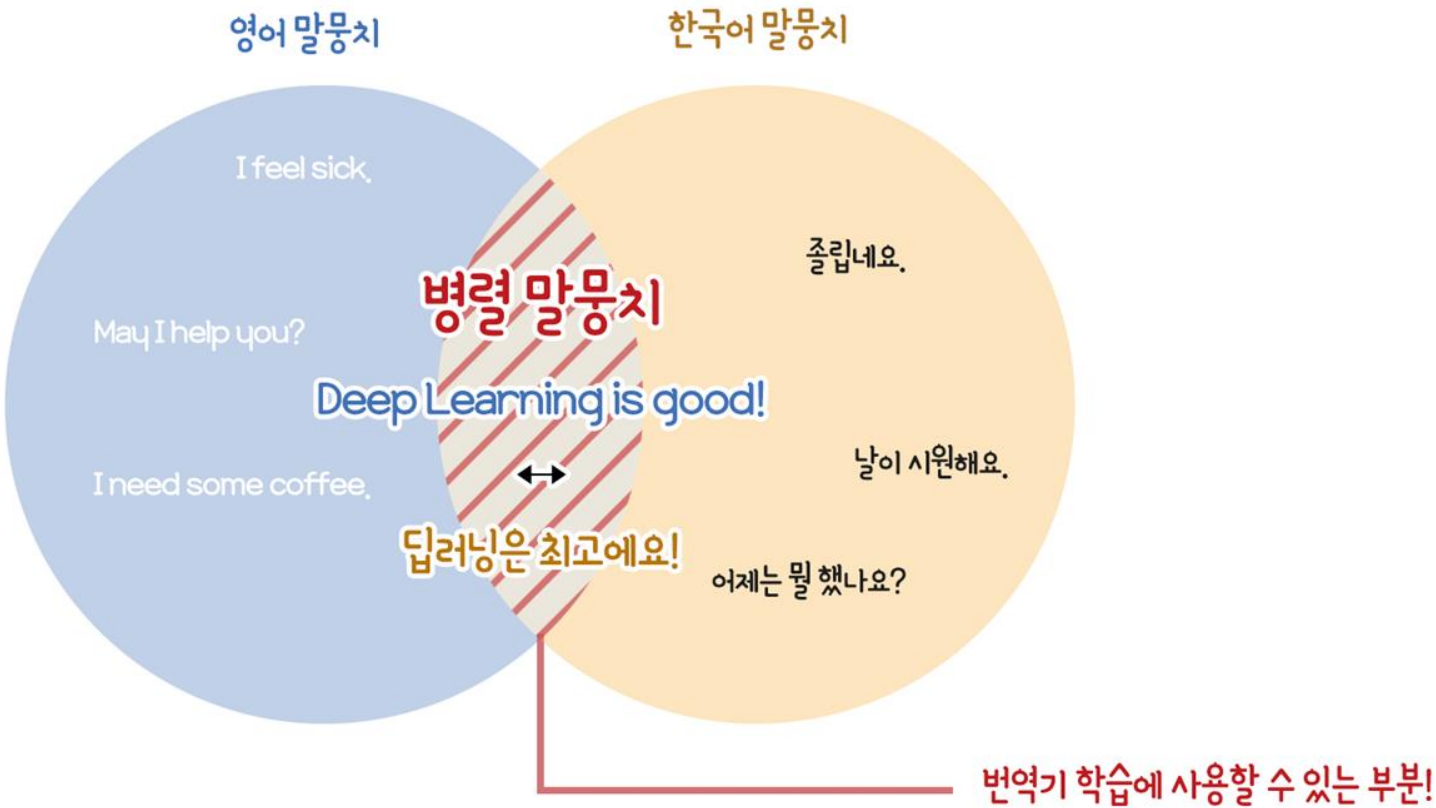
BACK TRANSLATION

: Understanding Back-translation at Scale

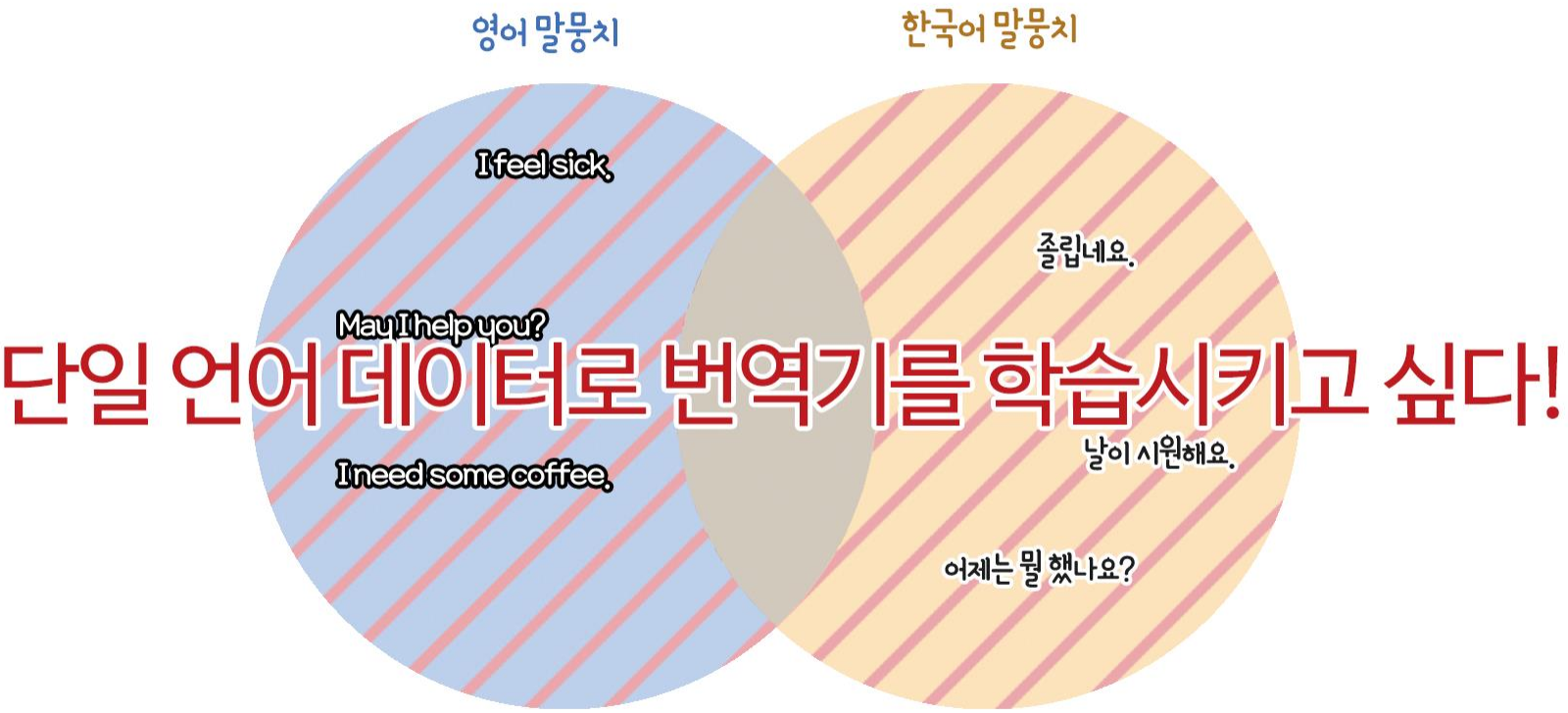
2020. 01. 10

문성원, 윤동준, 최재규, 주현진

Intro


영어 말뭉치
한국어 말뭉치
I feel sick.
May I help you?
I need some coffee.
병렬 말뭉치
Deep Learning is good!
딥러닝은 최고에요!
졸립네요.
날이 시원해요.
어제는 뭘 했나요?
번역기 학습에 사용할 수 있는 부분!

Intro


영어 말뭉치
한국어 말뭉치
I feel sick.
Mau I help you?
I need some coffee.
졸립네요.
날이 시원해요.
어제는 뭘 했나요?
단일 언어 데이터로 번역기를 학습시키고 싶다!

기계 번역

Machine Translation

사전 준비

기계 번역

Machine Translation

규칙 기반

Ex) *I'm looking for someone.*

→ i/PRP be/VBP look/VBG
for/IN someone/NN

→ 나/NP 는/FX 찾/VB
고있/IN 누군가/NN

→ *나는 누군가를 찾고 있습니다.*

사전 준비

기계 번역

Machine Translation

규칙 기반

Ex) *I'm looking for someone.*

→ i/PRP be/VBP look/VBG
for/IN someone/NN

→ 나/NP 는/FX 찾/VB
고있/IN 누군가/NN

→ *나는 누군가를 찾고 있습니다.*

통계 기반

Ex) *I'm looking for someone.*

→ 'I' 가 나오면 **대체로** '나'로 시작

→ '나' 다음엔 '는'이 **많이** 나오고...

→ 끝 단어가 **주로** 지금 나오니
'나는 누군가'...

→ (중략)

→ *나는 누군가를 찾고 있습니다.*

사전 준비

기계 번역

Machine Translation

규칙 기반

Ex) *I'm looking for someone.*

→ i/PRP be/VBP look/VBG
for/IN someone/NN

→ 나/NP 는/FX 찾/VB
고있/IN 누군가/NN

→ *나는 누군가를 찾고 있습니다.*

통계 기반

Ex) *I'm looking for someone.*

→ 'I' 가 나오면 **대체로** '나'로 시작

→ '나' 다음엔 '는'이 **많이** 나오고...

→ 끝 단어가 **주로** 지금 나오니
'나는 누군가'...

→ (중략)

→ *나는 누군가를 찾고 있습니다.*

신경망 기반

I'm looking for someone.

↓

(엄청난 번역 모델)

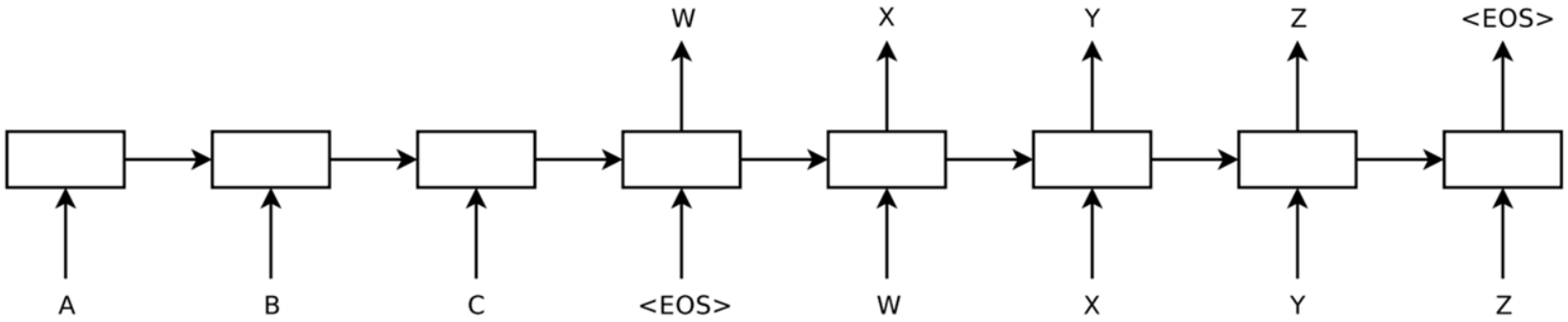
↓

나는 누군가를 찾고 있습니다.

사전 준비

신경망 모델

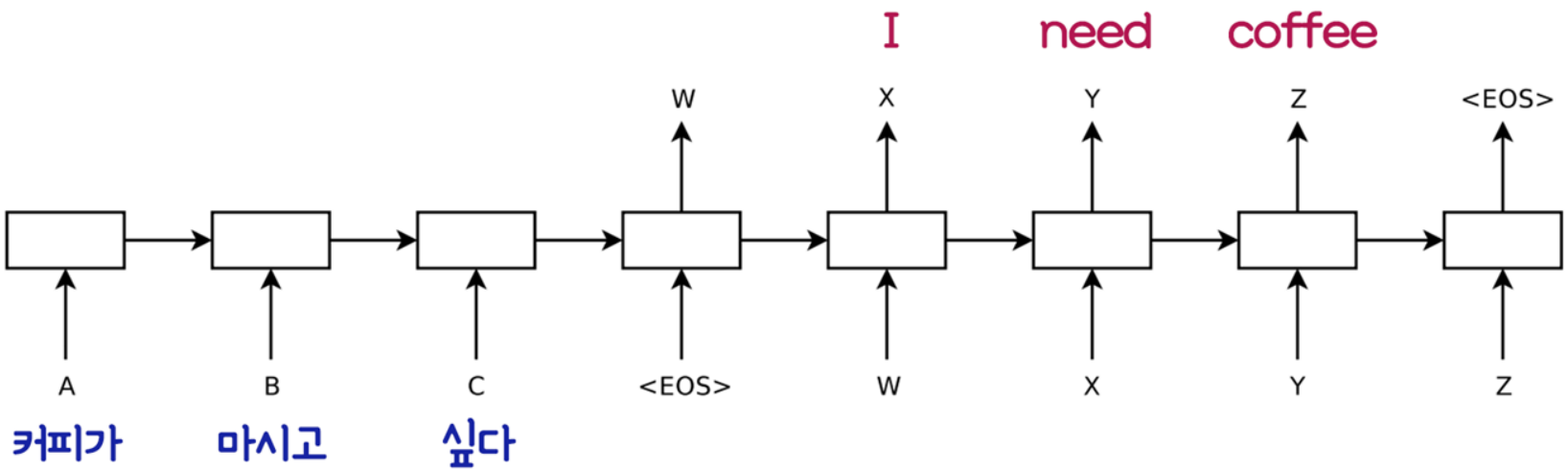
Vanila RNN



사전 준비

신경망 모델

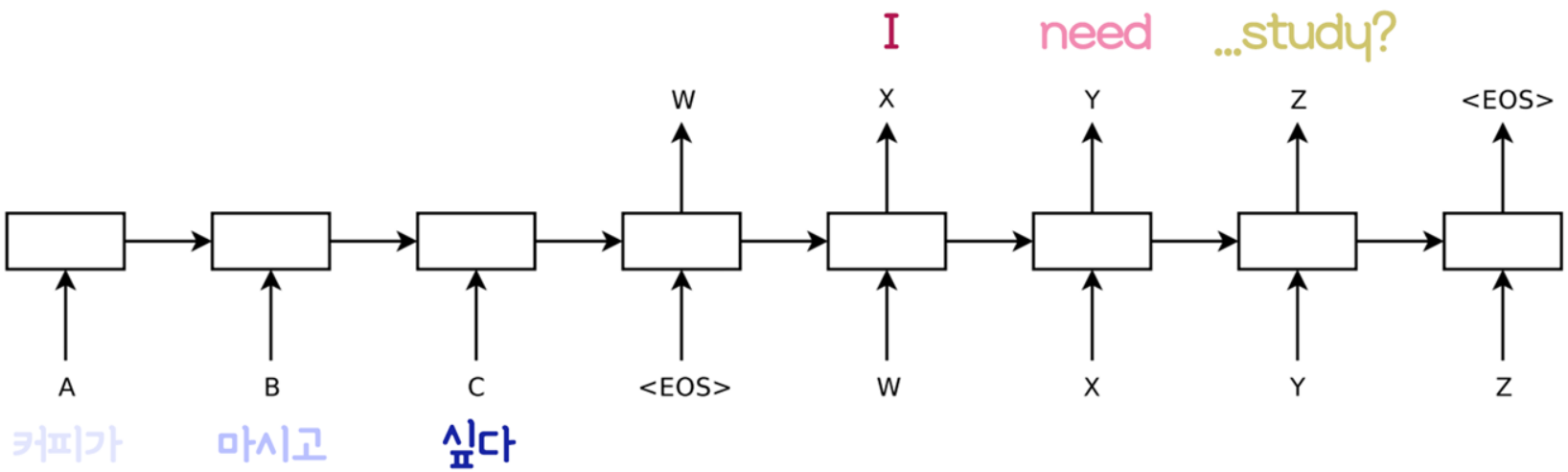
Vanila RNN



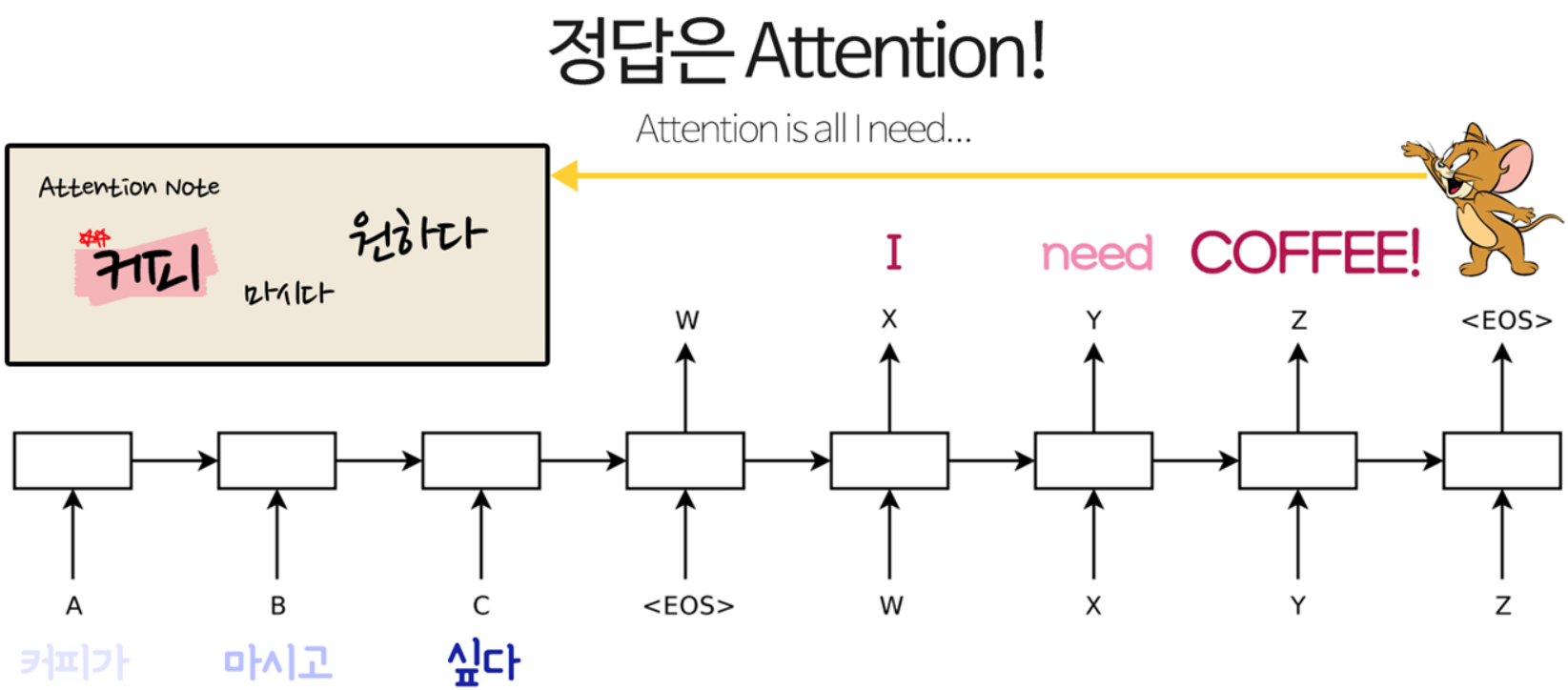
사전 준비

문제점

problem



사전 준비

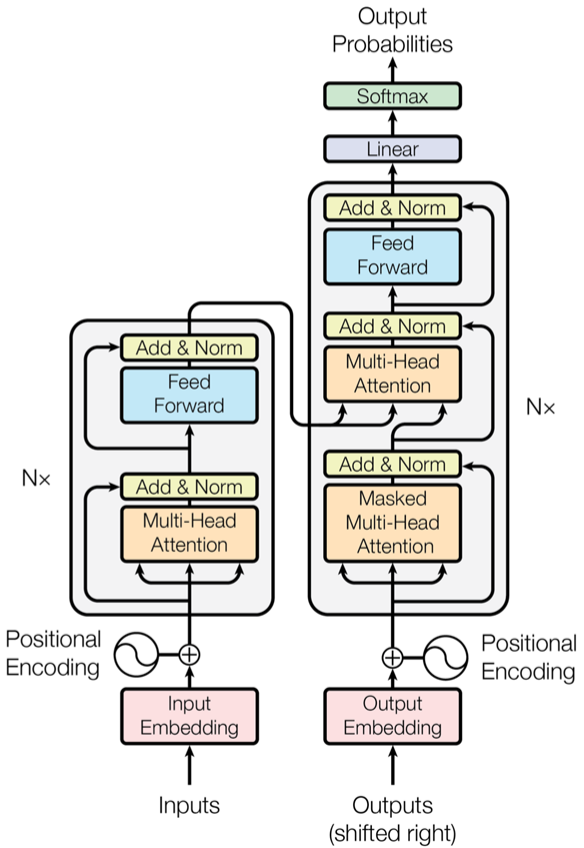

정답은 Attention!
Attention is all I need...
Attention Note
커피
원하다
마시다
I
need
COFFEE!
W
X
Y
Z
<EOS>
A
B
C
<EOS>
W
X
Y
Z
커피가
마시고
싶다

사전 준비

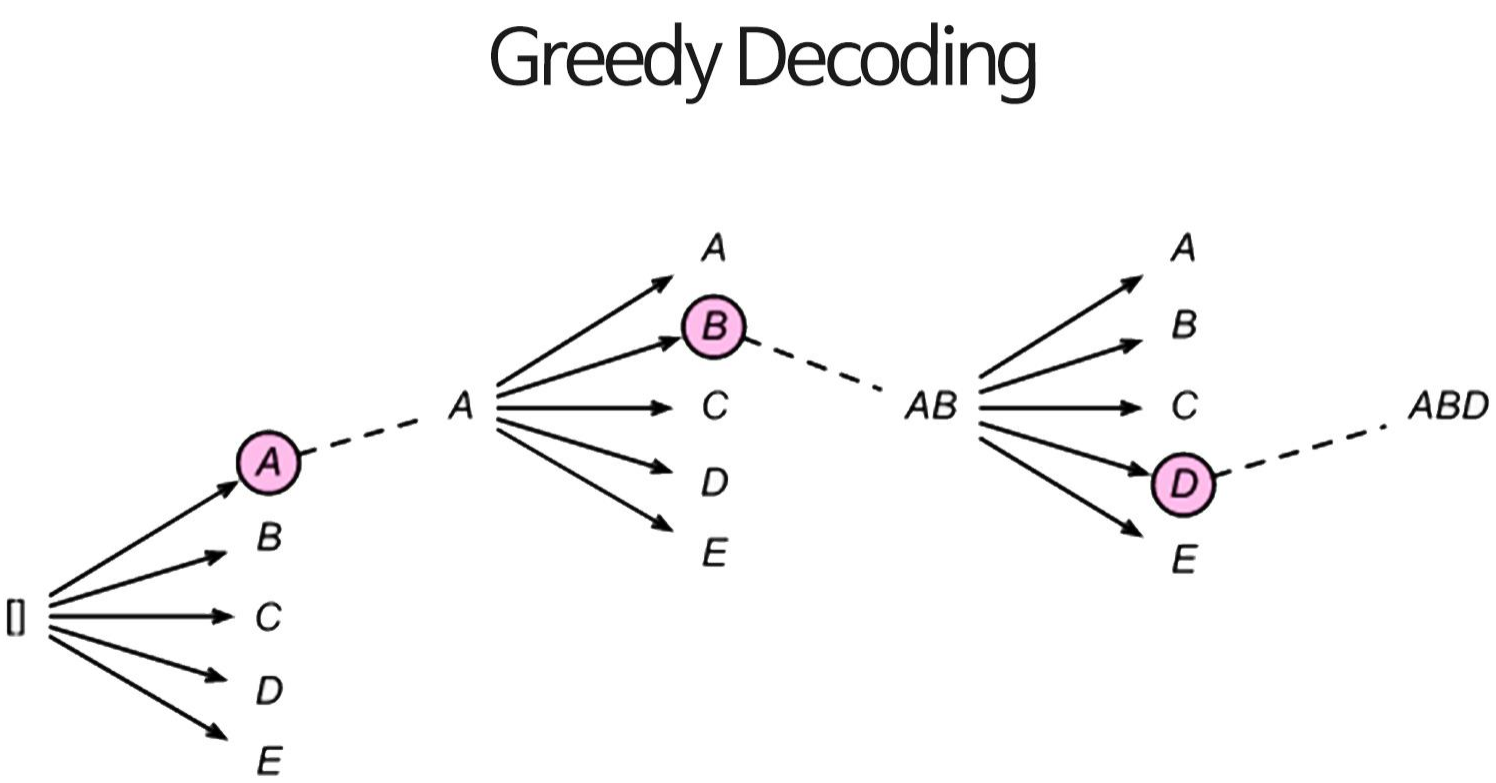
트랜스포머

Transformer

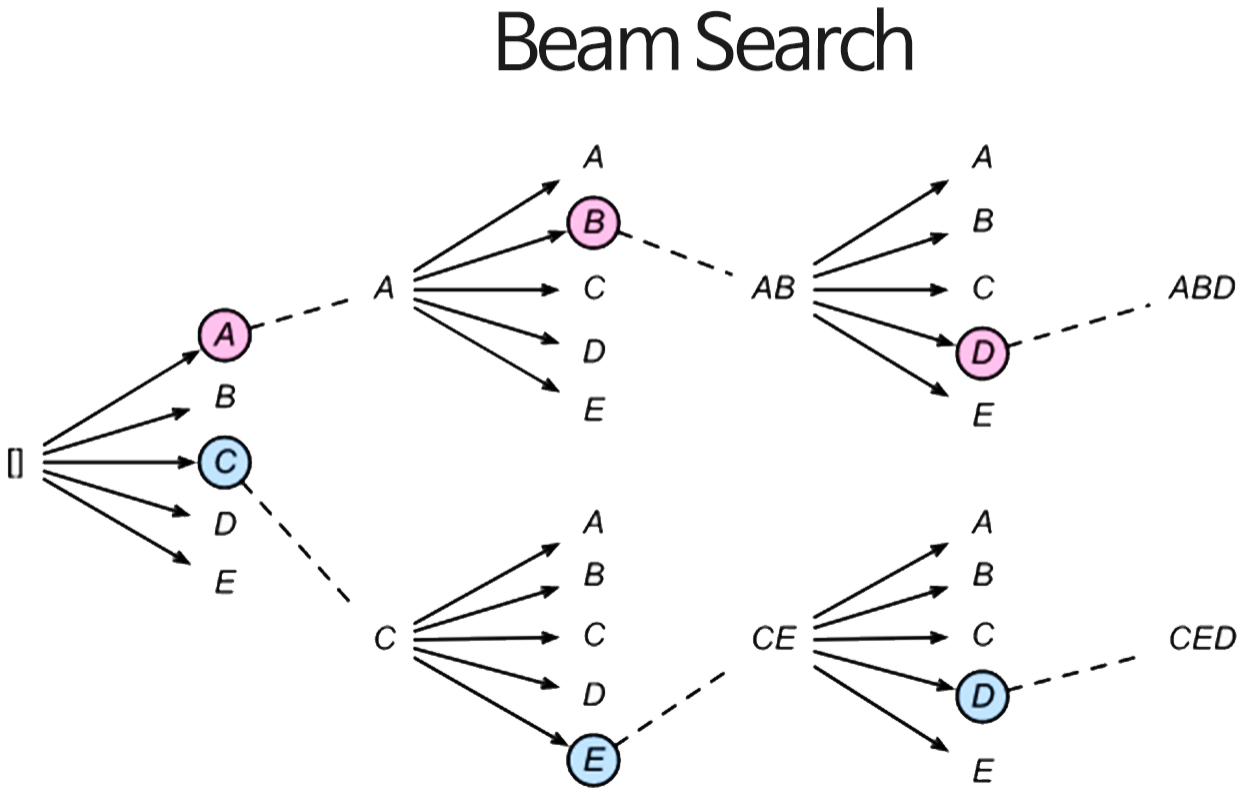
<Attention Is All You Need>



사전 준비



사전 준비



사전 준비

BLEU Score

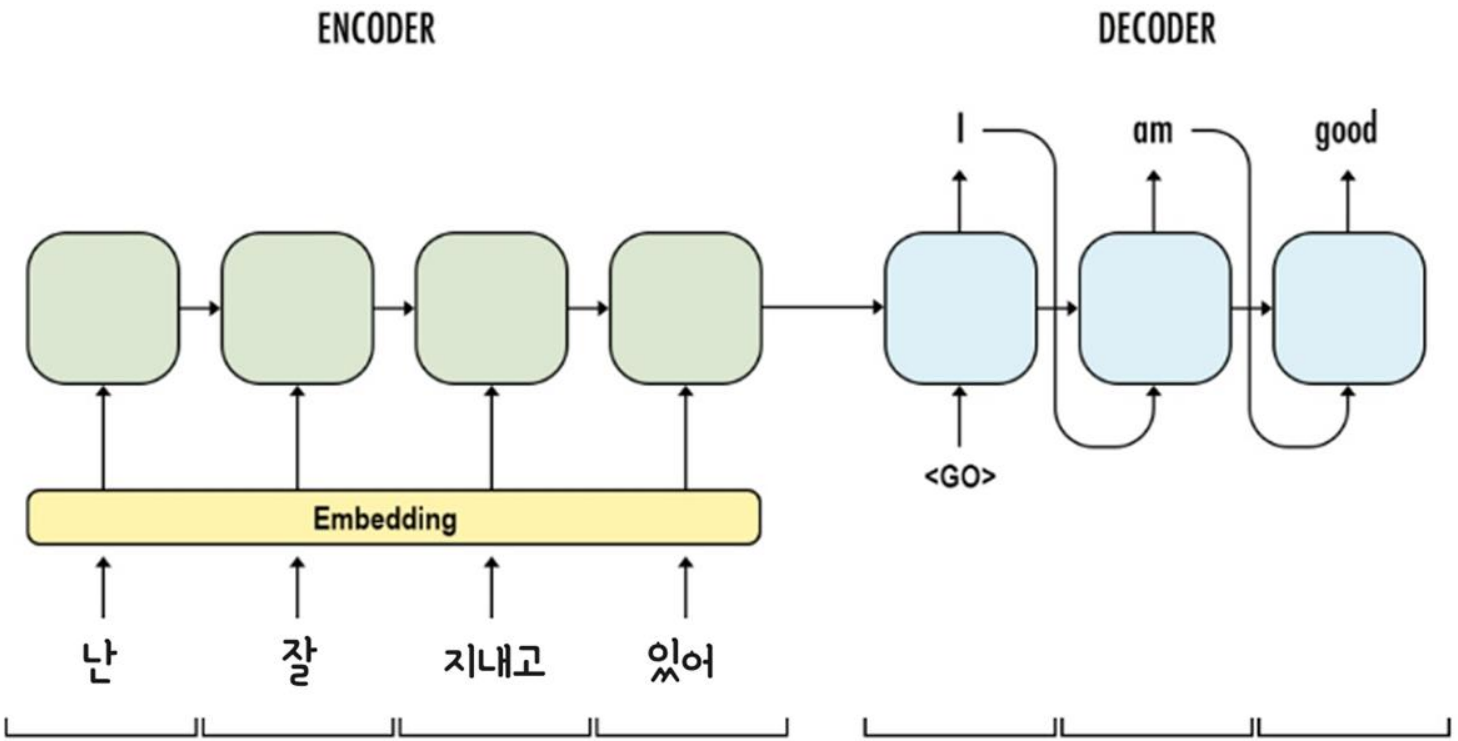
https://wikidocs.net/31695

```
import nltk.translate.bleu_score as bleu


candidate = 'It is a guide to action which ensures that the military always obeys the commands of the party'
references = [
    'It is a guide to action that ensures that the military will forever heed Party commands',
    'It is the guiding principle which guarantees the military forces always being under the command of the Party',
    'It is the practical guide for the army always to heed the directions of the party'
]

print(bleu_score(candidate.split(),list(map(lambda ref: ref.split()
# 이번 챕터에서 구현한 코드로 계산한 BLEU 점수
print(bleu.sentence_bleu(list(map(lambda ref: ref.split(), referenc
# NLTK 패키지 구현되어져 있는 코드로 계산한 BLEU 점수
```

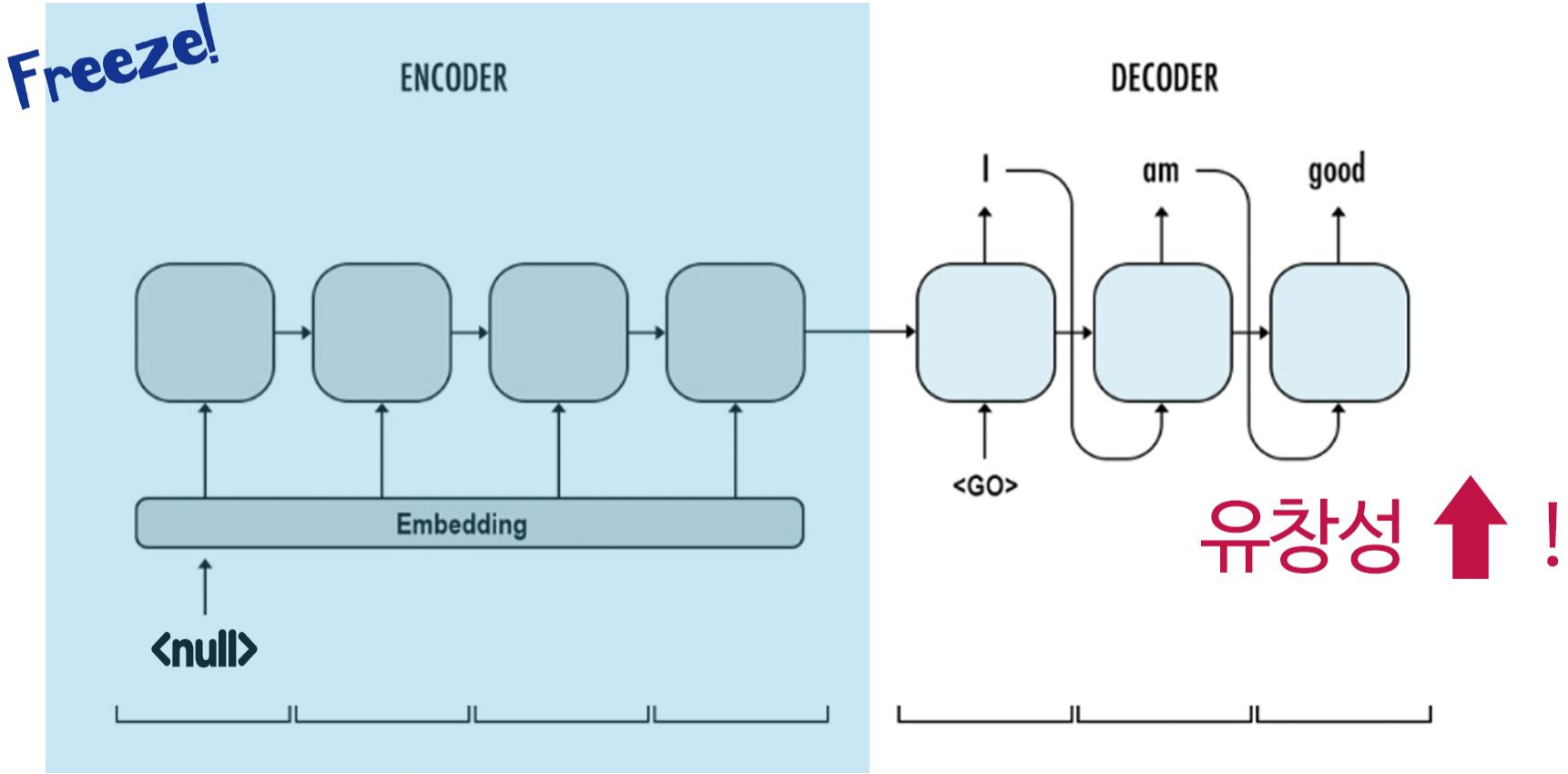
```
0.5045666840058485
0.5045666840058485
```

"Yes!" 한 단어 맞추고 만점을 받거나,
"the the the the the..." 등 나옴직한 같은 단어를
반복해 높은 점수를 얻거나 하는
불상사를 방지하기에 가장 좋은 지표!

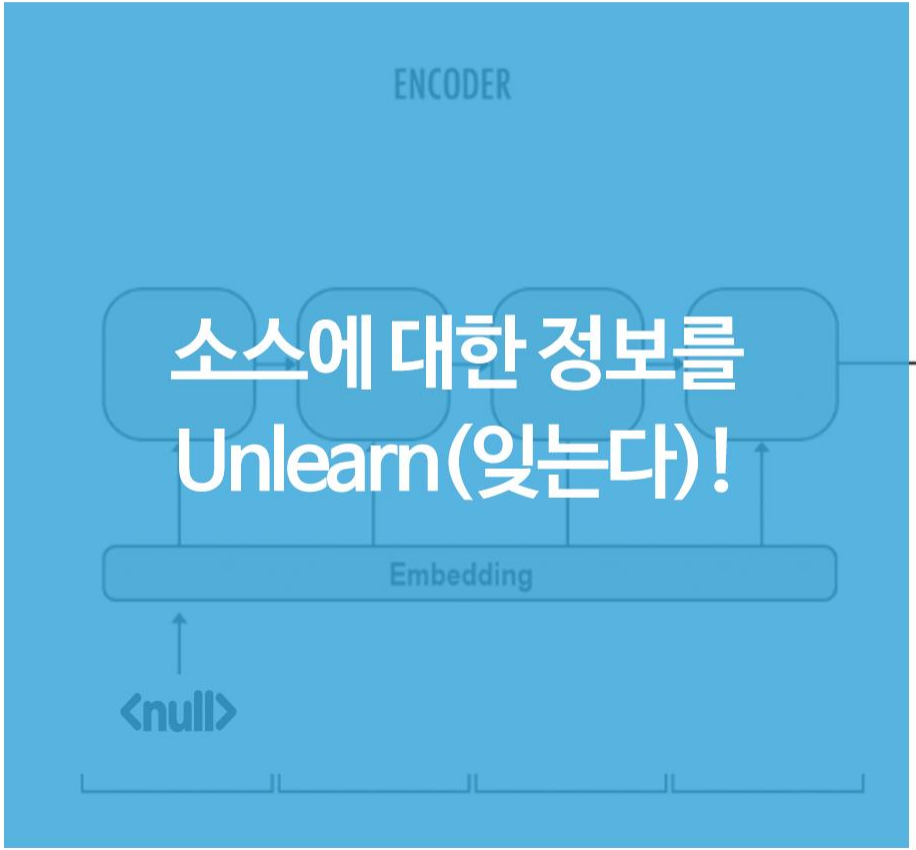
Dummy Source Sentences

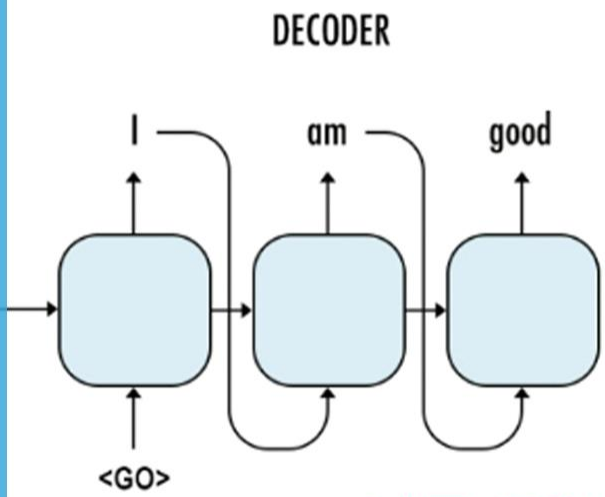


Dummy Source Sentences



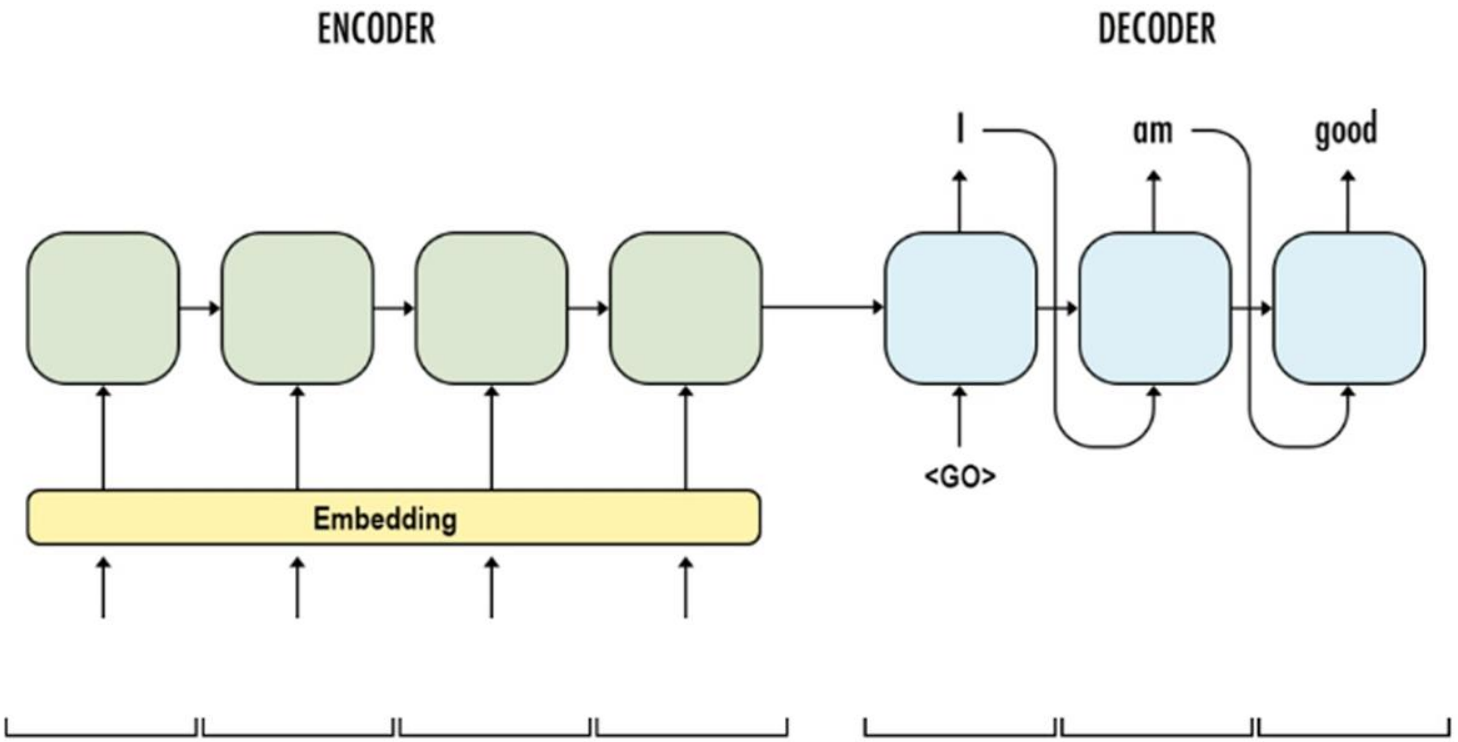
만약 단일 데이터가 너무 많으면...



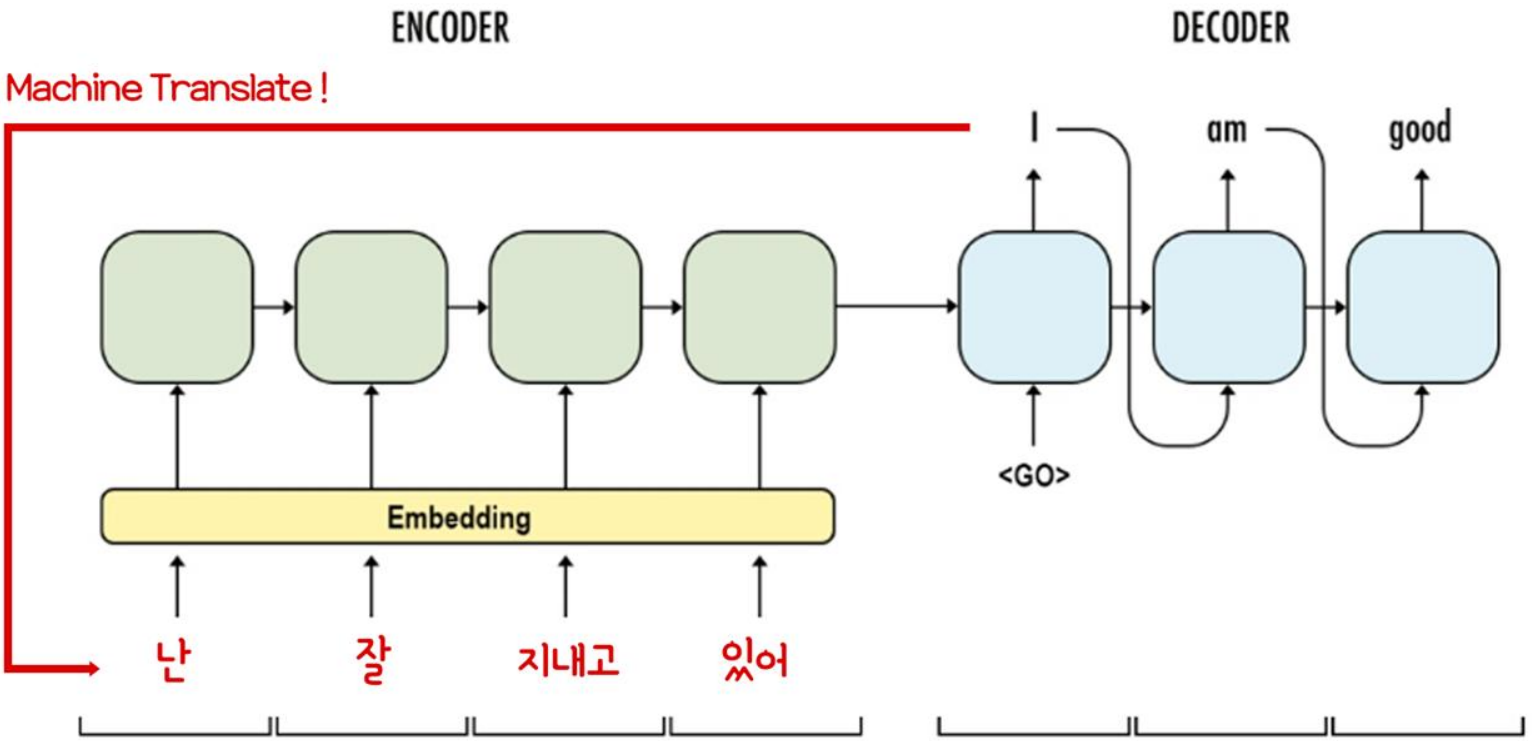


너무 유창해...

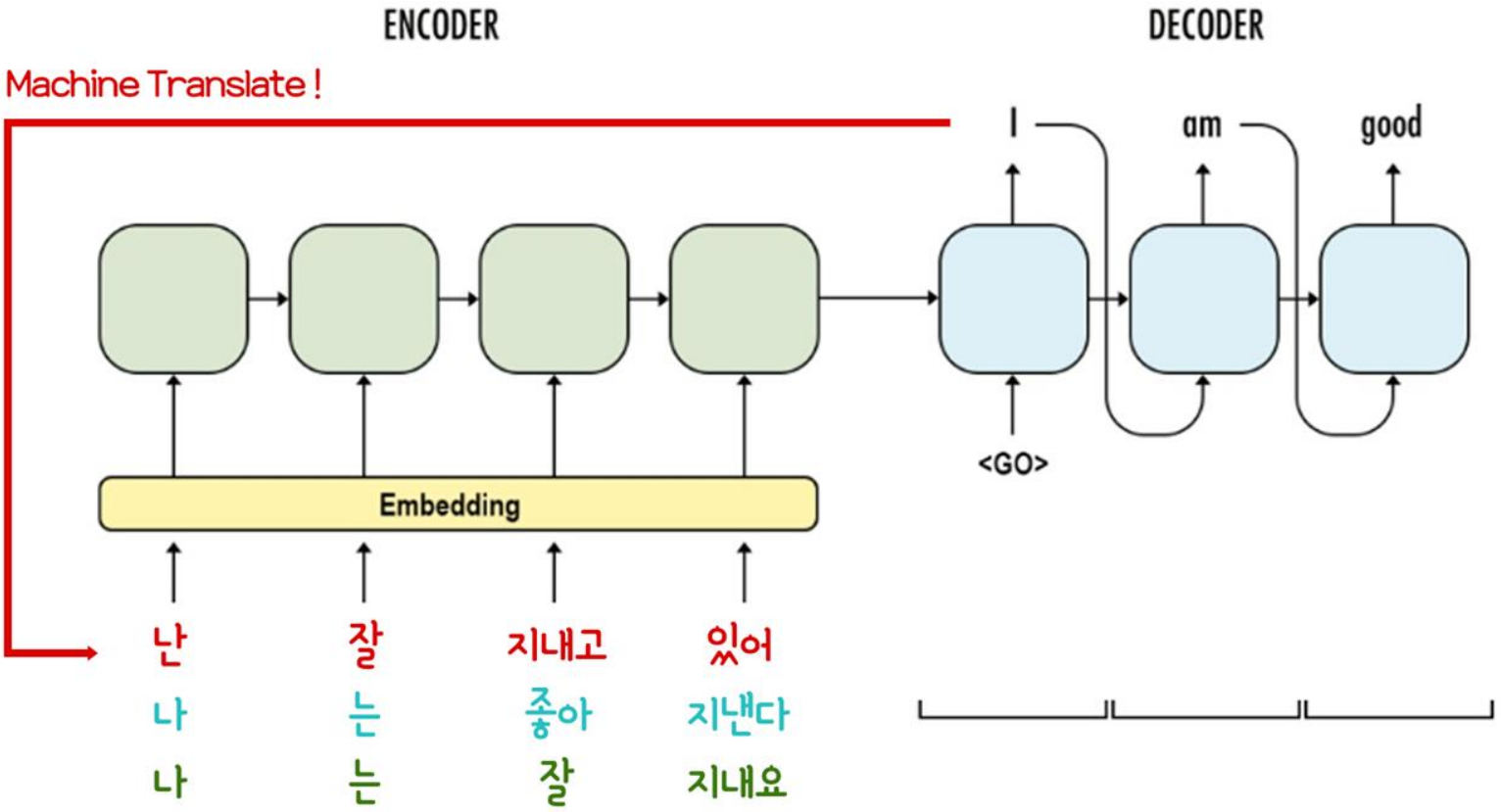
Synthetic Source Sentences



Synthetic Source Sentences



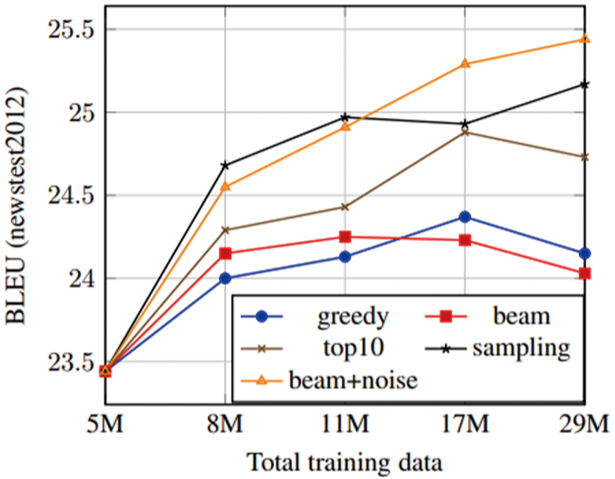
Synthetic Source Sentences



여기까지가

〈Improving Neural Machine Translation Models with Monolingual Data〉

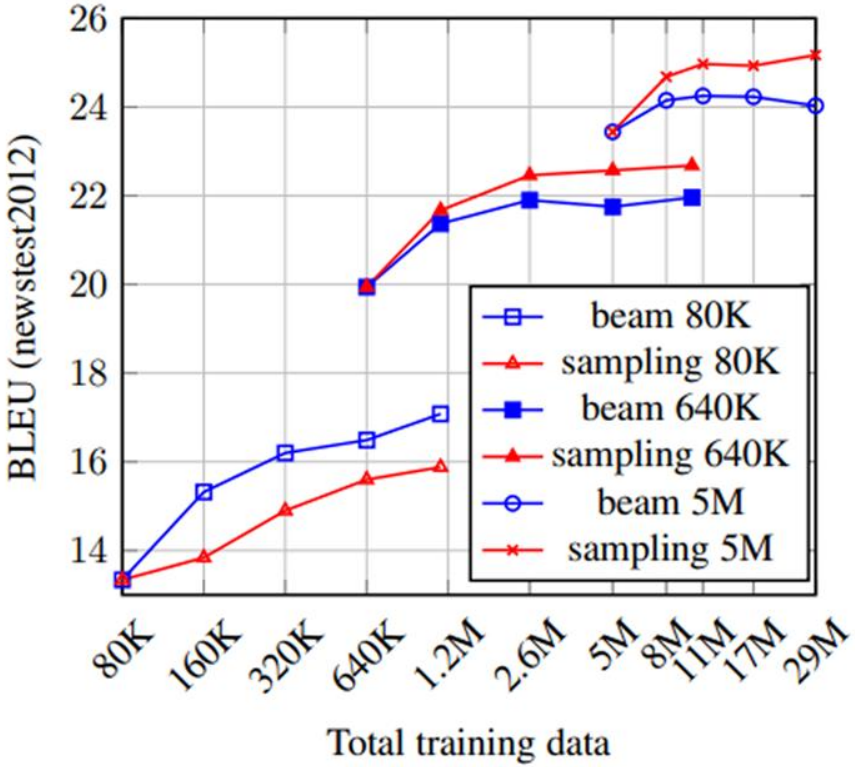
Result



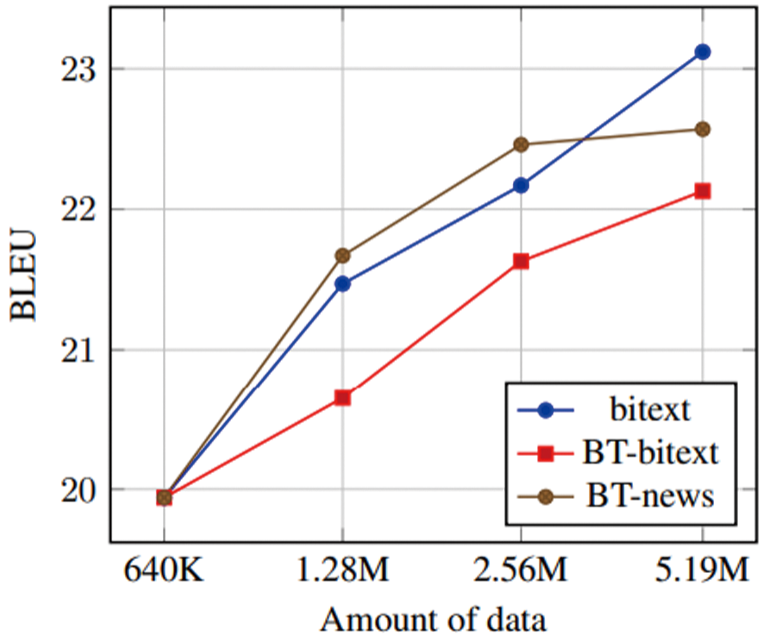
	Perplexity
human data	75.34
beam	72.42
sampling	500.17
top10	87.15
beam+noise	2823.73

Reference	사람 과 컴퓨터 가 의사소통 을 합니다.
Beam	사람 과 컴퓨터 가 의사소통 을 하기 합니다.
Random Sampling	역사 사람 은 과 전산 입니다 입니다 자동차 기술
Top-10 Sampling	사람 은 컴퓨터 는 기술 및 통신 이다.
Beam + Noise	과 〈BLANK〉 컴퓨터 가 의사소통 〈BLANK〉 합니다.

Result


26
24
22
20
18
16
14
BLEU (newstest2012)
beam 80K
sampling 80K
beam 640K
sampling 640K
beam 5M
sampling 5M
80K
160K
320K
640K
1.2M
2.6M
5M
8M
11M
17M
29M
Total training data

Result



(a) newstest2012

Source Sentence

Test:

사람의 목소리와 얼굴을 똑같이 합성해

국내 최초로 'AI 뉴스 앵커'를 탄생시켰다.

BT-bitext: (Machine Generated)

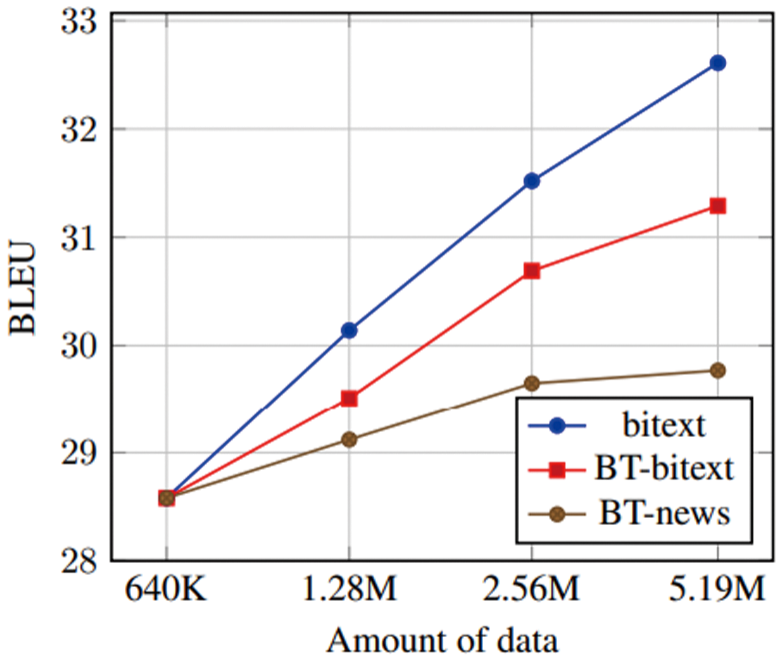
난 연애는 소질 없어, 만인의 연인이니까.

BT-news: (Machine Generated)

펭수가 그룹 방탄소년단 정국, 뷔를 만난

인증샷을 공개했다.

Result



(b) valid-mixed

Source Sentence

Test:

세금 할인해주는 네X버 고지서 신청하고

백만원에 도전하세요!

BT-bitext: (Machine Generated)

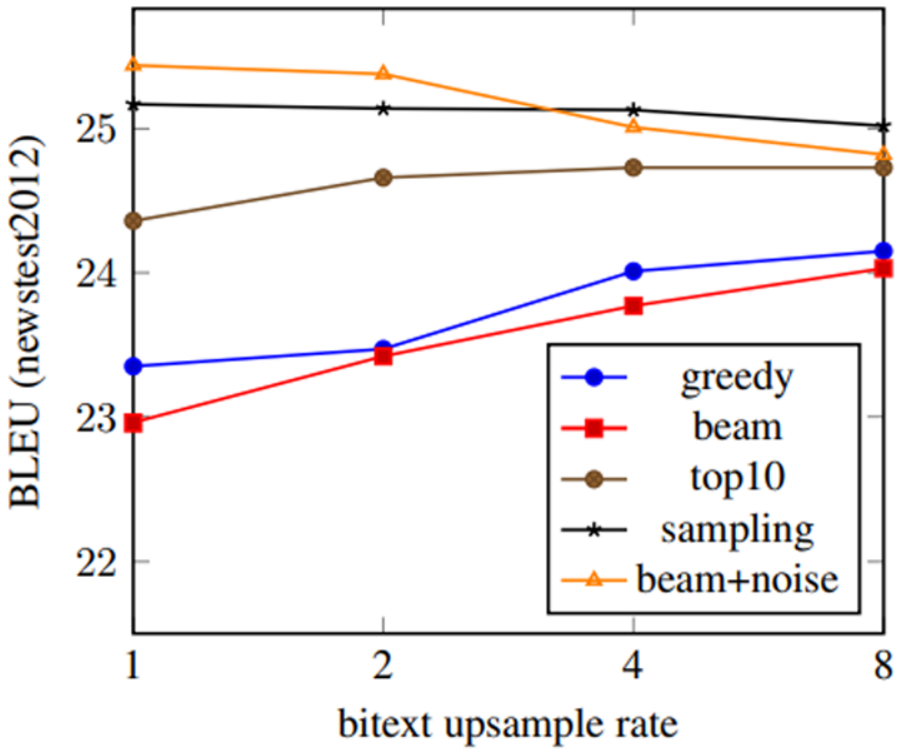
난 연애는 소질 없어, 만인의 연인이니까.

BT-news: (Machine Generated)

펭수가 그룹 방탄소년단 정국, 뷔를 만난

인증샷을 공개했다.

Result



Upsampling:
단일 데이터를 학습하는 도중,
병렬 데이터(실제 데이터)를 방문하는 빈도.

Result

	En–De	En–Fr
a. Gehring et al. (2017)	25.2	40.5
b. Vaswani et al. (2017)	28.4	41.0
c. Ahmed et al. (2017)	28.9	41.4
d. Shaw et al. (2018)	29.2	41.5
DeepL	33.3	**45.9**
Our result	**35.0**	45.6
detok. sacreBLEU[3]	*33.8*	*43.8*