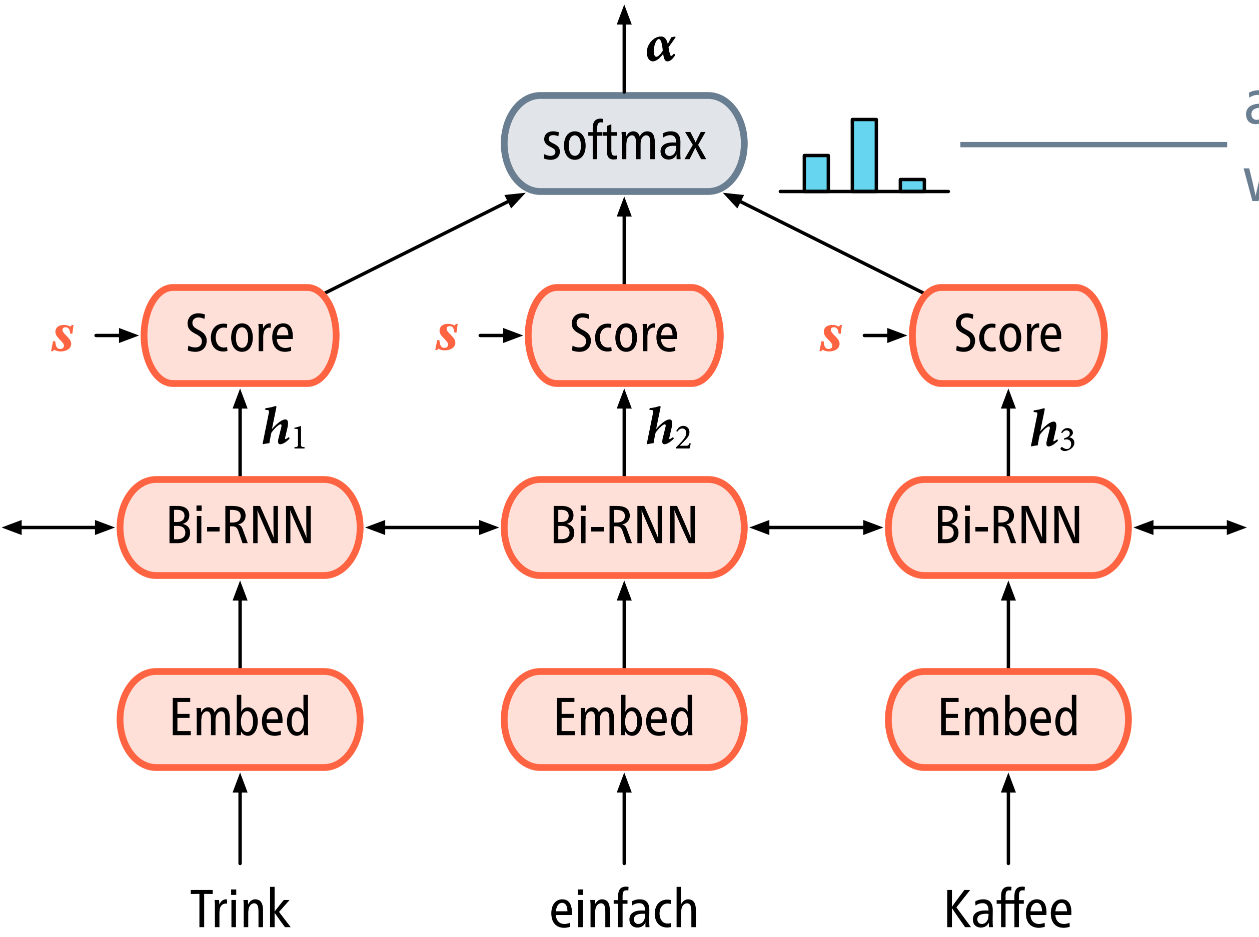
Attention for translation



Bahdanau et al. (2015)

Attention for translation

Just drink coffee

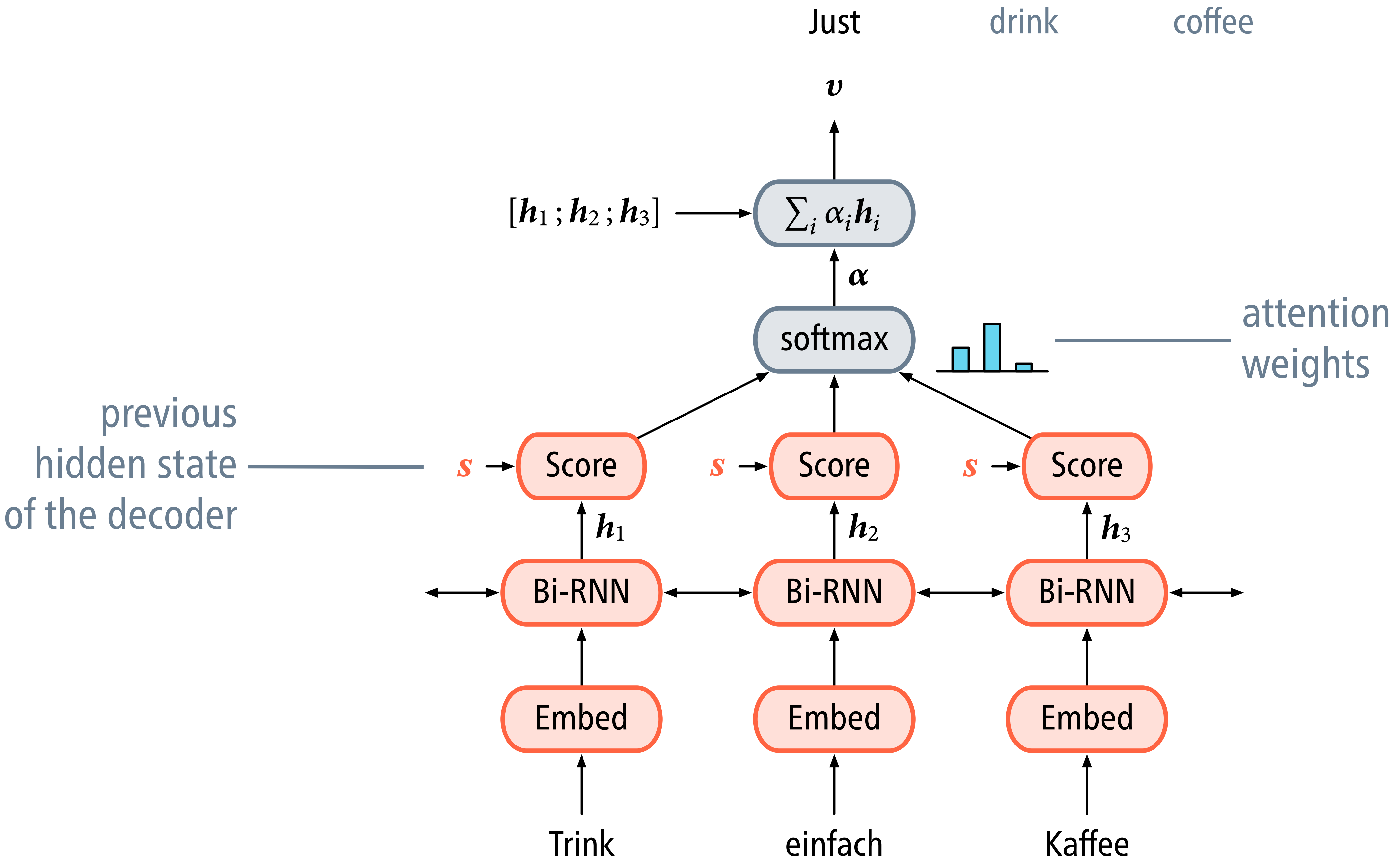


attention weights

previous hidden state of the decoder

Bahdanau et al. (2015)

Attention for translation



Bahdanau et al. (2015)

A general characterisation of attention

- In general, attention can be described as a mapping from a query $\boldsymbol{q}$ and a set of key–value pairs $\boldsymbol{k}_i$, $\boldsymbol{v}_i$ to an output.

- The output is the weighted sum of the $\boldsymbol{v}_i$, where the weight of each $\boldsymbol{v}_i$ is given by the affinity between $\boldsymbol{q}$ and $\boldsymbol{k}_i$:

$$\text{attention}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\big(\underbrace{a(\boldsymbol{q}, \boldsymbol{K})}_{\text{attention score}}\big)\boldsymbol{V}$$

$\boldsymbol{q} \in \mathbb{R}^{d_K}, \boldsymbol{K} \in \mathbb{R}^{n \times d_K}, \boldsymbol{V} \in \mathbb{R}^{n \times d_V}$

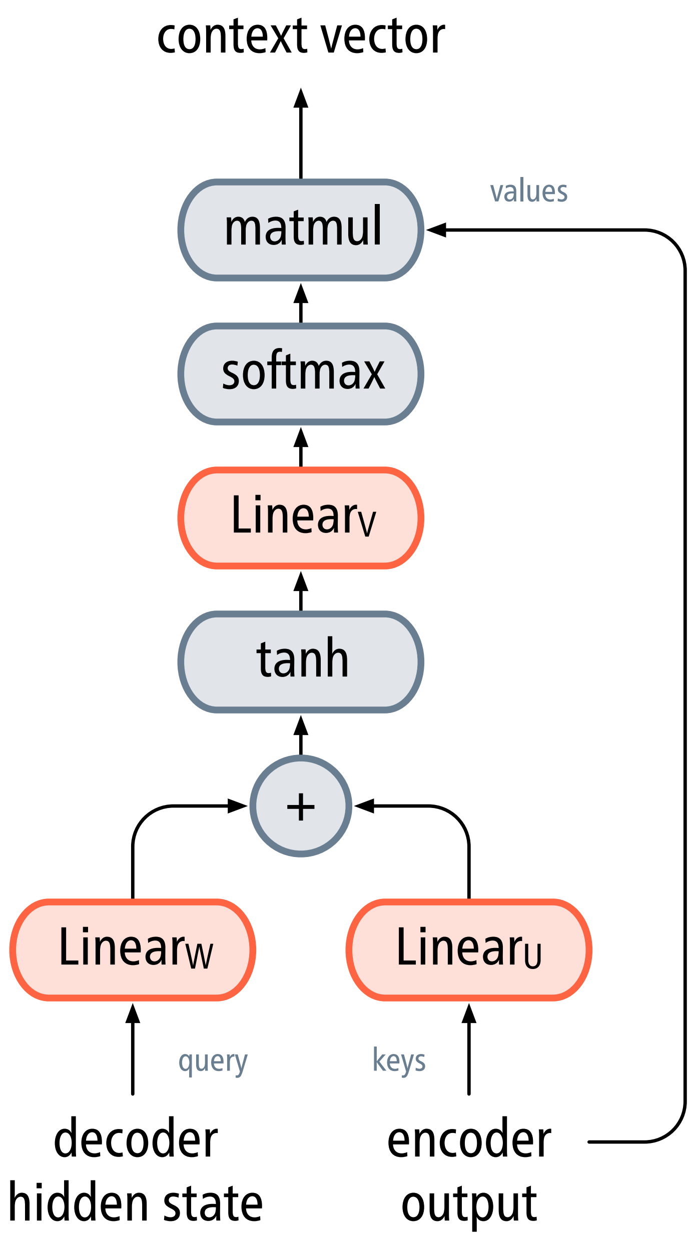
Vaswani et al. (2017)

Bahdanau attention

previous decoder
hidden state

$$a(\boldsymbol{s}_{i-1}, \boldsymbol{h}_j) = \boldsymbol{v}^\top \tanh(\boldsymbol{W}\boldsymbol{s}_{i-1} + \boldsymbol{U}\boldsymbol{h}_j)$$

encoder hidden
state at position j

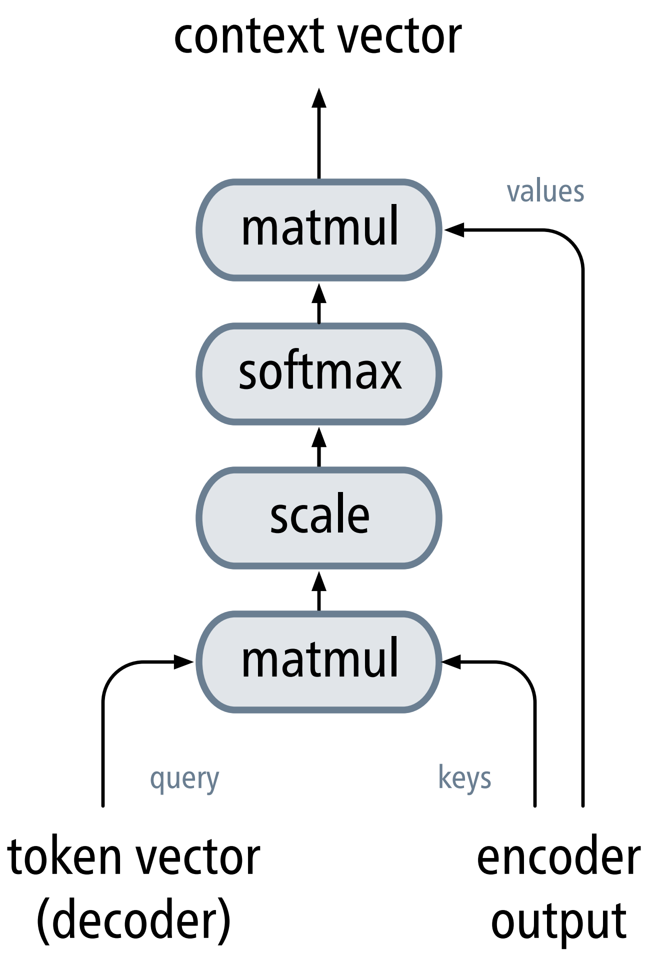


Scaled dot-product attention

token vector
at position i

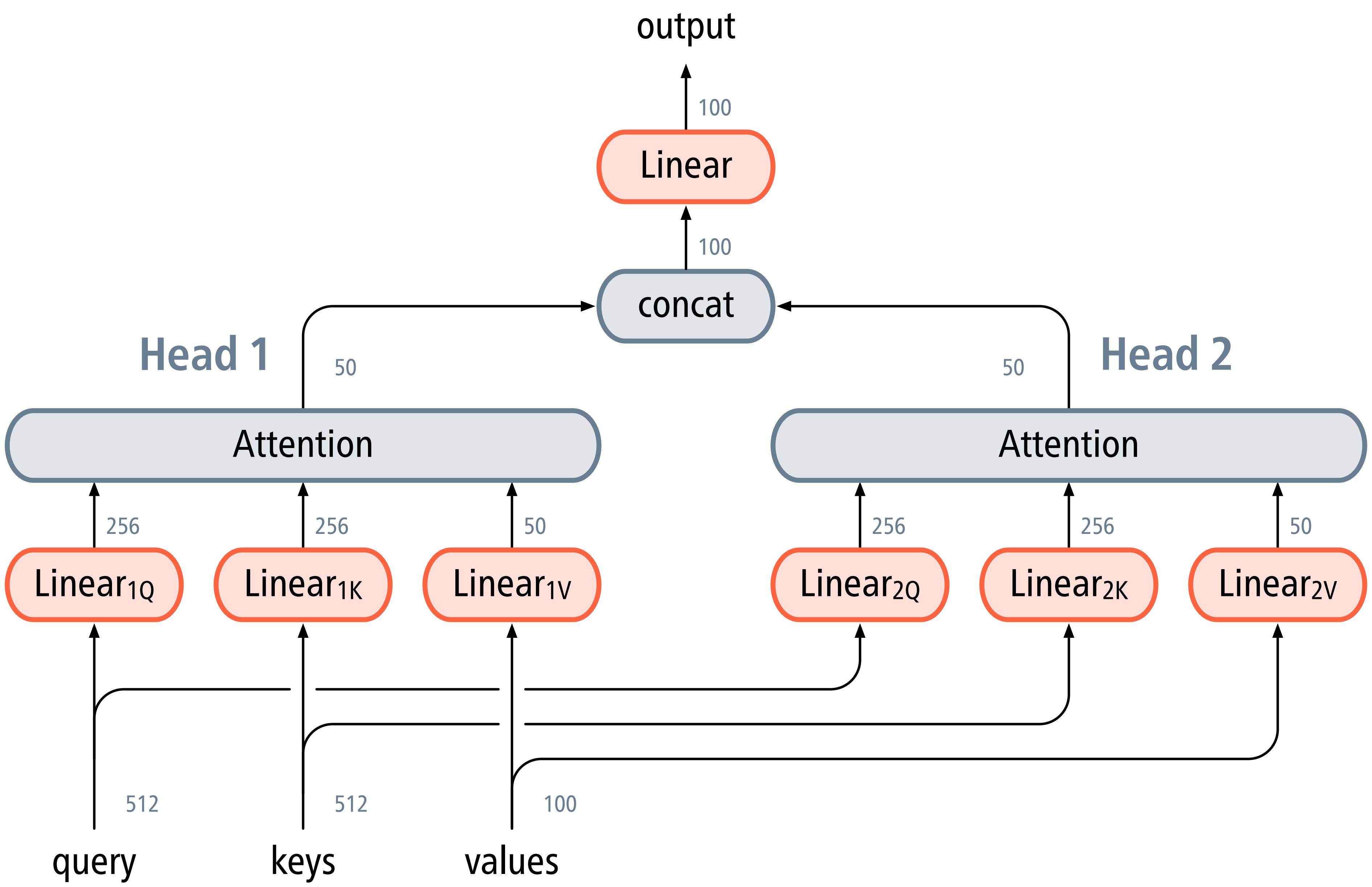
$$a(\boldsymbol{h}_i, \boldsymbol{h}_j) = \frac{\boldsymbol{h}_i \boldsymbol{h}_j^\top}{\sqrt{d_K}}$$

encoder output
at position j

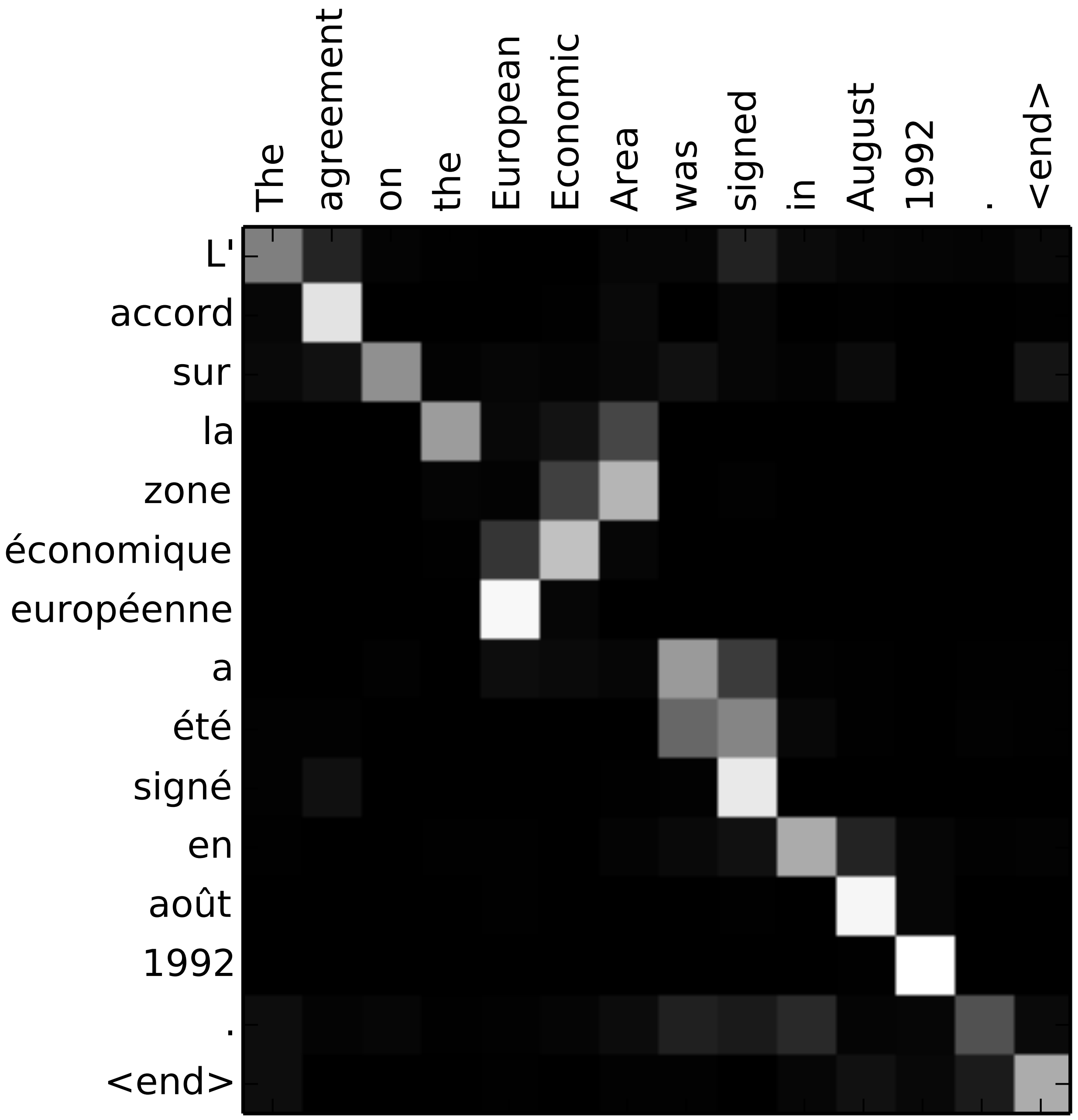


Vaswani et al. (2017)

Multi-head attention



Attention as word alignments



In the context of the encoder–decoder architecture for neural machine translation, attention weights resemble soft word alignments.

Image source: Bahdanau et al. (2015)