신입생 Deep Learning 기초 교육

5회: Seq2seq & Seq2seq with attention

Multimodal Language Cognition Lab,
Kyungpook National University

2023.02.10

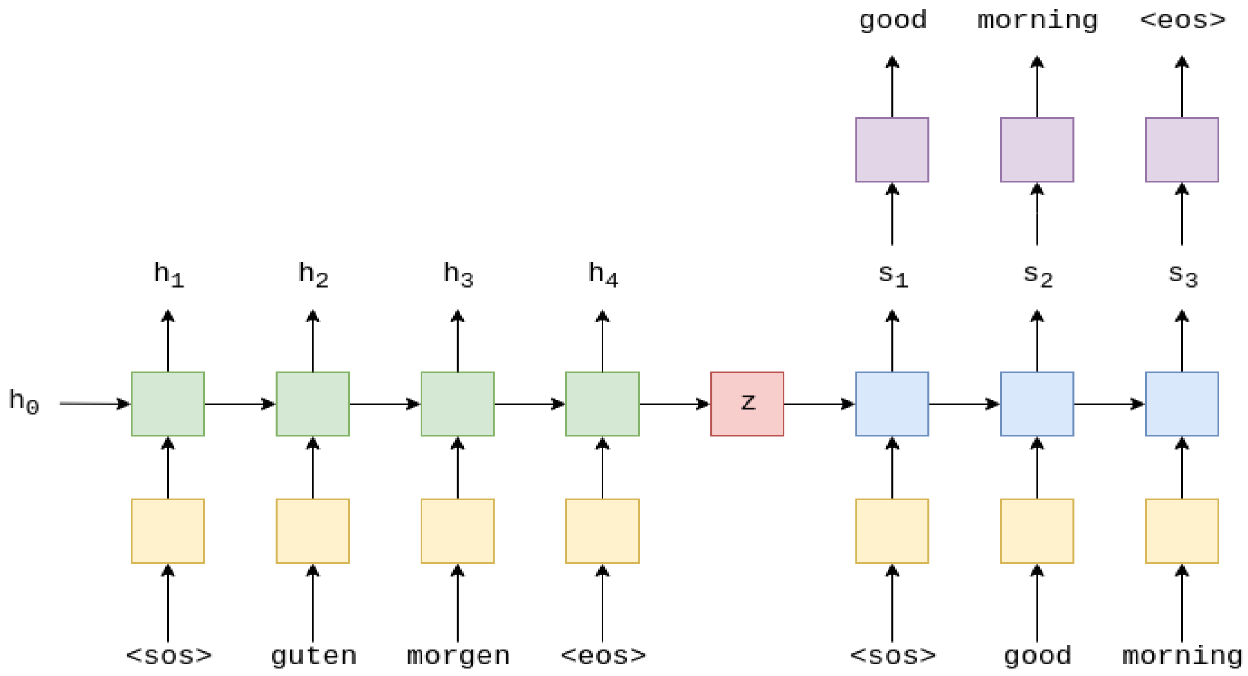
Sequence to Sequence	Sequence to Sequence Learning with Neural Networks (NeurIPS 2014)
Bahdanau Attention	Neural machine translation by jointly learning to align and translate (ICLR 2015)
Luong Attention	Effective Approaches to Attention-based Neural M achine Translation (EMNLP 2015)
Transformer	Attention is all you need (NeurIPS 2017)
GPT-1	Improving Language Understanding by Generative Pre-Training (2018)
BERT	Bert: Pre-training of deep bidirectional transformers for language understanding (2018)

Seq2Seq & Seq2Seq with attention

Machine translation is a major use-case of sequence-to-sequence, is improved by attention

- New Task: machine translation
- New architecture: sequence-to-sequence
- New technique: attention

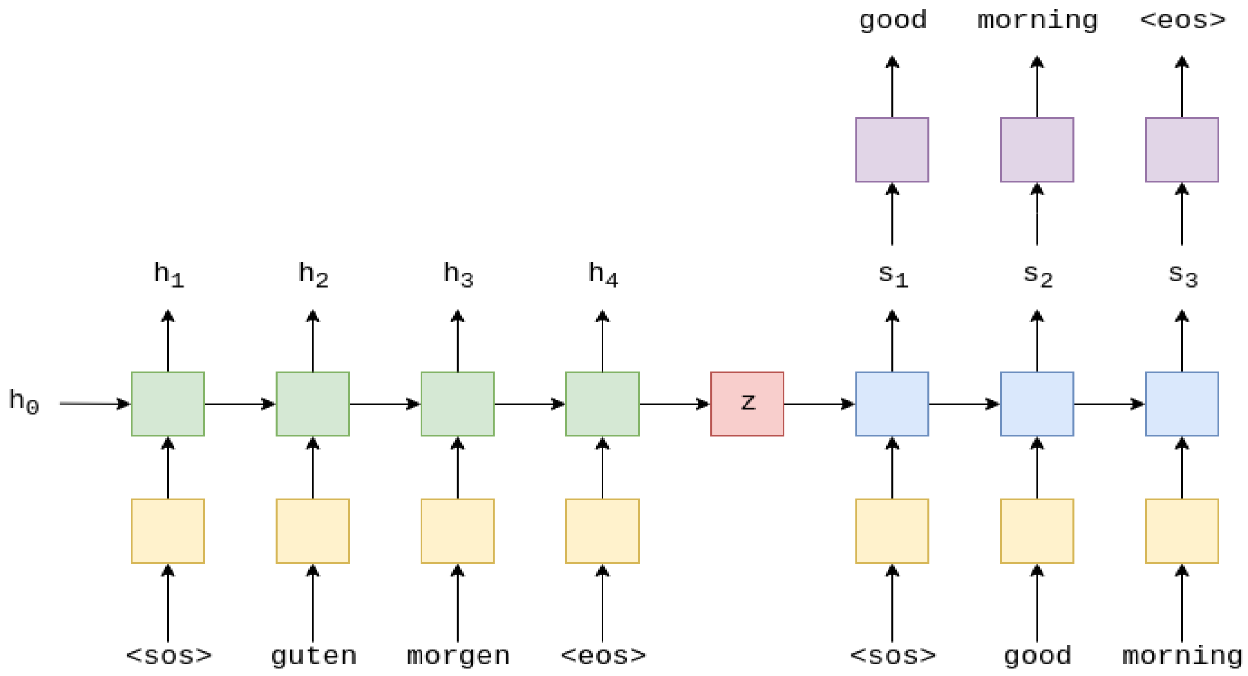
Sequence-to-sequence (Seq2Seq)



- Sequence-to-sequence learning (Seq2Seq) is about training models to convert sequences from one domain (e.g., sentences in English) to sequences in another domain (e.g., the same sentences translated to French)

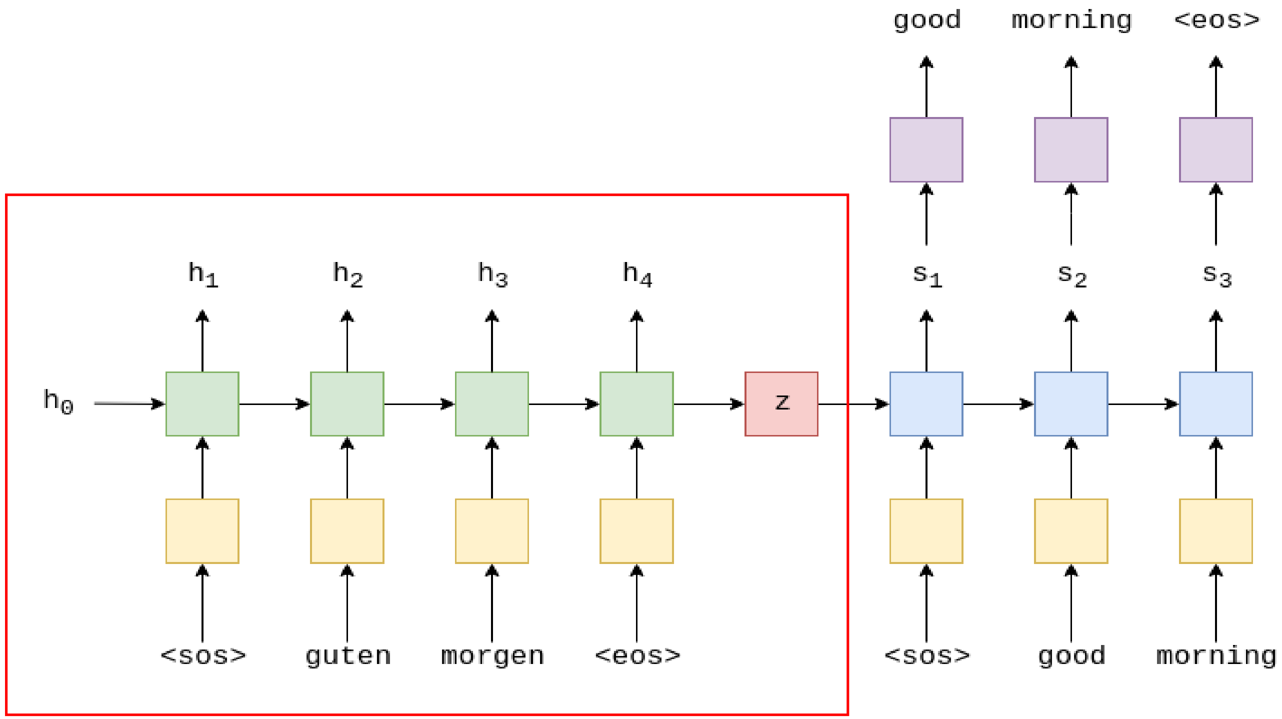
- Applications
  - Machine translation
  - Question answering
  - Speech recognition

Sequence-to-sequence (Seq2Seq)



- input sequences and output sequences have **different lengths** (e.g., machine translation)
- The idea is to use one LSTM to read the input sequence, one timestep at a time, to obtain large fixed dimensional **vector representation**, and then to use another LSTM to extract **the output sequence from that vector**

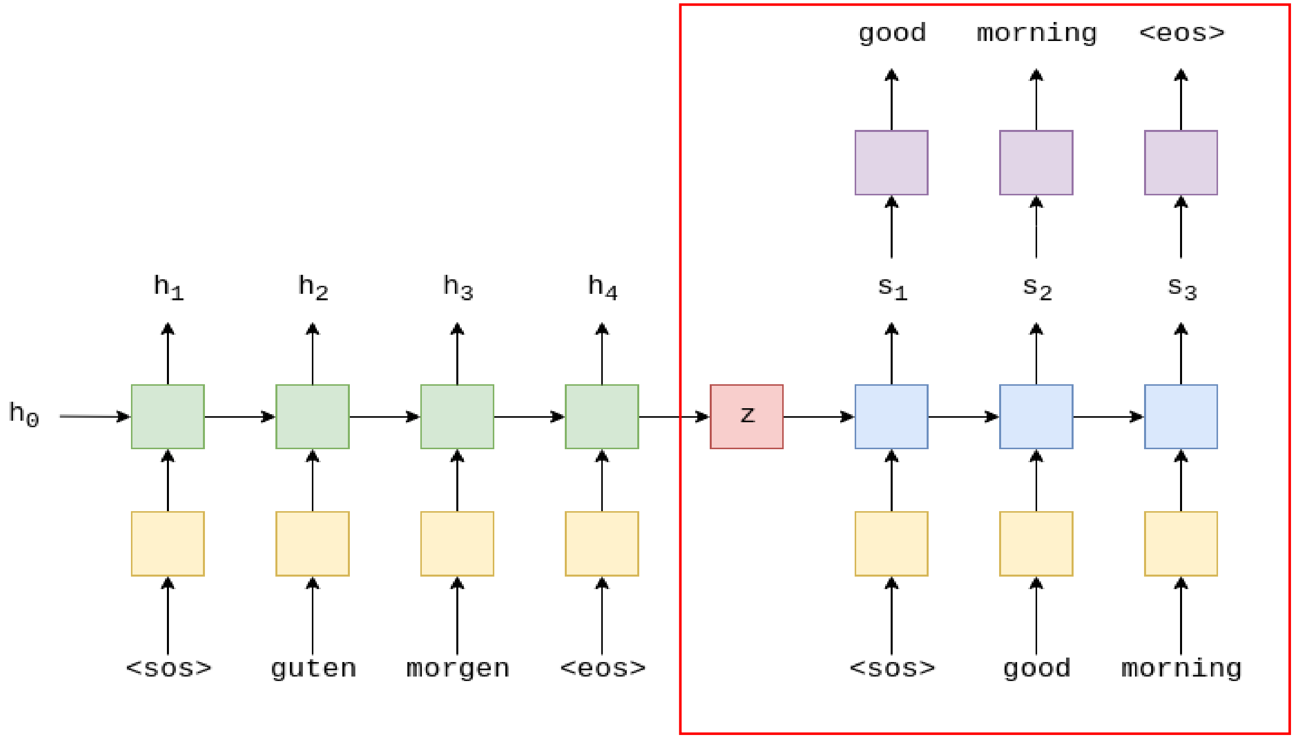
Sequence-to-sequence (Seq2Seq)



- A LSTM layer (or stack thereof) acts as **"encoder"**
  - It processes the **input sequence** and returns its **last hidden state(Z)**
  - Z is the fixed dimensional representation of input sequence
  - Z will serve as the "context", or "conditioning", of the decoder in the next step

Kyungpook National University

Sequence-to-sequence (Seq2Seq)

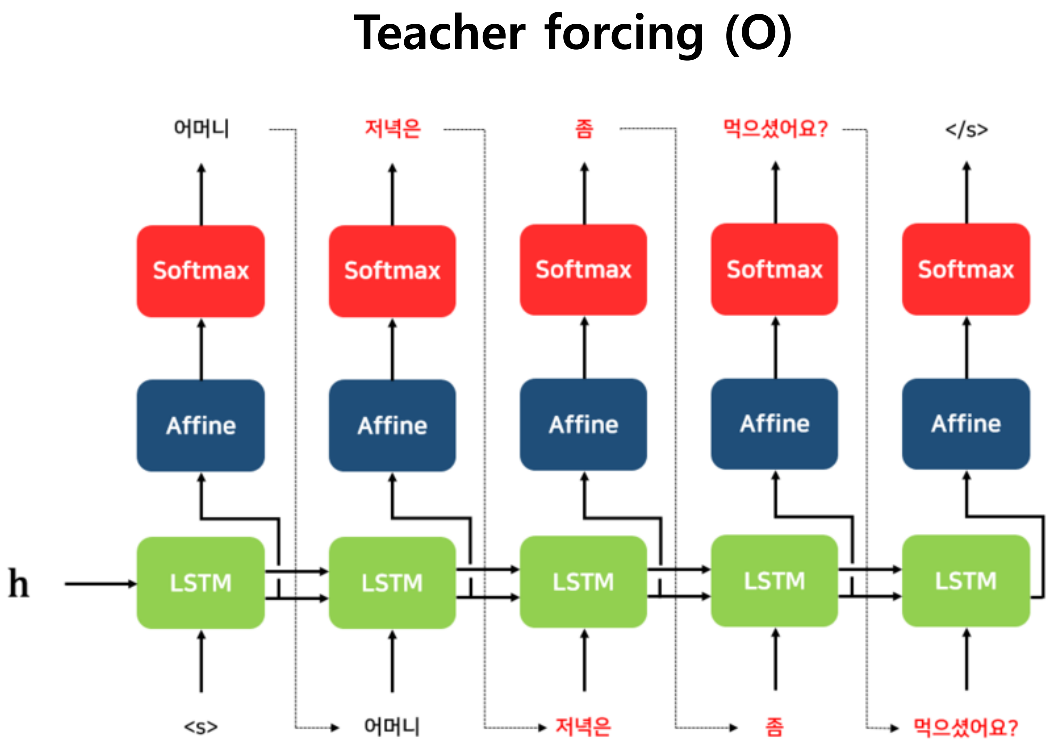


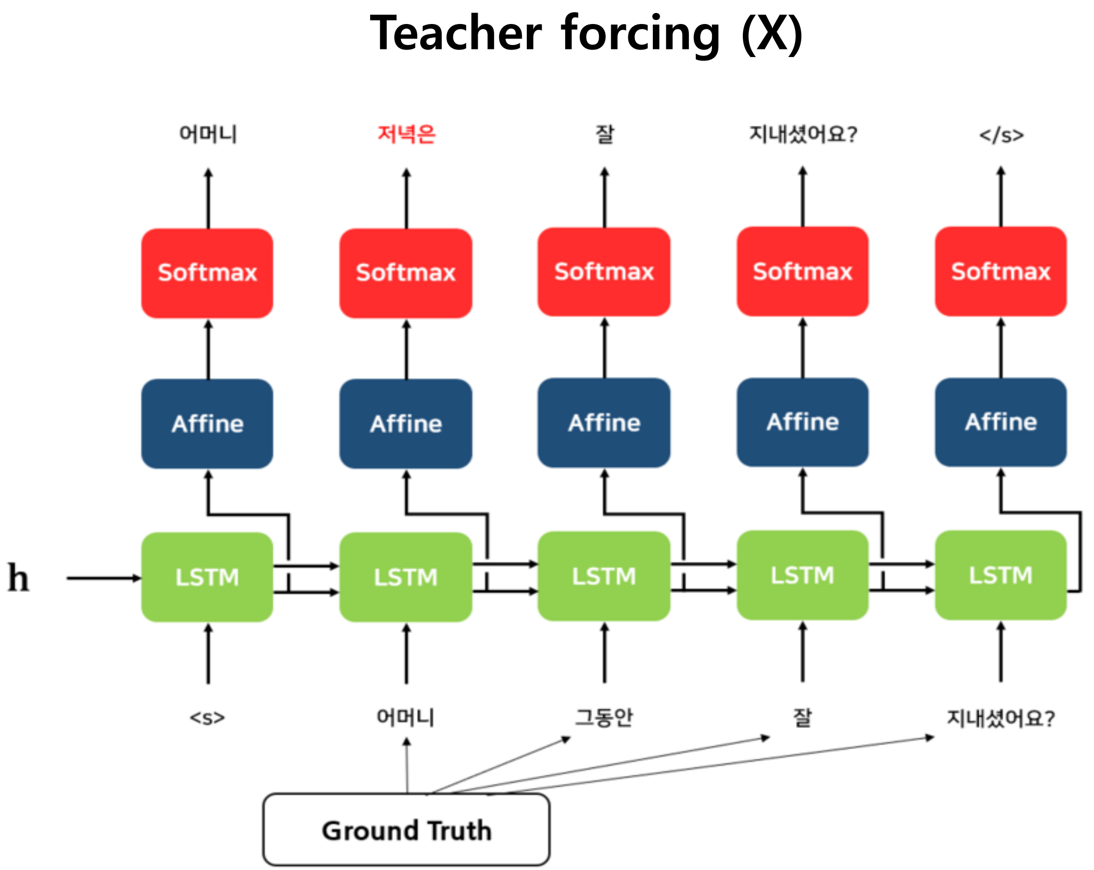
- A LSTM layer (or stack thereof) acts as **"decoder"**
  - **Language Model** because the decoder is predicting the next word of the target sentence y
  - **Conditional** because its predictions are also conditioned on the source sentence x
  - $p(y_1, \cdots, \ y_{T'} | x_1, \cdots, x_T) = \prod_{t=1}^{T} p(y_t \mid Z, y_1, \cdots, y_{t-1}),$

    $\{x_1, \ldots, x_T\}$ *is source sequence*, $\{y_1, \ldots, y_{T'}\}$ *is target sequence*

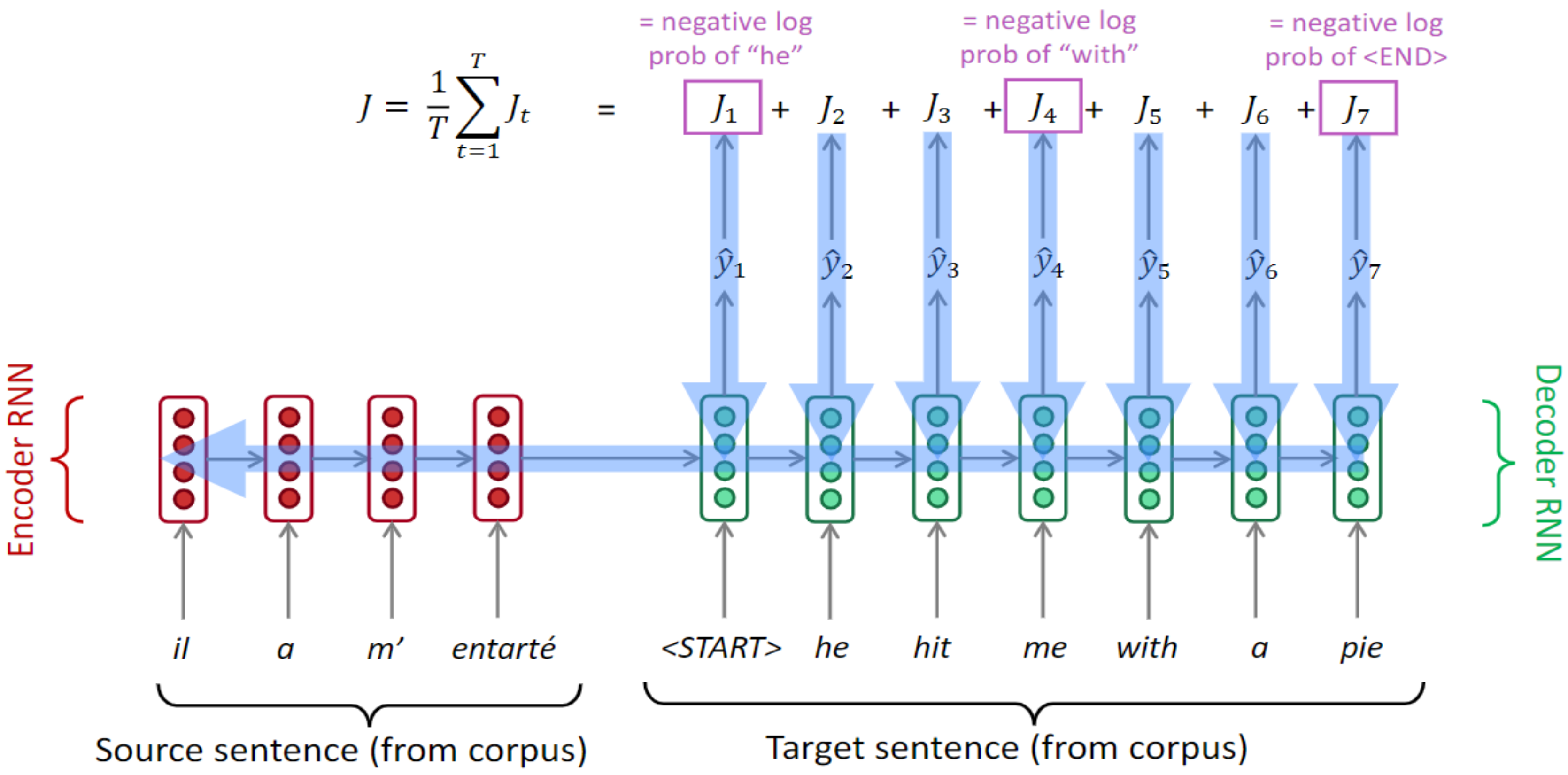
MLCL Kyungpook National University

Sequence-to-sequence (Seq2Seq)



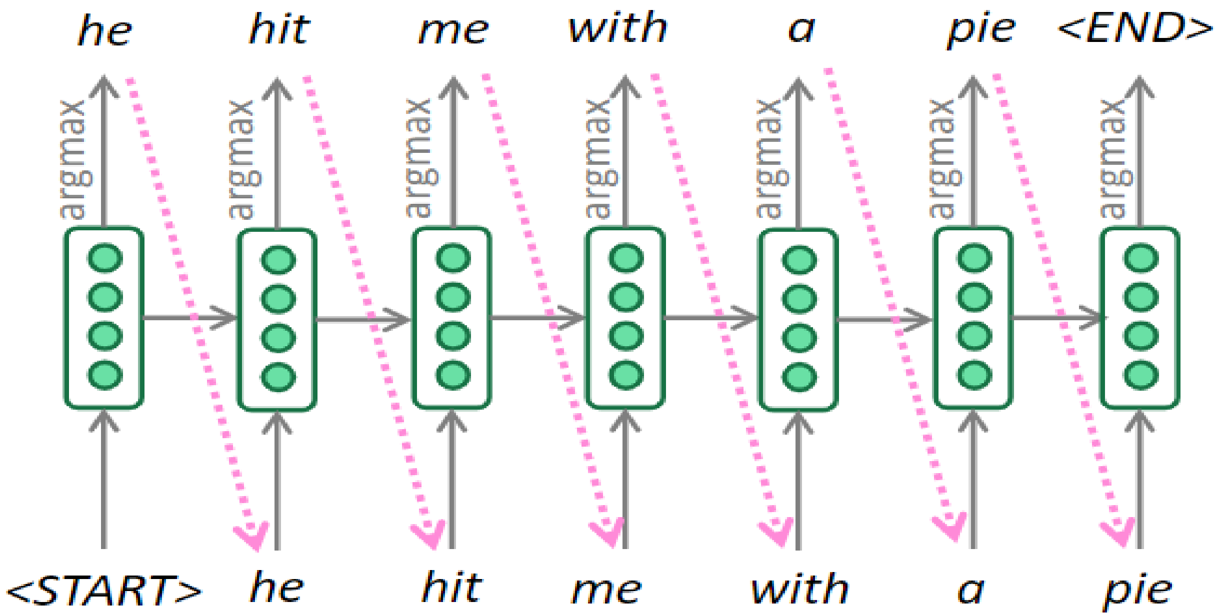


Sequence-to-sequence (Seq2Seq)



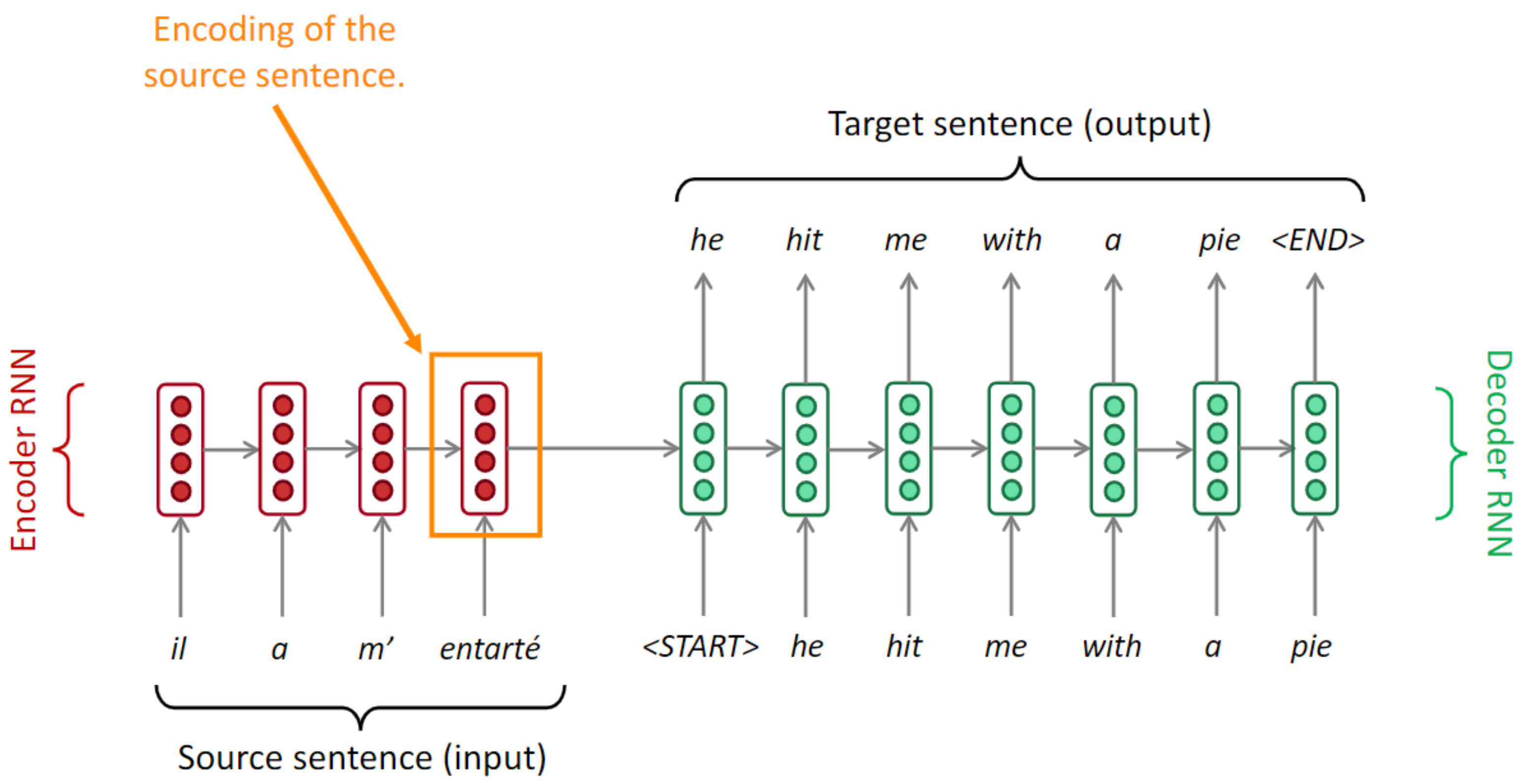
- Seq2seq is optimized as a single system
- Backpropagation operates "end-to-end"

Sequence-to-sequence (Seq2Seq)



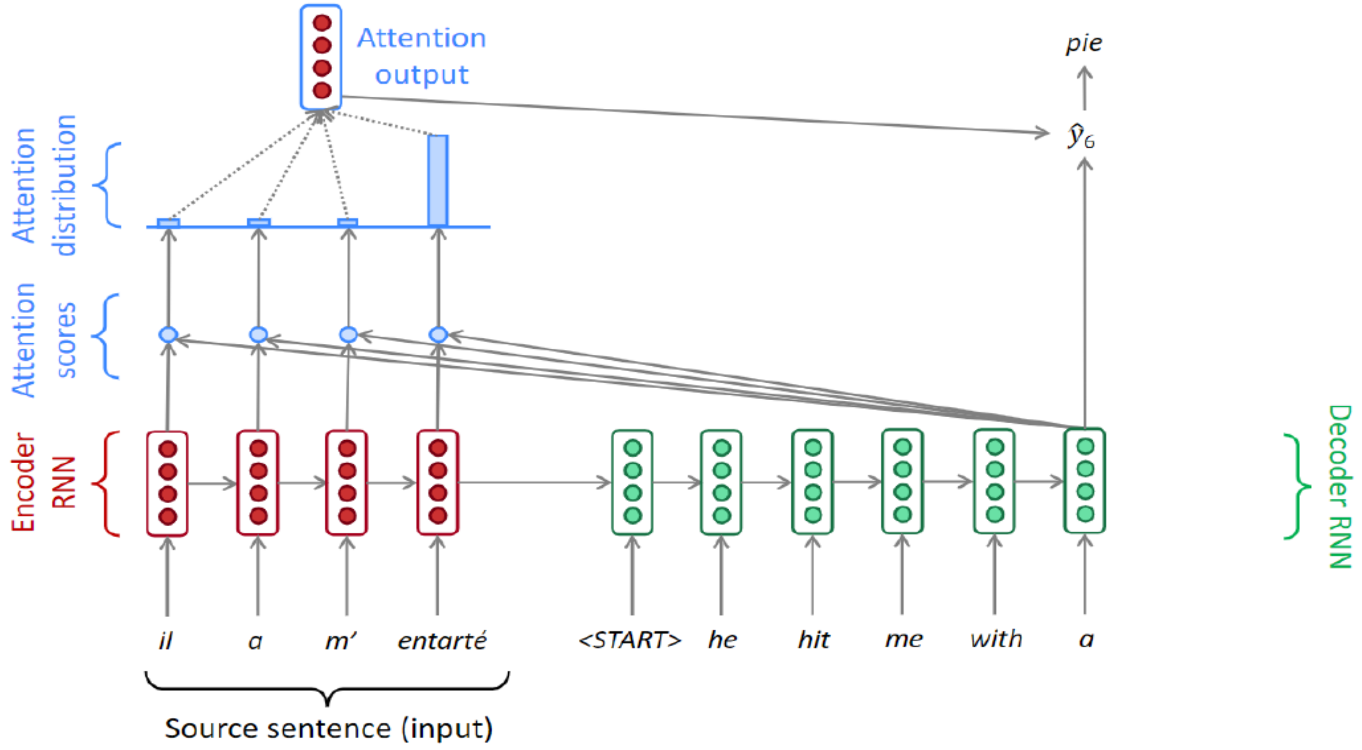
- **Greedy Decoding**
  - Let's see how to generate the target sentence by taking **argmax** on each step of the decoder
  - It takes most probable word on each step → Greedy decoding

Sequence-to-sequence with attention



Problems with Seq2Seq ?

Sequence-to-sequence with attention

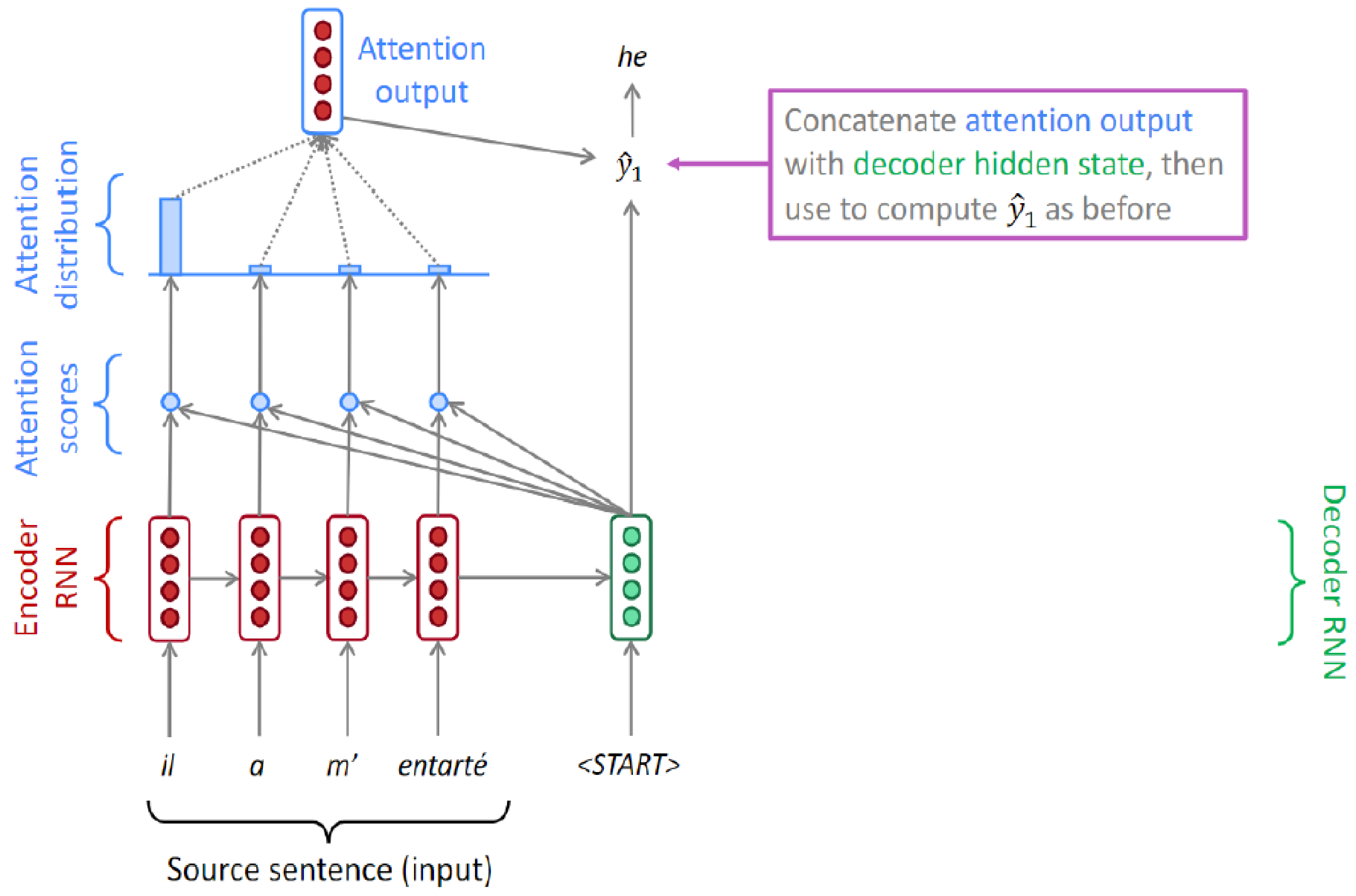


- **Attention** provides a solution to the bottleneck problem
- Core idea

  On each step of the decoder, use direct connection to the encoder to focus on a particular part of the source sequence

MLCL Kyungpook National University

Sequence-to-sequence with attention



- $encoder\ hidden\ state = \{h_1, h_2, h_3, h_4\}$
- $decoder\ hidden\ state = \{s_1, s_2, s_3, s_4, s_5\}$
- $decoder\ time\ step\ 1,$

  1) $Attention$ scores(e_1)

  $$score(s_1, h_t) = s_1^T h_t$$

  $$e_1 = [s_1^T h_1, s_1^T h_2, s_1^T h_3, s_1^T h_4]$$

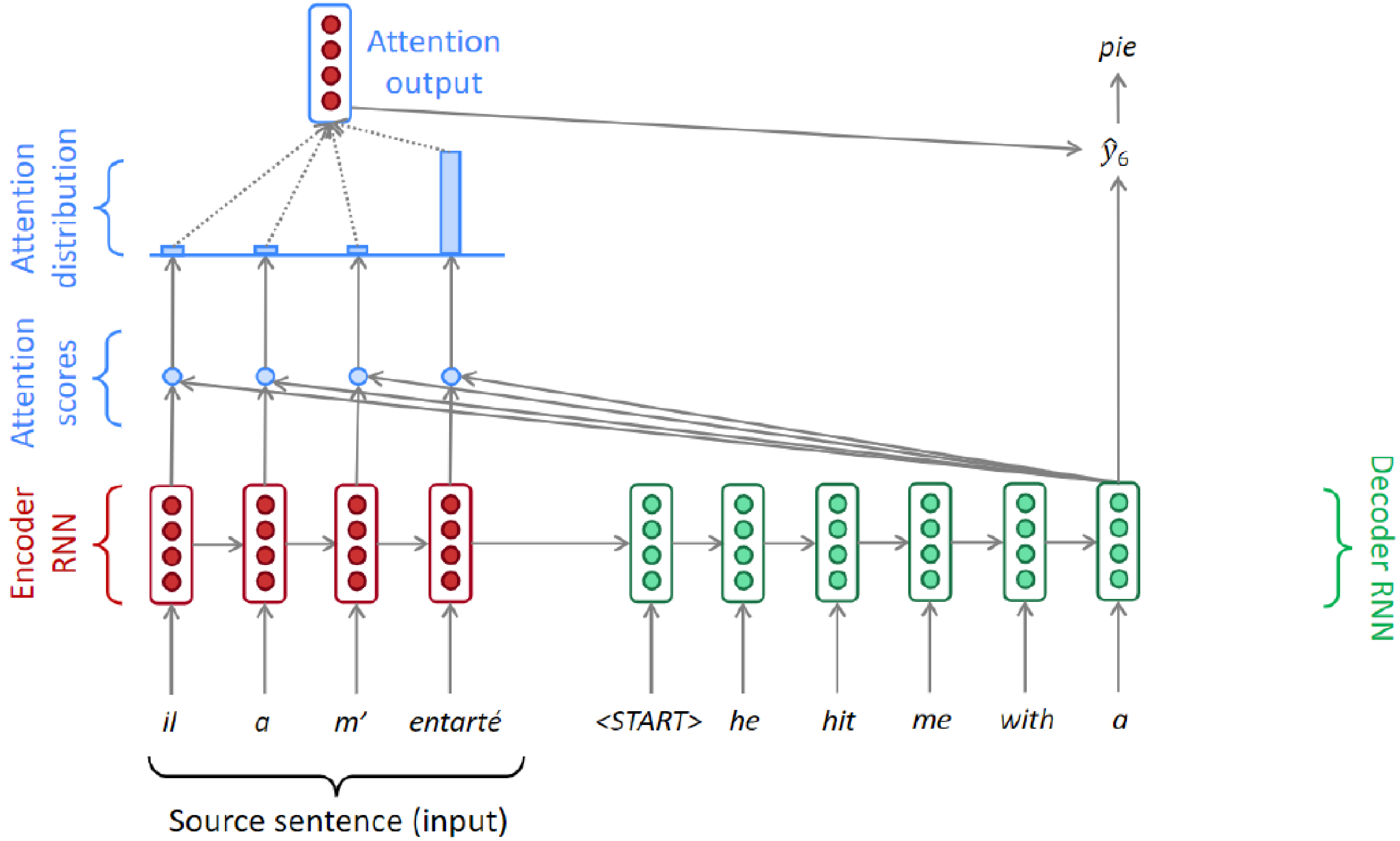
  2) $Attention\ distribution(e_1)$

  $$\alpha_1 = softmax(e_1)$$

  3) $Attention\ outputs\ (a_1)$

$$a_1 = \sum_{i=1}^{4} \alpha_1^i h_i$$

Sequence-to-sequence with attention



Sequence-to-sequence with attention

Name	Attention score	Defined by
Dot-product	$s_t^T h_i$	Luong et al. (2015)
Scaled dot	$\frac{s_t^T h_i}{\sqrt{n}}$	Vaswani et al. (2017)
Additive attention	$W_a^T \tanh(W_b s_t + W_c h_i)$	Bahdanau et al. (2015)