

# Sequence to Sequence Learning with Neural Networks

---

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le.  
*Advances in neural information processing systems* 27 (2014).

2024.02.19  
유하영



# JAPPU!

자연어처리 논문을 읽는 스터디입니다!


readme.md



## Basic-Course for NLP

- 자연어 처리 스터디의 발표 ppt 및 논문 구현 자료를 업로드하는 공간입니다.

### Study Info

- Goal : 논문을 통한 자연어처리 이해 및 기술 구현
- Participants : 유하영, 황현태
- Start Date : 2024.01.02
- Meeting Date : 매주 일요일 08:00.PM
- Location :  Discord

- **04 : Sequence to Sequence Learning with Neural Networks**

[Paper](#), [Presentation](#)

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014.

- Keywords : seq2seq
- Date : 2024.01.18
- Presenter : 유하영

- 이전 연구의 한계

### DNN

레이블이 지정된 **training set**에서는 유용하지만,  
sequence와 같이 **가변적인 길이의 데이터를 처리하는 데에는 제한적**  
(입력 시퀀스와 출력 시퀀스의 길이가 다른 경우.  
ex) speech recognition, machine translation)

### RNN

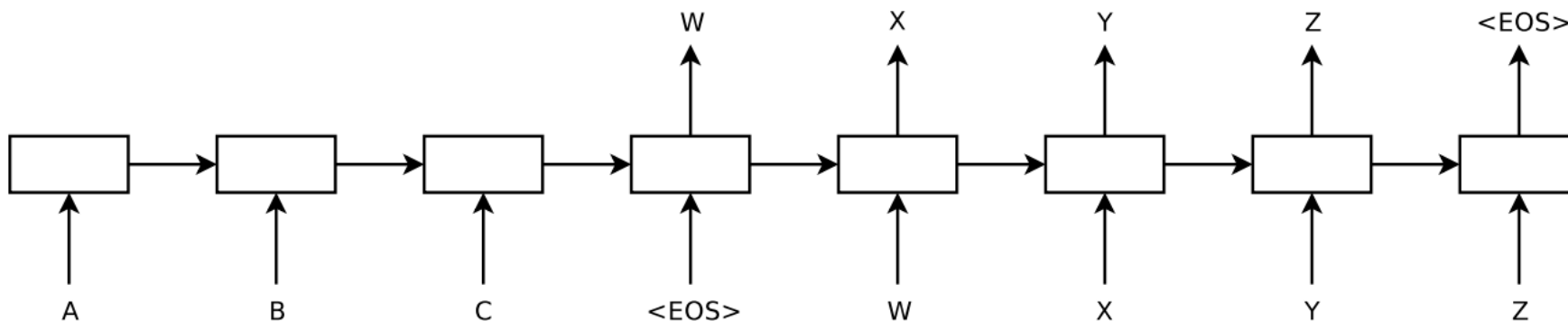
시퀀스를 처리할 수 있는 DNN의 **generalization version**  
timestep  $t$ 에서 RNN은 입력  $x_t$ 와 이전 timestep의 hidden state  $h_{t-1}$ 을 함께 고려하여  $y_t$ 를 계산

이 역시 **가변적인 길이의 데이터 처리에는 제한적**

- RNN based seq2seq

두 개의 RNN을 연결하는 Encoder-Decoder 구조로 **가변길이의 입,출력 문제 해결**이 가능하다고 봄

- encoder를 통하여 입력 시퀀스  $(x_1, \dots, x_T)$ 를 고정된 길이의 벡터  $C$ (context vector, 입력 데이터의 요약)로 표현하고 모델의 다음 부분인 decoder로 전달
- decoder에서는 고정된 길이의 벡터  $C$ 를 출력 시퀀스  $(y_1, \dots, y_{T'})$ 로 mapping한다.
- 각 입출력 시퀀스의 종료는 보통  $\langle \text{EOS} \rangle$  토큰을 기준으로 정해짐



- RNN based seq2seq

Encoder-Decoder 구조는 일반적인 언어모델을 사용하여 다음과 같은 조건부 확률로 모델링할 수 있다.

(우변은 모든 vocabular에 대한 softmax probability로서 표현된다.)

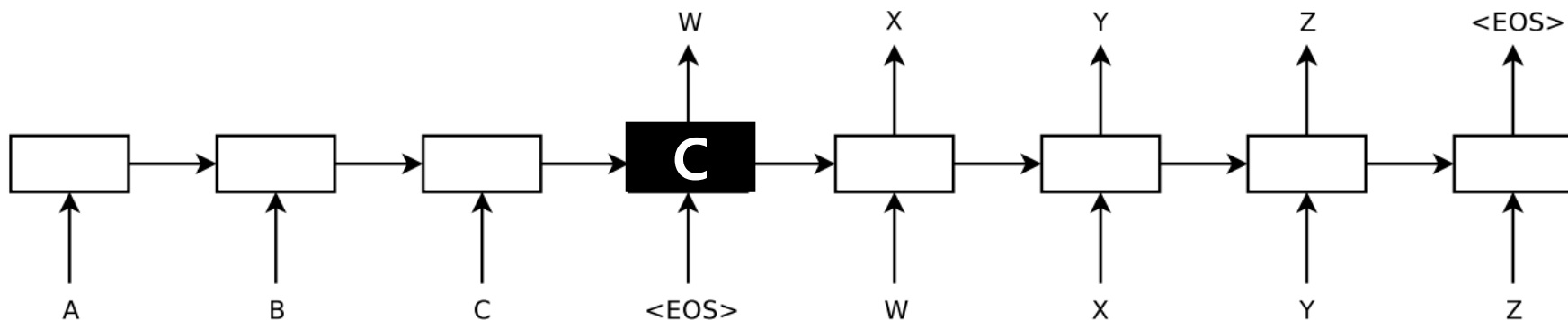
$$(x_1, \dots, x_T) \rightarrow C \rightarrow (y_1, \dots, y_{T'})$$

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

- 본 논문은 기존 RNN based seq2seq 모델의 성능을 다음 세 가지 방법으로 개선한다.
- LSTM based seq2seq
  1. 입력과 출력에 대한 서로 다른 2개의 LSTM을 사용한다.
  2. Deep LSTM을 사용하여 Shallow LSTM 보다 좋은 성능을 제공한다.
  3. 입력의 순서를 뒤집어서 제공한다. (reversed order Input sequence)

# 1. LSTM 사용

- LSTM은 RNN의 한계인 **long term dependency**을 개선할 수 있다.  
시퀀스가 길어져 **timestep**의 간격이 멀어질수록 **input sequence**의 정보를 모델이 제대로 반영하지 못하는 점





## 2. Deep LSTM

- single layer 대신 Deep LSTM 사용
- 본 논문에서는 layer 4 deep LSTM을 사용하였다.

### 3. Reversed order Input sequence

- Input sequence의 순서를 거꾸로 하여 LSTM Encoder의 새로운 입력으로 사용

A B C -> D E F

C B A -> D E F

- 입력과 출력 시퀀스 간의 평균거리를 유지하면서, 시퀀스의 각 요소 간의 최소 거리는 훨씬 줄어든다.
- 특히, 입력 시퀀스와 출력 시퀀스 간의 직접적인 대응 관계가 있는 경우(예: 기계번역)에 유용함

Ex) I am happy → 나는 행복하다  
happy am I → 나는 행복하다

- 즉, 시퀀스의 시작 부분과 끝 부분 사이의 거리를 줄이는 것이 이 기법의 핵심

## ??? ‘양방향 LSTM’과 ‘Reversed order Input sequence’와의 차이

- Reversed order Input sequence는 단순히 입력 시퀀스의 순서를 바꾸는 것
- 즉, 여전히 단방향 LSTM을 사용함

**A B C → D E F    ->    C B A → D E F**

- 반면, Bi-LSTM은 각 시점에서 입력 시퀀스의 양방향(정방향, 역방향) 정보를 동시에 고려

**A B C → D E F    A 시점 – 정방향 LSTM: “A”까지의 정보만 고려**  
– 역방향 LSTM: “C B A”의 정보를 고려  
(두 정보는 A 시점에서 결합)

**B 시점 – 정방향 LSTM: “A B”까지의 정보를 고려**  
– 역방향 LSTM: “C B”까지의 정보를 고려  
(두 정보는 B 시점에서 결합)

**C 시점 – 정방향 LSTM: “A B C”까지의 정보를 고려**  
– 역방향 LSTM: “C”까지의 정보를 고려  
(두 정보는 C 시점에서 결합)

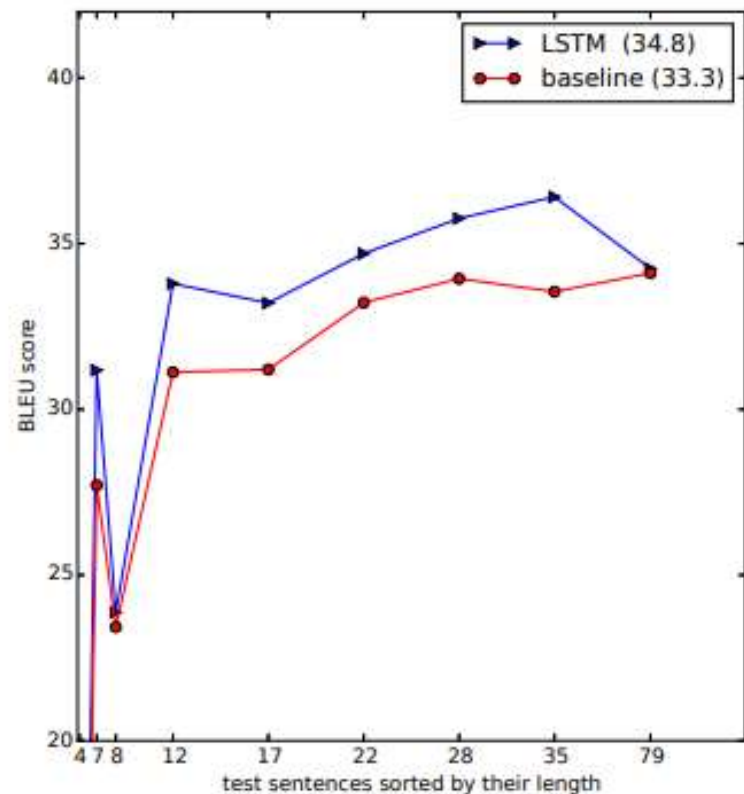
# Experimental Results

- 각 LSTM layer 마다 서로 다른 초기화 과정을 거치고 minibatch를 랜덤하게 섞어 학습한 LSTM 앙상블로부터 best result를 얻었다.

English-French 번역작

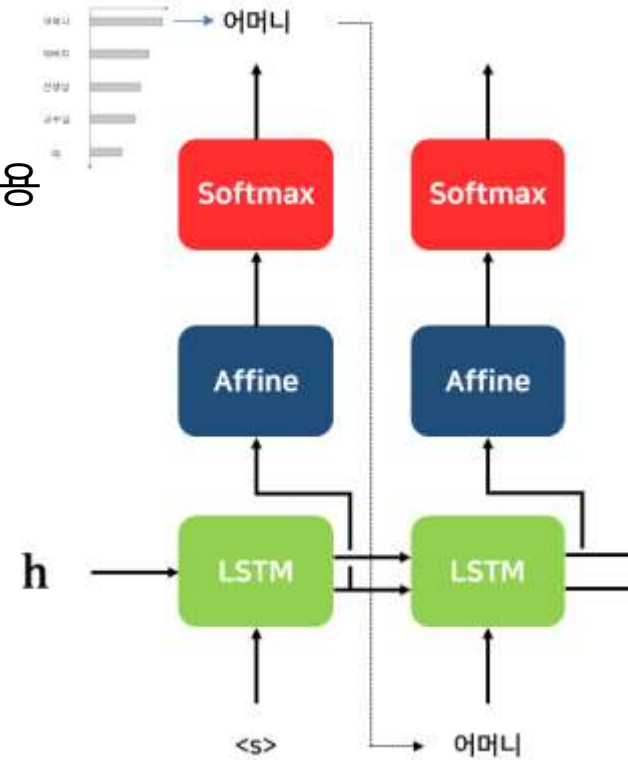
표

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>



### \*\* Greedy Decoding

해당 시점에서 **가장 확률이 높은 후보를 선택**하는 것  
한 번이라도 틀린 예측이 나오게 되면 치명적인 문제로 작용



### \*\* Beam search

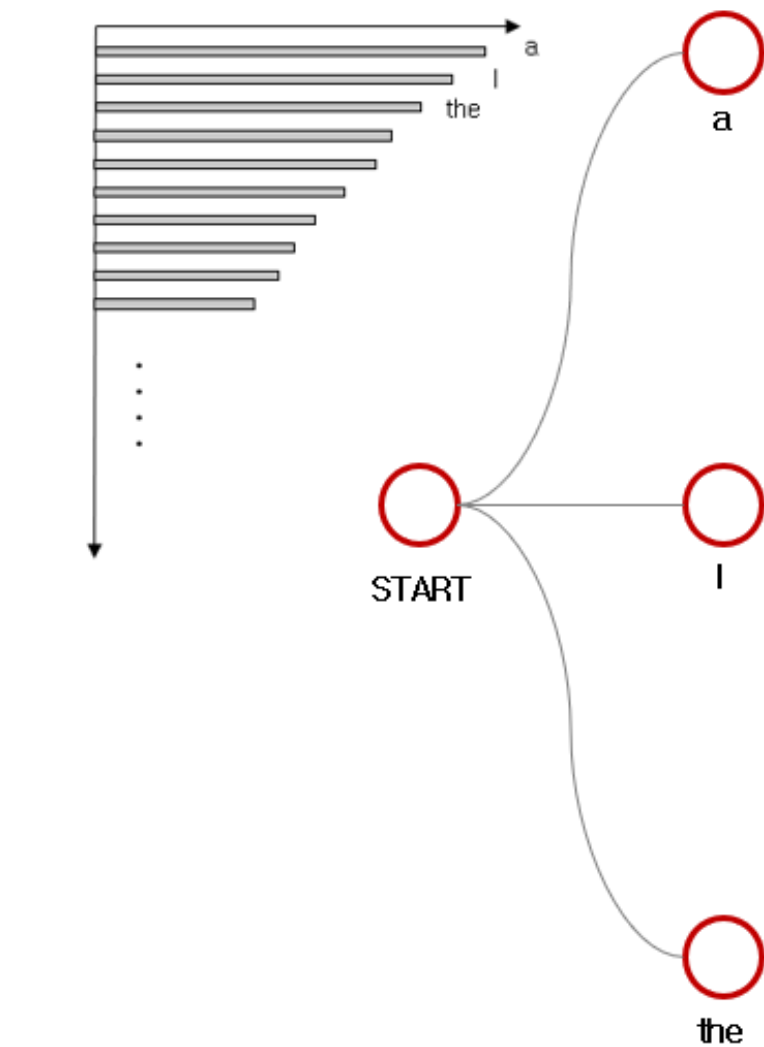
- Beam search : 번역이나 텍스트 생성 과정에서 **최적의 출력 시퀀스를 찾기 위해 사용**
- Beam size : 탐색과정에서 한 번에 고려하는 후보 시퀀스의 수
  - 최상위 후보만을 유지하면서 **가장 가능성이 높은 출력 시퀀스를 찾아냄**

Beam size가 크면 더 많은 후보 시퀀스를 고려할 수 있어. 탐색이 보다 포괄적으로 됨

## \*\* Beam search

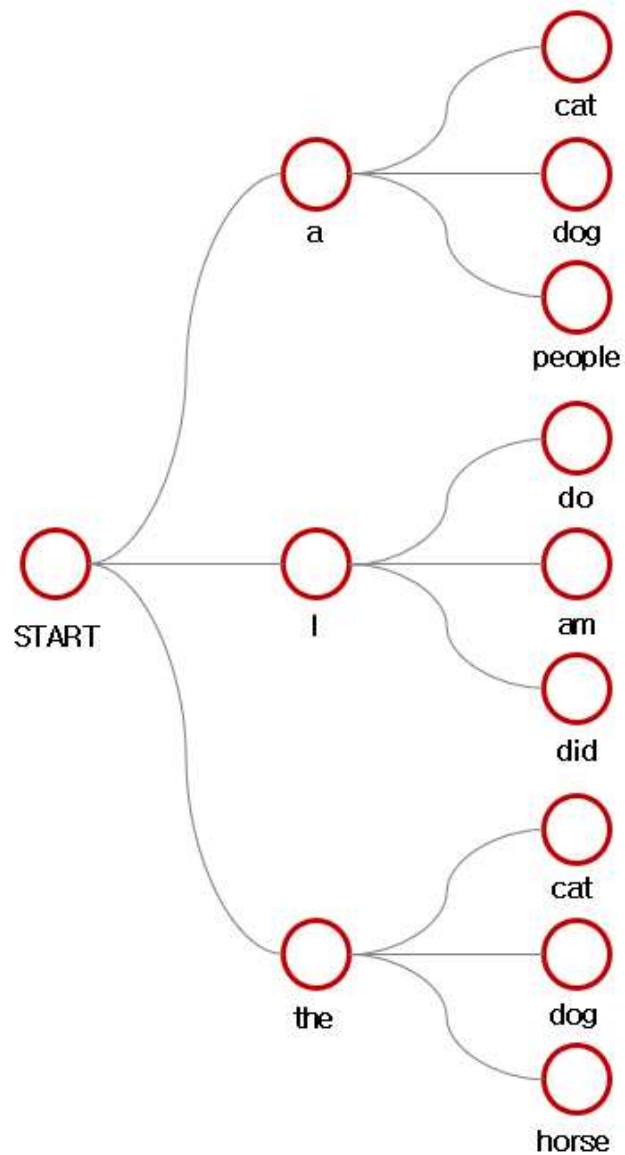


Step 1

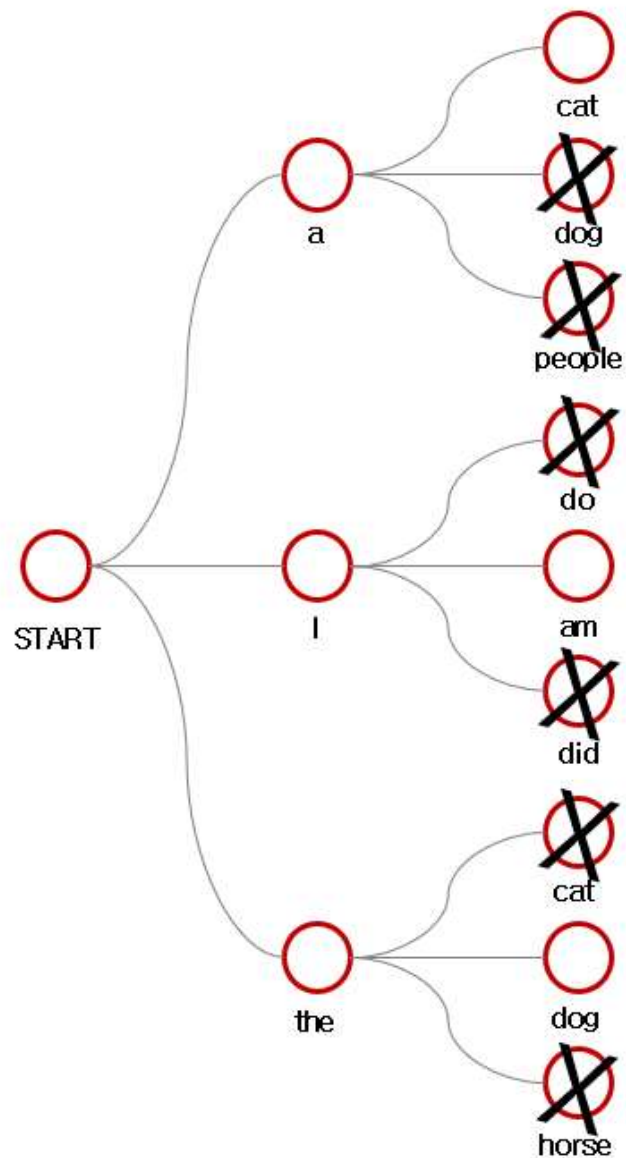


# Experimental Results

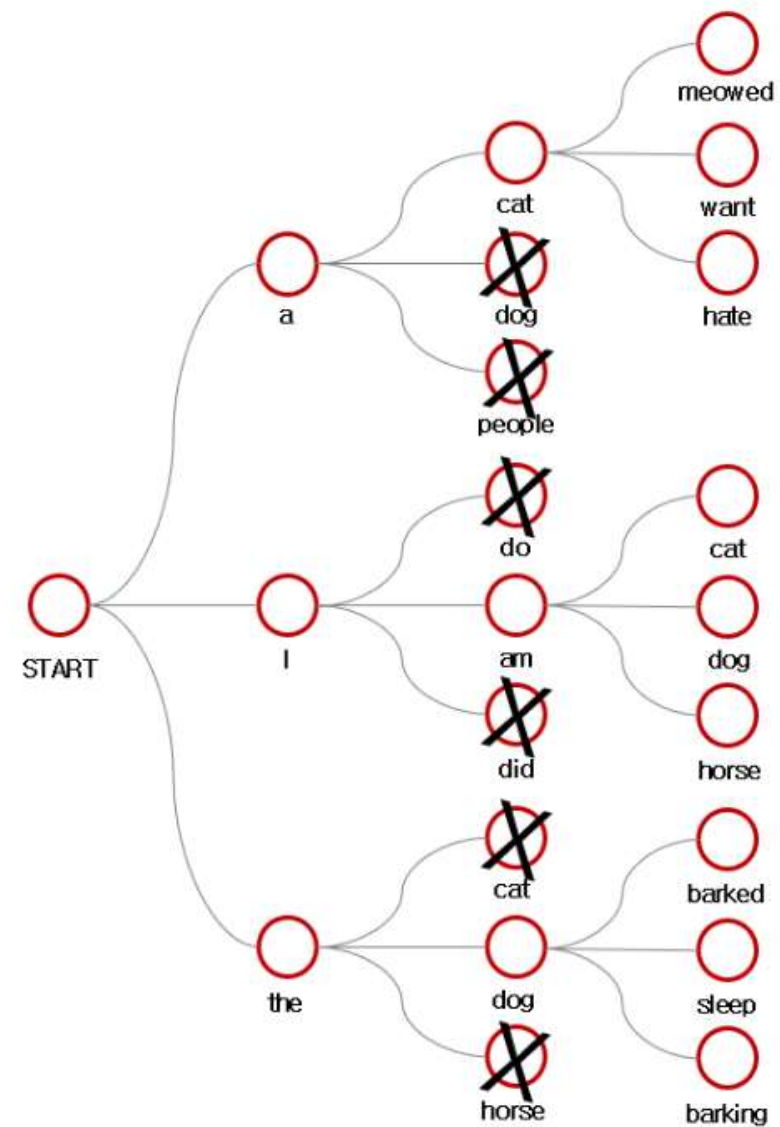
Step 2



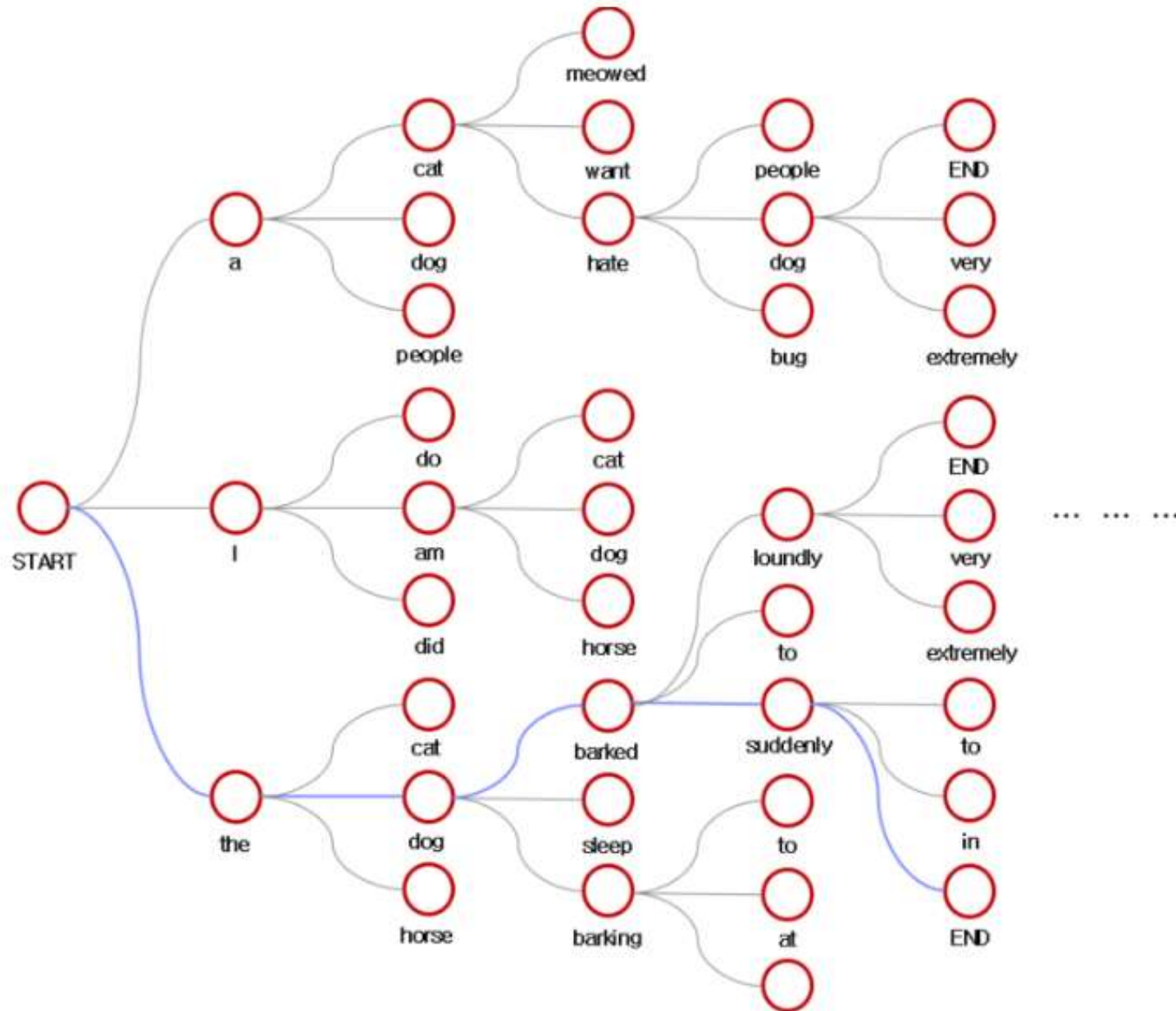
Step 3



Step 4



# Experimental Results



## CANDIDATES

1. The dog barked suddenly
2. A cat hate dog
3. The dog barking to me



- 앙상블 모델 사용

앙상블 모델에서 각각의 LSTM은 같은 입력 데이터에 대해 독립적으로 번역을 수행하고, 이후 다양한 방법으로 결합

- BLEU Score

(Bilingual Evaluation Understudy Score)

: 기계 번역 결과와 실제 인간의 번역과 얼마나 유사한지를 측정, 번역의 자연스러움과 정확성을 수치로 나타냄

- 언어에 구애받지 않고 사용가능

- BLEU(Bilingual Evaluation Understudy)

기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지 비교하여 번역에 대한 성능을 측정하는 방법

- 대규모 **deep LSTM**이 대규모 기계 번역 작업에 있어 무제한의 어휘록을 가진 **standard SMT(Statistical Machine Translation)** 기반 시스템 보다 더 높은 성능을 발휘함
- **source sentences**의 단어를 역순으로 배치하는 것이 더 높은 성능을 보임
- **LSTM**은 매우 긴 문장도 거의 올바르게 번역하였다.  
(**but**, 아주 긴 문장을 역순으로 배치하여 학습할 때는 아직 한계가 보임)

- **SMT(Statistical Machine Translation)**

: 통계적 기법을 사용하여 한 언어에서 다른 언어로 텍스트를 번역하는  
기계 번역의 한 형태