Attention Mechanism

Machine Translation

- Before 2014: Statistical Machine Translation (SMT)
  - Extremely complex systems that require massive human efforts
  - Separately designed components
  - A lot of feature engineering
  - Lots of linguistic domain knowledge and expertise

- Before 2016:
  - Google Translate is based on statistical machine learning

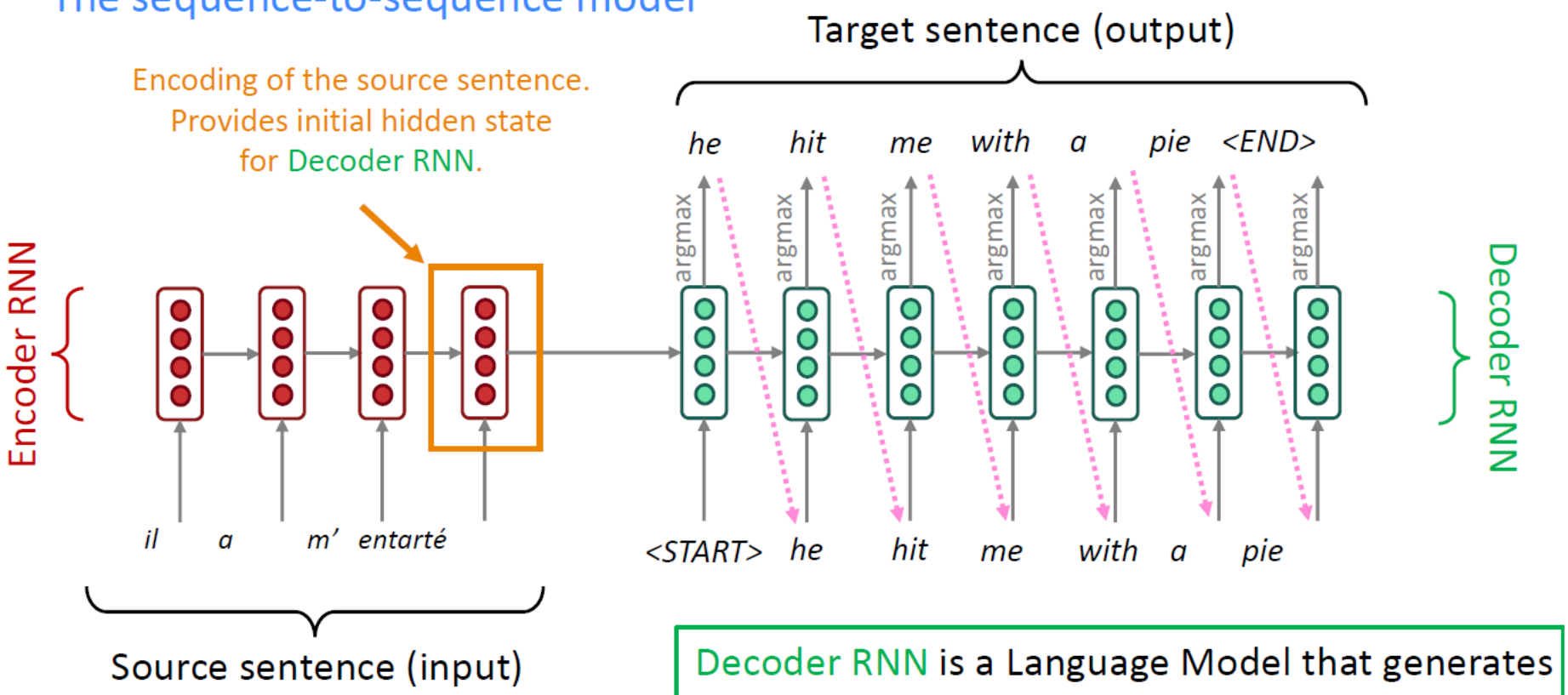
- What happened in 2014?
  - Neural machine translation (NMT)

Sequence to Sequence Model

- Neural Machine Translation (NMT)
  - Learning to translate via a **single end-to-end** neural network.
  - Source language sentence X, target language sentence $Y = f(X; \theta)$

- Sequence to Sequence Model (Seq2Seq, Sutskever et al. , '14)
  - Two RNNs: f_{enc} and f_{dec}
  - Encoder f_{enc}:
    - Takes X as input, and output the initial hidden state for decoder
    - Can use bidirectional RNN
  - Decoder f_{dec}:
    - It takes in the hidden state from f_{enc} to generate Y
    - Can use autoregressive language model

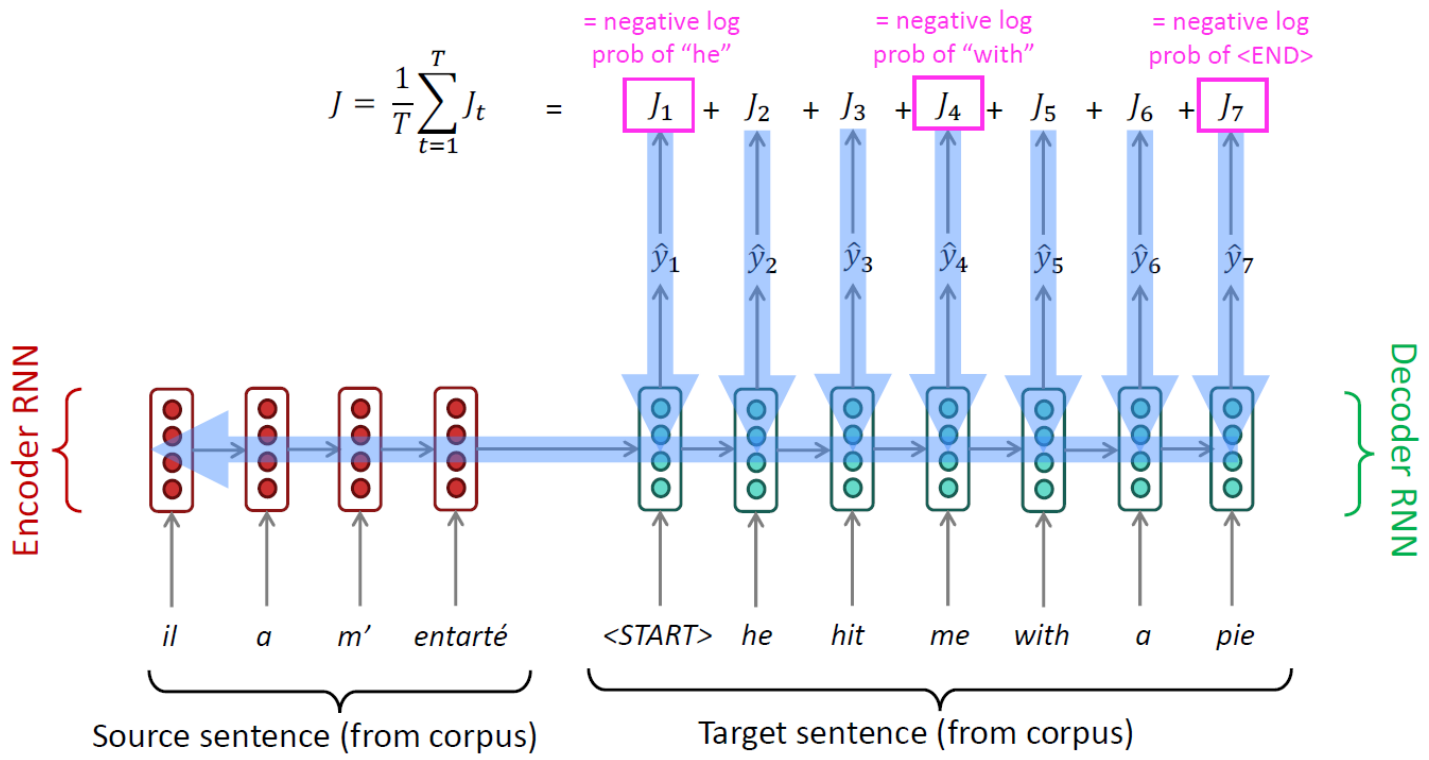
Sequence to Sequence Model



Training Sequence to Sequence Model

- Collect a huge paired dataset and train it end-to-end via BPTT
- Loss induced by MLE $P(Y|X) = P(Y|f_{enc}(X))$

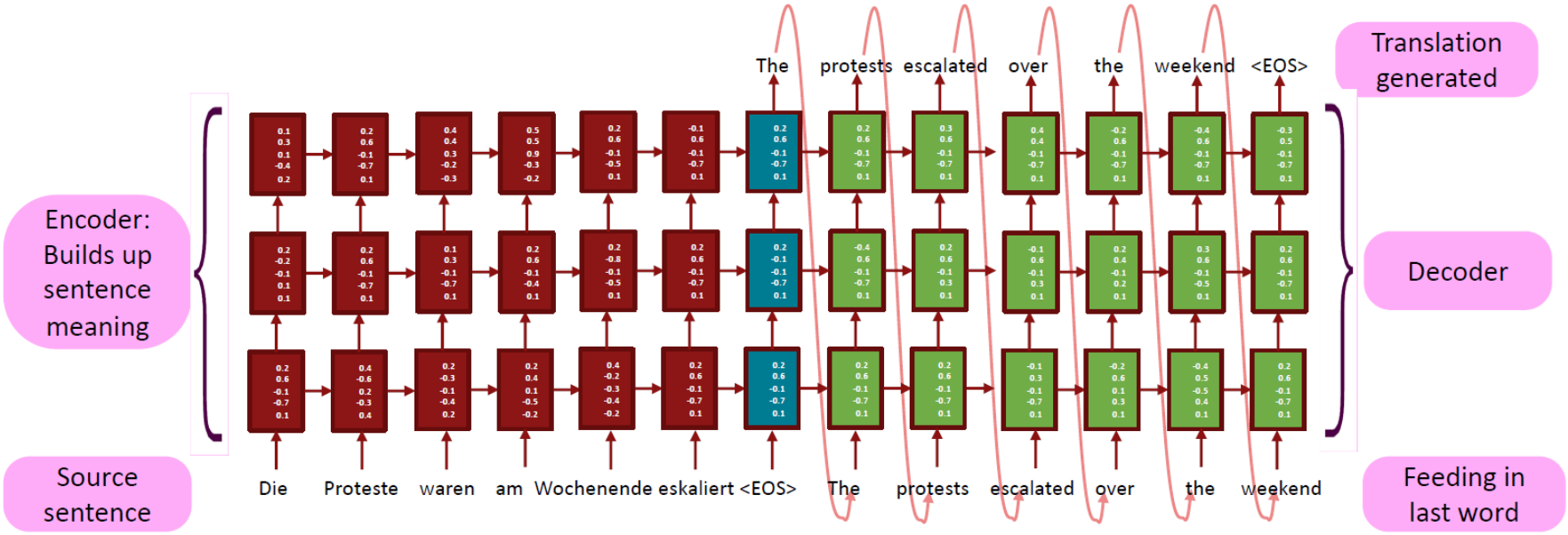
Back Propagation Through Time



Seq2seq is optimized as a **single system.** Backpropagation operates *"end-to-end"*.

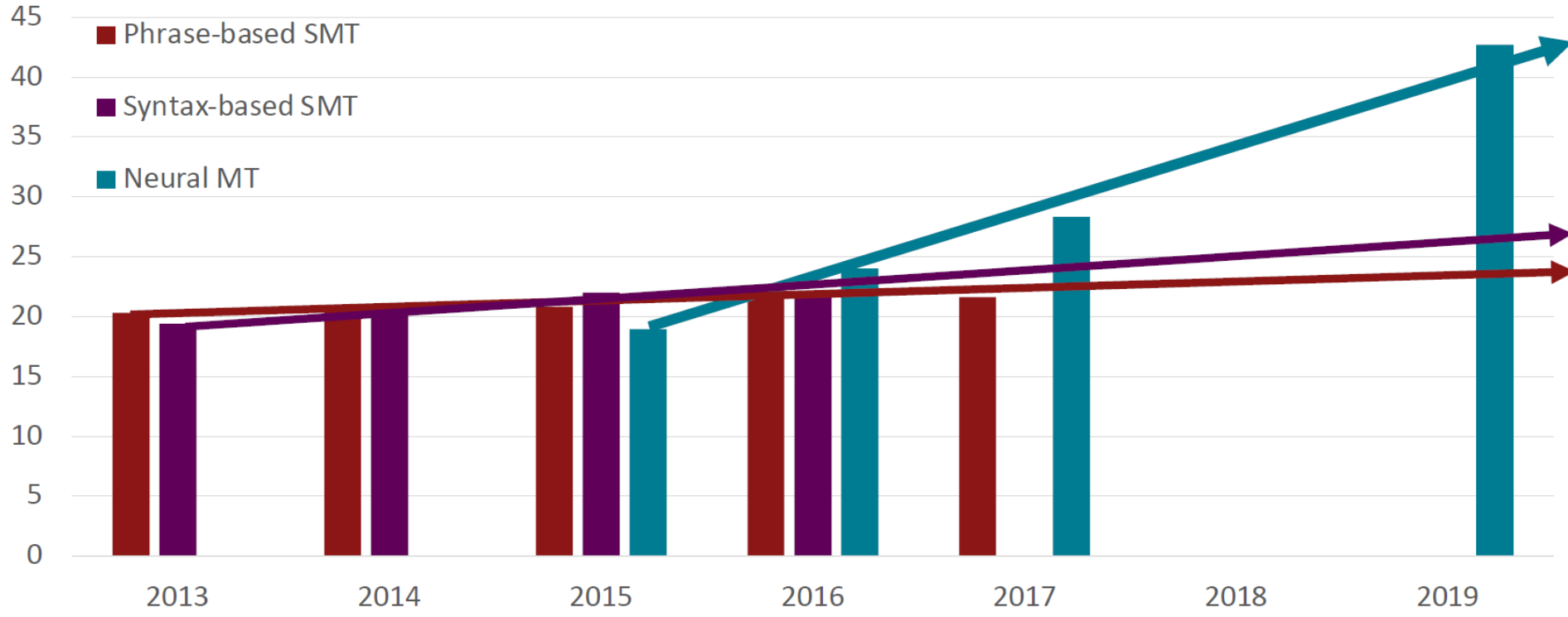
Deep Sequence to Sequence Model

- Stacked seq2seq model



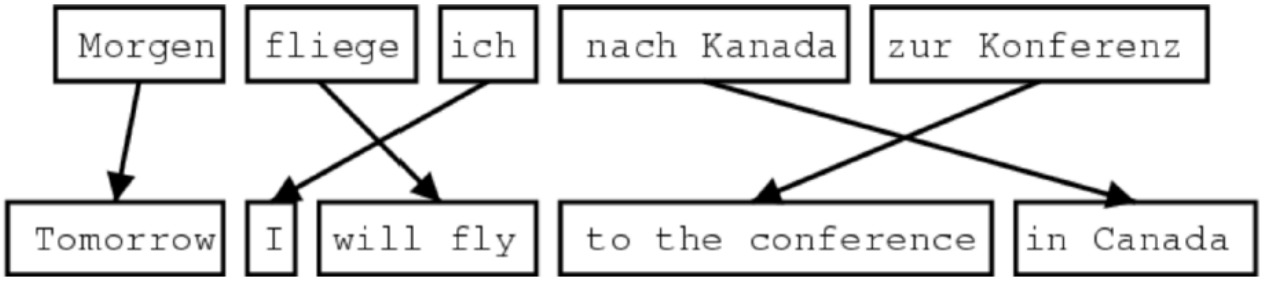
Machine Translation

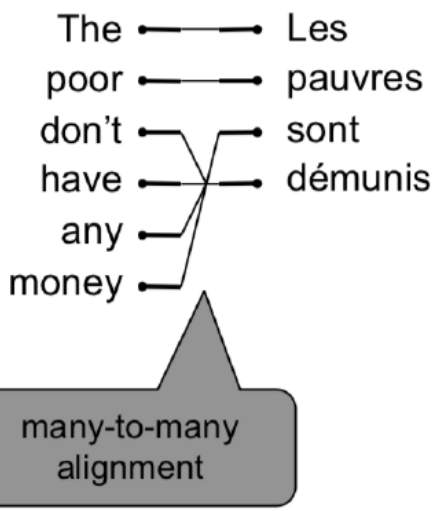
- 2016: Google switched Google Translate from SMT to NMT

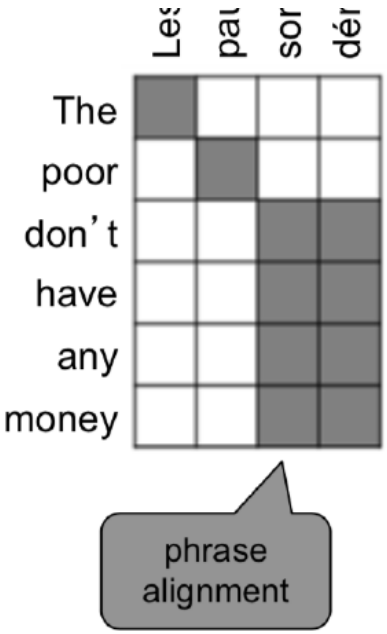


Alignment

- Alignment: the word-level correspondence between X and Y
- Can have complex long-term dependencies

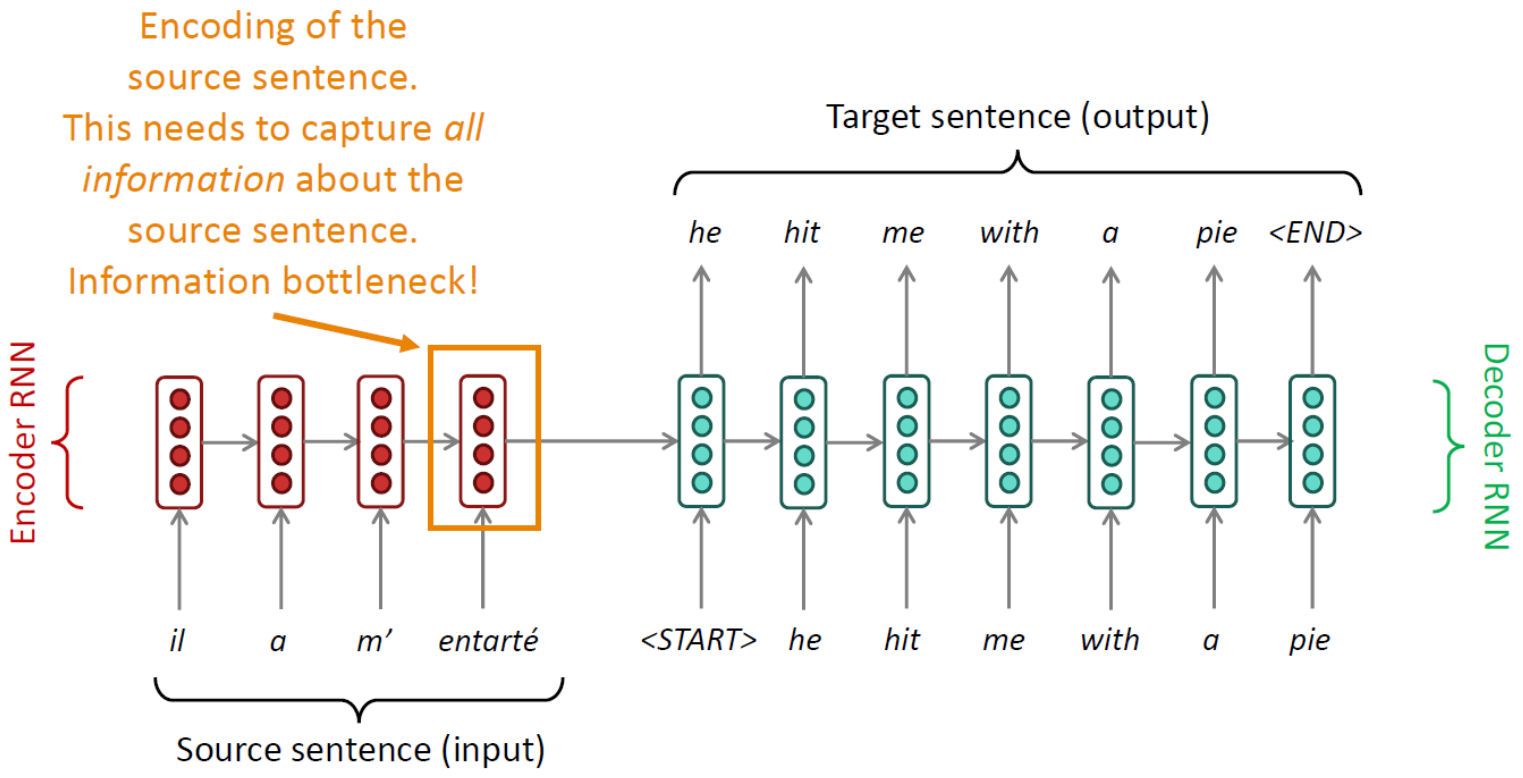






Issue in Seq2Seq

- Alignment: the word-level correspondence between X and Y
  - The information bottleneck due to the hidden state h
  - We want each Y_t to also focus on some X_i that it is aligned with



Attention;

or "Neural Machine Translation by Jointly Learning to Align and Translate", Bahdanau, Cho, Bengio. ICLR '15.

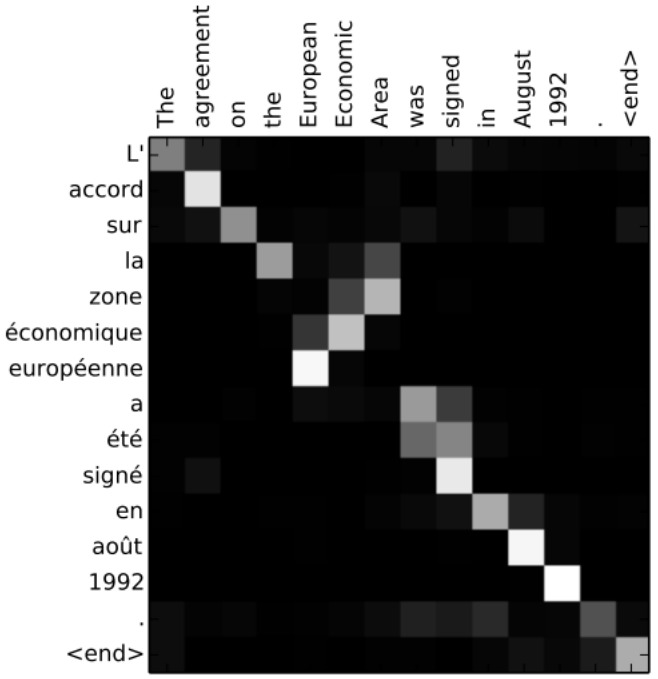
- Context c_i used for predicting the next word is a weighted sum of attention weights α_{ij} * encoded words h_j

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j .$$

- Attention weights are a softmax over learned attention heads

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)},$$

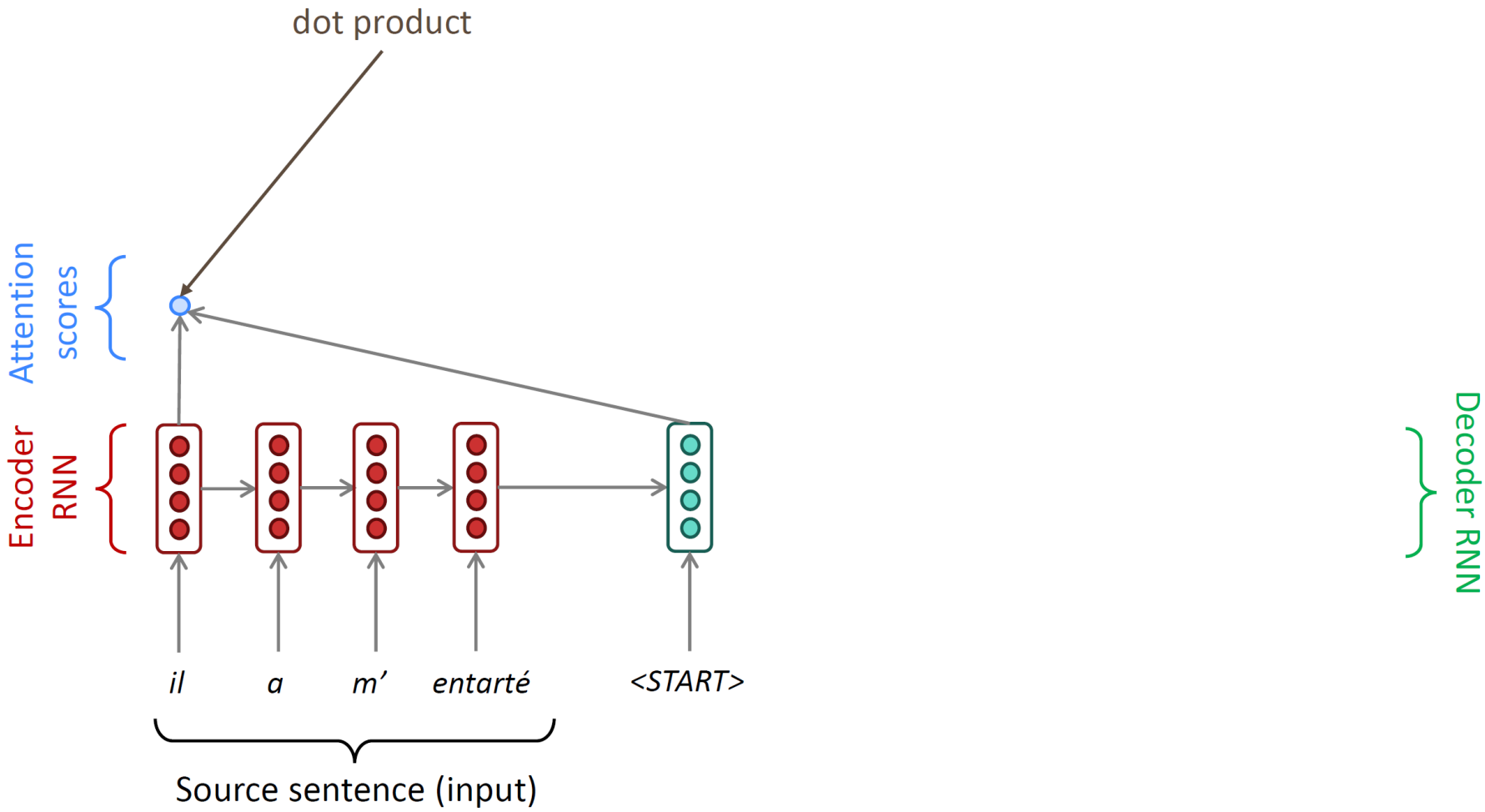
$$e_{ij} = a(s_{i-1}, h_j)$$



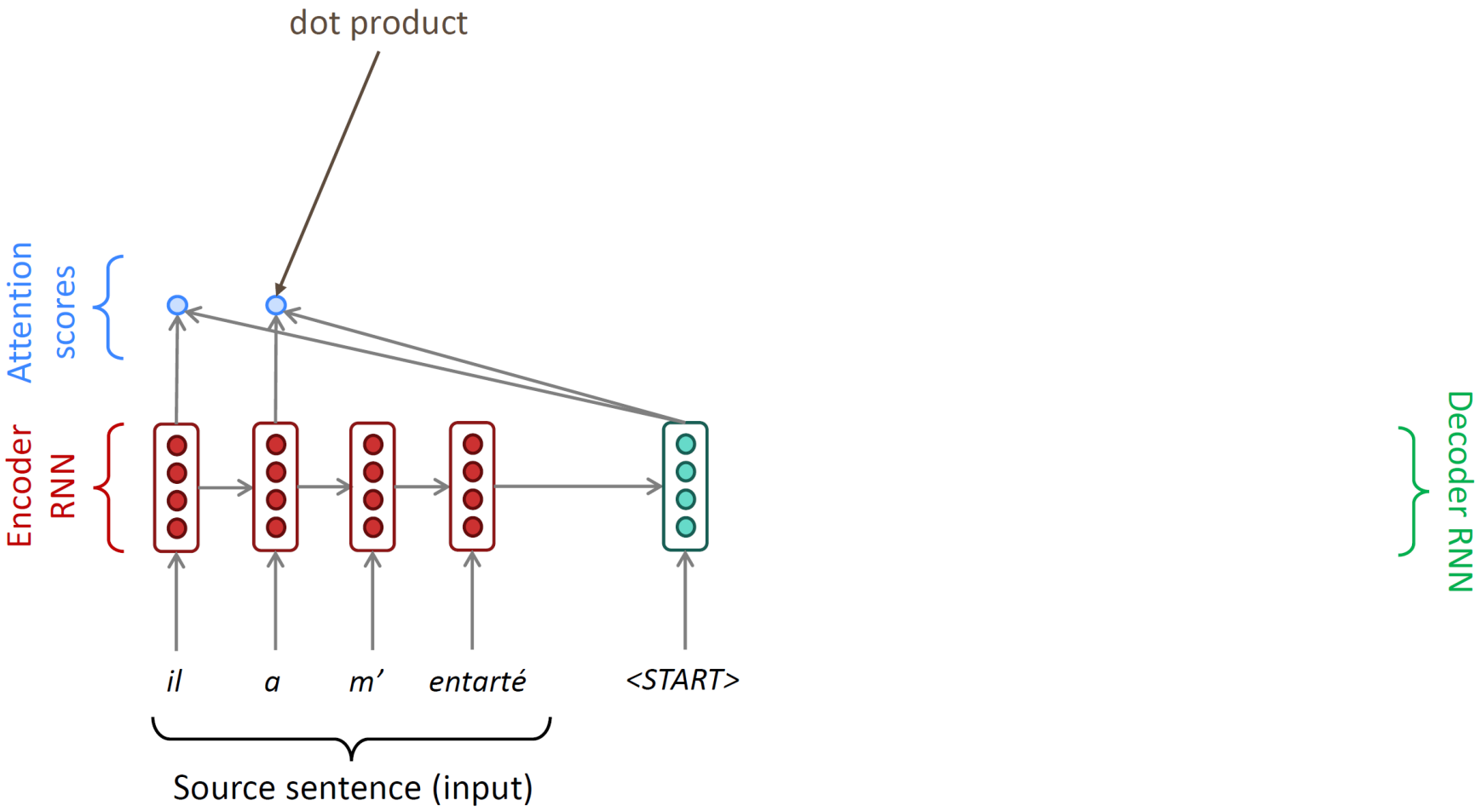
Seq2Seq with Attention

- NMT by jointly learning to align and translate (Bahdanau, Cho, Bengio, '15)
- Core idea:
  - When decoding Y_t, consider both hidden states and alignment:
    - Hidden state: $h_t = f_{dec}(Y_{i<t})$
    - Alignment: connect to a portion of X
  - When portion of X to focus on?
    - Learn a softmax weight over X: attention distribution P_{att}
    - $P_{att}(X_i \mid h_t)$: how much attention to put on word X_i
    - Attention output $h_{att} = \sum_i f_{enc}(X_i \mid X_{j<i}) \cdot P_{att}(X_i \mid h_{t-1})$
    - Use h_{t-1} and h_{att} to compute Y_t

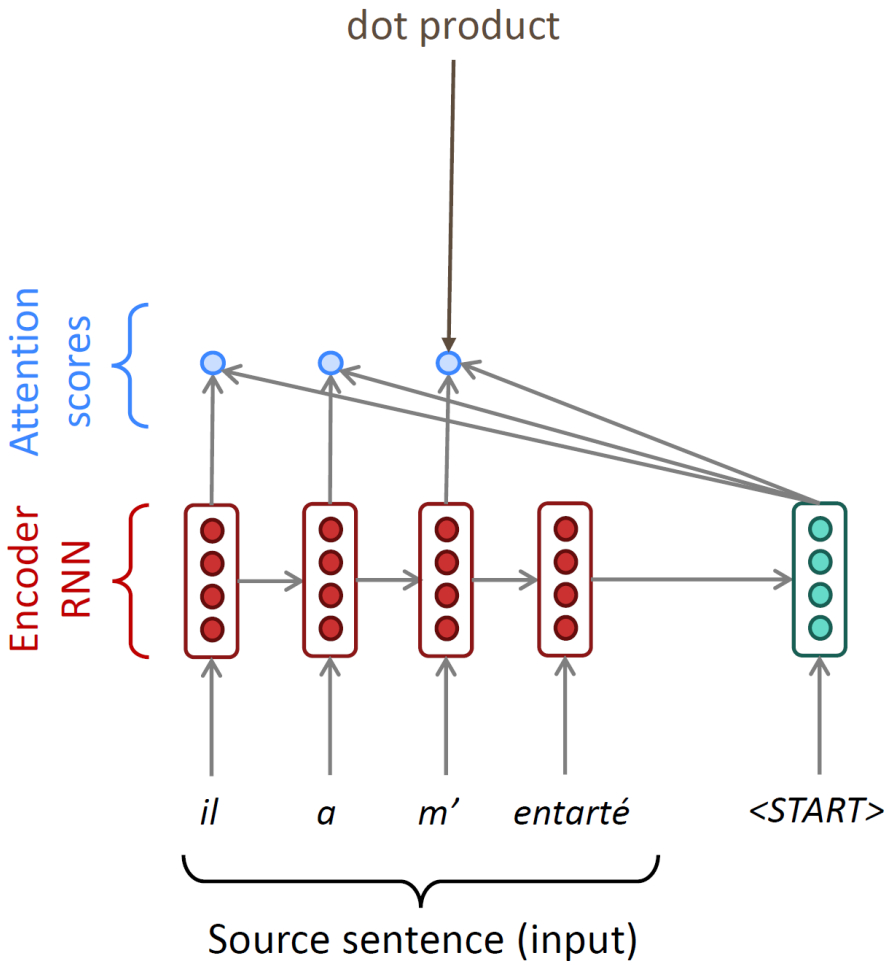
Seq2Seq with Attention



Seq2Seq with Attention

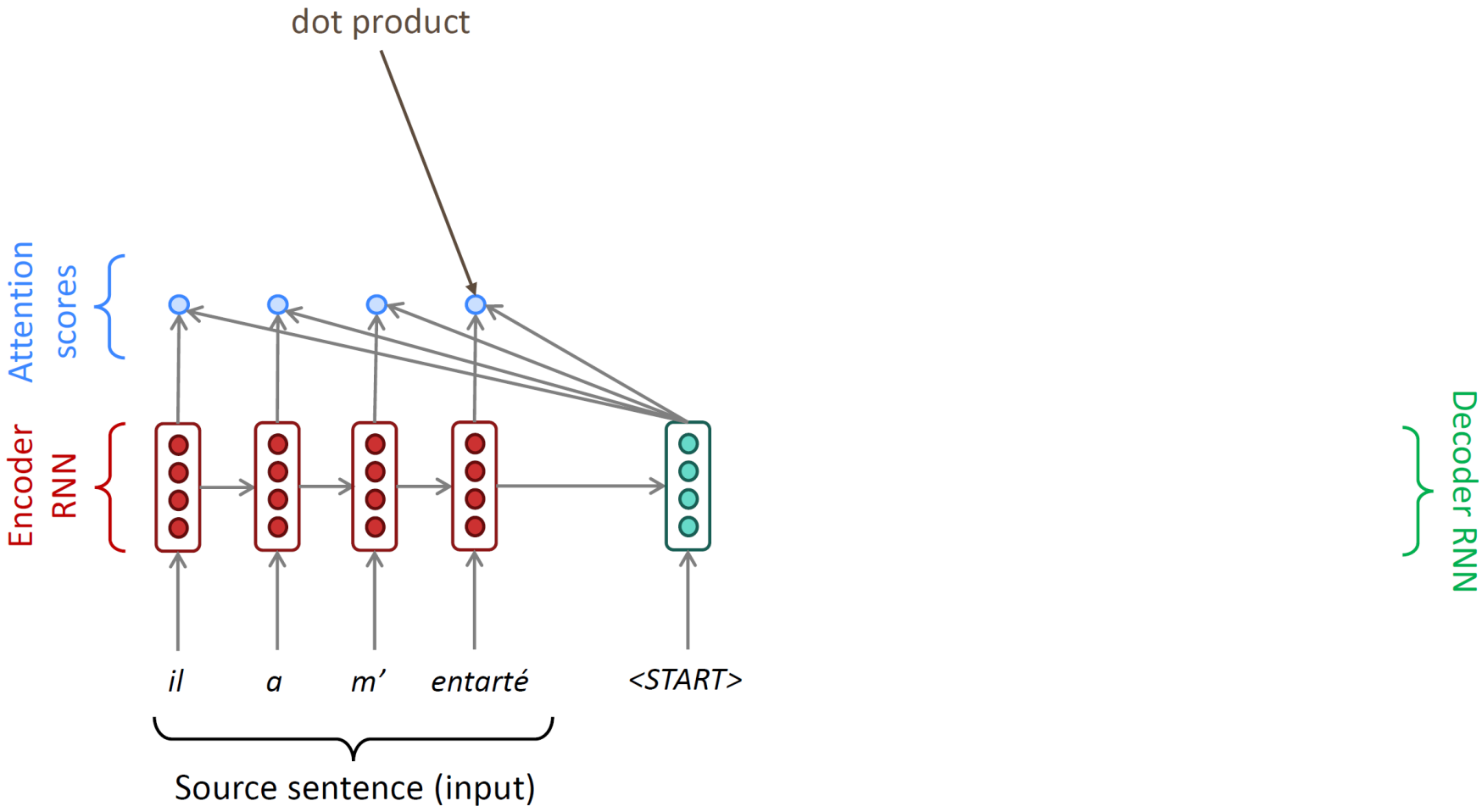


Seq2Seq with Attention

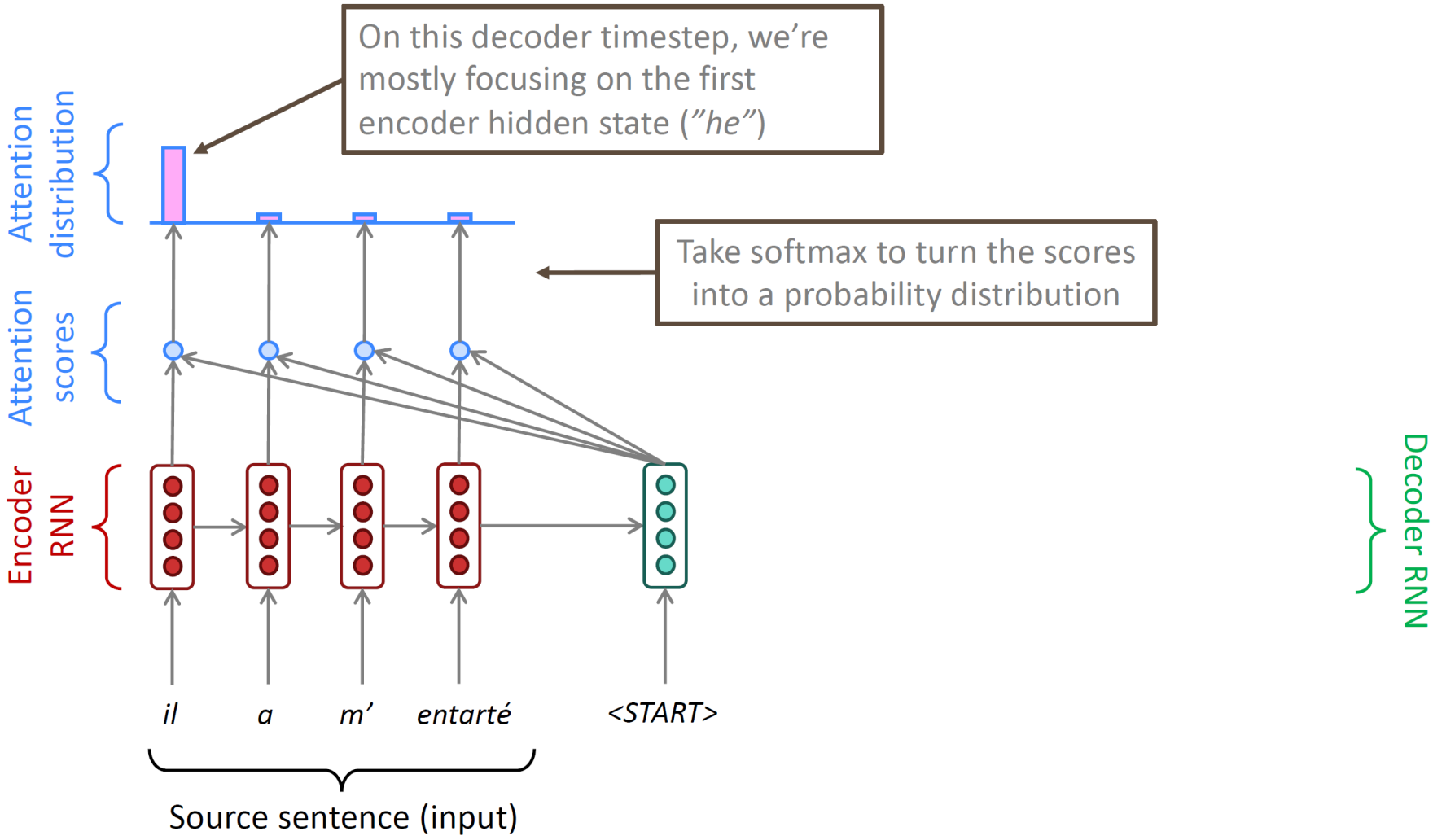


Decoder RNN

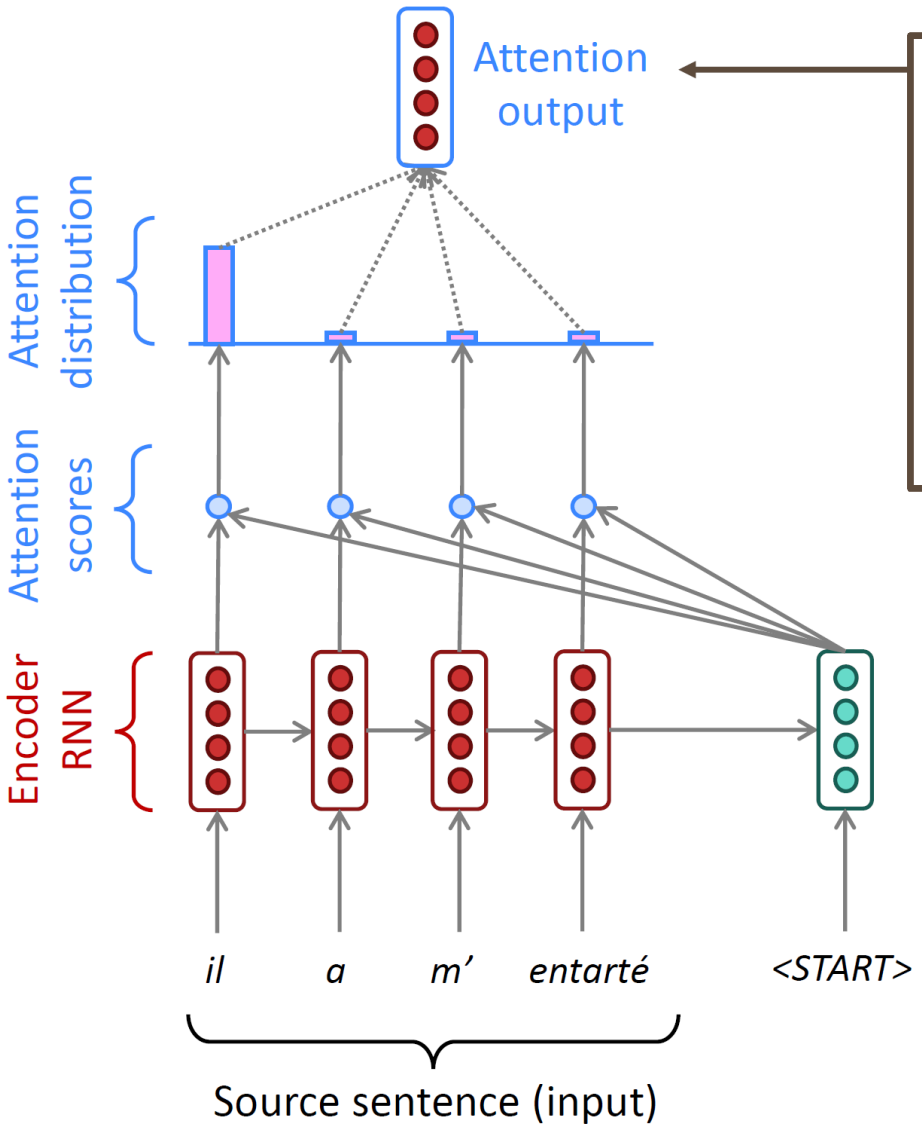
Seq2Seq with Attention



Seq2Seq with Attention



Seq2Seq with Attention

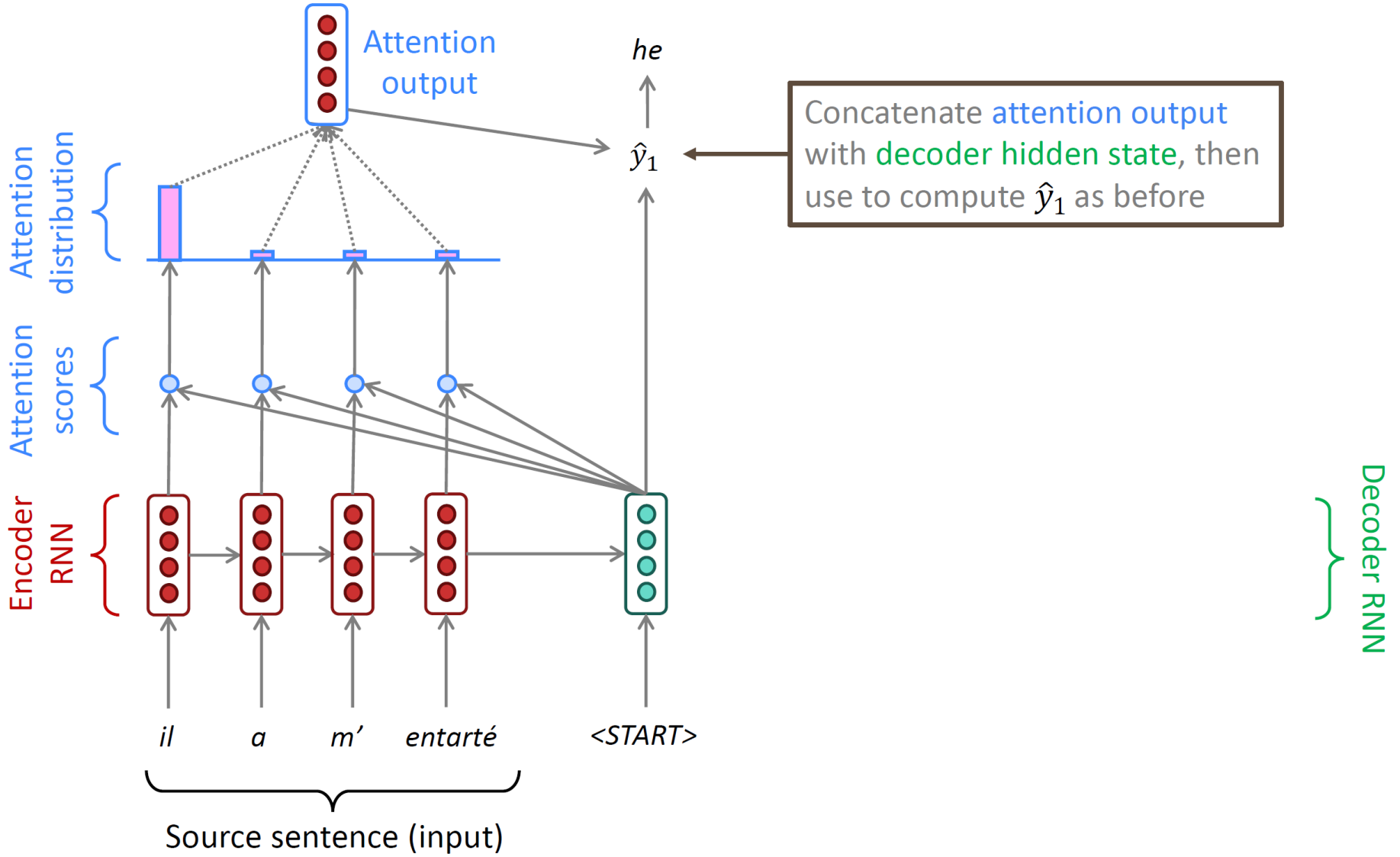


Use the attention distribution to take a **weighted sum** of the encoder hidden states.

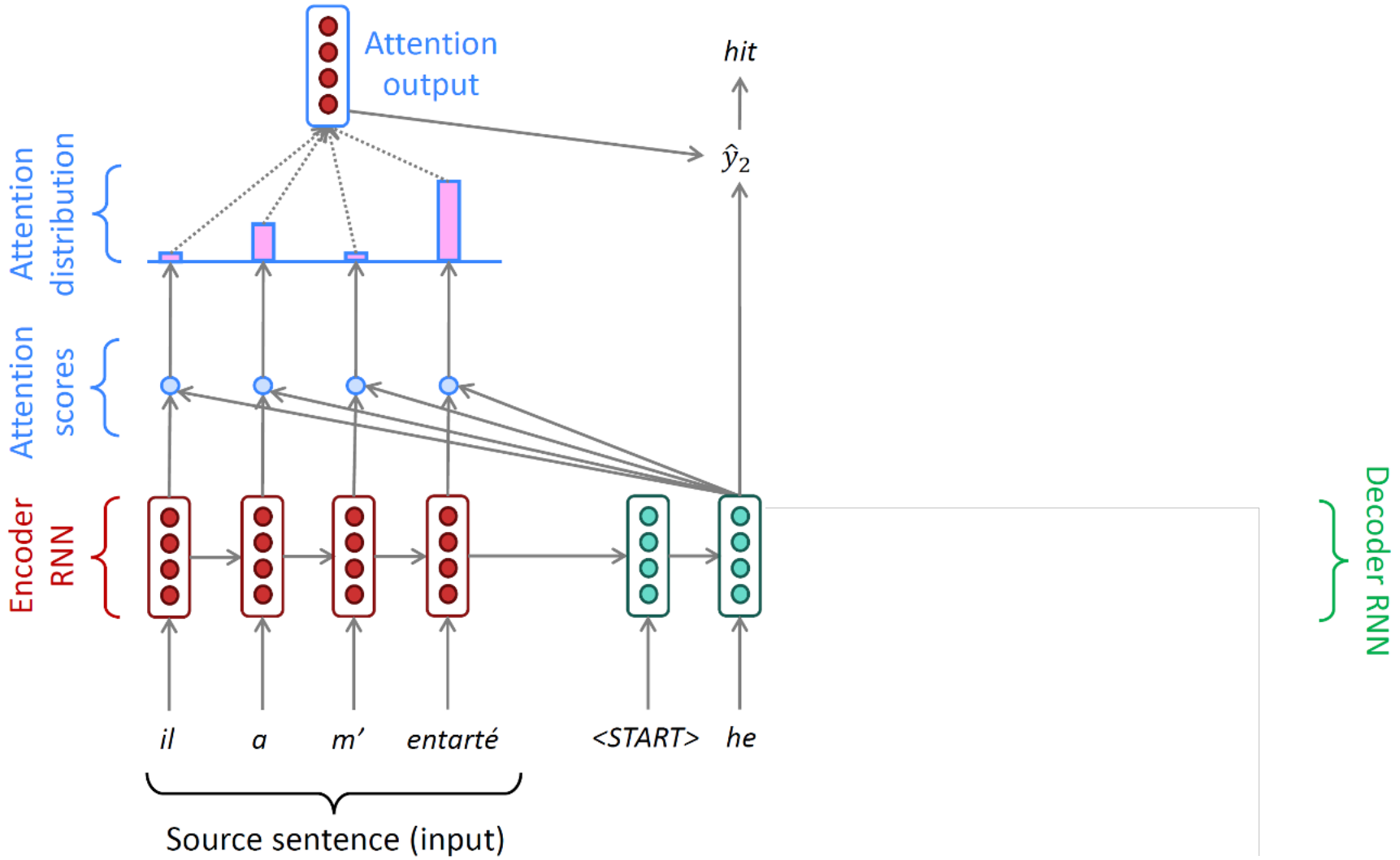
The attention output mostly contains information from the hidden states that received high attention.

Decoder RNN

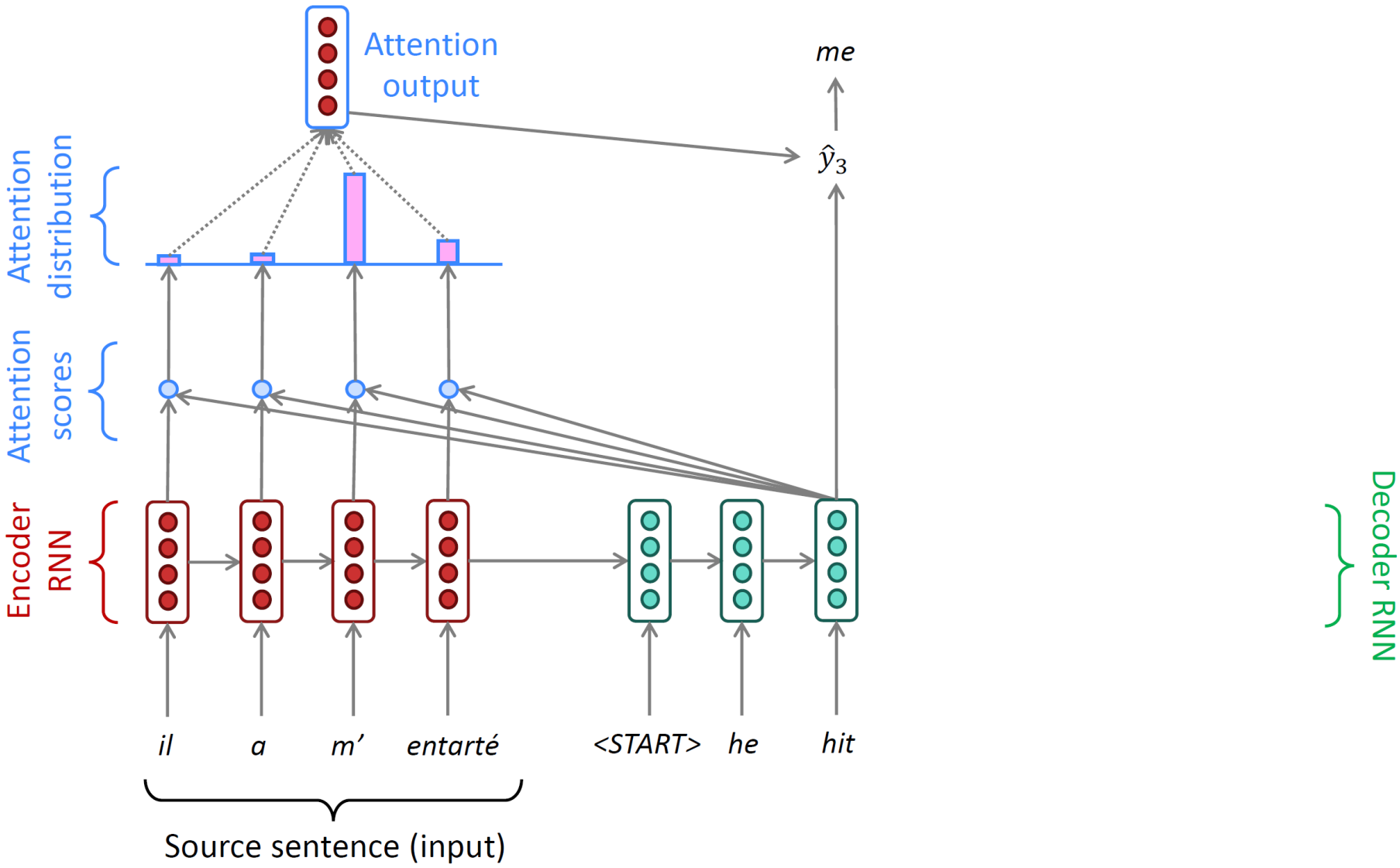
Seq2Seq with Attention



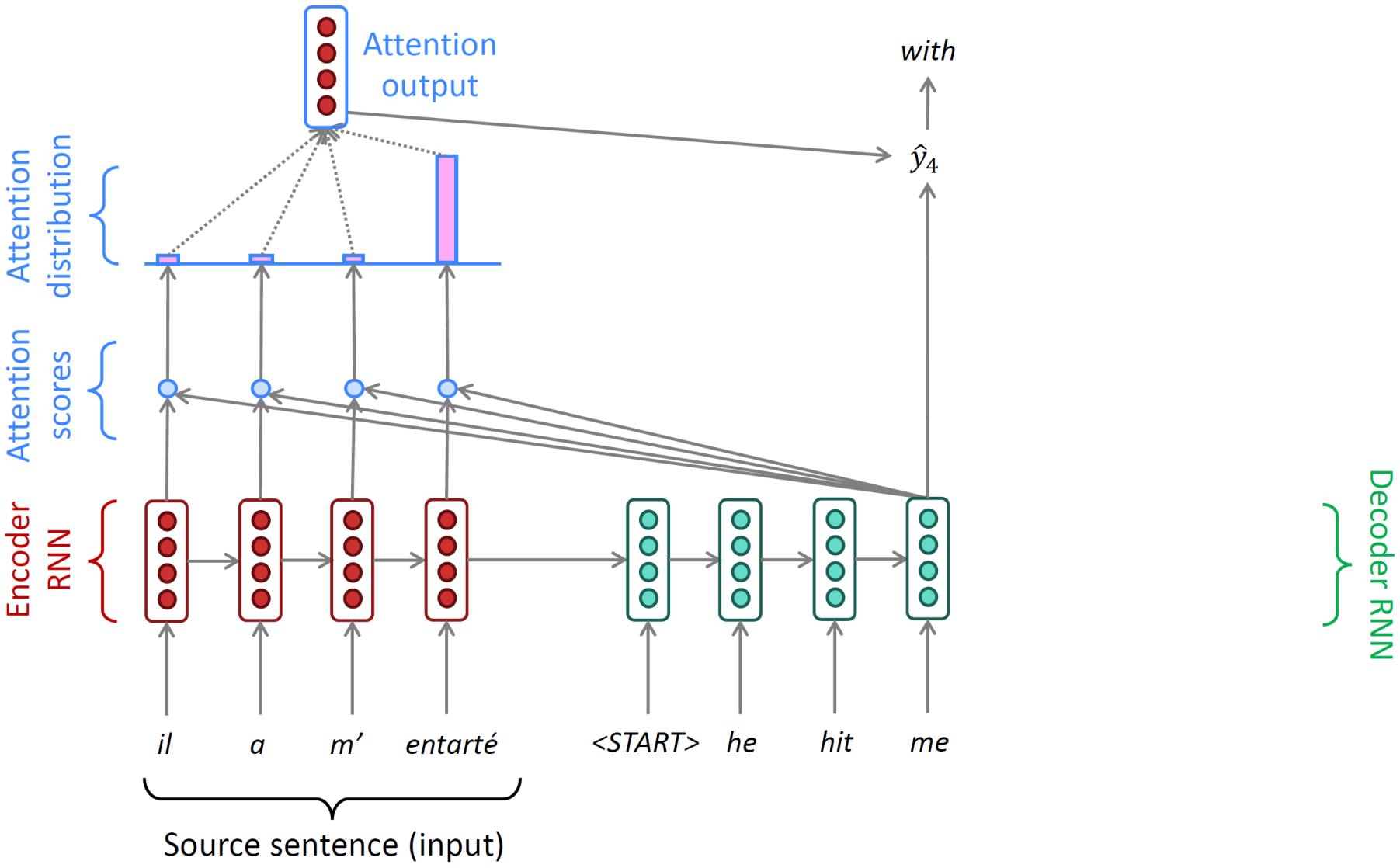
Seq2Seq with Attention



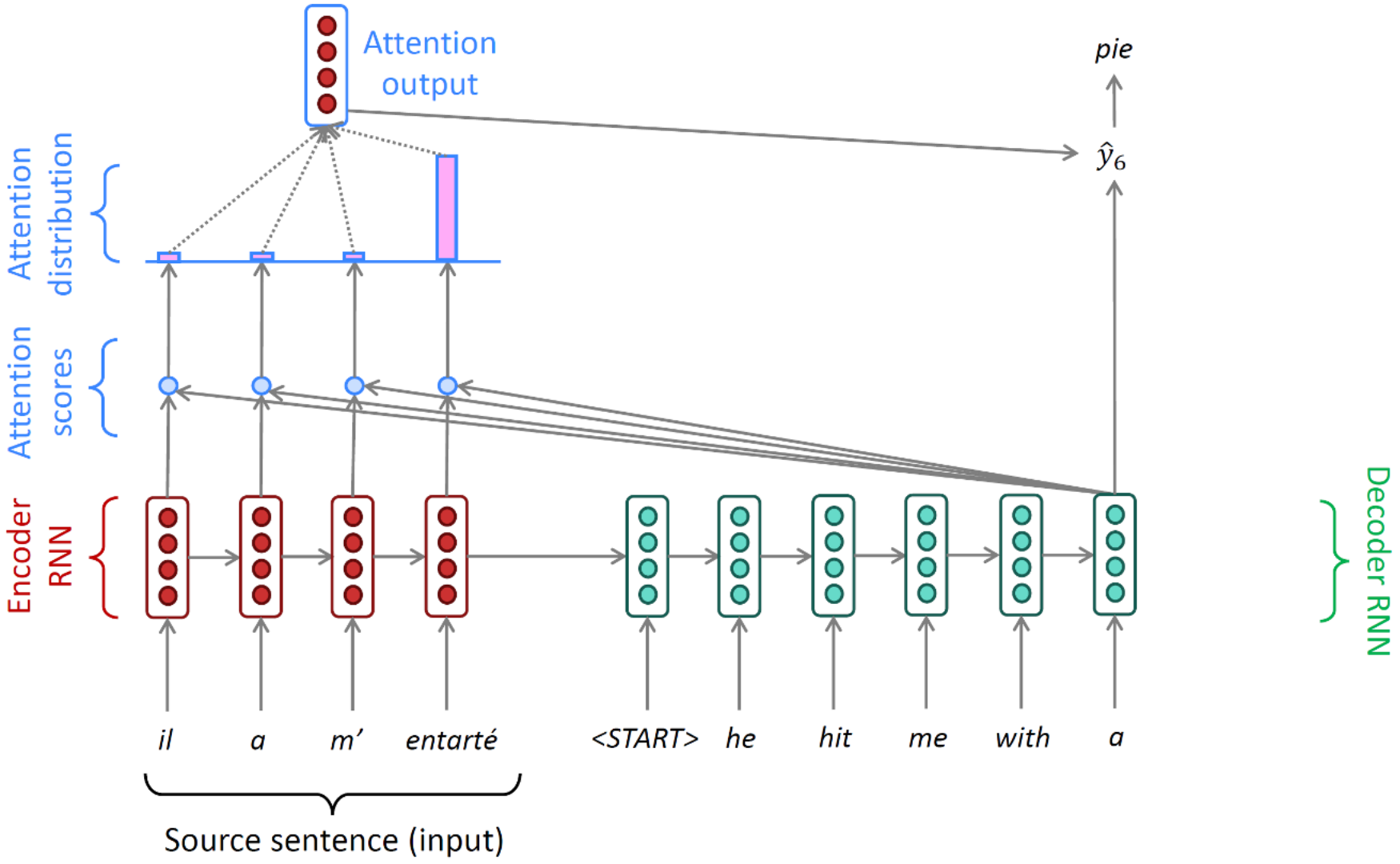
Seq2Seq with Attention



Seq2Seq with Attention



Seq2Seq with Attention



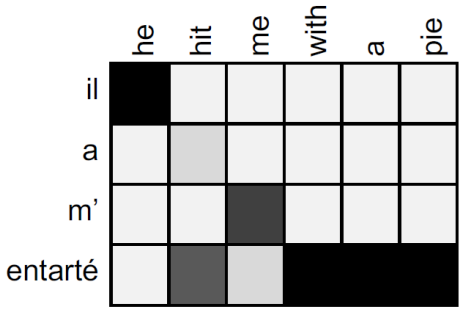
Seq2Seq with Attention

Summary

- Input sequence X, encoder f_{enc}, and decoder f_{dec}
- $f_{enc}(X)$ produces hidden states $h_1^{enc}, h_2^{enc}, \ldots, h_N^{enc}$
- On time step t, we have decoder hidden state h_t
- Compute attention score $e_i = h_t^\top h_i^{enc}$
- Compute attention distribution $\alpha_i = P_{att}(X_i) = \text{softmax}(e_i)$
- Attention output: $h_{att}^{enc} = \sum_i \alpha_i h_i^{enc}$
- $Y_t \sim g(h_t, h_{att}^{enc}; \theta)$
  - Sample an output using both h_t and h_{att}^{enc}

Attention

- It significantly improves NMT.
- It solves the bottleneck problem and the long-term dependency issue.
- Also helps gradient vanishing problem. # Why?
- Provides some interpretability
  - Understanding which word the RNN encoder focuses on



- Attention is a general technique
  - Given a set of vector values V_i and vector query q
  - Attention computes a weighted sum of values depending on q

Other use cases:

- Attention can be viewed as a module.
- In encoder and decoder (more on this later)
- A representation of a set of points
  - Pointer network (Vinyals, Forunato, Jaitly '15)
  - Deep Sets (Zaheer et al., '17)
- Convolutional neural networks
  - To include non-local information in CNN (Non-local network, '18)

Attention

- Representation learning:
  - A method to obtain a fixed representation corresponding to a query q from an arbitrary set of representations $\{V_i\}$
  - Attention distribution: $\alpha_i = \text{softmax}(f(v_i, q))$
  - Attention output: $v_{att} = \sum_i \alpha_i v_i$

- Attent variant: $f(v_i, q)$
  - Multiplicative attention: $f(v_i, q) = q^\top W h_i$, W is a weight matrix
  - Additive attention: $f(v_i, q) = u^\top \tanh(W_1 v_i + W_2 q)$