

# Sequence to Sequence Learning with Neural Networks

Sutskever., et. al., 2014, Advances in neural information processing systems

A table of contents.

---

- 1 Introduction
- 2 Vanilla RNN & LSTM
- 3 Seq2Seq Model & Beam Search Algorithm
- 4 Details & Experimental Results & Conclusion

# 1. Introduction

---



DNN은 고정된 길이의 Input, Output Vector에 대해서만 적용할 수 있다는 한계가 존재한다.



그러나 실제로는 고정된 Input, Output Vector가 없는 문제들이 상당히 많다.

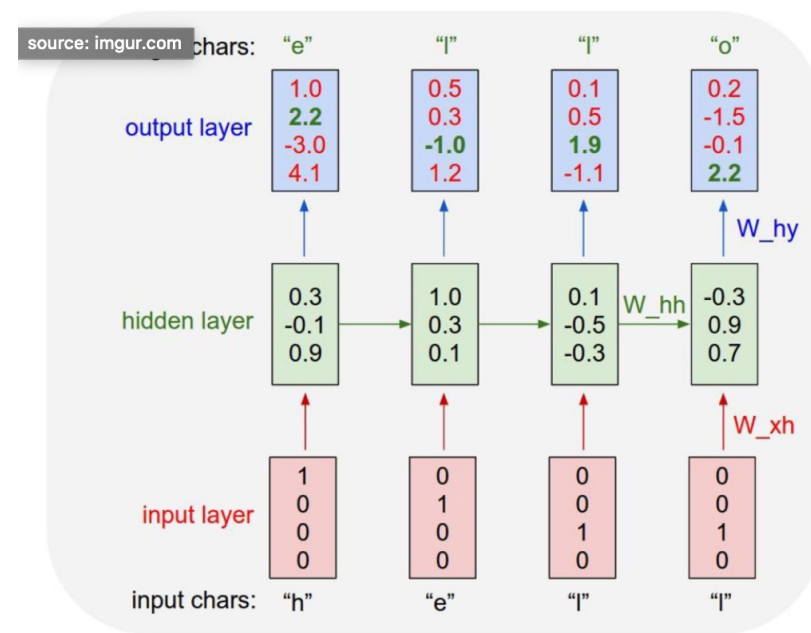
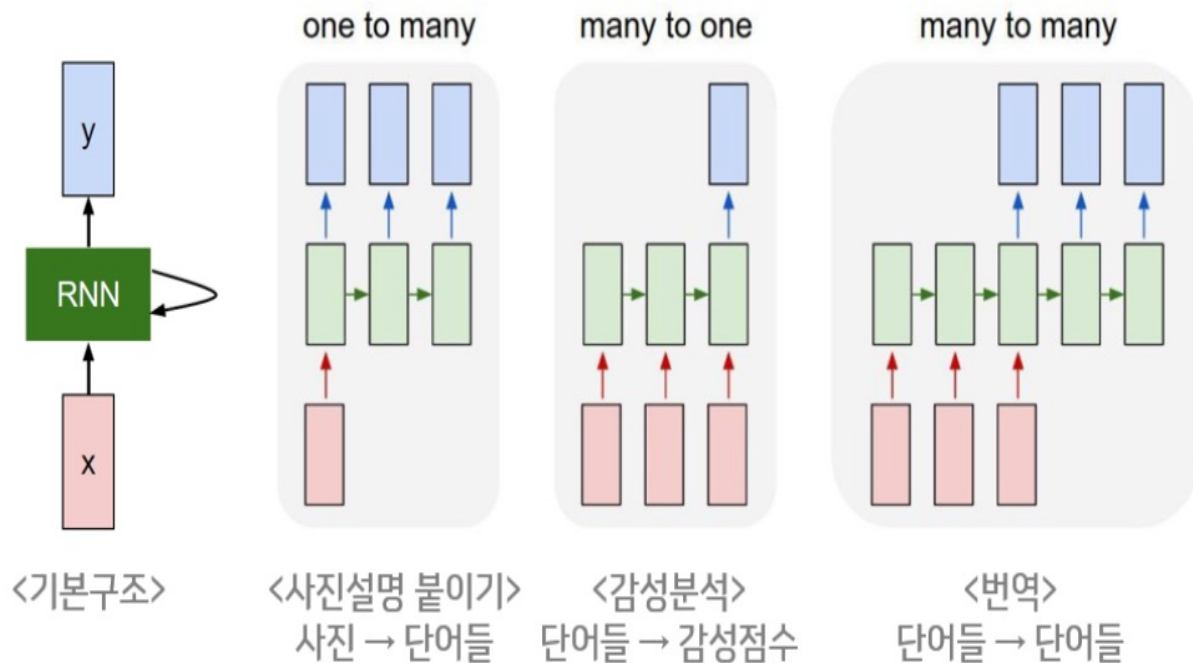
Speech Recognition, Machine Translation, Question & Answering



Introduce a domain-independent **method that learns to map sequences to sequences** based on LSTM architecture "Seq2Seq with Neural Networks"

## 2. Vanilla RNN(Recursive Neural Network) – What is RNN?

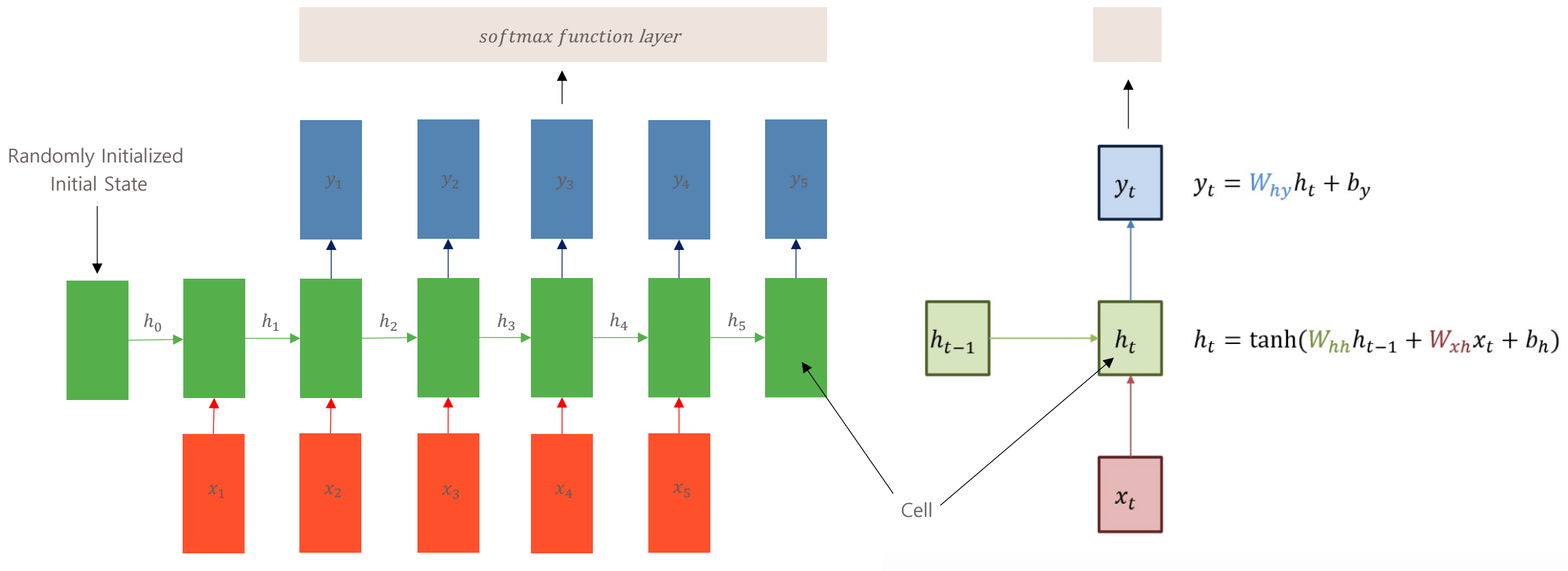
### ✓ Structure of RNN



- Sequence Data, Time-Series Data를 활용해 Modeling할 때 많이 사용하는 구조이다.

## 2. Vanilla RNN – What is RNN?

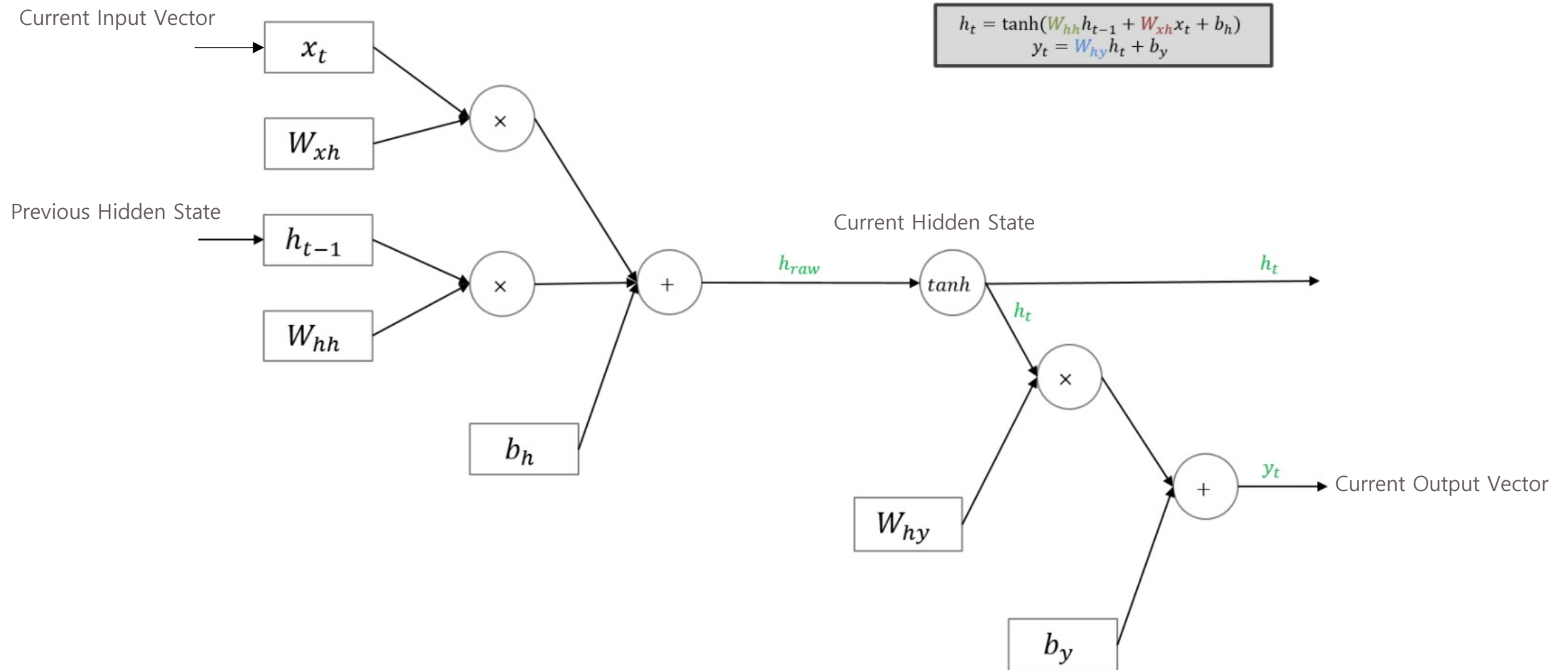
### ✓ Forward Propagation of RNN (Many to Many)



- 실제 위의 Model에서의 Cell은 1개이며, Cell은 Recursive하게 Current State를 Next State에 넘겨준다.

## 2. Vanilla RNN – What is RNN?

### ✓ Forward Propagation of RNN



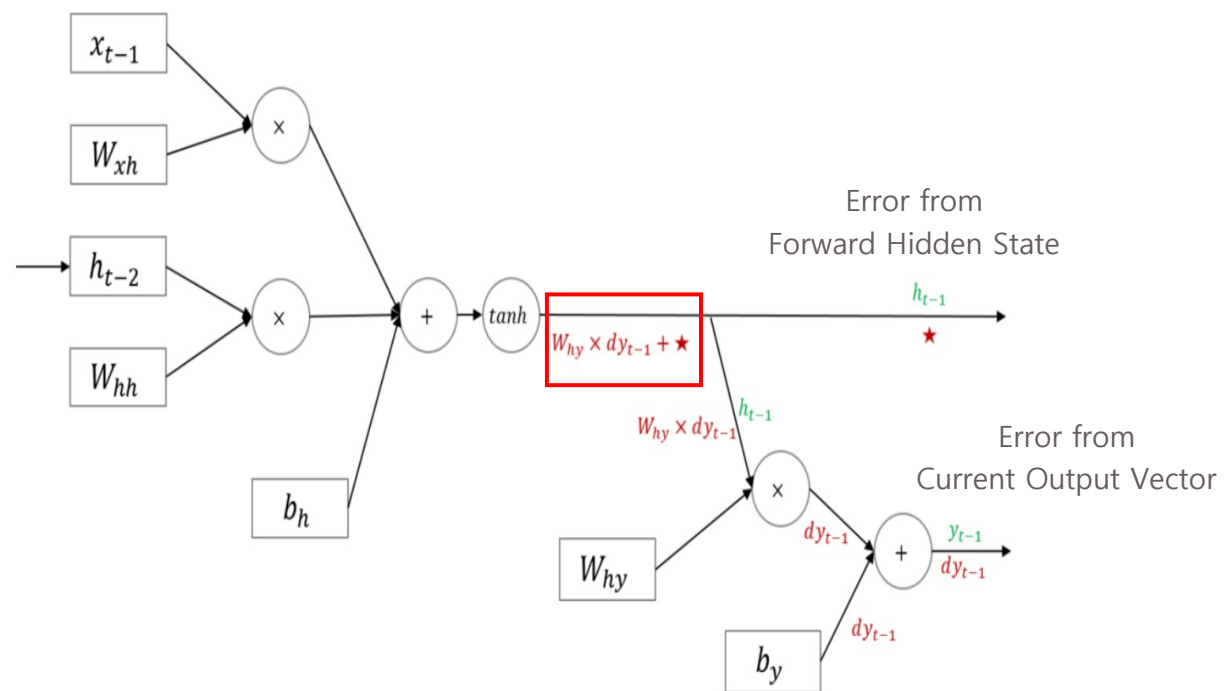
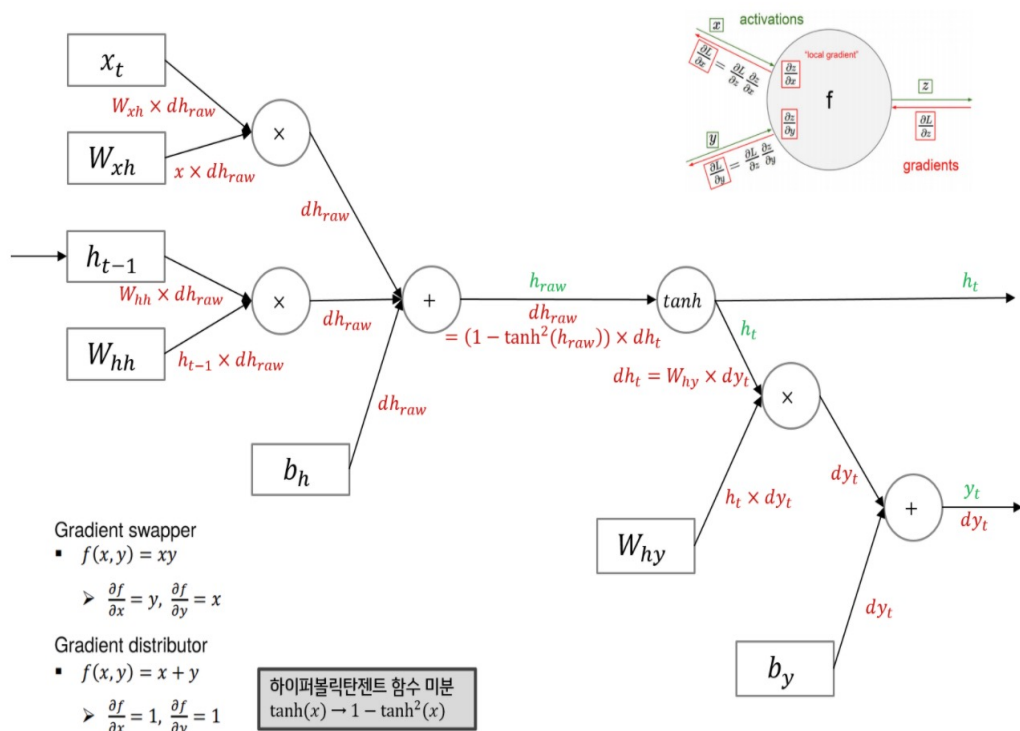
## 2. Vanilla RNN – What is RNN?



### Back Propagation of RNN

\* RNN은 해당 시점에서의 Error은 **Current State Output의 Error**, **Forward Hidden State의 Error**를 동시에 반영해야한다.

- RNN 에서 Chain Rule을 활용한 Back Propagation 계산 과정



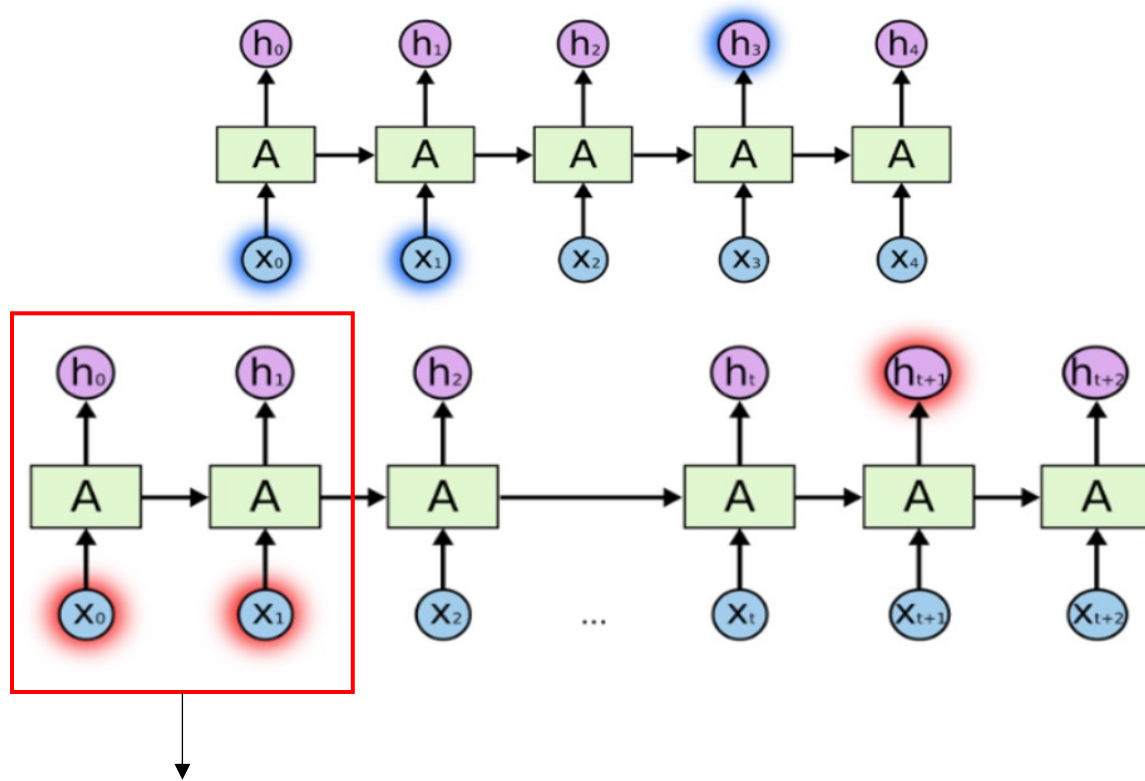
## 2. Vanilla RNN – Problem of Vanilla RNN



### Long range temporal dependency problem of chain rule

\* 기본적인 RNN의 가정은 **Current Hidden State**는 **Current Input Vector**을 가장 잘 반영한다는 것이다.

\* Input Sentence가 길 경우, Chain Rule의 특성상 앞 쪽에 위치한 정보를 반영하지 못해 **Vanishing/Exploding Gradient Problem** 발생한다.



Skip Connection, Dropout, ...  
, GRU, "LSTM"

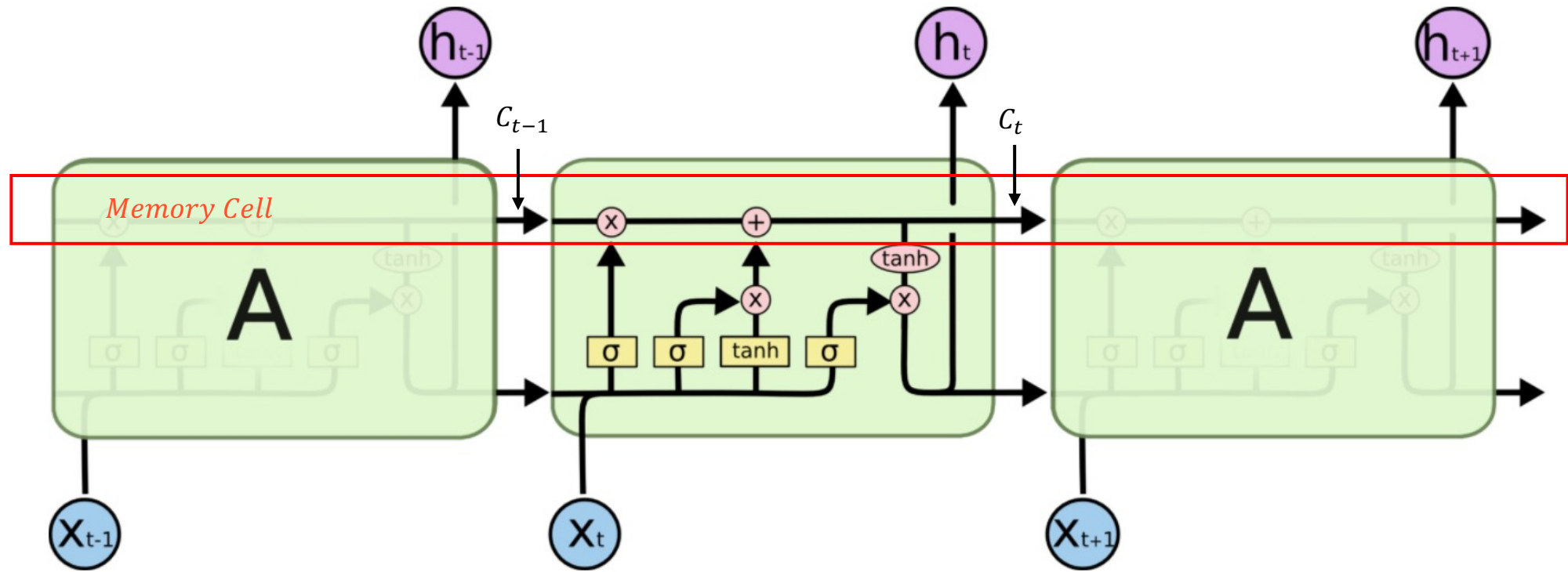
$$\theta_{new} = \theta_{old} - \alpha \cdot \Delta\theta_{old}$$
$$\Delta\theta_{old} \approx 0$$



## 2. LSTM(Long Short Term Memory) – What is LSTM?

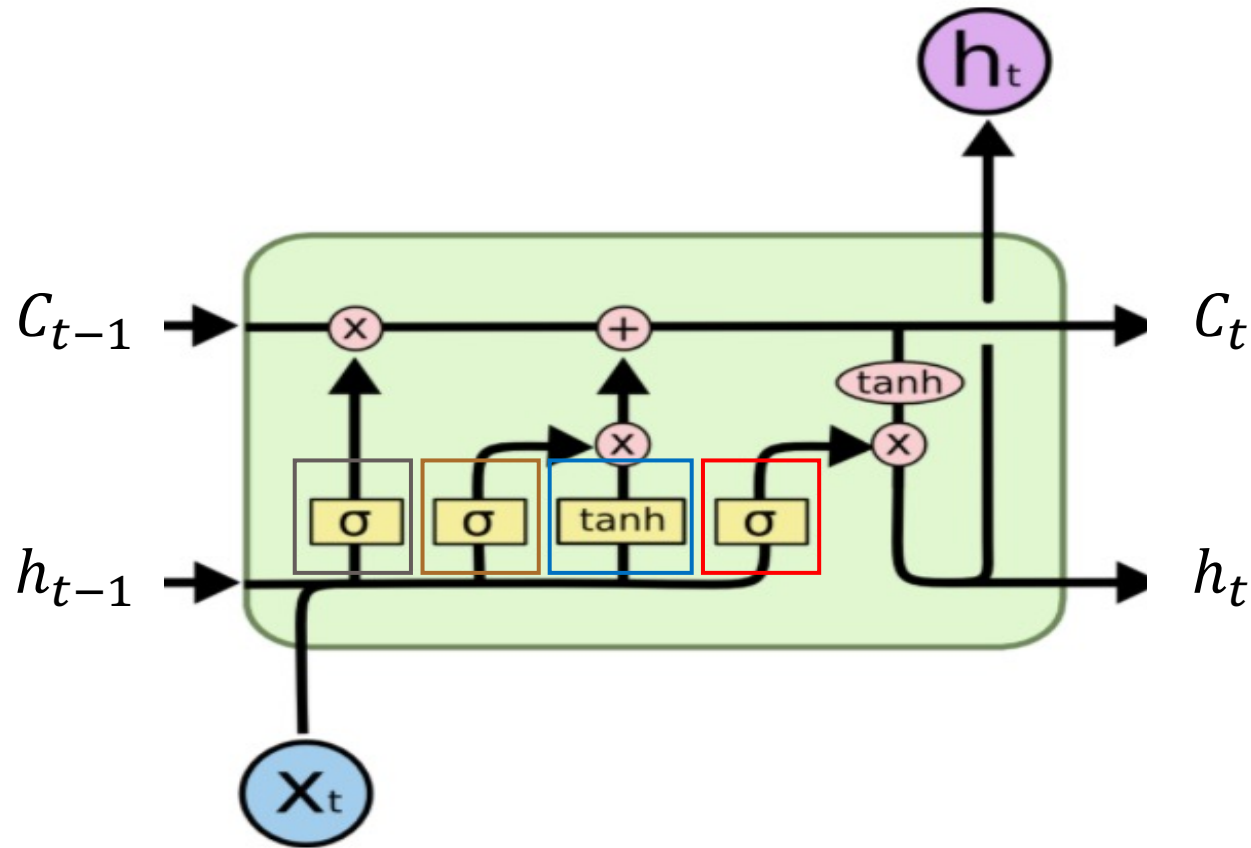
### ✓ Structure of LSTM

- Memory Cell( $c_t$ )을 활용해 과거의 정보를 일부 활용해 Current Hidden State을 계산한다.



## 2. LSTM – What is LSTM?

### ✓ Structure of LSTM

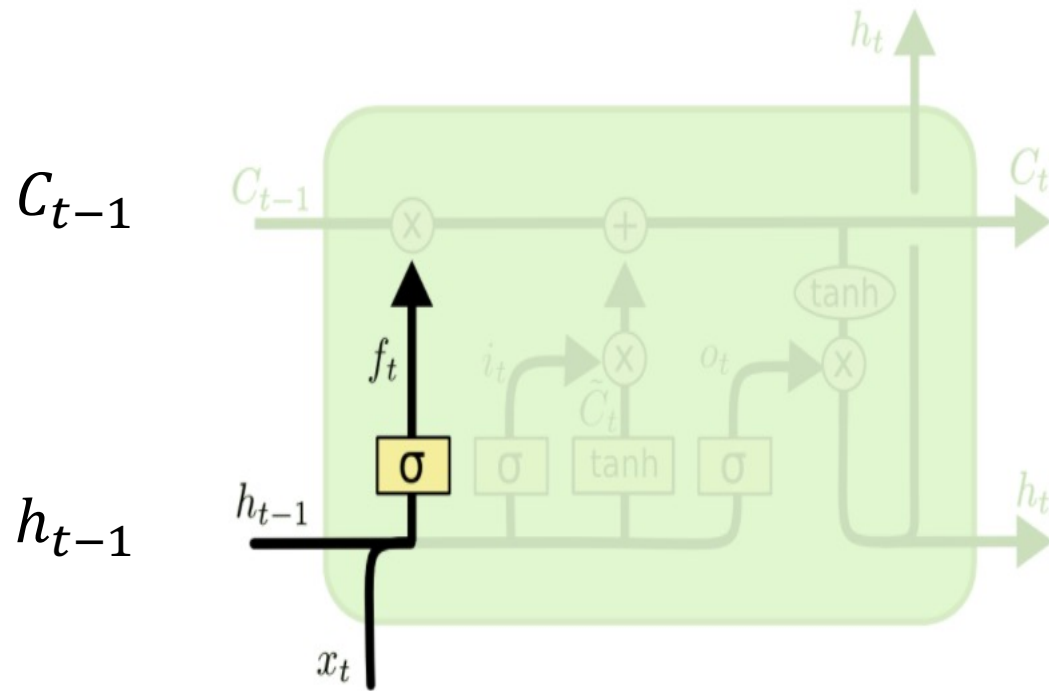


$$\begin{aligned} f_t &= \sigma(W_{xh_f}x_t + W_{hh_f}h_{t-1} + b_{h_f}) \\ i_t &= \sigma(W_{xh_i}x_t + W_{hh_i}h_{t-1} + b_{h_i}) \\ o_t &= \sigma(W_{xh_o}x_t + W_{hh_o}h_{t-1} + b_{h_o}) \\ g_t &= \tanh(W_{xh_g}x_t + W_{hh_g}h_{t-1} + b_{h_g}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

## 2. LSTM – What is LSTM?

### ✓ Structure of LSTM - Forget

\* Previous Hidden State, Current Input Vector를 활용해 과거의 정보의 일정 부분만 기억하자.



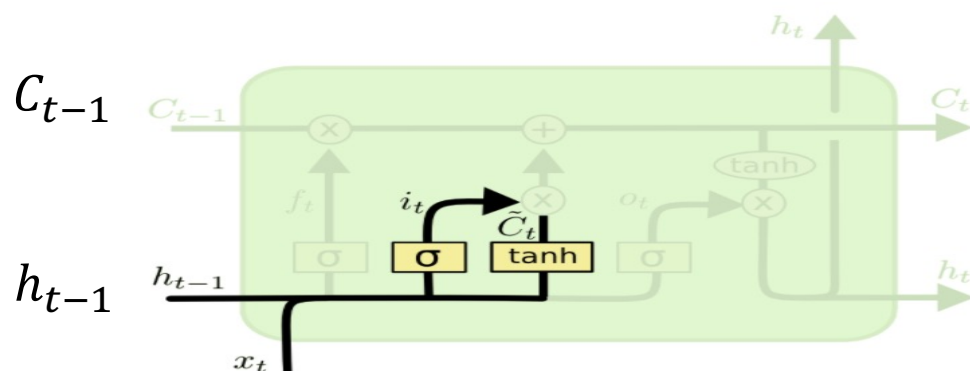
$$f_t = \sigma(W_{xh\_f}x_t + W_{hh\_f}h_{t-1} + b_{h\_f})$$

## 2. LSTM – What is LSTM?



### Structure of LSTM – Update Memory Cell Information

\* Previous Hidden State, Current Input Vector를 통해 현재 정보를 활용해 Memory Cell에 계속 기억해야할 정보를 Update 해주자.



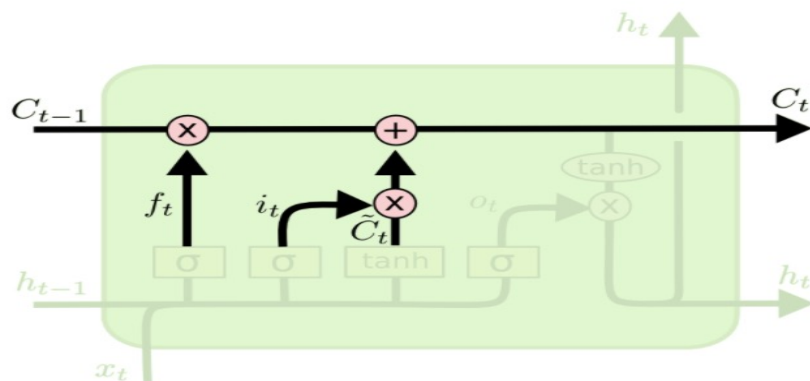
현재 정보의 일정 %를 활용하자.

$$i_t = \sigma(W_{xh\_i}x_t + W_{hh\_i}h_{t-1} + b_{h\_i})$$

현재 정보로 새로운 방향성을 만들어주자.

$$g_t = \tanh(W_{xh\_g}x_t + W_{hh\_g}h_{t-1} + b_{h\_g})$$

$$(\tilde{c}_t = g_t)$$

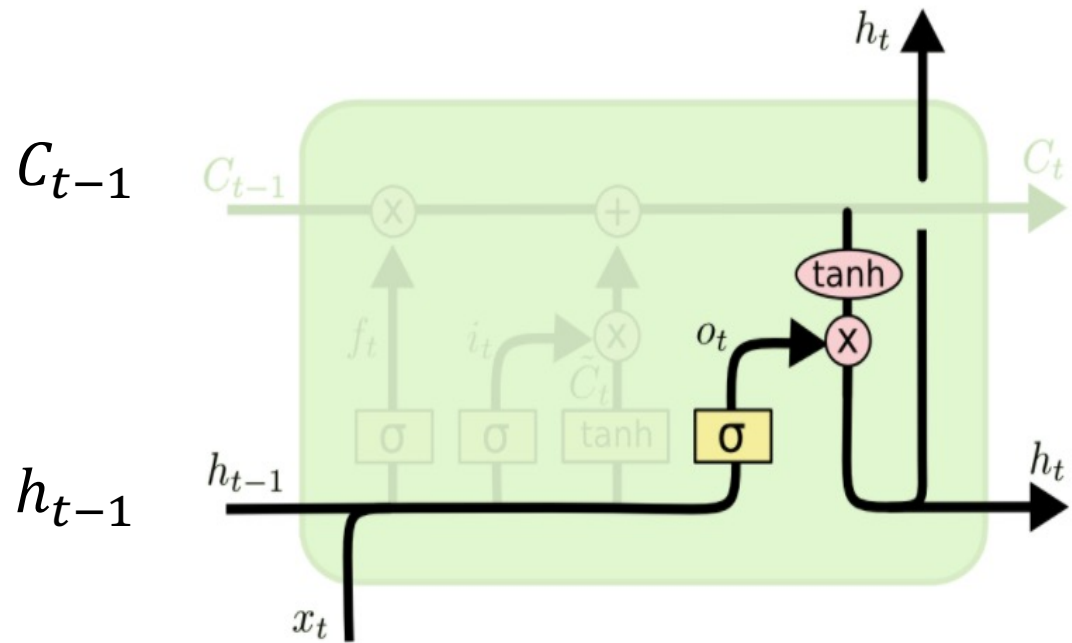


Memory Cell Update

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

## 2. LSTM – What is LSTM?

- ✓ Structure of LSTM – Calculate Current Hidden State

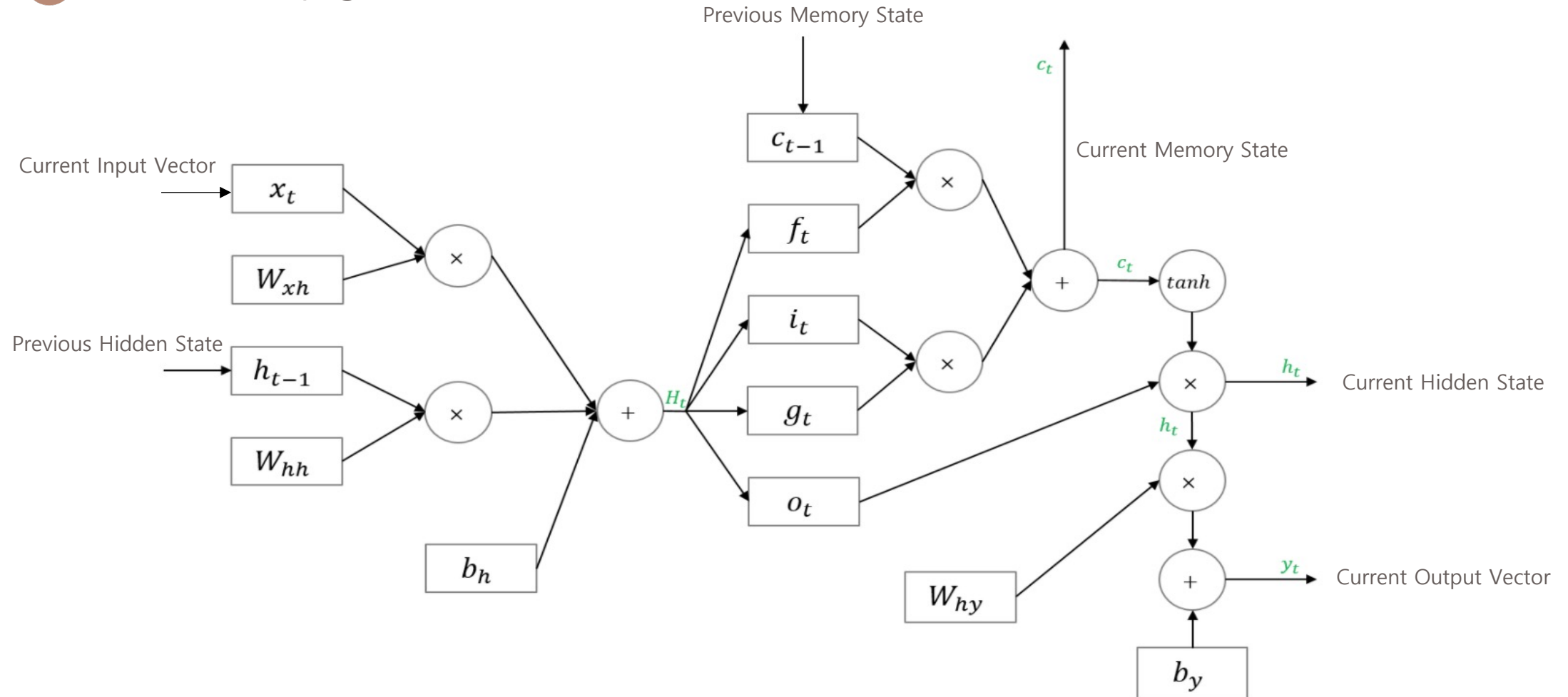


$$o_t = \sigma(W_{xh_o}x_t + W_{hh_o}h_{t-1} + b_{h_o})$$

$$h_t = o_t \odot \tanh(c_t)$$

## 2. LSTM – What is LSTM?

### ✓ Forward Propagation of LSTM

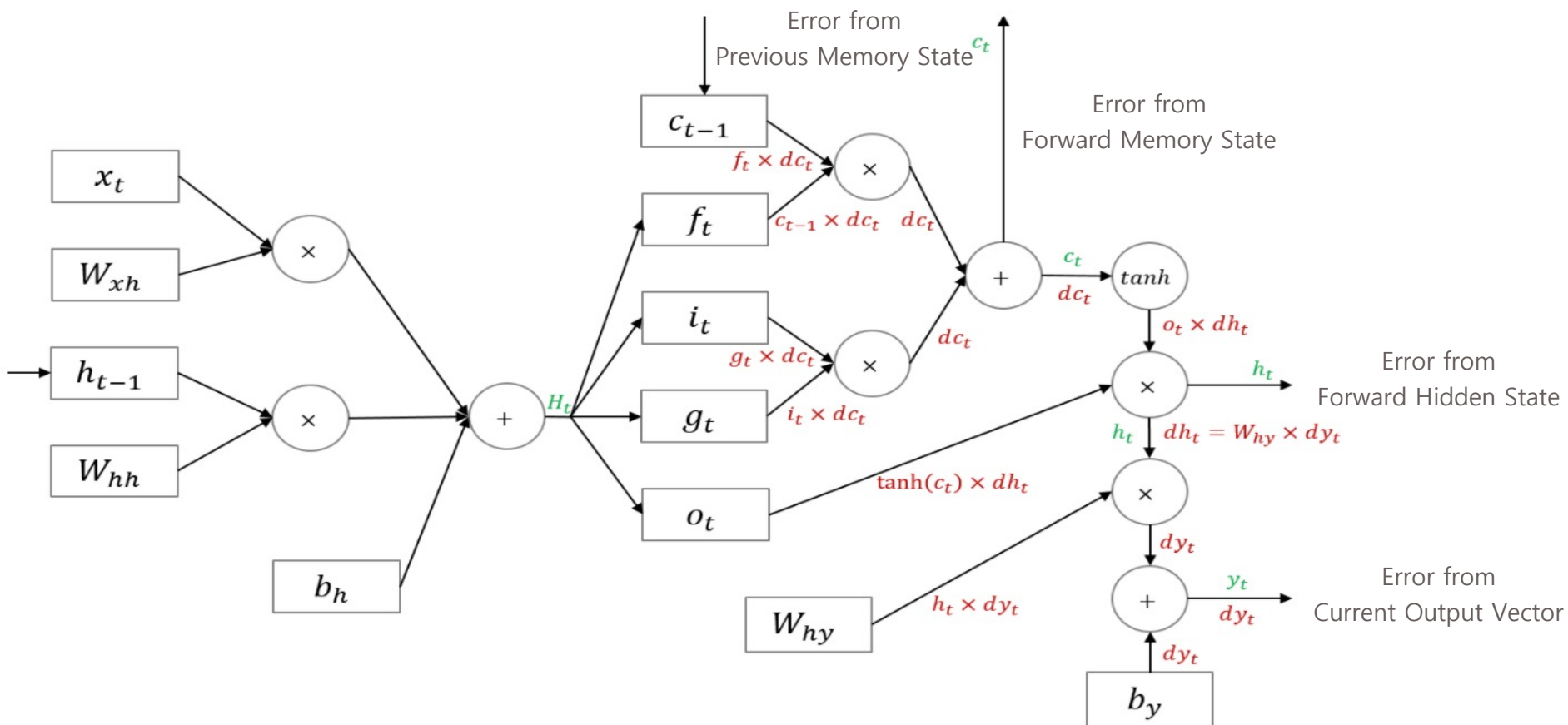


## 2. LSTM – What is LSTM?



### Back Propagation of LSTM

\* LSTM은 Current State Output의 Error, Forward Hidden State의 Error, Forward & Previous Memory State의 Error를 동시에 반영해야한다.

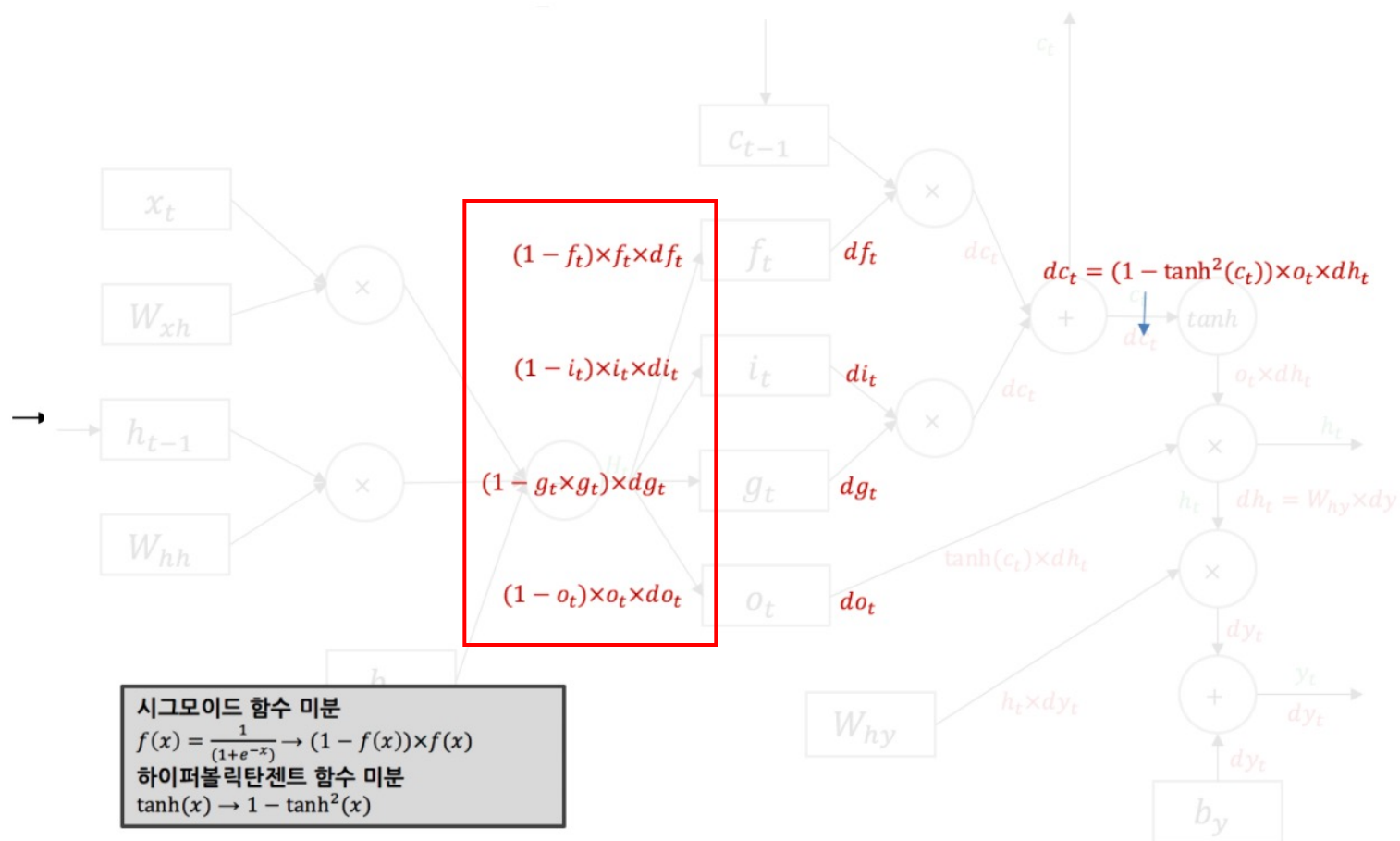


## 2. LSTM – What is LSTM?



### Back Propagation of LSTM

\* LSTM은 Current State Output의 Error, Forward Hidden State의 Error, Forward & Previous Memory State의 Error를 동시에 반영해야한다.

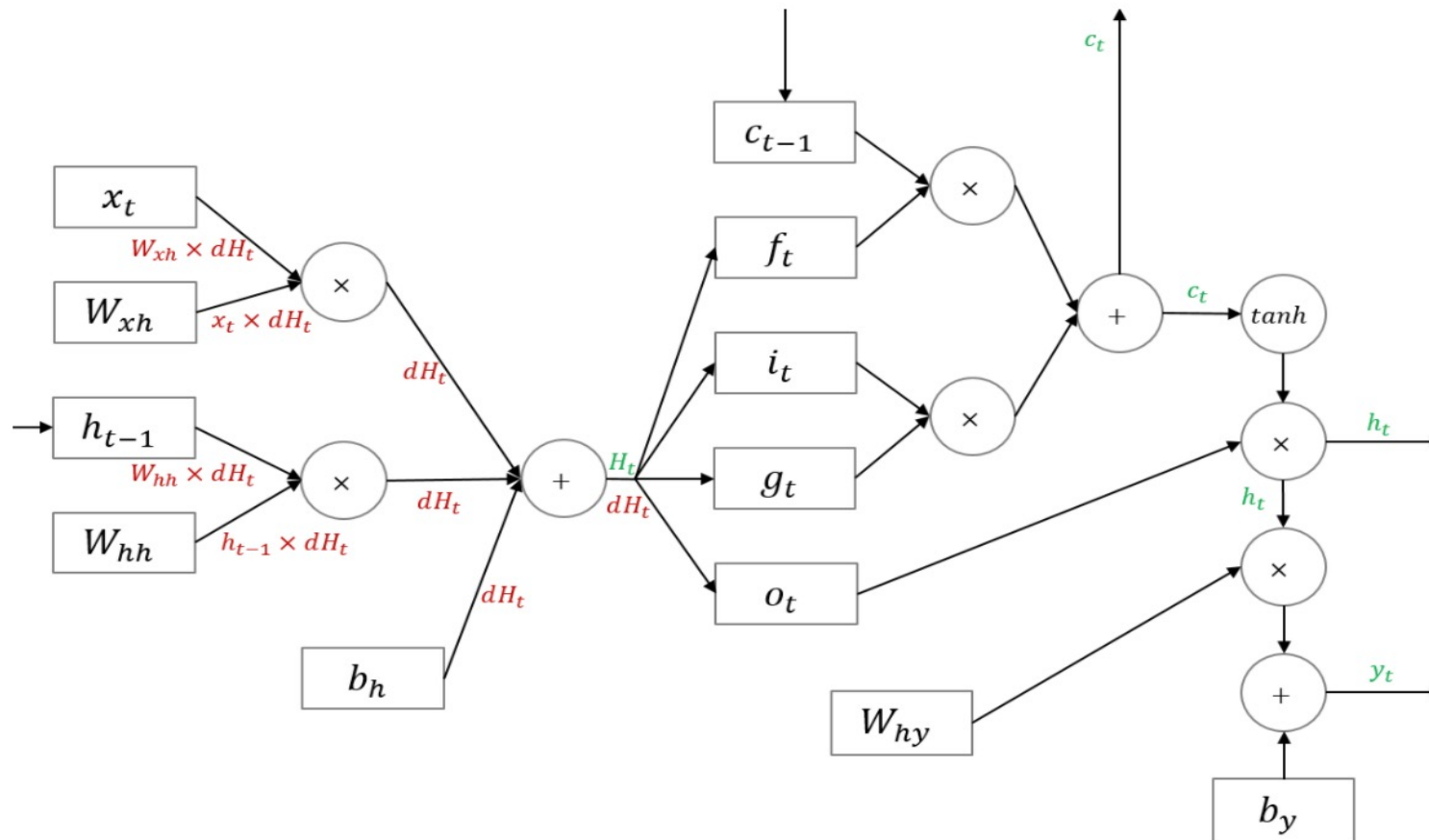




## 2. LSTM – What is LSTM?

### ✓ Back Propagation of LSTM

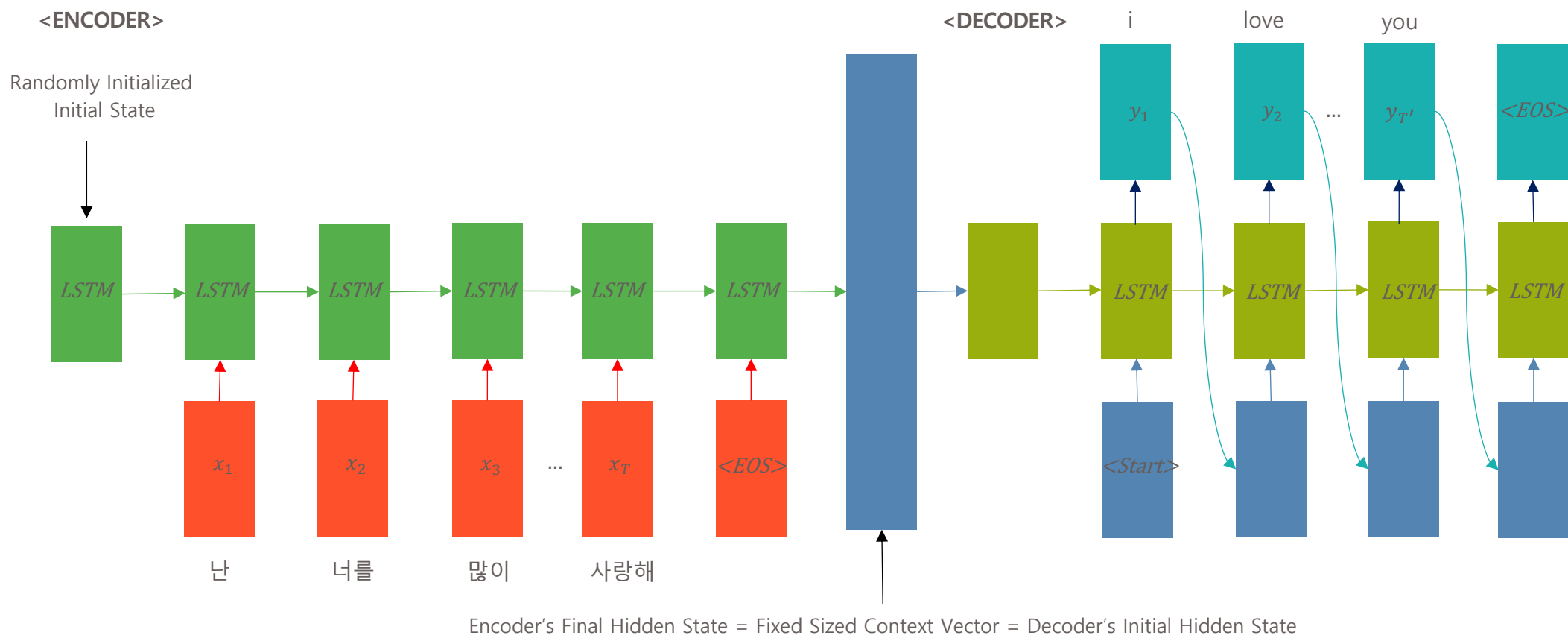
\* LSTM은 Current State Output의 Error, Forward Hidden State의 Error, Forward & Previous Memory State의 Error를 동시에 반영해야한다.



### 3. Sequence to Sequence Model

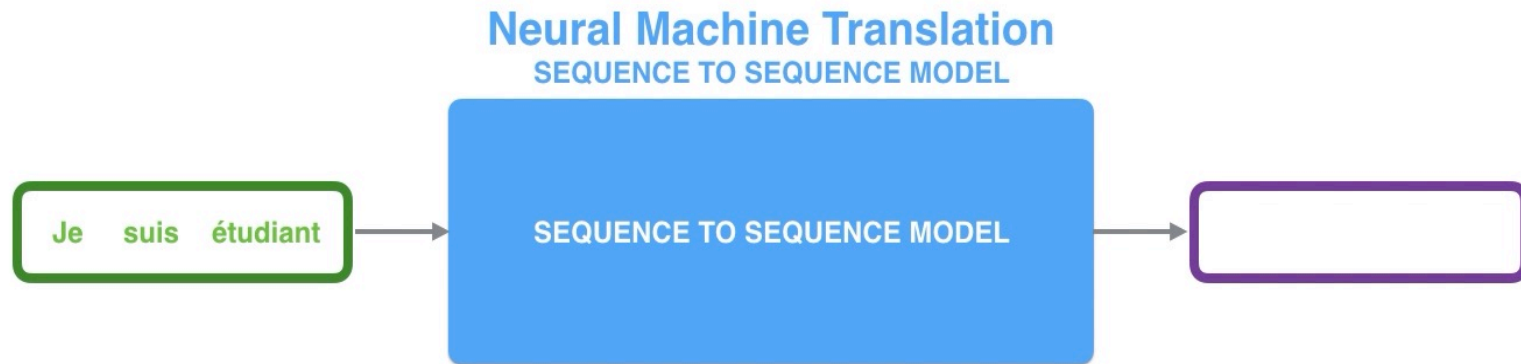
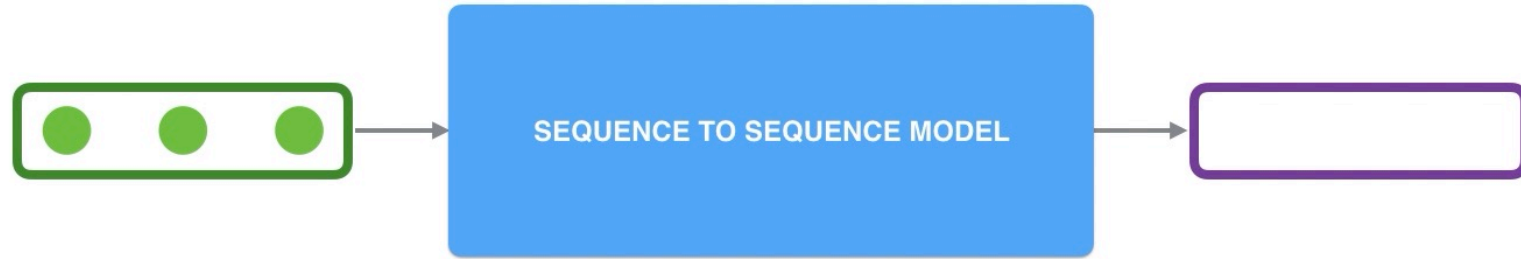
- ✓ LSTM을 통해 Input Sentence를 하나의 고정된 길이의 Context Vector로 Mapping 한 이후, 또다른 LSTM을 통해 Context Vector를 Output Sentence로 Mapping 하는 모델을 만들자!

#### - Structure of Seq2Seq Model



### 3. Sequence to Sequence Model

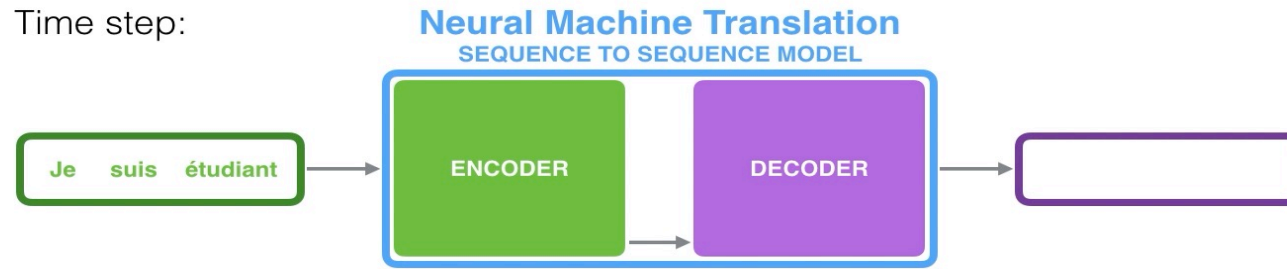
---



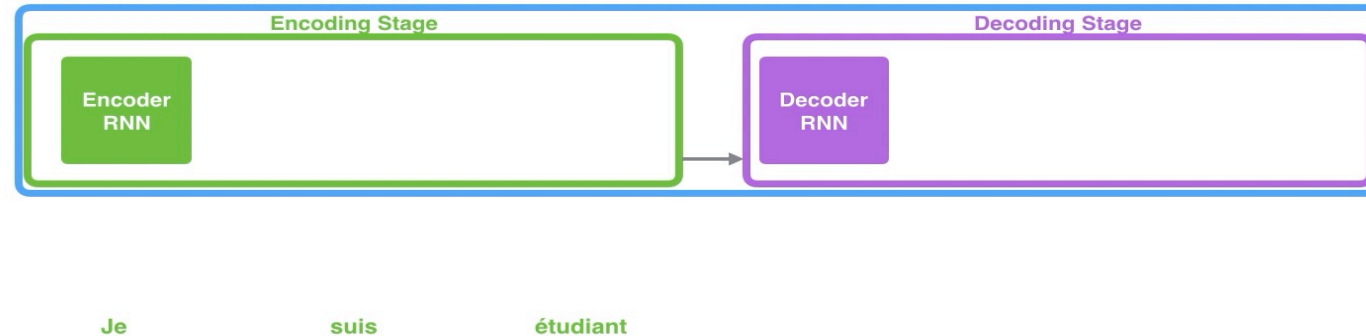
### 3. Sequence to Sequence Model

---

Time step:



**Neural Machine Translation**  
SEQUENCE TO SEQUENCE MODEL



### 3. Sequence to Sequence Model

---

- ✓ English > French Translation Data Set을 통해 Model을 학습시켰다.
- ✓ 160,000개의 영어 단어, 80,000개의 불어 단어를 활용했고, 1,000 차원 공간에 Embedding 하였다.
- ✓ Embedding Space에 없는 단어는 "UNK" Token으로 처리했다.
- ✓ 문장이 끝나면 <EOS> Token이 등장하도록 했다.
- ✓ LSTM Cell을 활용해 Long term dependency 문제를 보완했다.
- ✓ 예측할 때는 Beam Search Algorithm을 통해 가장 높은 확률의 번역을 선택했다.
- ✓ Source 문장을 역순으로 입력했을 경우, Model Performance가 더 좋았다.

### 3. Sequence to Sequence Model

---

#### ✓ LSTM Formulation

$$p(y_1, y_2, \dots, y_{T'} | x_1, x_2, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, y_2, \dots, y_{t-1})$$

where  $(x_1, x_2, \dots, x_T)$  is input sentence,  $(y_1, y_2, \dots, y_{T'})$  is output sentence,  $v$  is last hidden state of Encode LSTM

#### ✓ Objective Function

$$\max 1/|\mathcal{S}| \sum_{(T,S) \in \mathcal{S}} \log p(T|S)$$

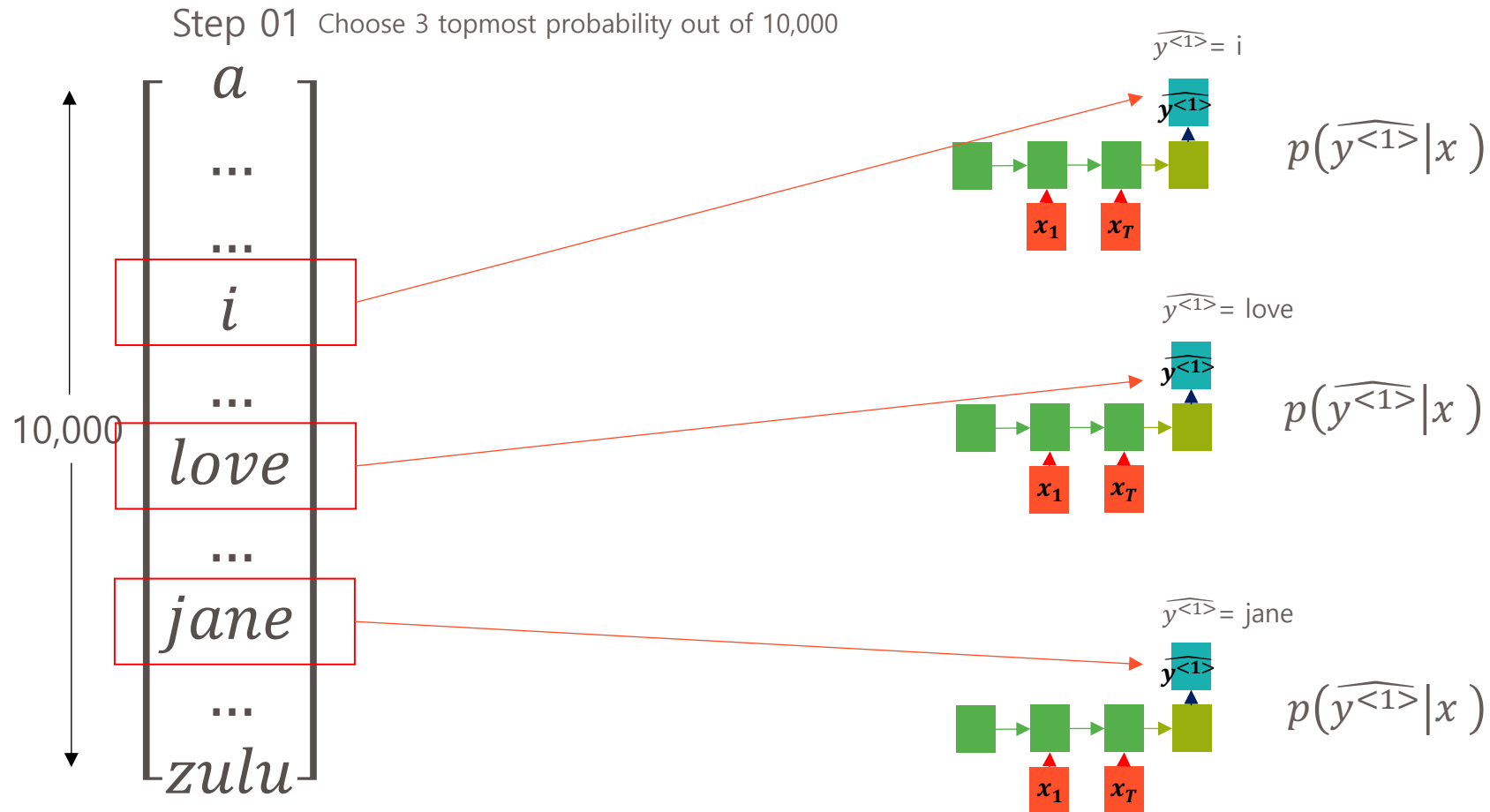
where  $\mathcal{S}$  is training set

#### ✓ Find most likely translation according to the LSTM with following equation through Beam search decoder with Beam Width "B"

$$\hat{T} = \arg \max_T p(T|S)$$

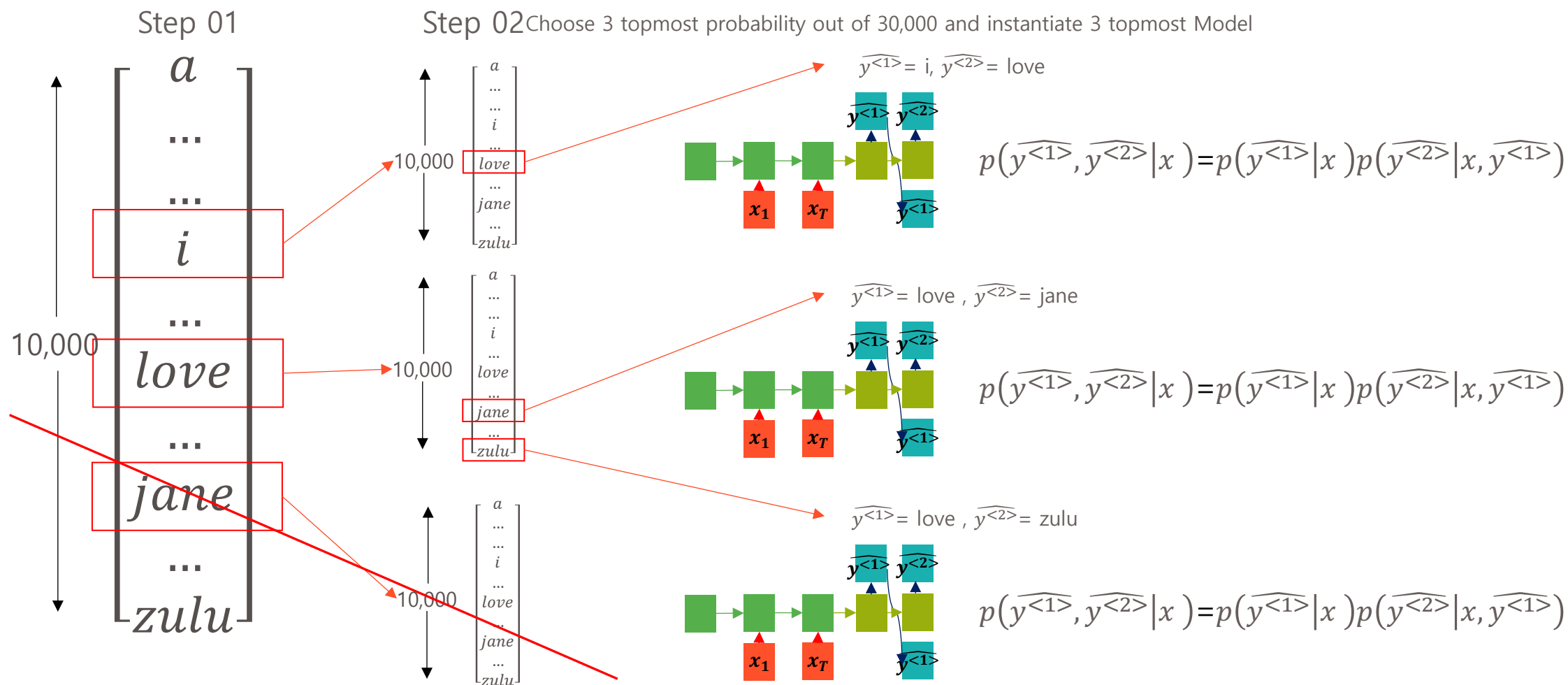
### 3. Beam Search Algorithm - Consider Multiple Probability

- Beam Search with Beam width = 3 (If Beam width = 1, it is Greedy Search Algorithm)



### 3. Beam Search Algorithm

### - Beam Search with Beam width = 3

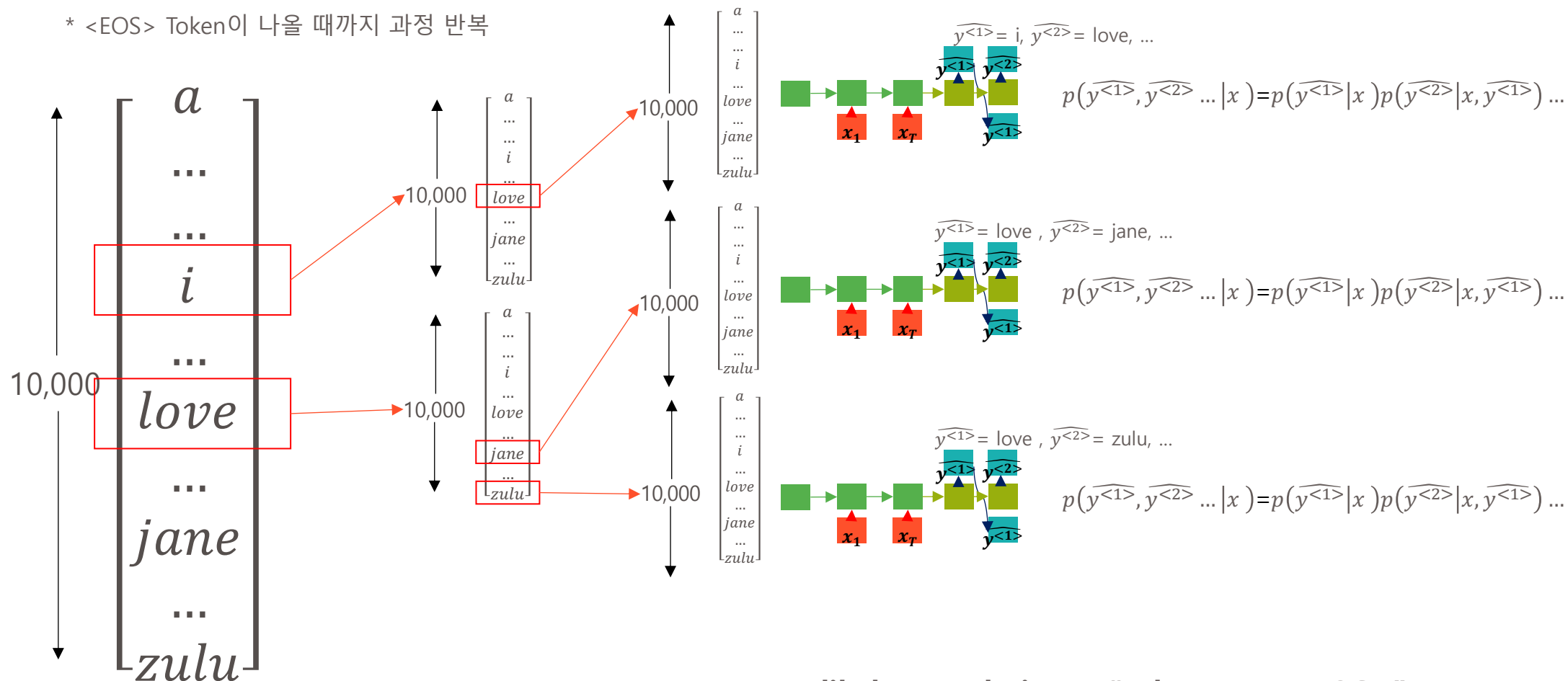




### 3. Beam Search Algorithm

#### - Beam Search with Beam width = 3

\* <EOS> Token이 나올 때까지 과정 반복



Most likely Translation : " I love you <EOS> "

### 3. Beam Search Algorithm

---

- ✓ Objective Function 때문에, Beam Search Algorithm은 짧은 Output 문장을 Return할 확률이 높음

$$\arg \max_{y_t} \prod_{t=1}^{T'} p(y_t | v, y_1, y_2, \dots, y_{t-1}) \rightarrow \arg \max_{y_t} \sum_{t=1}^{T'} \log p(y_t | v, y_1, y_2, \dots, y_{t-1})$$

↓

- ✓ Length Normalization을 통해 긴 Output 문장에 대한 Penalty 완화

$$\frac{1}{(T')^\alpha} \sum_{t=1}^{T'} \log p(y_t | v, y_1, y_2, \dots, y_{t-1}), \text{ where } 0 \leq \alpha \leq 1$$

- ✓ Large Beam Width "B" : Better Result, Slower Computation
- ✓ Small Beam Width "B" : Worse Result, Faster Computation
- ✓ Beam Search Algorithm는 Heuristic-based method이므로, BFS나 DFS처럼 Optimal Solution을 보장하진 못함


## 4. Details – Reversing the Source Sentences

---

- ✓ Source 문장을 역순으로 입력했을 경우, Model Performance가 더 좋았다.  
(Perplexity 5.08 -> 4.7 BLEU Score 25.9 -> 30.6)

$$(a, b, c) \rightarrow (\alpha, \beta, \gamma) > (c, b, a) \rightarrow (\alpha, \beta, \gamma)$$

- ✓ 이론적으로 증명하지는 못했지만, Source 문장을 역순으로 입력하더라도, Output 문장과 대응되는 단어의 평균거리는 변하지 않으나, Source 문장의 첫 몇 단어들에 대한 대응 거리가 짧아져서 Source 문장과 Output 문장 간의 관계를 더 잘 고려하지 않았나라고 연구진은 추측한다.

$$(a, b, c) \rightarrow (\alpha, \beta, \gamma) > (c, b, a) \rightarrow (\alpha, \beta, \gamma)$$


- ✓ 실험 결과, Source 문장의 첫 몇 단어를 넘어 전체 문장에 대해서 성능이 향상됨을 확인하였다 .

## 4. Details – Training details

---

- ✓ Independent LSTM for Encoder and Decoder with 4 layers and 1000 cells
- ✓ Parameters initialized with the uniform distribution between  $-0.08 \sim 0.08$
- ✓ SGD with learning rate 0.7, after 5 epochs, halving the learning rate every half epochs (Total 7.5 epochs)
- ✓ 128 sequence for batch size
- ✓ No suffering from vanishing gradient problem  
But for vanishing exploding problem enhance penalty if  $L_2 \text{ norm} > 5$
- ✓ Made sure all sentences within a minibatch were roughly of the same length

## 4. Experimental Results

---



Performance on WMT'14 English to French Data Set (Baseline System : SMT Model)

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	<b>34.81</b>



Model with SMT System is close to SOTA

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	<b>37.0</b>
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	<b>36.5</b>
Oracle Rescoring of the Baseline 1000-best lists	~45

## 4. Experimental Results

---

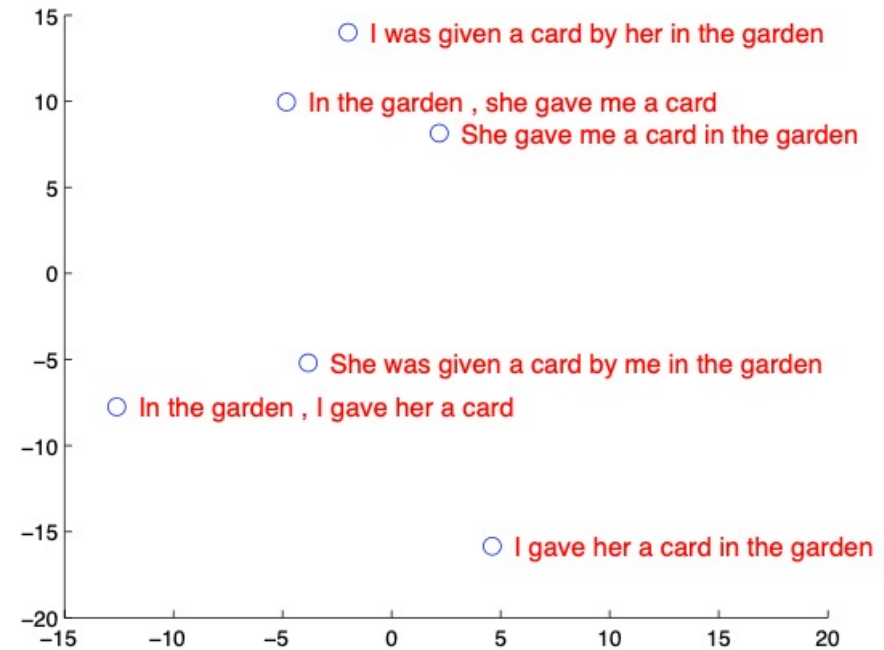
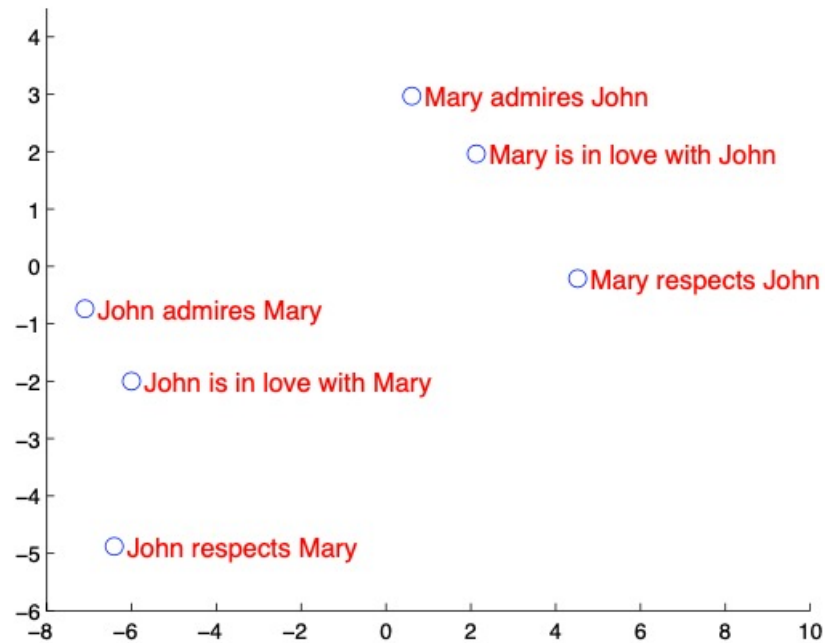


Model works well on long sentences

Type	Sentence
<b>Our model</b>	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
<b>Truth</b>	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
<b>Our model</b>	“ Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air ” , dit UNK .
<b>Truth</b>	“ Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord ” , a déclaré Rosenker .
<b>Our model</b>	Avec la crémation , il y a un “ sentiment de violence contre le corps d' un être cher ” , qui sera “ réduit à une pile de cendres ” en très peu de temps au lieu d' un processus de décomposition “ qui accompagnera les étapes du deuil ” .
<b>Truth</b>	Il y a , avec la crémation , “ une violence faite au corps aimé ” , qui va être “ réduit à un tas de cendres ” en très peu de temps , et non après un processus de décomposition , qui “ accompagnerait les phases du deuil ” .

## 4. Experimental Results

- ✓ 2-dimensional PCA projection of LSTM Hidden State

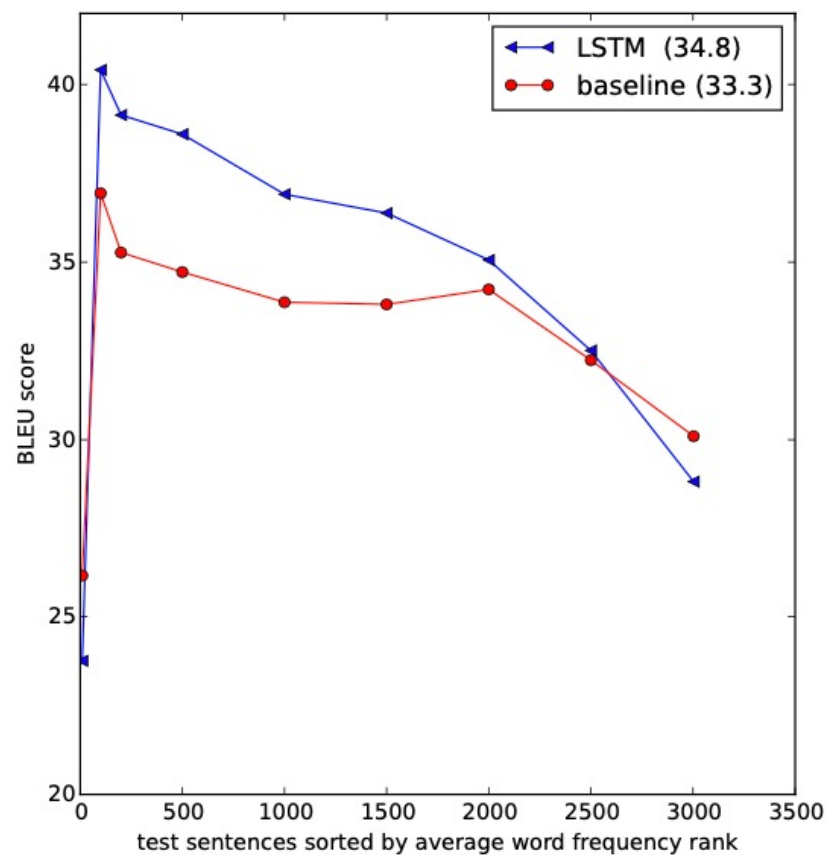
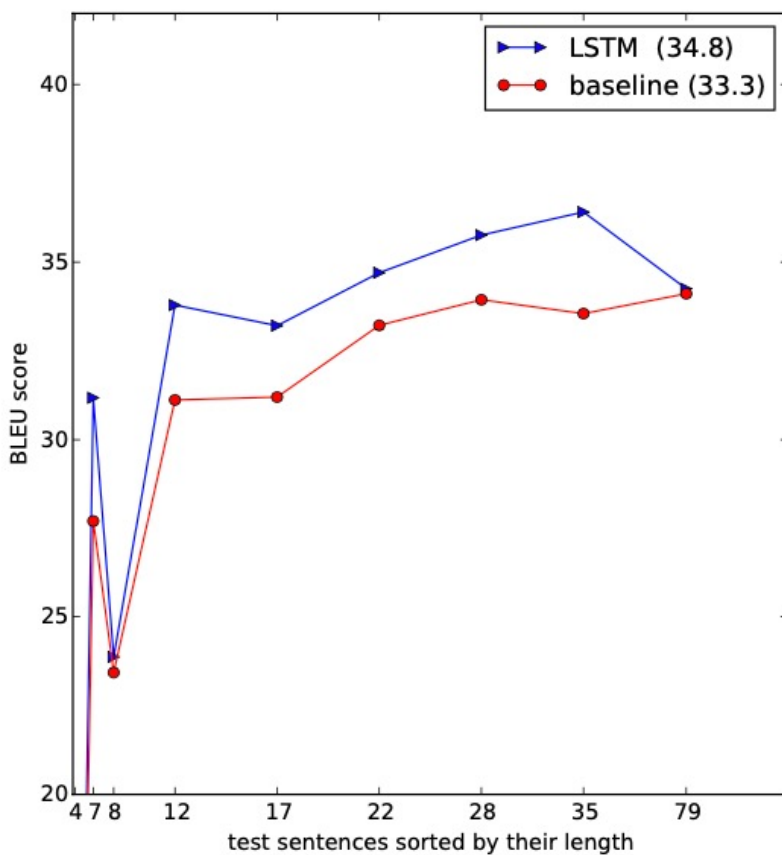


- 단어 순서에 따라 비슷한 문장들 간에 군집을 이루는 것을 확인할 수 있다.

## 4. Experimental Results



Performance of Model as a function of sentence length and average word frequency rank





## 5. Conclusion

---

- ✓ Sequence 2 Sequence 문제 처리에 대한 초석을 마련했다.
- ✓ Source 문장을 역순으로 입력해 Model Performance를 향상시켰다.
- ✓ 예상보다 긴 문장을 잘 번역했다.
- ✓ But we need Attention!

# Reference

---

- ✓ <https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/>
- ✓ <https://nlpinkorean.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/>
- ✓ <https://distill.pub/2016/augmented-rnns/>
- ✓ <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- ✓ <https://www.youtube.com/watch?v=RLWuzLLSIgw>
- ✓ <https://www.quantumdl.com/entry/5%EC%A3%BC%EC%B0%A82-Sequence-to-Sequence-Learning-with-Neural-Networks?category=691904>