<Paper Review>

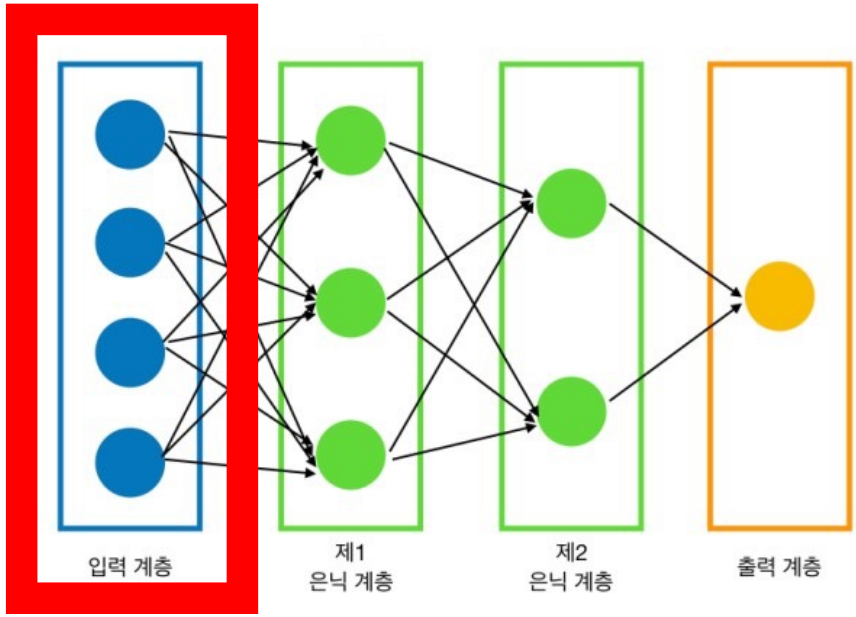
Sequence to Sequence Learning with Neural Networks

발표자: 최희정

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

DNN(Deep Neural Network)

→ Hidden Layer가 2개 이상 포함된 Neural Network



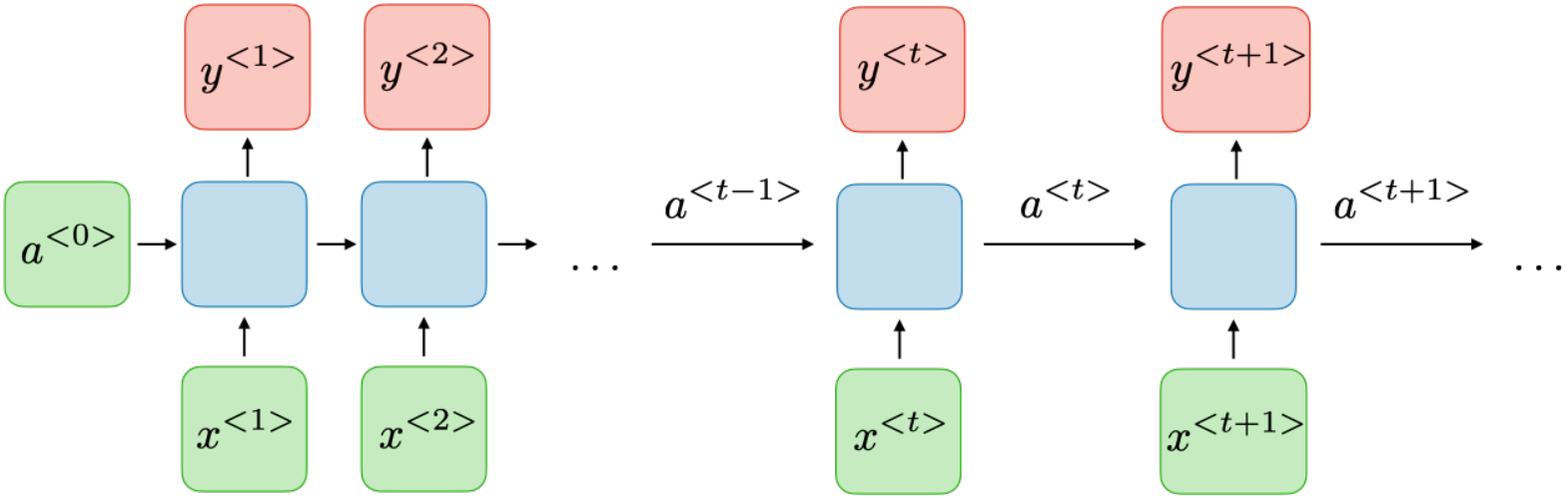
Fixed Dimensionality

DNN Structure – 김성진, 코딩셰프의 3분 딥러닝, 케라스맛, 한빛미디어

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

Feed Forward Neural Network (FFNL): hidden layer를 통과한 값은 output layer로만 들어감

Recurrent Neural Network (RNN): hidden layer를 통과한 값은 hidden layer와 output layer로 들어감

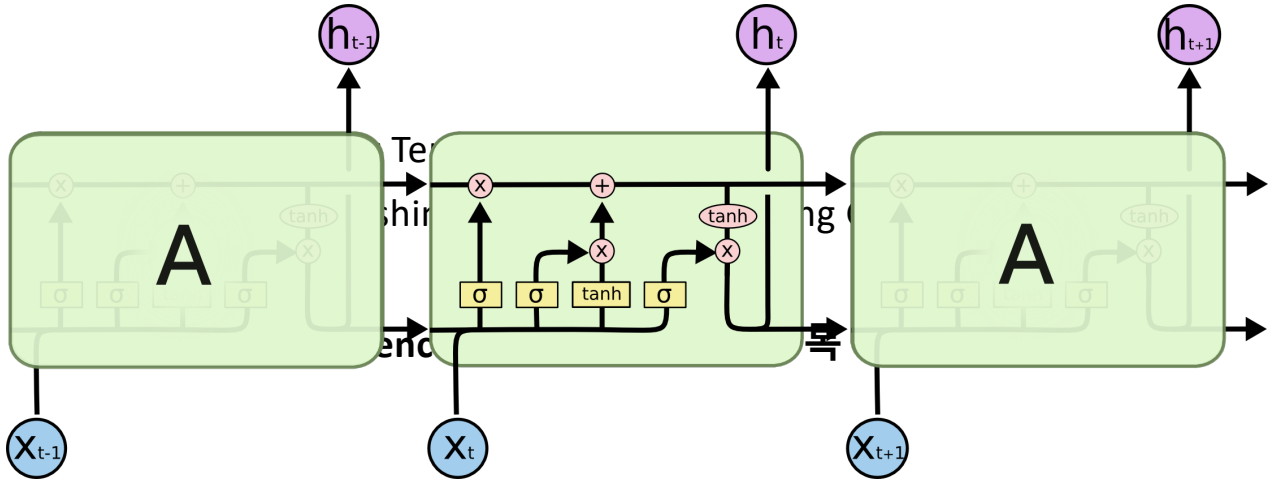


$$h_t = sigmoid\left(W^{hx}x_t + W^{hh}h_{t-1}\right)$$

$$y_t = W^{yh}h_t$$

https://stanford.edu/~shervine/l/ko/teaching/cs-230/cheatsheet-recurrent-neural-networks

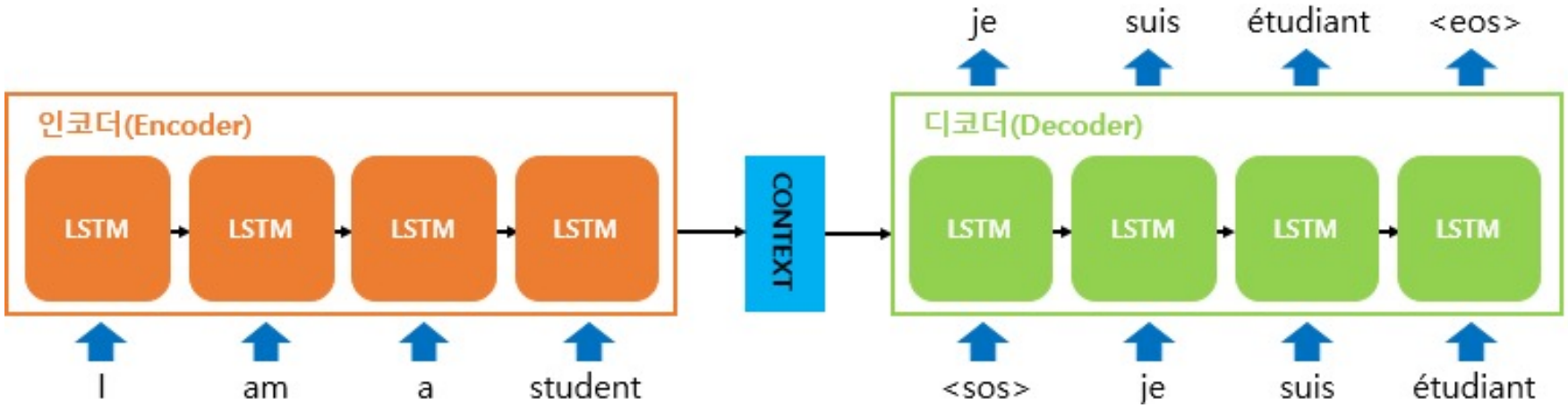
I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.



LSTM이 이러한 문제점을 해결함

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

LSTM을 활용한 모델 구조



Encoder: input sentence를 context vector로 만듦

Decoder: encoder가 만든 context vector를 target sequence로 만듦

https://wikidocs.net/24996

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

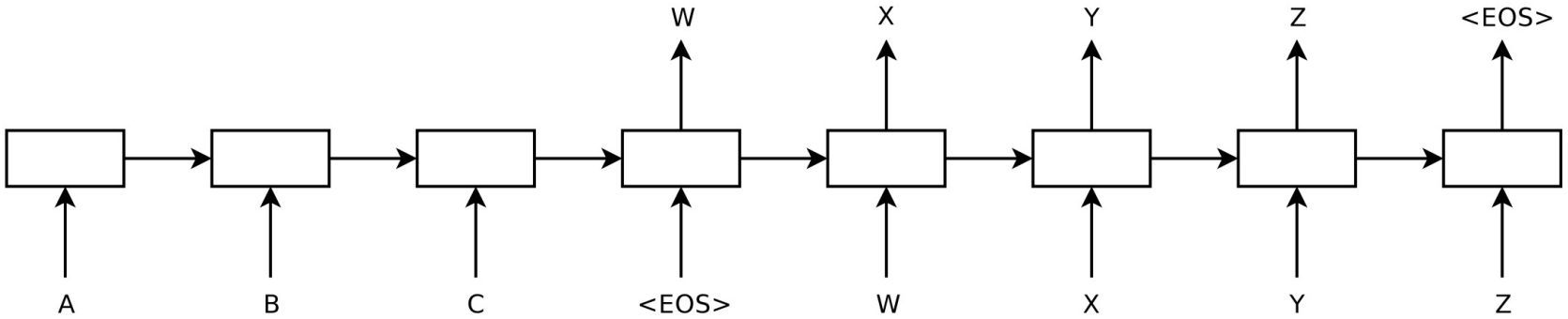
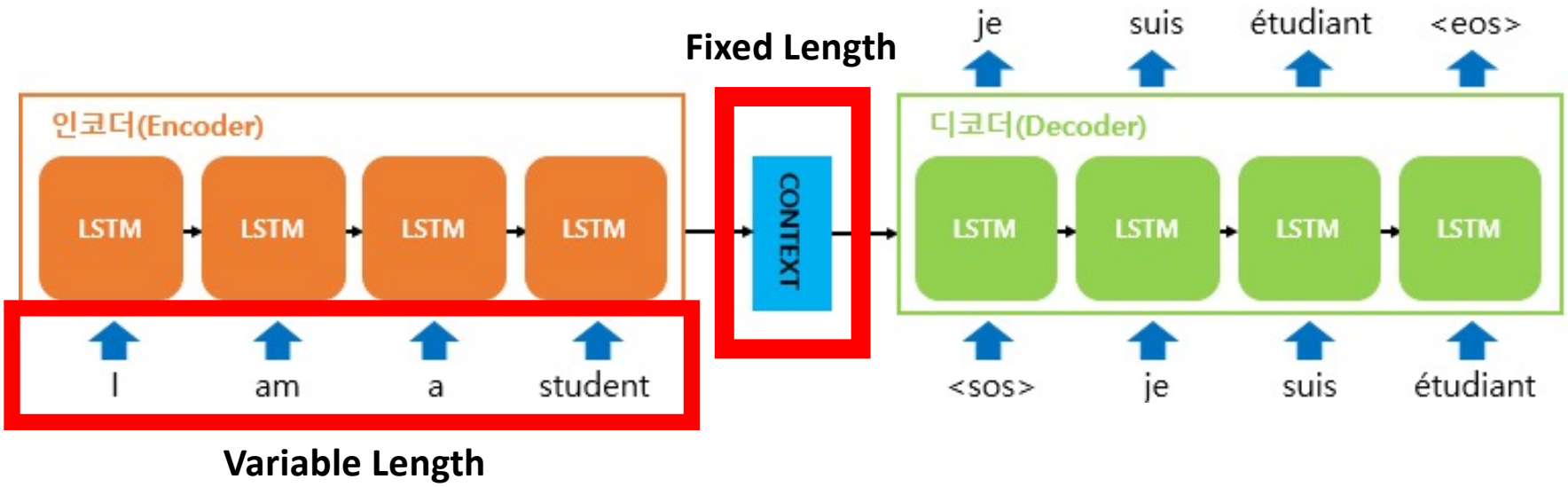


Figure 1. input 시퀀스 “ABC”를 읽고, output 시퀀스 “WXYZ”를 만들어낸다.
end-of-sentence token(<EOS>)를 출력한 다음에 예측을 멈춘다.
LSTM은 input 시퀀스를 역으로 읽음에 주의해야 한다.
→이렇게 해서 데이터의 최적화 문제를 훨씬 쉽게 만드는, 많은 short term dependency를 만들어낸다.

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.



https://wikidocs.net/24996

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

Contribution

"First, we used **two different LSTMs**: **one for the input sequence and another for the output sequence**"

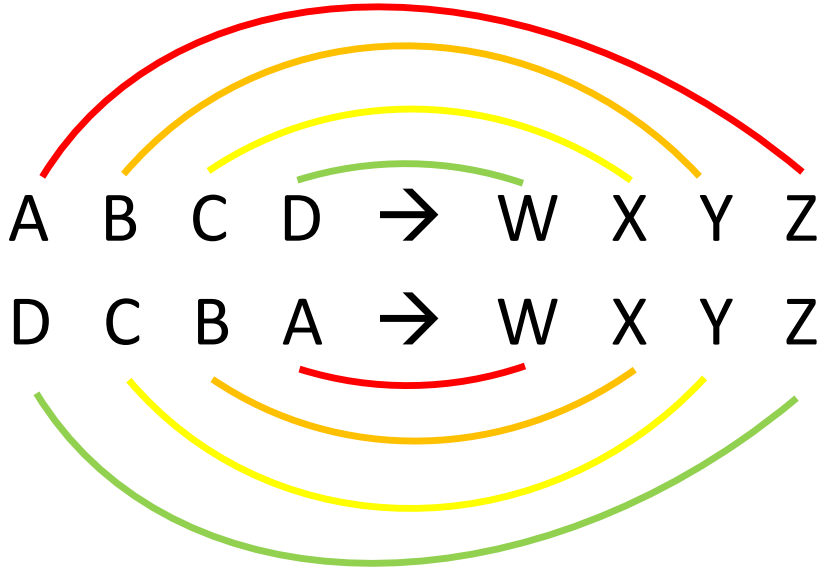
→ 모델의 파라미터 수가 증가하므로 더 많은 양의 학습 가능

"Second, we found that **deep LSTMs** significantly outperformed shallow LSTMs, so we chose **an LSTM with four layers**."

"Third, we found it extremely valuable to **reverse the order of the words of the input sentence**."

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

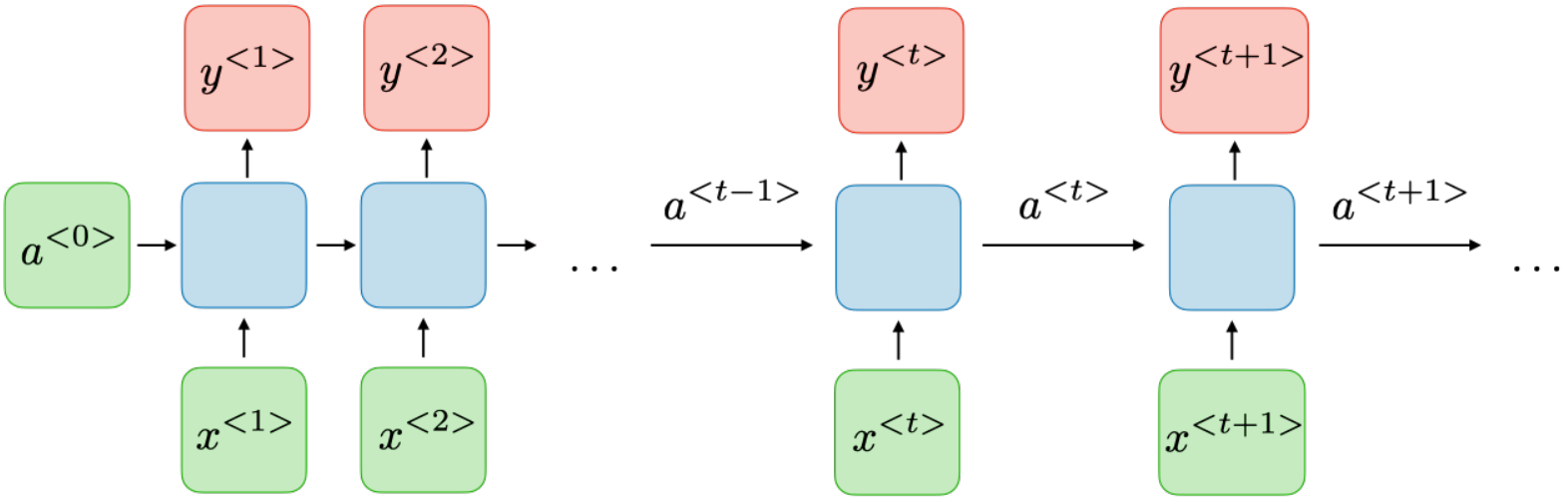
"Third, we found it extremely valuable to **reverse the order of the words of the input sentence**."



A와 W 사이의 거리는 가까워졌지만, 단어 쌍 사이의 거리 평균은 동일하다.

Why?

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.



앞에 나온 데이터가 더 중요하다.

→ Source sentence의 앞쪽 데이터를 뒤쪽으로 옮겨주면 (reverse)

→ Source sentence의 중요한 데이터와 target word 사이의 거리가 가까워진다.

https://stanford.edu/~shervine/l/ko/teaching/cs-230/cheatsheet-recurrent-neural-networks

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

주어진 source sentence S에 대해, 맞는 번역 T의 log probability를 최대화

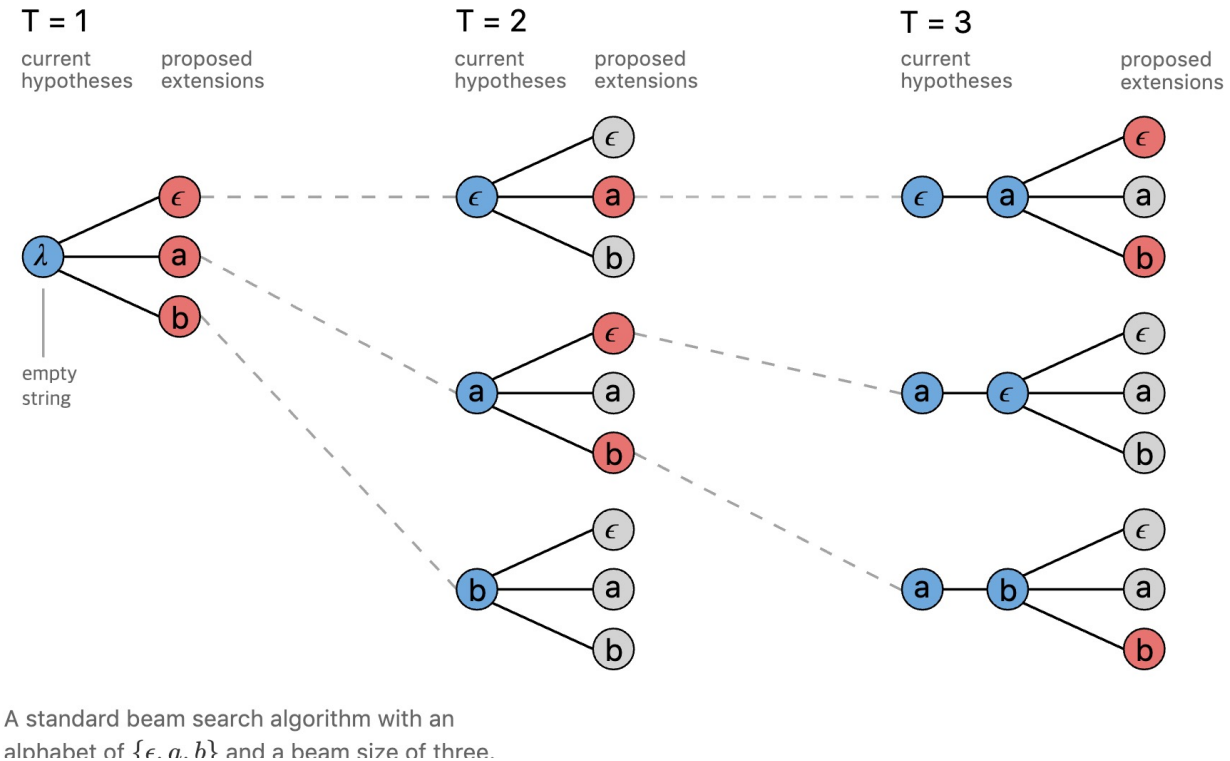
$$1/|S| \sum_{(T,S)\in S} \log p(T|S)$$

일단 학습이 한 번 끝나면, LSTM에 따라 가장 확률이 높은 translation을 찾는다.

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

Left-to-Right Beam Search Decoder

Beam size에서 가장 큰 확률을 계속해서 찾아나간다.



A standard beam search algorithm with an alphabet of $\{\epsilon, a, b\}$ and a beam size of three.

https://ratsgo.github.io/speechbook/docs/neuralam/ctc

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

BLUE score(Bilingual Evaluation Understudy Score)

$$BLUE = \min(1, \frac{output\ length(\text{예측 문장})}{reference\ length(\text{실제 문장})})(\prod_{i=1}^{4} precision)^{\frac{1}{4}}$$

- 1-gram precision: $\frac{\text{일치하는}1-gram\text{의 수(예측된 }sentence\text{중에서)}}{\text{모든}1-gram\text{쌍 (예측된 }sentence\text{중에서)}} = \frac{10}{14}$
- 2-gram precision: $\frac{\text{일치하는}2-gram\text{의 수(예측된 }sentence\text{중에서)}}{\text{모든}2-gram\text{쌍 (예측된 }sentence\text{중에서)}} = \frac{5}{13}$
- 3-gram precision: $\frac{\text{일치하는}3-gram\text{의 수(예측된 }sentence\text{중에서)}}{\text{모든}3-gram\text{쌍 (예측된 }sentence\text{중에서)}} = \frac{2}{12}$
- 4-gram precision: $\frac{\text{일치하는}4-gram\text{의 수(예측된 }sentence\text{중에서)}}{\text{모든}4-gram\text{쌍 (예측된 }sentence\text{중에서)}} = \frac{1}{11}$

$$(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} = (\frac{10}{14} \times \frac{5}{13} \times \frac{2}{12} \times \frac{1}{11})^{\frac{1}{4}}$$

https://donghwa-kim.github.io/BLEU.html

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

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**

- WMT'14 English to French test set (ntst14)
- Beam size가 2인 5개의 LSTM의 앙상블이 beam size가 12인 1개의 LSTM보다 성능이 더 좋음에 주목하라.

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
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	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

- WMT'14 English to French test set (ntst14)
- Statistical Machine Learning(SMT) system과 Neural Network를 함께 쓴 방법으로 BLUE score 측정

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.

Type	Sentence
Our model	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 .
Truth	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 .
Our model	" 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 .
Truth	" 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 .
Our model	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 " .
Truth	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 " .

I Sutskever. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 2014.