RNN Overview - 2

Encoder-Decoder and Attention

2021. 4. 21

Dajin Han

Index

Encoder and Decoder

Sequence to sequence

SMT

- Statistical Machine Translation
- **Translator**
- For each phrase, the probability of corresponding sentences is calculated
- Mathematical method

Encoder and Decoder

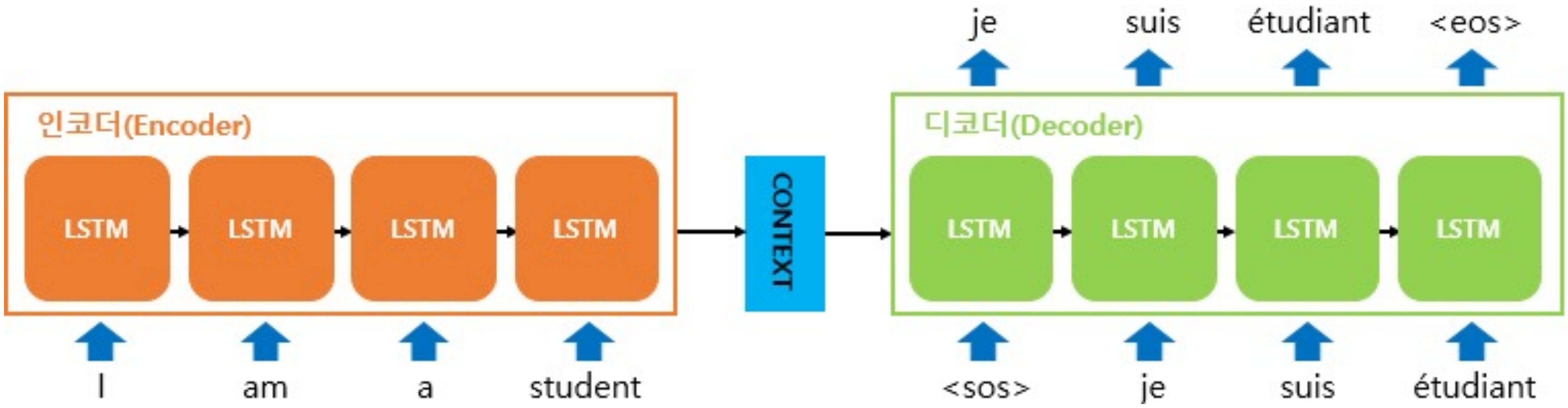
- Application of Simple RNN concepts

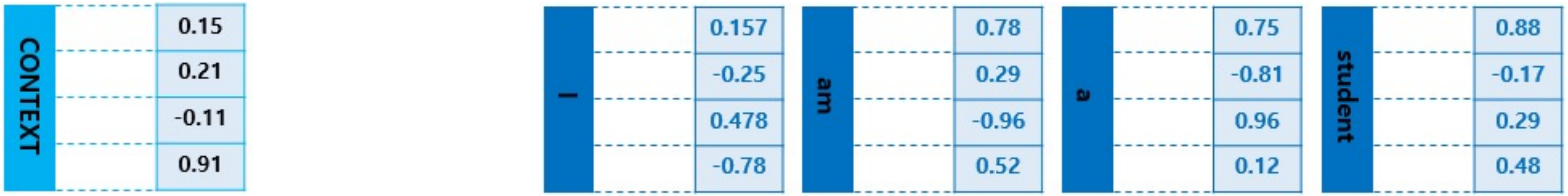
- **Translator (SMT)**

- **Encoder**: convert input sequence to vector
- **Decoder**: convert vector to output sequence

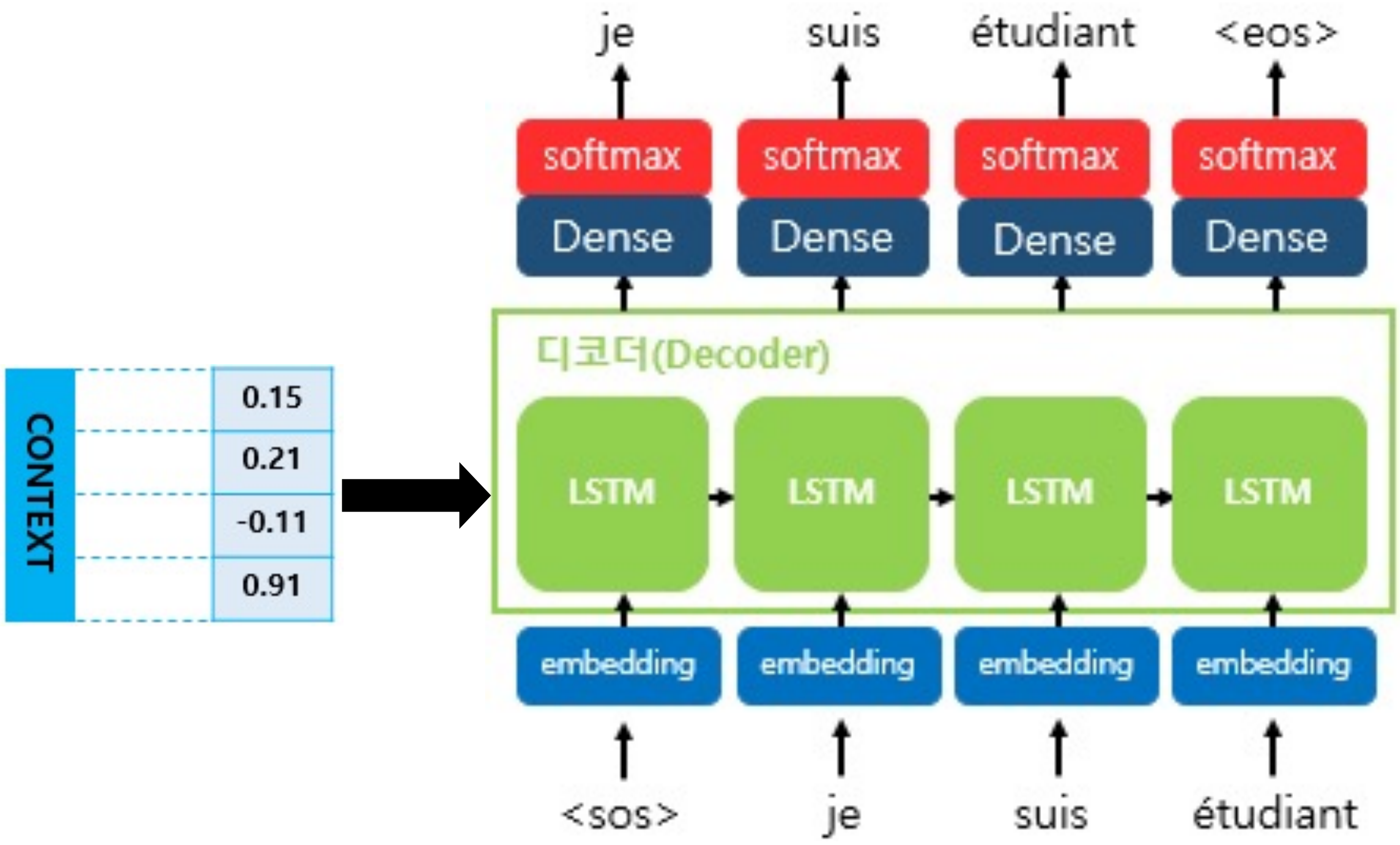
- Sequence to Sequence (**seq2seq**)

Standard Structure

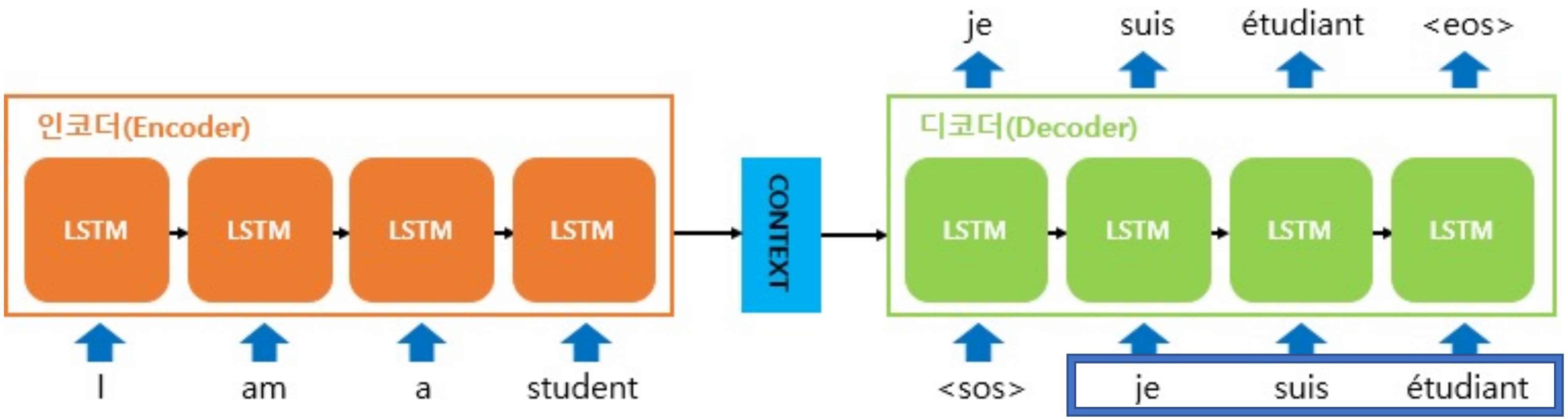




Standard Structure



Teaching forcing



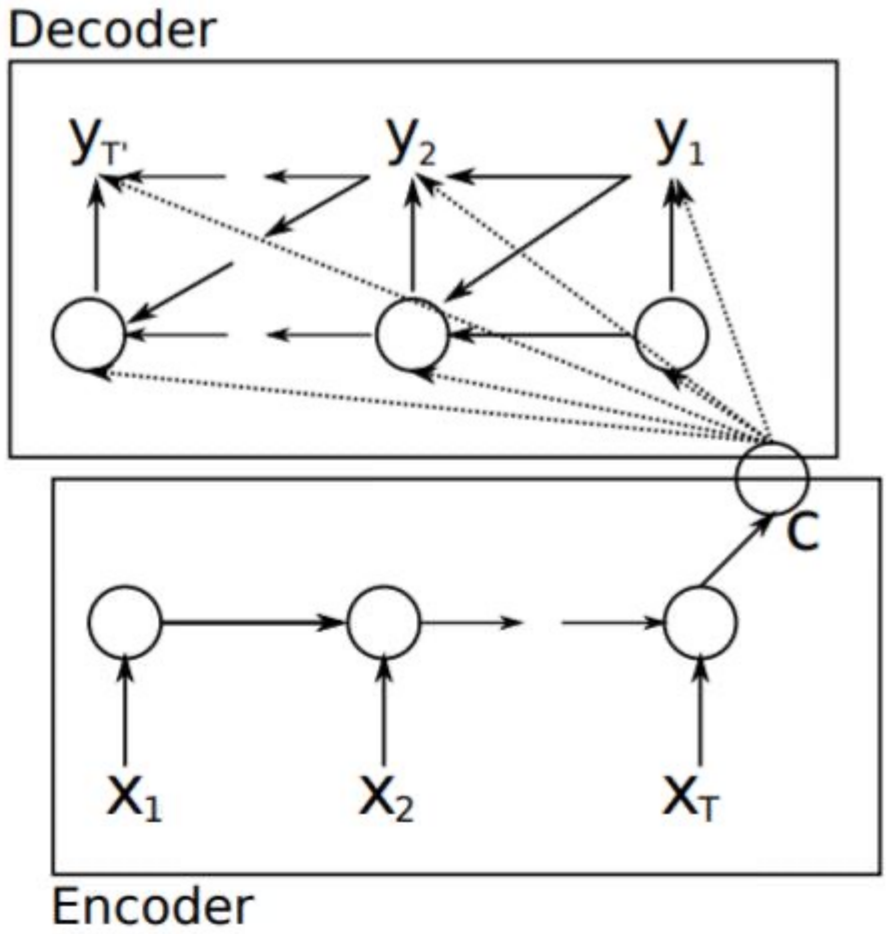
Predicted Value NO
Actual Value YES

1. Prediction of cell at the current time is also likely to go wrong.
2. A chain reaction makes it difficult to predict the entire decoder.
3. This leads to longer training time.

Encoder and Decoder

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

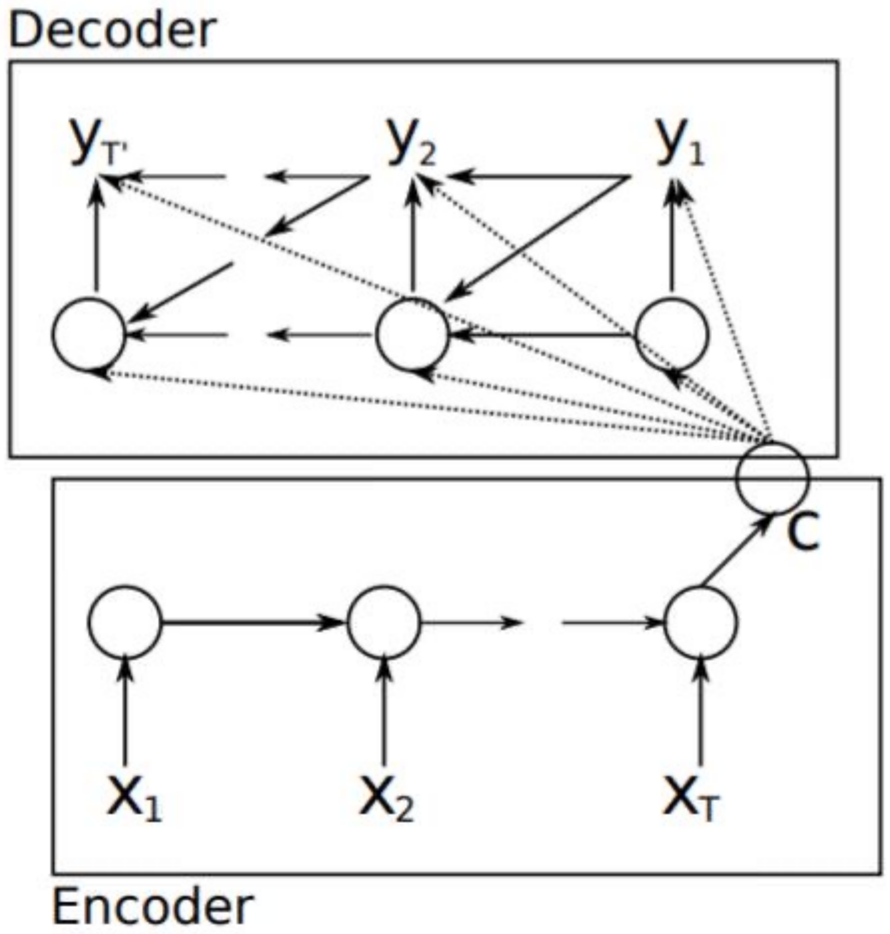
Encoder and Decoder Structure



Two RNN Models: Encoder and Decoder
-> Using GRU

Input: Variable length
vector C: fixed length
Output: Variable length

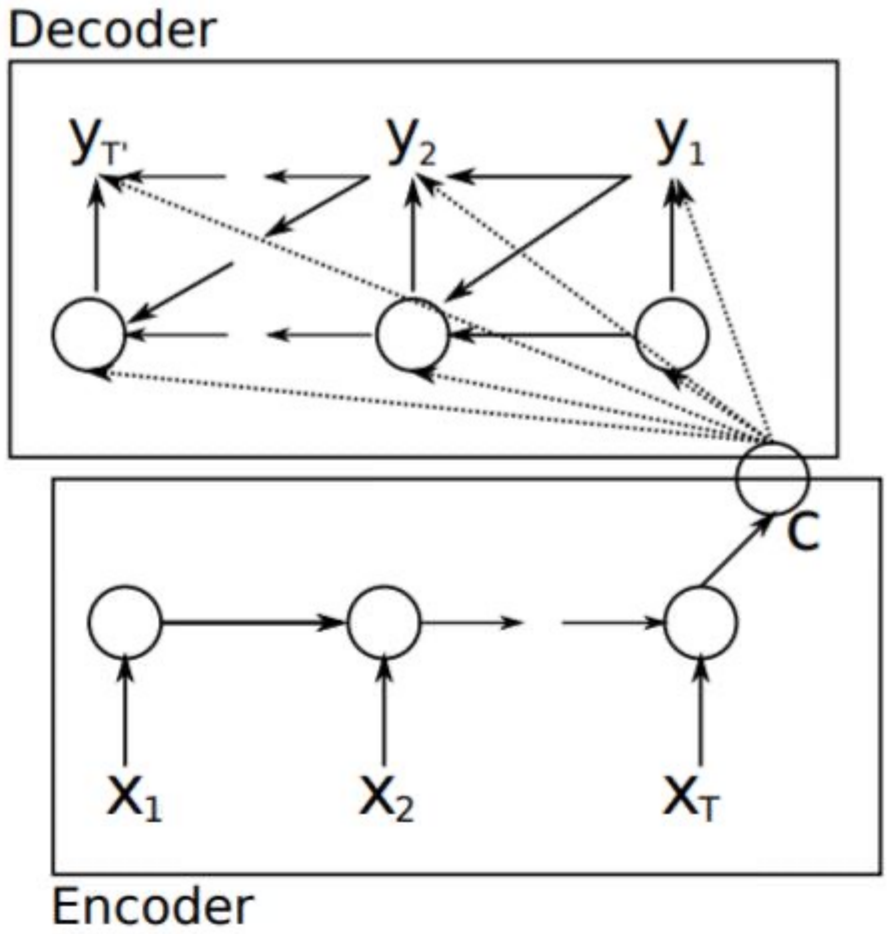
Encoder and Decoder Structure



$$p(y_1, \ldots, y_{T'} \mid x_1, \ldots, x_T)$$

Learning Probability Distribution

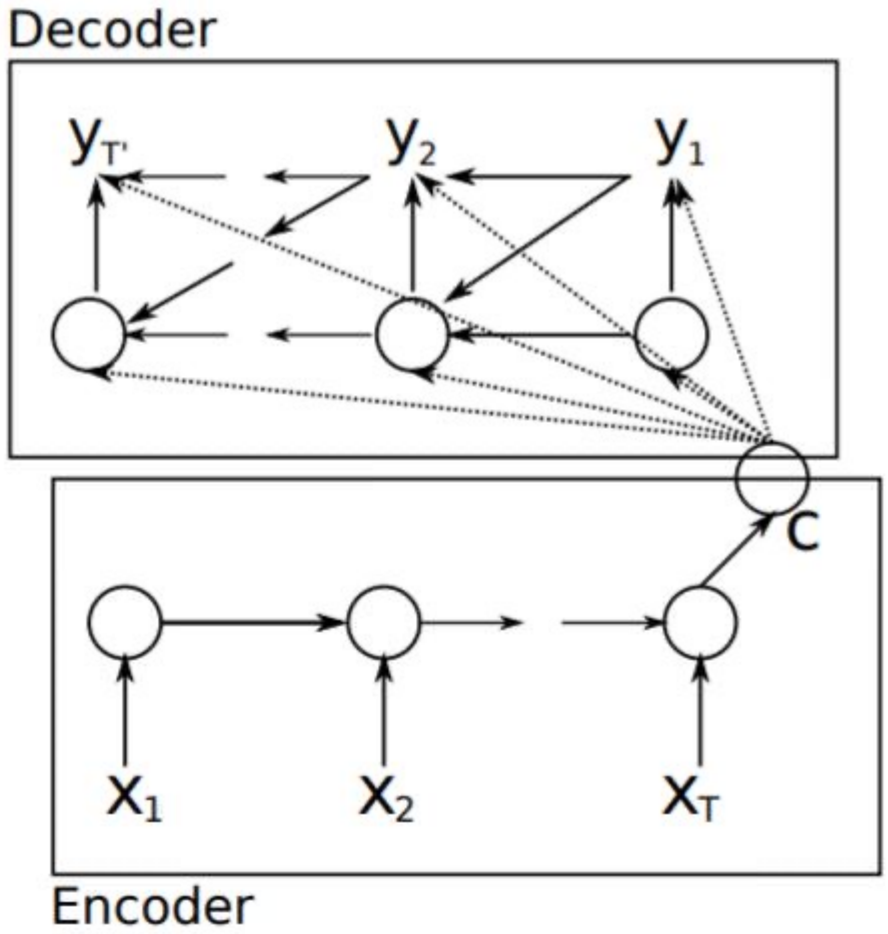
Encoder and Decoder Structure



$$\mathbf{h}_{<t>} = f(\mathbf{h}_{<t-1>}, y_{t-1}, \mathbf{c})$$

$$\mathbf{h}_{<t>} = f(\mathbf{h}_{<t-1>}, x_t)$$

Encoder and Decoder Structure



$$p(y_t | y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c}) = g(\mathbf{h}_{<t>}, y_{t-1}, \mathbf{c})$$

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}(\mathbf{y}_n | \mathbf{x}_n)$$

(**x**,**y**) is pair of input and output sequence of training data

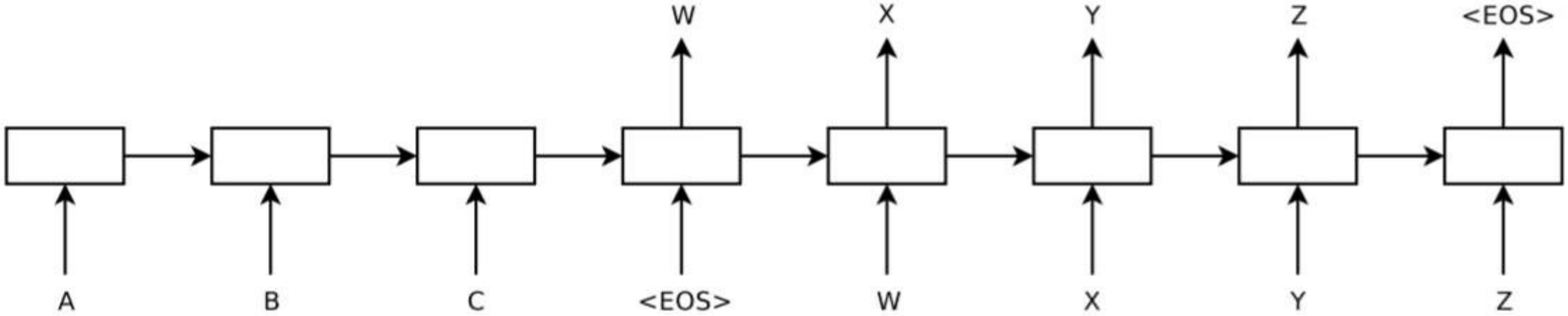
Encoder and Decoder Results

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64
CSLM + RNN + WP	31.50	34.54

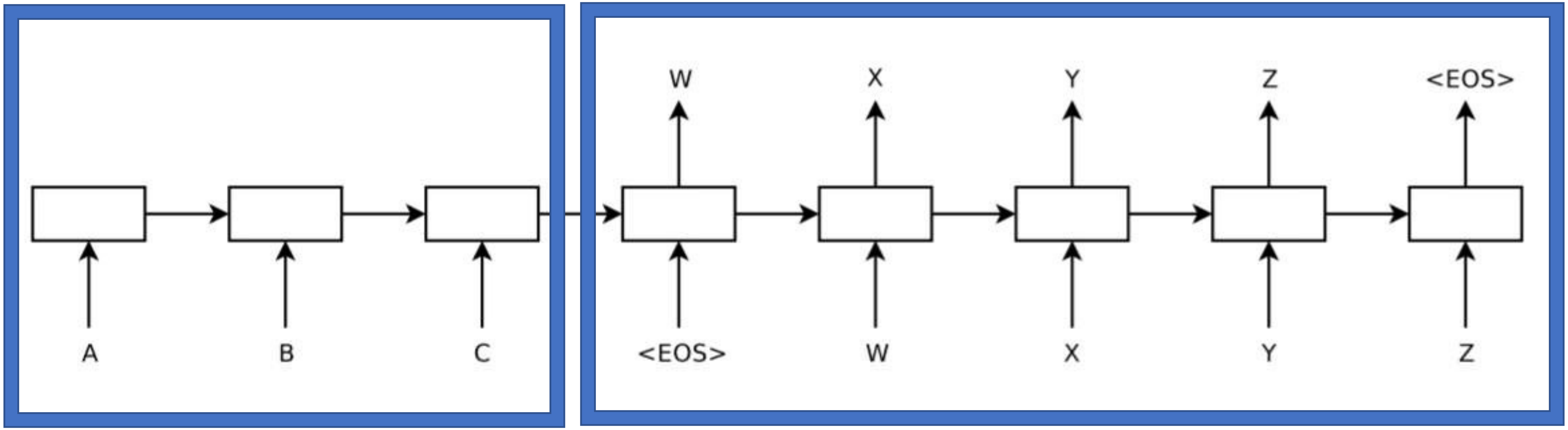
Seq2seq

Sequence to Sequence with Neural Network

Seq2seq Structure



Seq2seq Structure



1. With LSTM
2. Reversing the source sentence

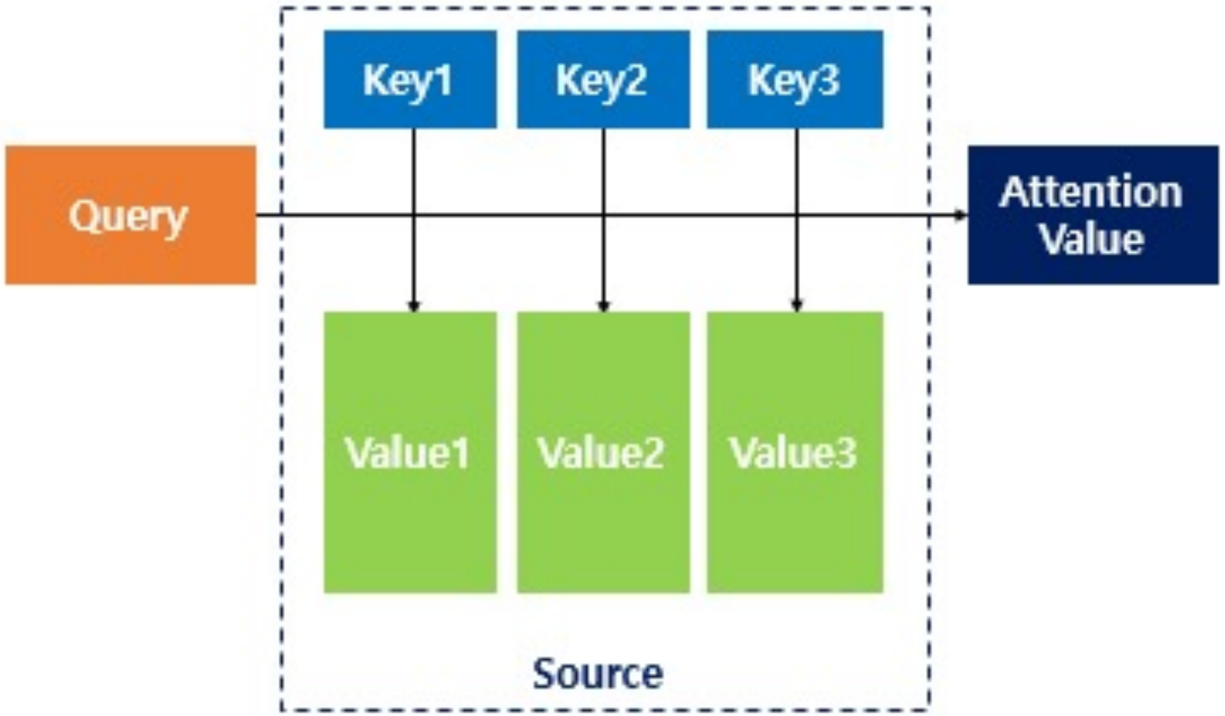
Seq2seq Results

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Attention Mechanism

Improve accuracy of seq2seq model

Concept of Attention

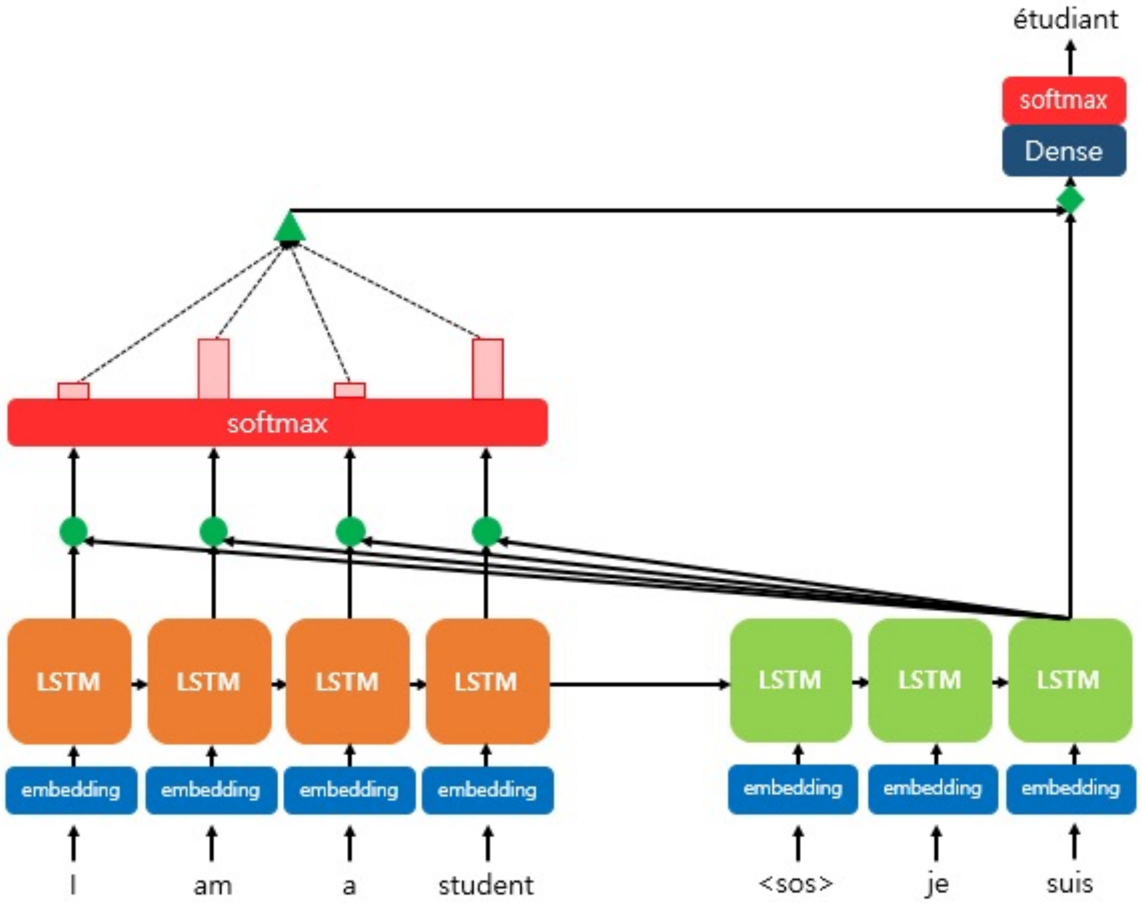


Attention(Q, K, V) = Attention Value

Query : t 시점의 디코더 셀에서의 은닉 상태
Keys : 모든 시점의 인코더 셀의 은닉 상태들
Values : 모든 시점의 인코더 셀의 은닉 상태들

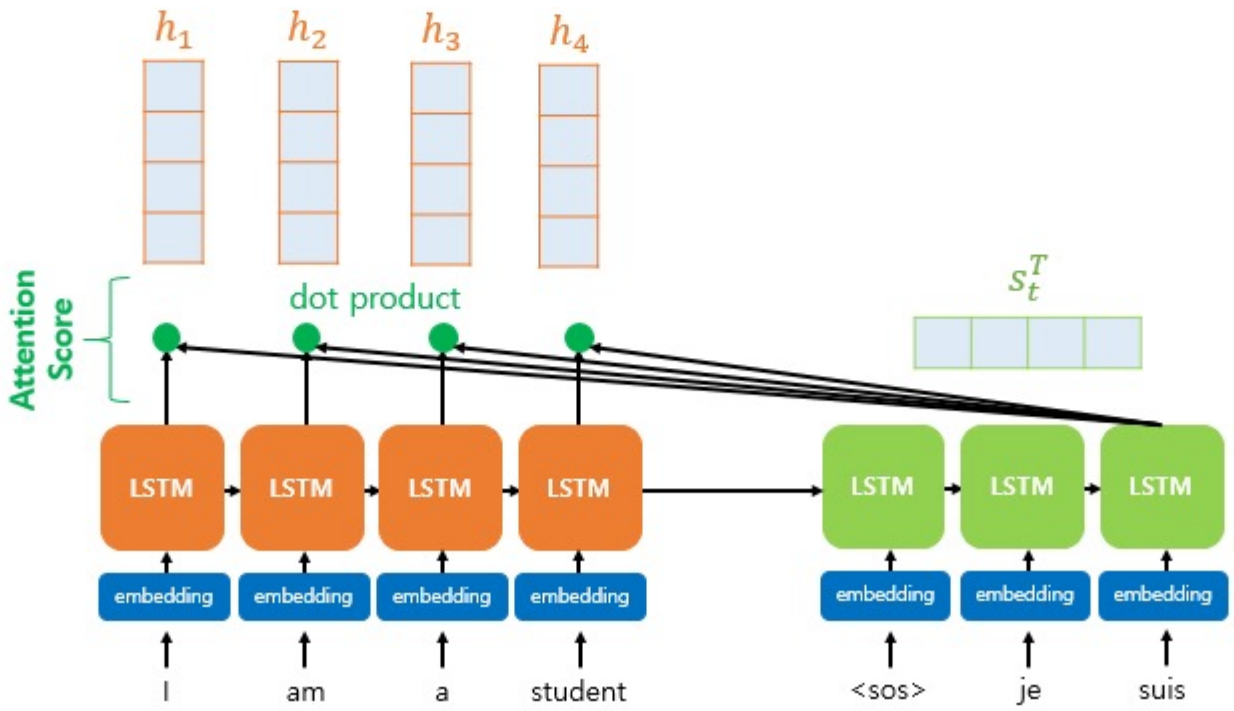
Assist seq2seq model

Dot-Product Attention

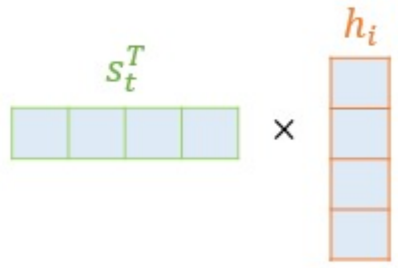


Dot-Product Attention

1. Get Attention Scores



similarity determination

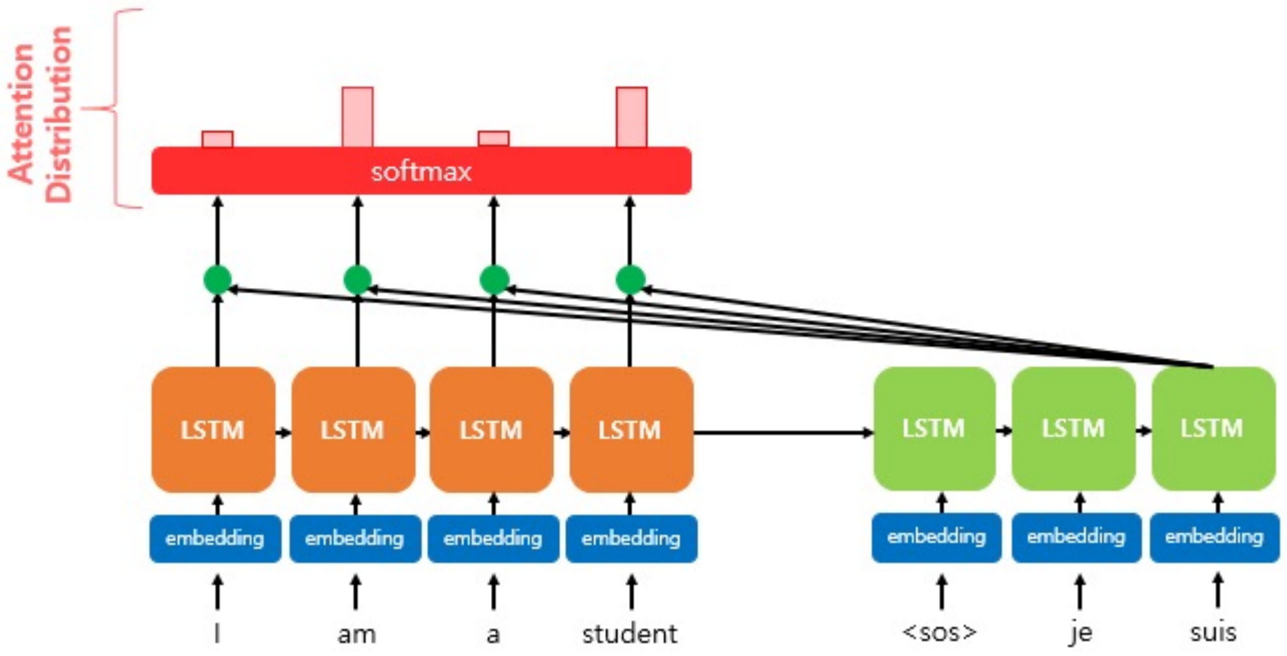


$$score(s_t,\ h_i) = s_t^T h_i$$

$$e^t = [s_t^T h_1, \ldots, s_t^T h_N]$$

Dot-Product Attention

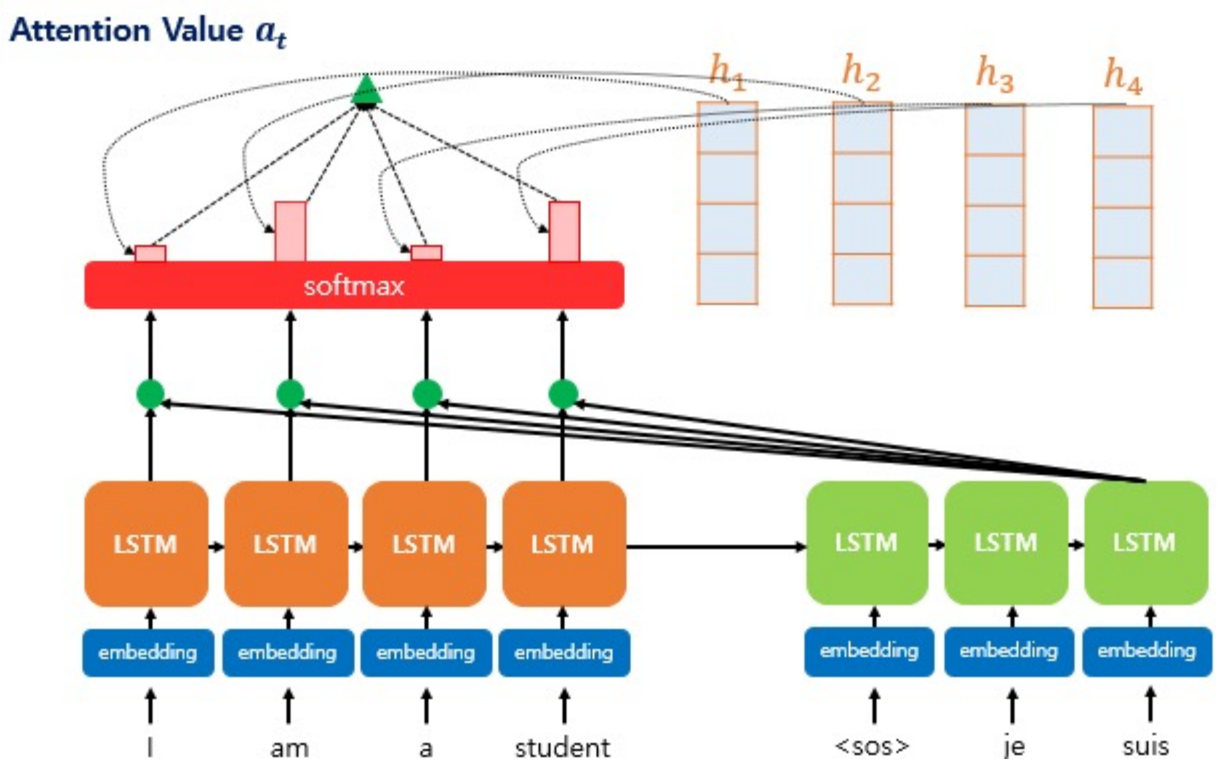
2. Get Attention Distribution via softmax function



$$\alpha^t = softmax(e^t)$$

Dot-Product Attention

3. Get Attention Value via weighted sum

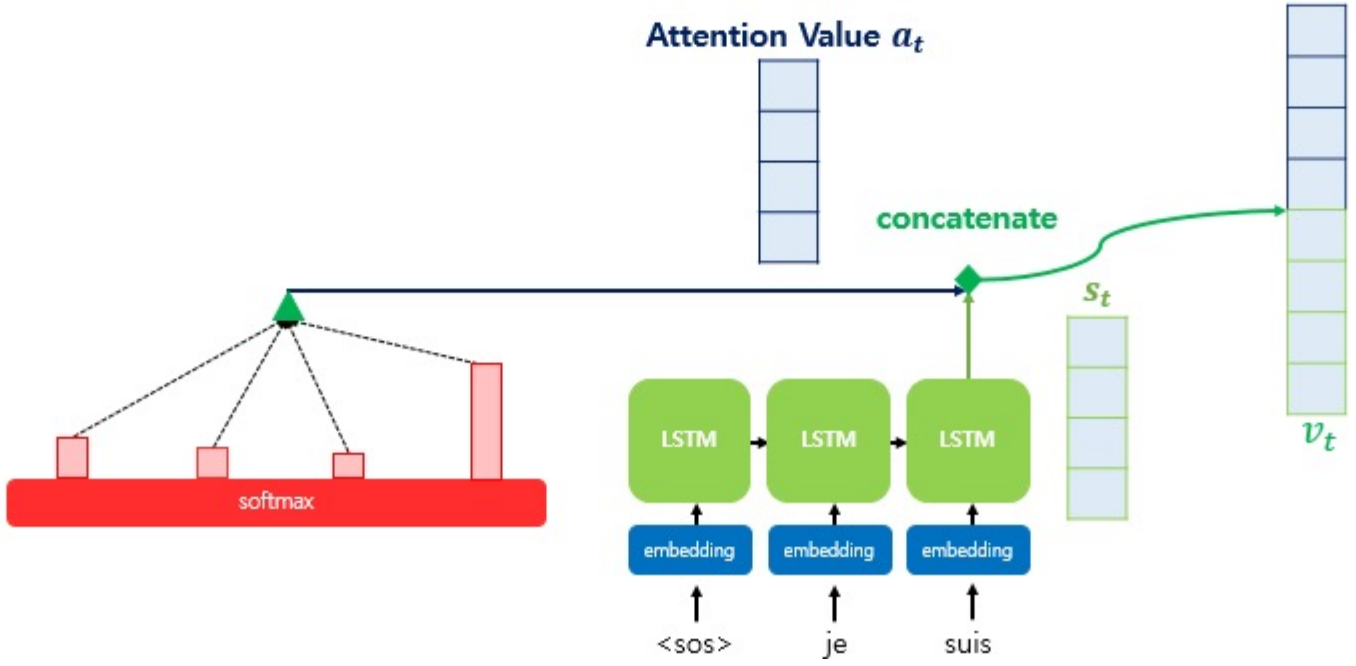


$$a_t = \sum_{i=1}^{N} \alpha_i^t h_i$$

Context Vector

Dot-Product Attention

4. Concatenate attention value and hidden state at the time t



Concatenated vector is used for input

Various Attentions

이름	스코어 함수
dot	$score(s_t,\ h_i) = s_t^T h_i$
$scaled\ dot$	$score(s_t,\ h_i) = \frac{s_t^T h_i}{\sqrt{n}}$
$general$	$score(s_t,\ h_i) = s_t^T W_a h_i$ // 단, W_a는 학습 가능한 가중치 행렬
$concat$	$score(s_t,\ h_i) = v_a^T\ tanh(W_a[s_t; h_i])$
$location - base$	$\alpha_t = softmax(W_a s_t)$ // α_t 산출 시에 s_t만 사용하는 방법.

Next Presentation

- Transformer
- BERT

Next Presentation

- Transformer
- BERT

References

References

- Kyunghyun, C., Bart, V. M., Caglar, G., Dzmitry. B. Fethi, B., Holger S., Yoshua B. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
- Sequence to Sequence with Neural Network
- Encoder-decoder paper summary
  - https://reniew.github.io/31/
- seq2seq paper summary
  - https://reniew.github.io/35/
- seq2seq
  - https://wikidocs.net/24996
- Attention mechanism
  - https://wikidocs.net/22893