ICT이노베이션스퀘어 AI복합교육 고급 언어과정

자연어처리를 위한 Machine Translation

현청천

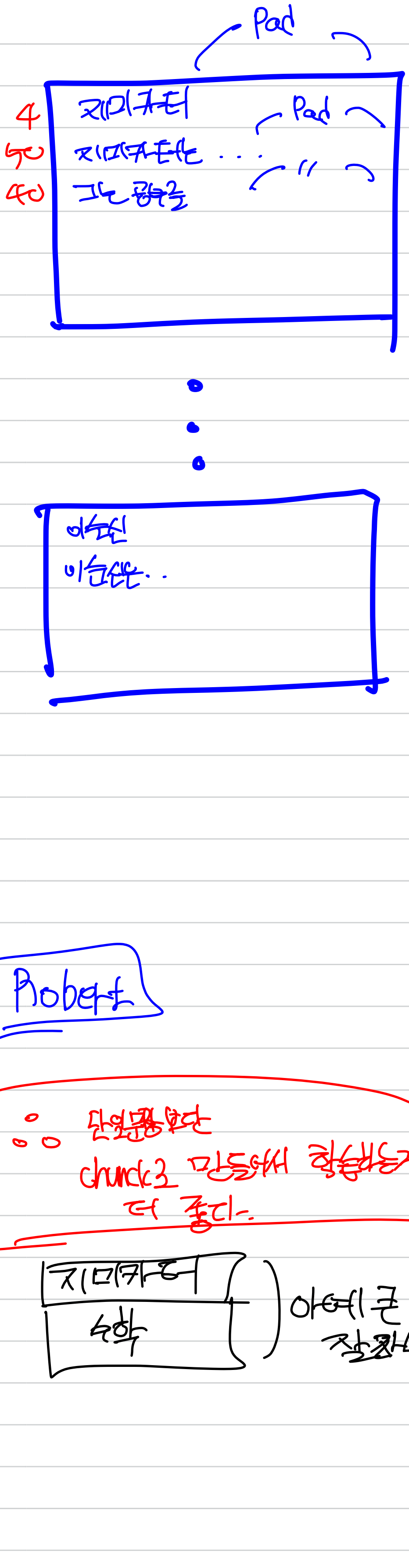
2021.04.19

ex)

지미카터. 지미카터는 미국//대통령이다. 그는
시작 시작(x)

같은 주제 (동일)

4 + 50 + 40 + 150



① ② ③ ④ . . ⑩⑩ ⇒

128

①+②+③//...... 자르고 버리고.

④+⑤+⑥+⑦+⑧//.... 자르고 버리고..

- 시작점은 같지만
- 뒷쪽은 자르자.
- but) 최대한 많은 문장들을 합쳐서 긴 문장을 학습시킬수있는 장점이 있다.

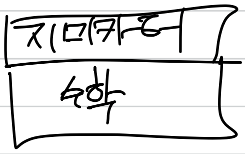
이때 자르는 단위 ⇒ 청크 = chunck

⇒ 128

⇒ 문장대로 끊는게 더 안 좋다라는... 논문에서 ⇒ Robert

- segment-Pair + NSP < 512 ⇒ 우리가 쓰는거
- sentence-Pair + NSP
  - 한줄씩. 한 문장씩만
- Full sentence

∴ 단일문장보단 chunk로 만들어서 학습하는게 더 좋다.



아예 큰 꼴로 봐서 잘라내자.

What is Machine Translation

인간이 사용하는 자연 언어를 컴퓨터를 사용하여 다른 언어로 번역하는 것

Source (x)

Target (y)

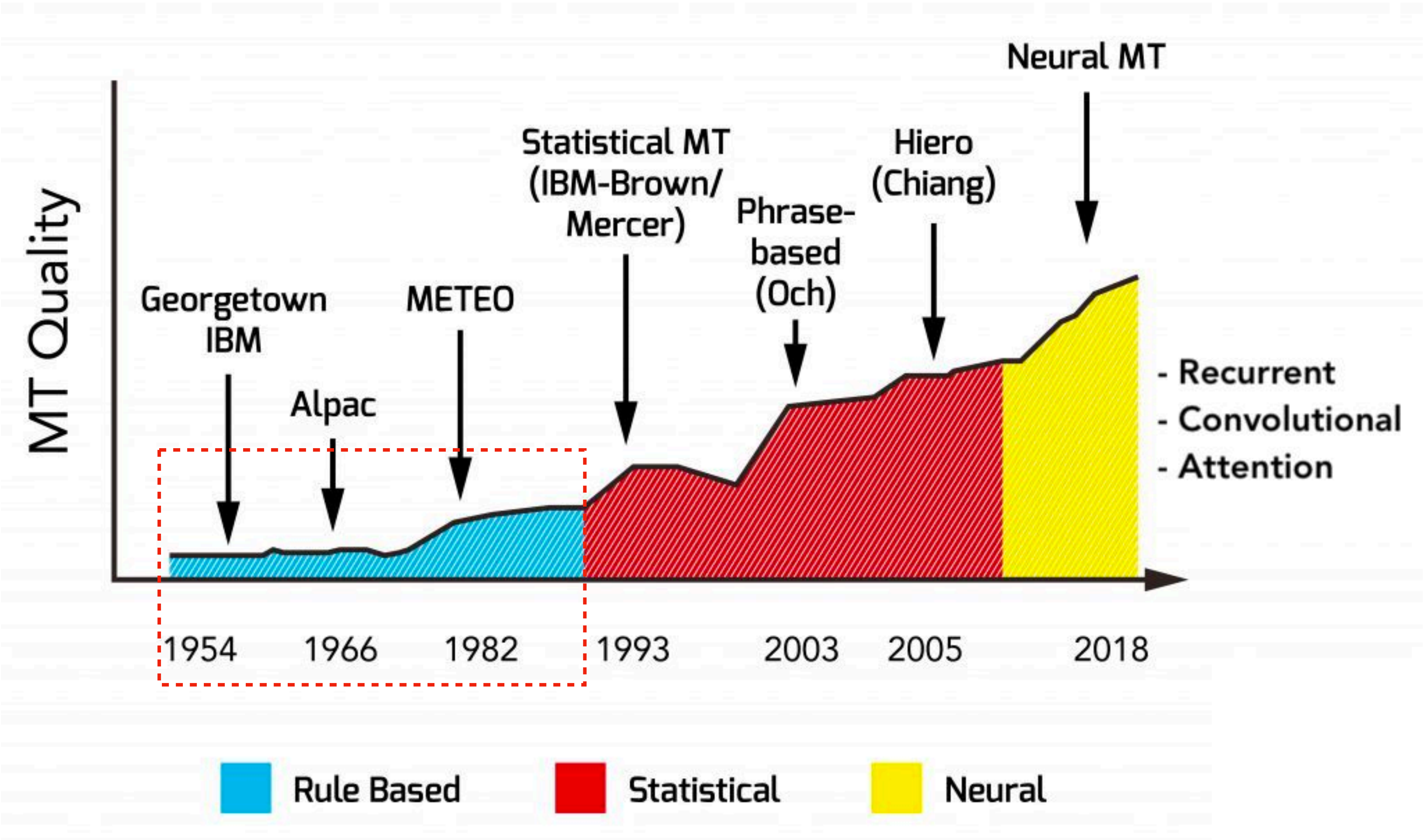
Education is the most powerful weapon we can use to change the world.

↓

교육은 세상을 바꿀 수 있는 가장 강력한 무기이다.

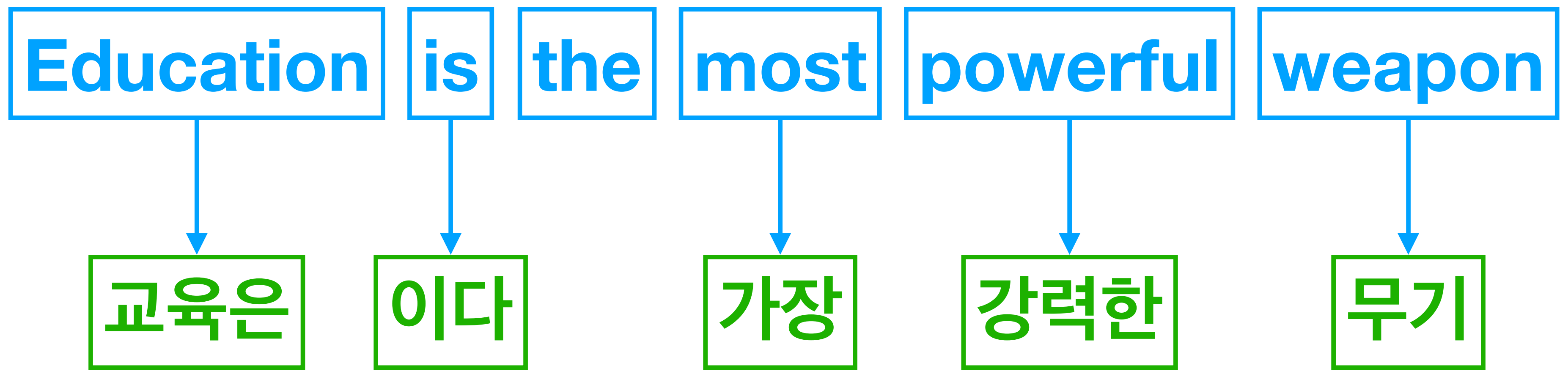
출처: https://ko.wikipedia.org/wiki/%EA%B8%B0%EA%B3%84_%EB%B2%88%EC%97%AD

What is Machine Translation (history)



Rule-based Machine Translation

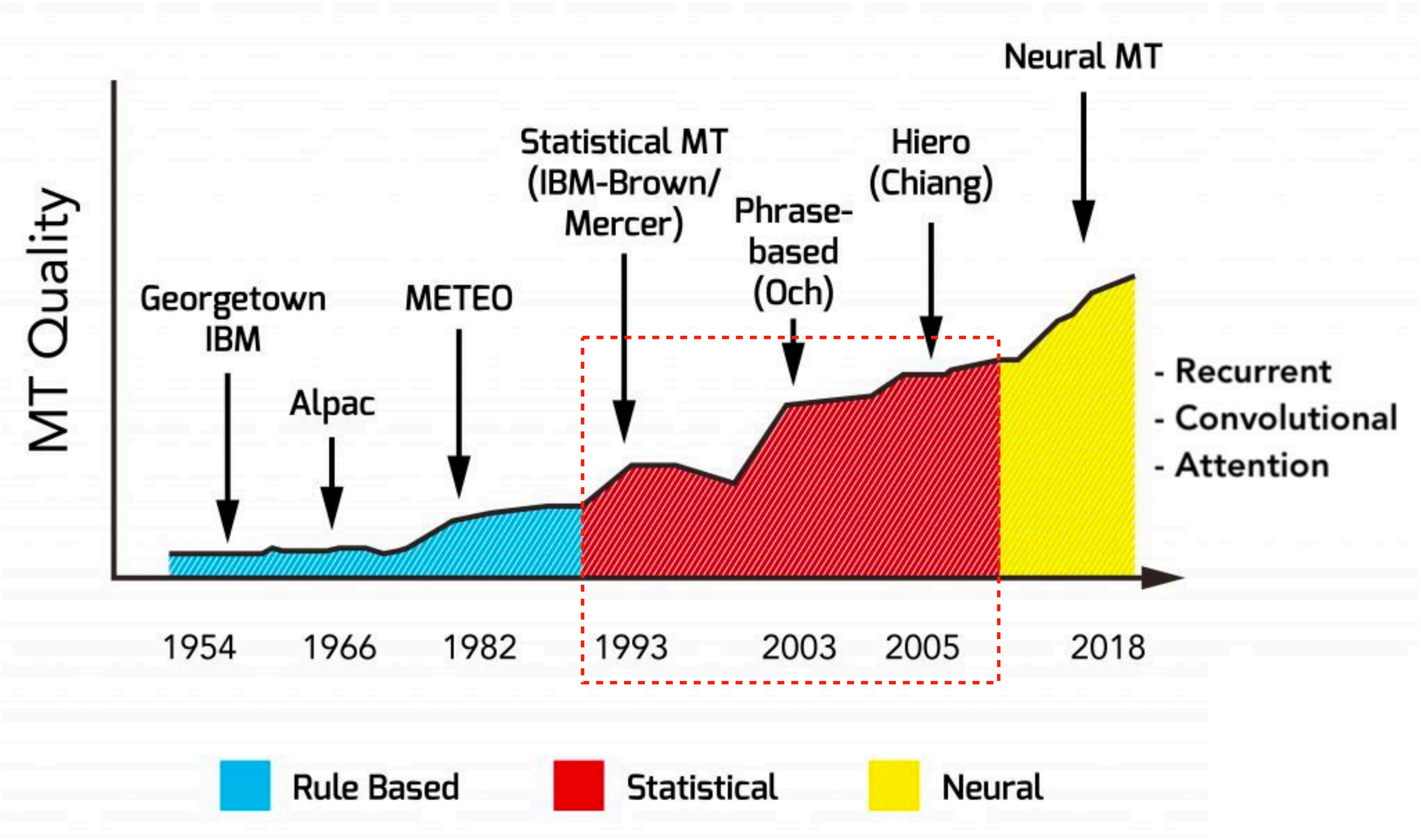
- Bilingual dictionary
- Linguistic rules for each language



I saw a man on a hill with a telescope?

이미지: https://iconictranslation.com/what-we-do/neural-machine-translation/
참조: https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

What is Machine Translation (history)



Statistical Machine Translation

- Language pair로부터 패턴 학습
- 데이터가 많을 수록 좋은 결과
- $argmax_y P(y|x)$

input.

$x =$ **'Education is the most powerful weapon'**

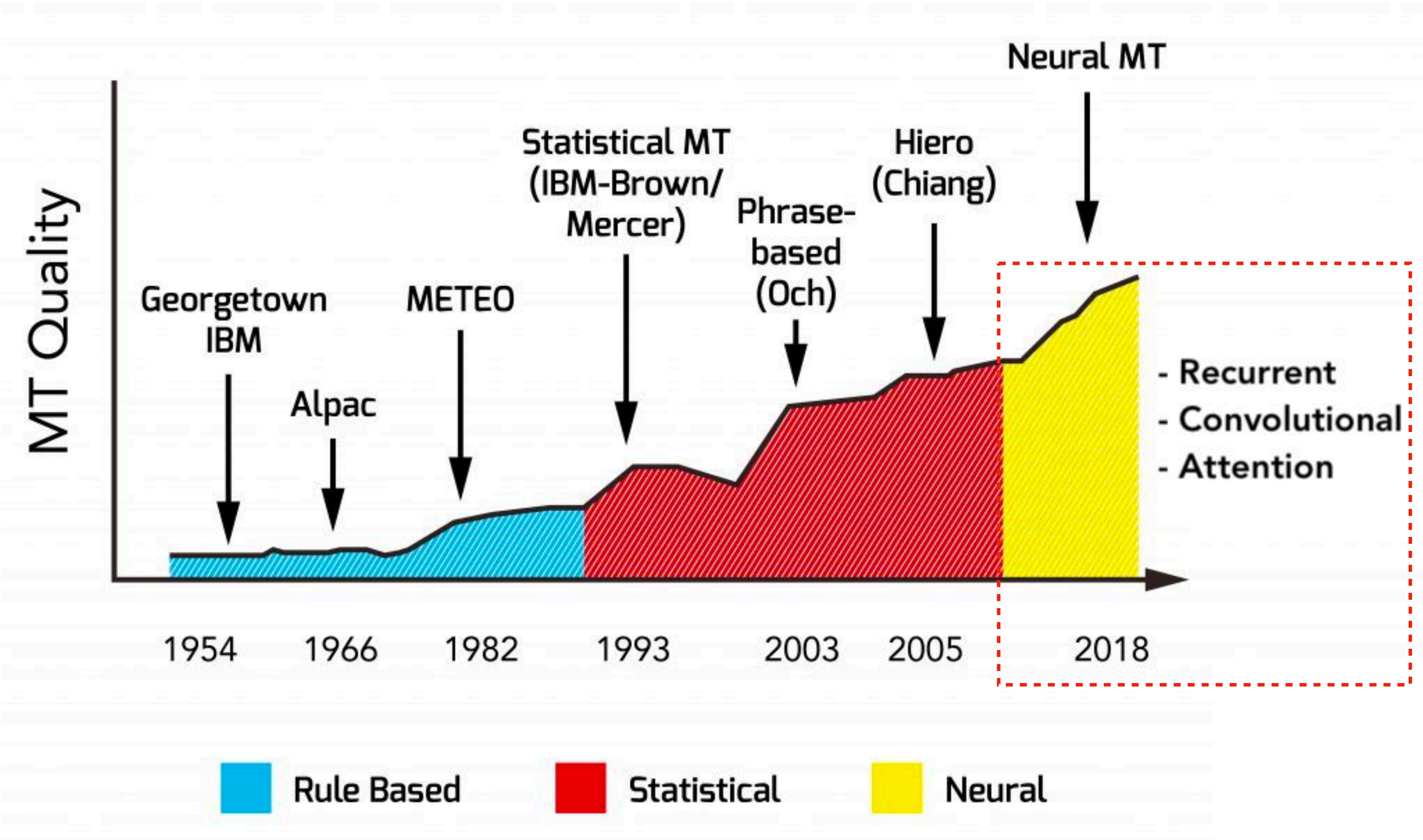
'교육은' $= argmax_y P(y|x)$

'가장' $= argmax_y P(y|x,$ '교육은'$)$

'강역한' $= argmax_y P(y|x,$ '교육은', '가장'$)$

이미지: https://iconictranslation.com/what-we-do/neural-machine-translation/
참조: https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

What is Machine Translation (history)

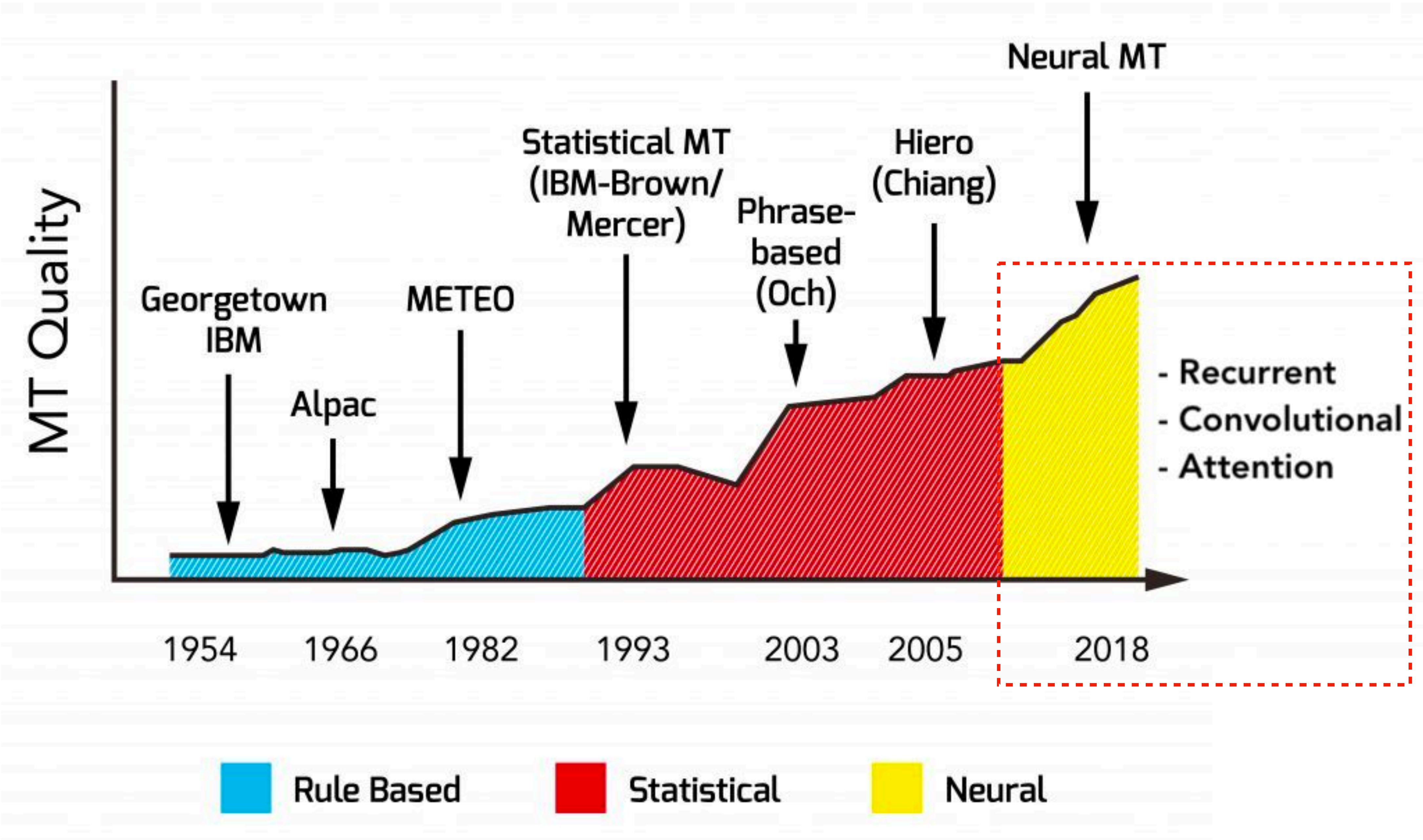


Neural Machine Translation

- 20~~11~~14년 sequence to sequence = seq2seq 모델 등장
- 데이터에서 Neural Network 학습
- $P(y|x;\theta)$
  - 어순 오류 감소
  - 어휘 오류 감소
  - 문법 오류 감소

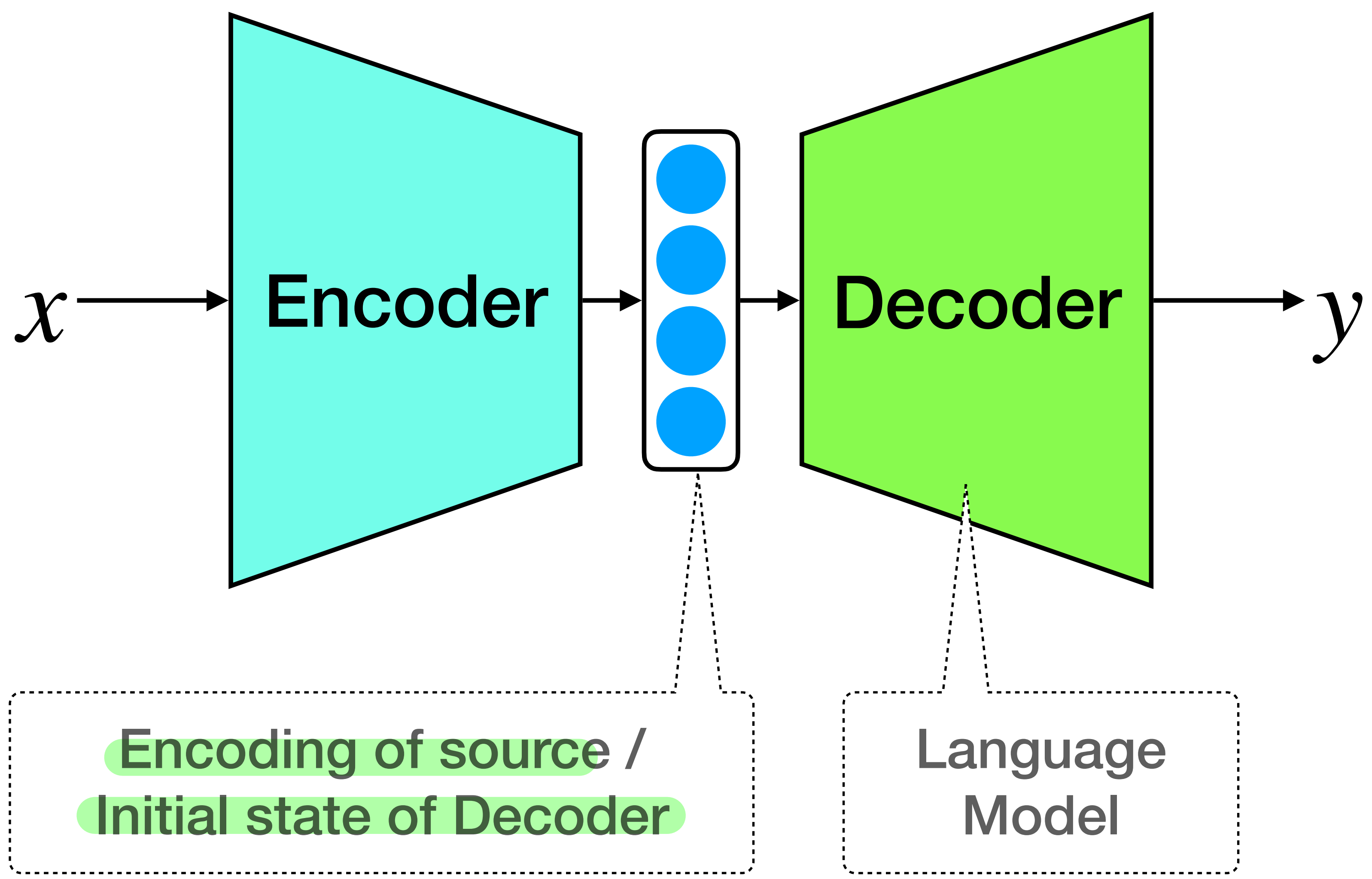
이미지: https://iconictranslation.com/what-we-do/neural-machine-translation/
참조: https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

What is Machine Translation (history)



Neural Machine Translation

sentence embedding



○ 디코더의 초기값

○ input을 Encoding 해서 Vector로 만들기!

이미지: https://iconictranslation.com/what-we-do/neural-machine-translation/

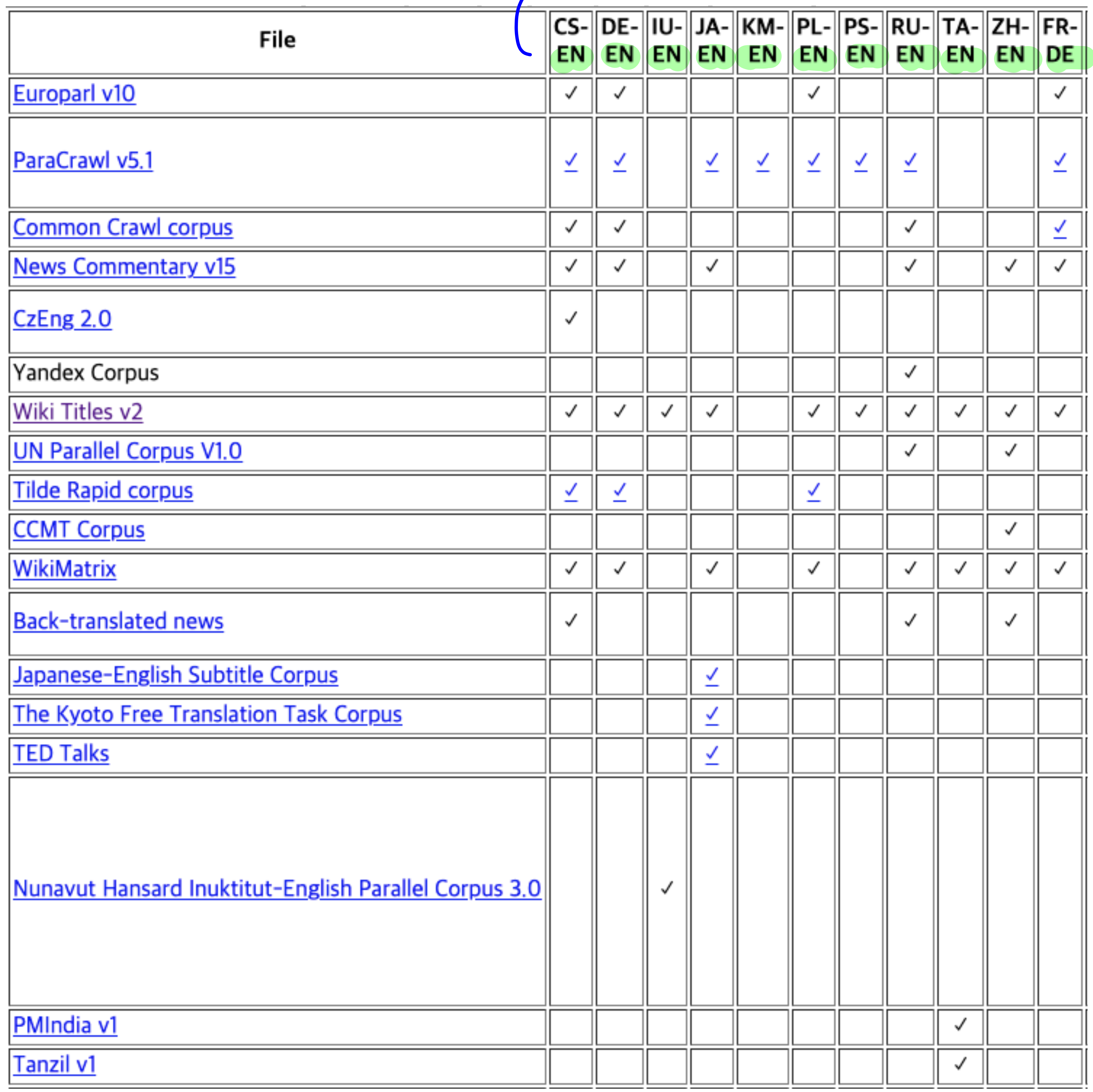
참조: https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/

Machine Translation DataSet

ex) Bible,,은 보통 다 전세계에 있으니까..

- WMT Dataset

* 다 영어 기반이네..

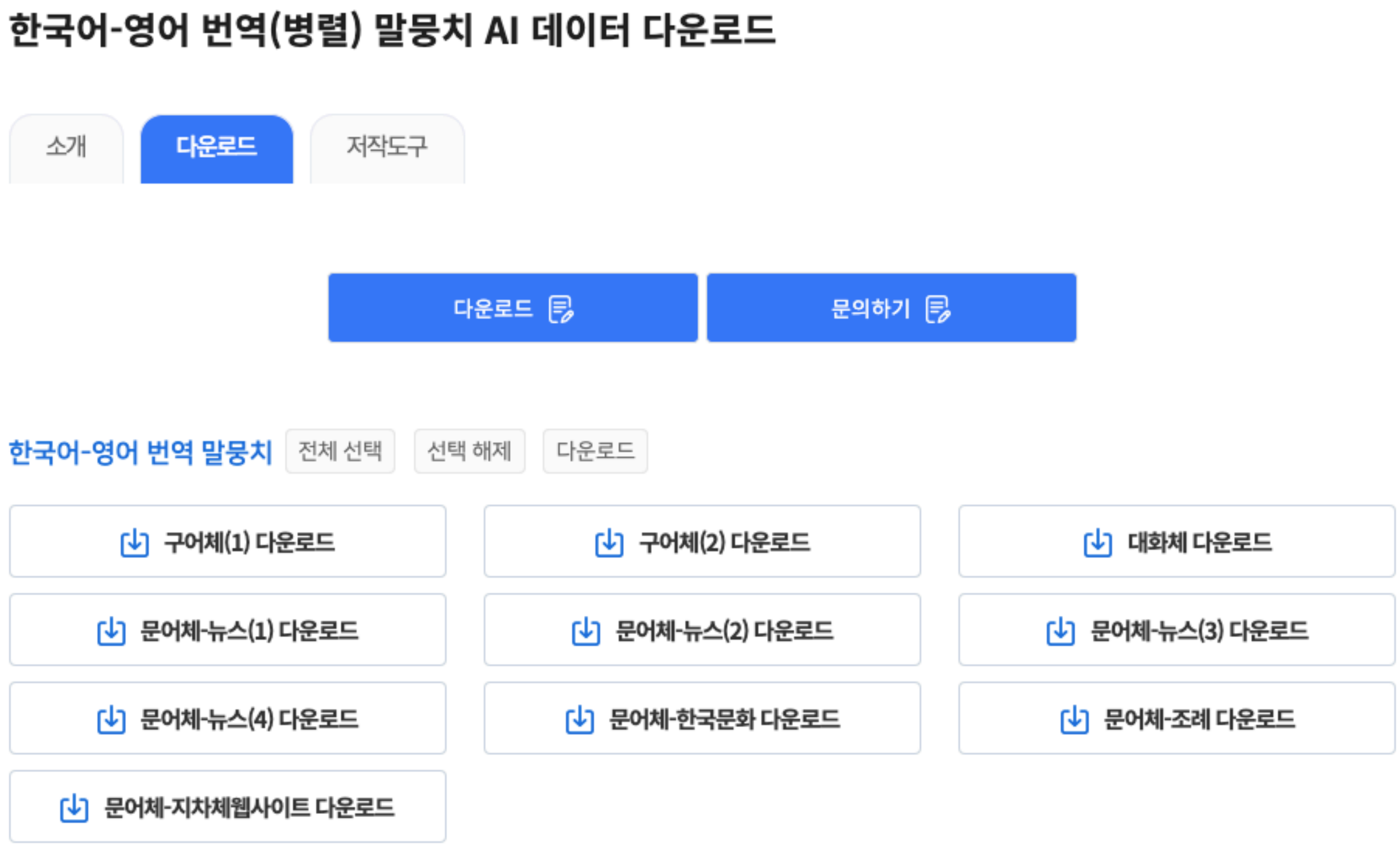


File	CS-EN	DE-EN	IU-EN	JA-EN	KM-EN	PL-EN	PS-EN	RU-EN	TA-EN	ZH-EN	FR-DE
Europarl v10	✓	✓				✓					✓
ParaCrawl v5.1	✓	✓		✓	✓	✓	✓	✓			✓
Common Crawl corpus	✓	✓						✓			✓
News Commentary v15	✓	✓		✓				✓		✓	✓
CzEng 2.0	✓										
Yandex Corpus								✓			
Wiki Titles v2	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
UN Parallel Corpus V1.0								✓		✓	
Tilde Rapid corpus	✓	✓				✓					
CCMT Corpus										✓	
WikiMatrix	✓	✓		✓		✓		✓	✓	✓	✓
Back-translated news	✓							✓		✓	
Japanese-English Subtitle Corpus				✓							
The Kyoto Free Translation Task Corpus				✓							
TED Talks				✓							
Nunavut Hansard Inuktitut-English Parallel Corpus 3.0			✓								
PMIndia v1									✓		
Tanzil v1									✓		

- Workshop on Statistical Machine Translation
- Bilingual Datasets
- English Based Datasets
- http://www.statmt.org/wmt20/translation-task.html

Machine Translation DataSet

- WMT Dataset
- AI-Hub 한국어-영어 병렬 말뭉치



한국어-영어 번역(병렬) 말뭉치 AI 데이터 다운로드

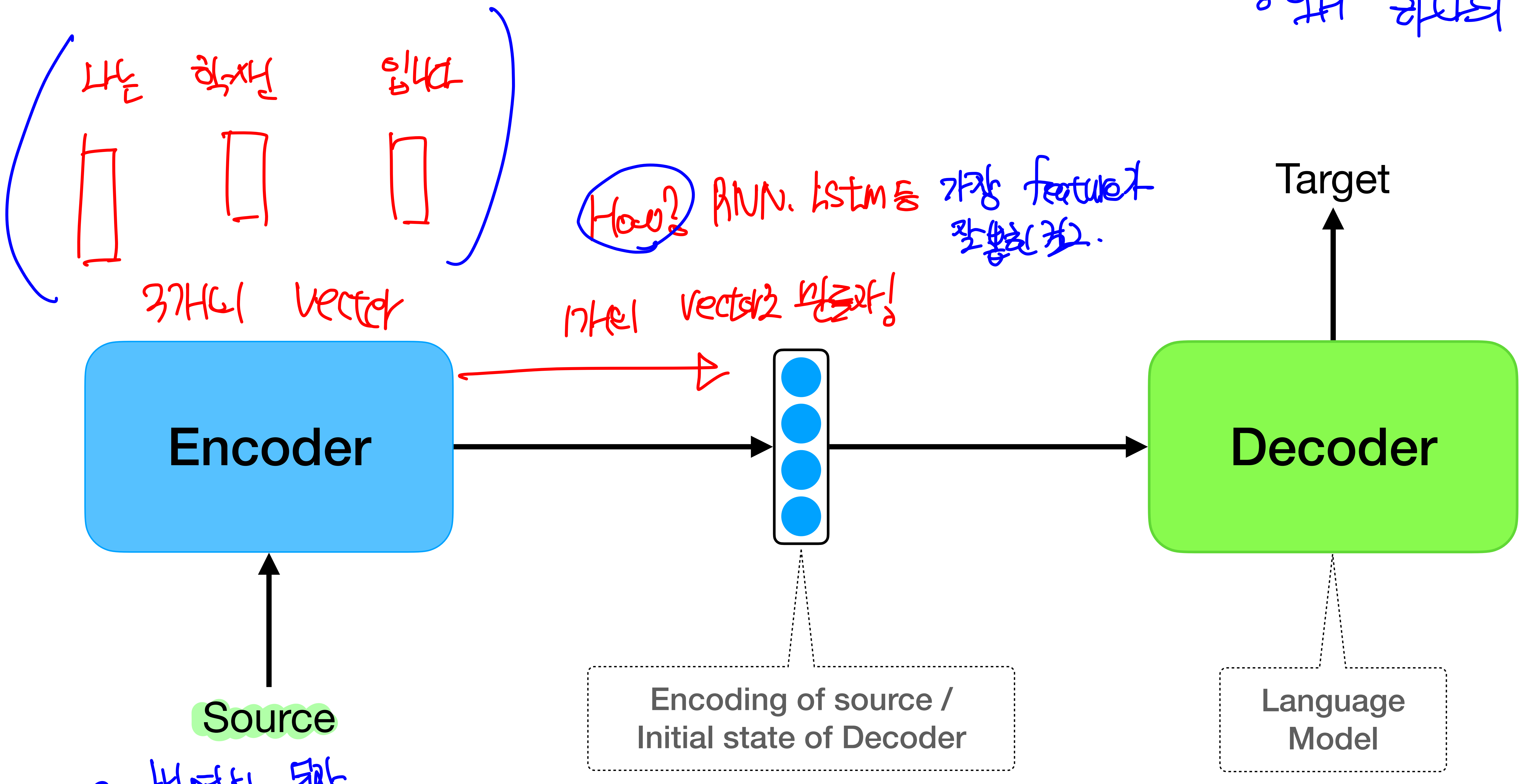
- AI-Hub 한국어-영어 데이터
- 회원 가입 및 별도의 서류제출 후 다운로드
- https://aihub.or.kr/aidata/87/download

ex) convolutional seq 2 seq

Machine Translation Model

○ 왜 하필 Encoder → Decoder form으로 만든거지?

○ 왜 하나의 vector로 만들어야되는거지?



나는 학생 입니다

3개의 vector

How? RNN, LSTM 등 가장 feature가 잘 뽑힌 것으로.

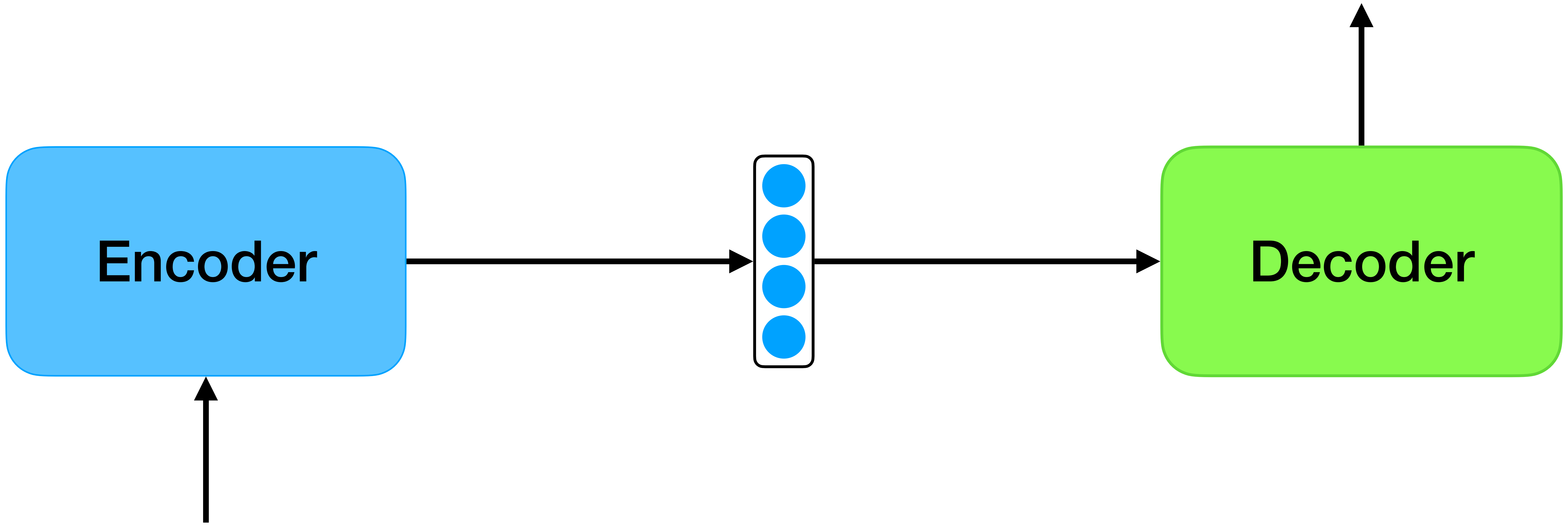
1개의 vector로 변환!

○ 번역할 문장

○ sentence embedding

○ 생성

Machine Translation Model



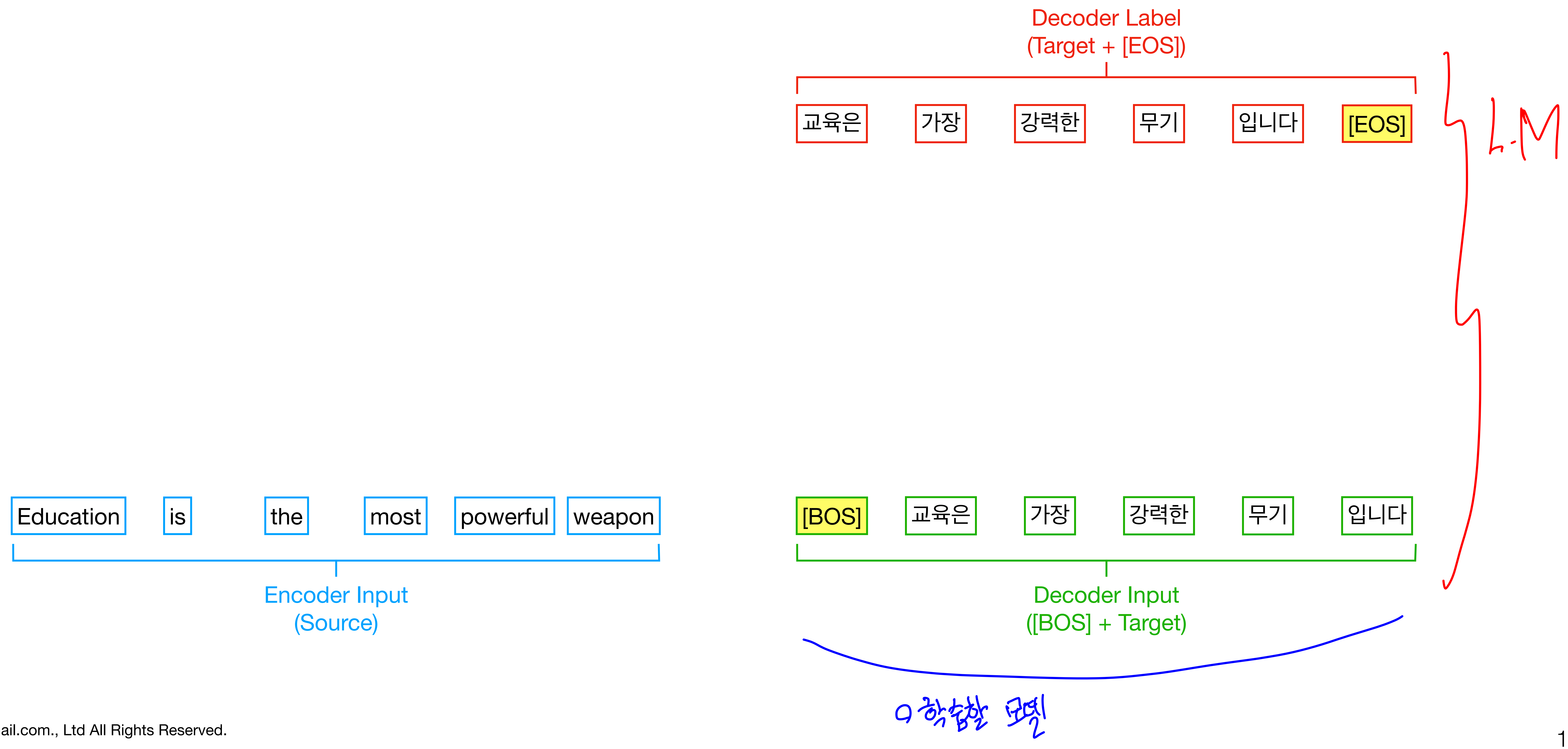
교육은 가장 강력한 무기이다

Source = Education is the most powerful weapon

같은 말임

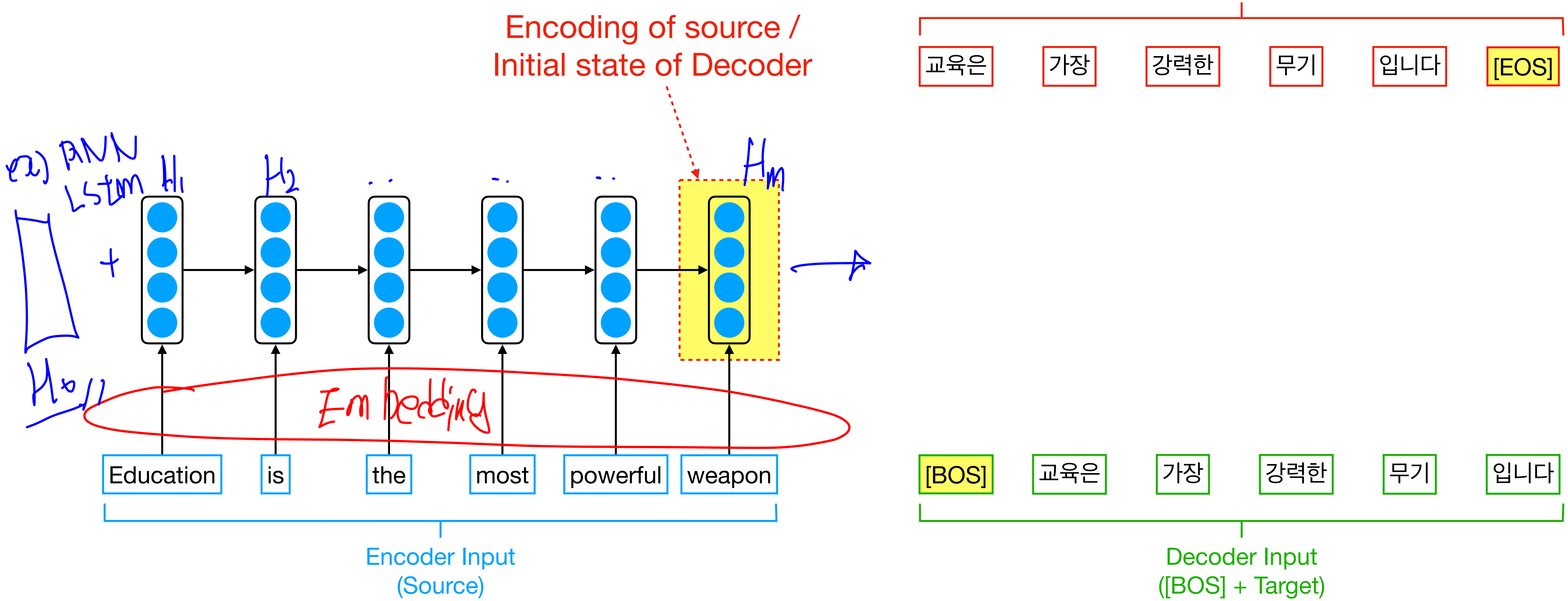
Encoder Decoder Architecture = / Sequence to Sequence

Machine Translation Model (Training)



Machine Translation Model (Training)

- 일반적으로 Lstm을 많이사용!
- (encoder_hidden_state
  encoder_cell_state)


Decoder Label
(Target + [EOS])

Encoding of source /
Initial state of Decoder

Encoder Input
(Source)

Decoder Input
([BOS] + Target)

Machine Translation Model (Training)

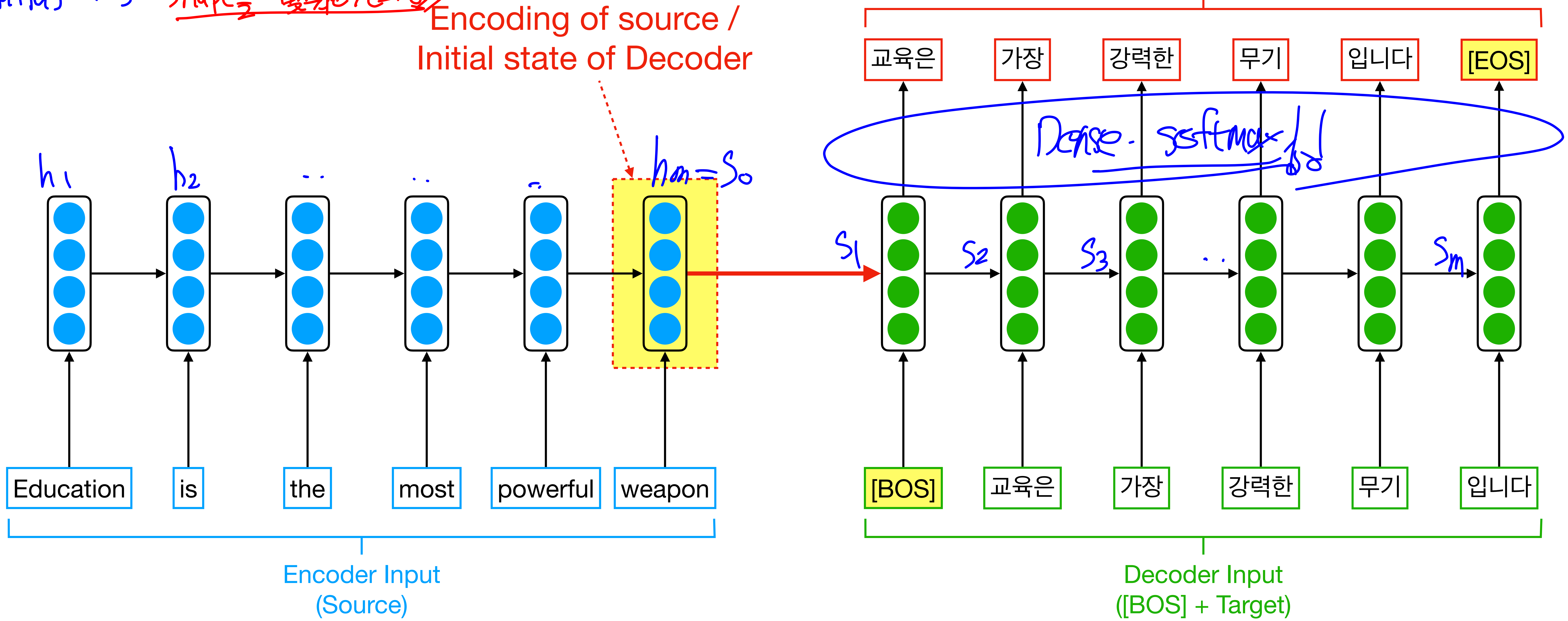
output

- Encoder (LSTM) → ① e-h
  ② e-h-state
  ③ e-c-state) concat ×2배로

(bs.5)
(bs.5)
(bs.5)) → concat ⇒ (bs,10)
= 'cat'이라 불러

∴ GRU (unit =10, inital_state = [cat])

Decoder (GRU) → shape를 맞춰줘야 해!

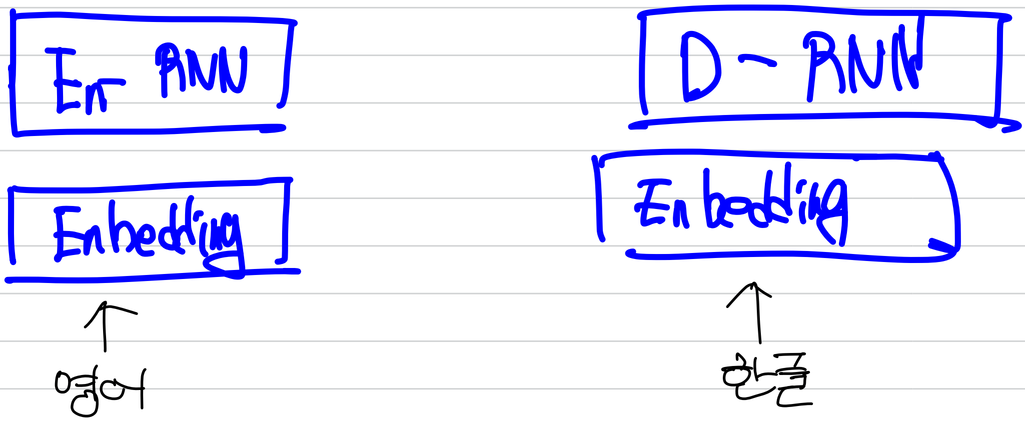


Encoding of source /
Initial state of Decoder

Decoder Label
(Target + [EOS])

Encoder Input
(Source)

Decoder Input
([BOS] + Target)

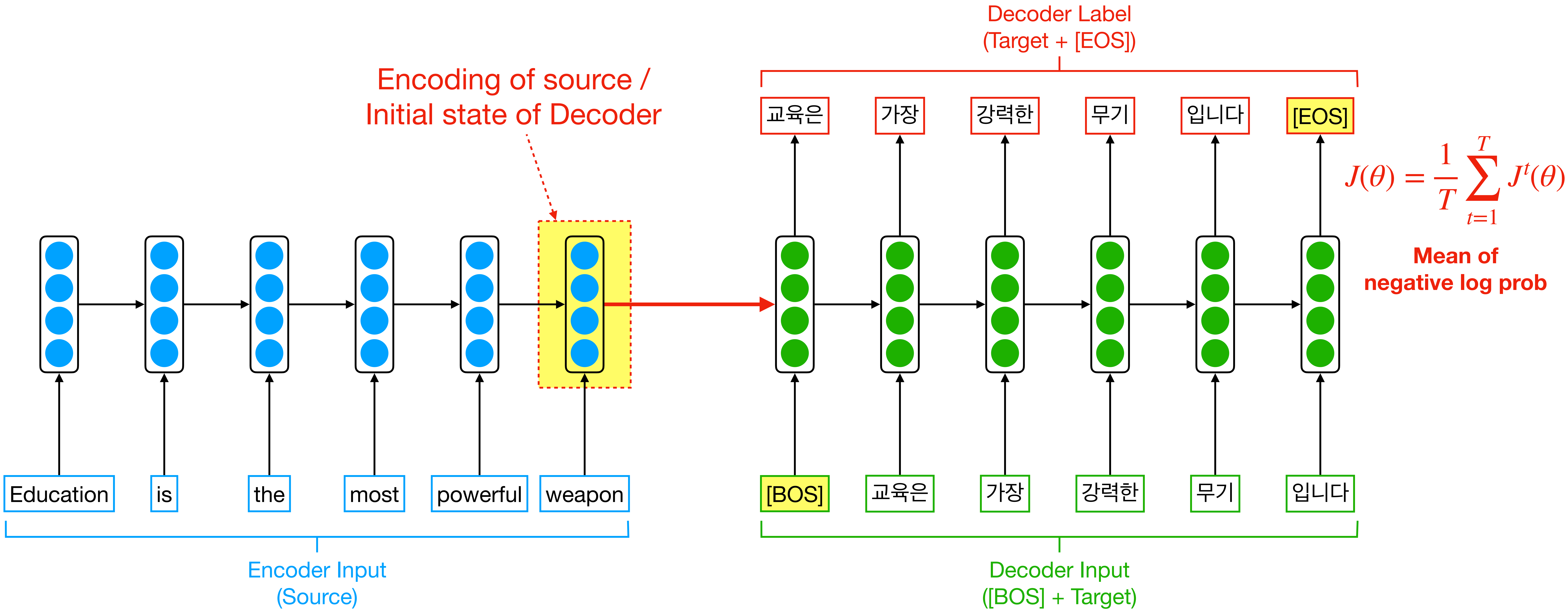


○ 영어, 한국 임베딩 나뉘는게 일반적

○ but) 한국, 한국임베딩 같은걸 사용!

그렇지 않으면 다른걸로 매칭되어서 학습될수있지.

Machine Translation Model (Training)



$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^t(\theta)$$

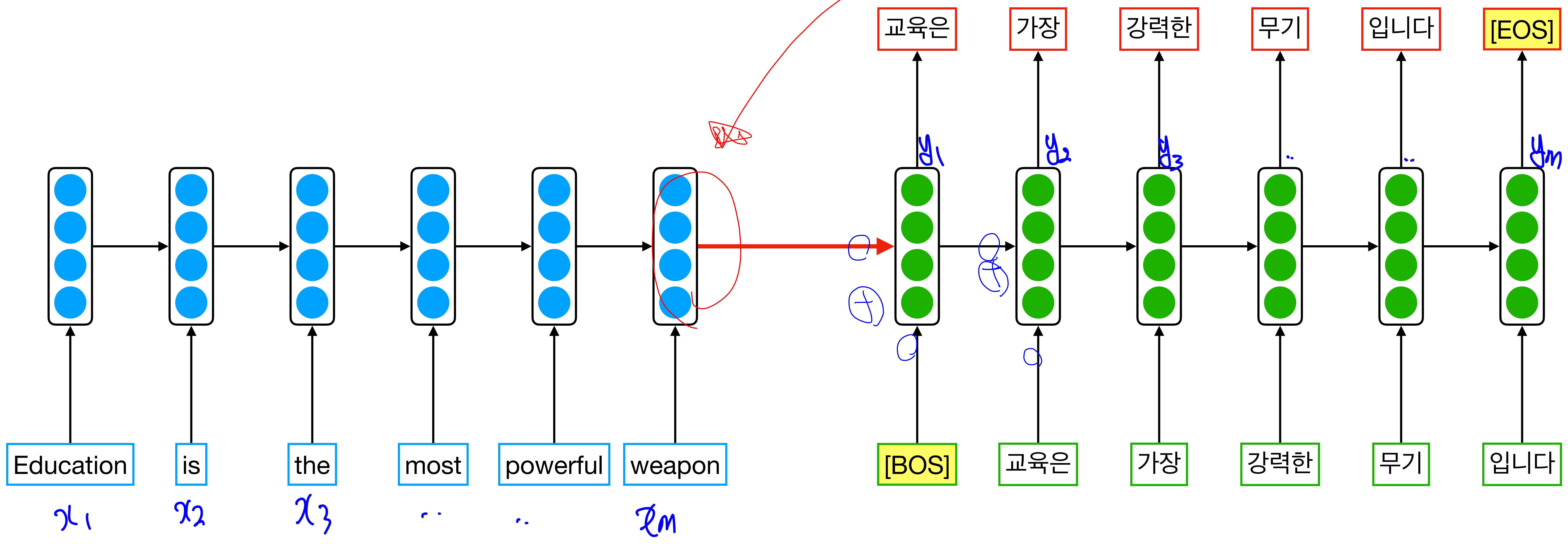
Mean of negative log prob

Machine Translation Model (Training)

output// input(2) 각각의 확률분포를 다 곱하기 Likelihood

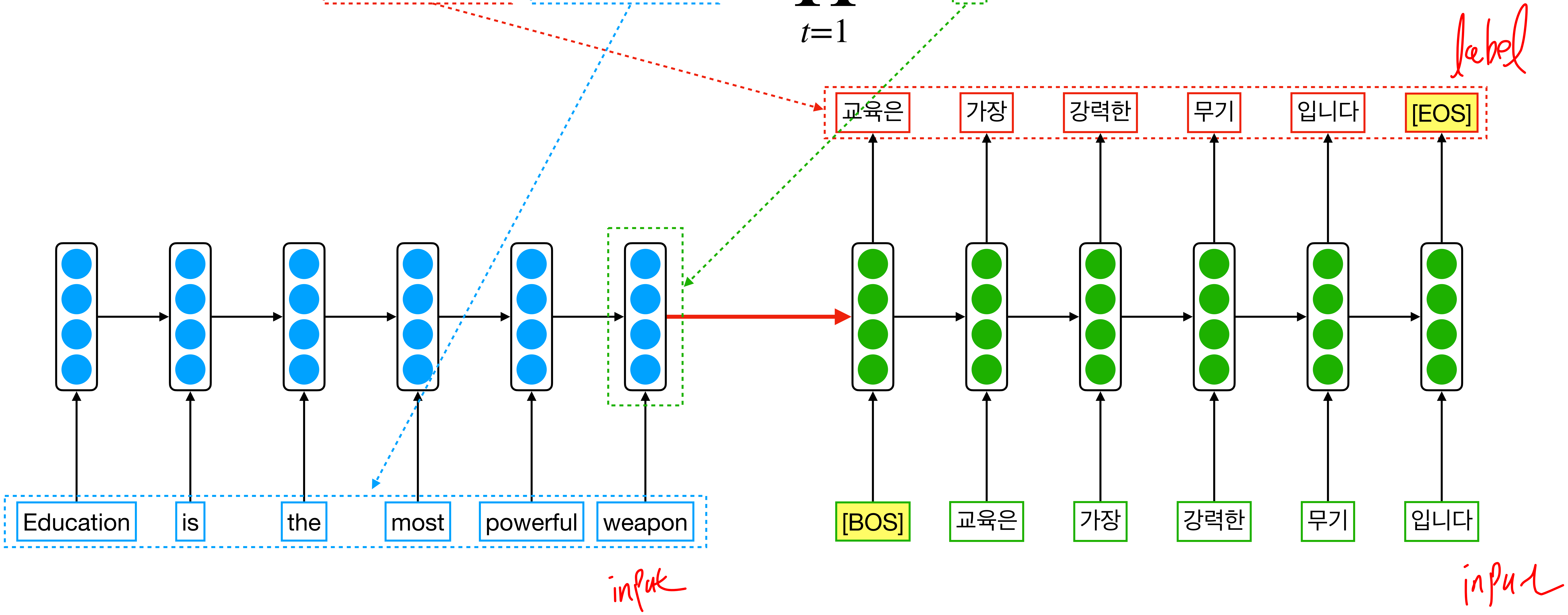
fixed vector

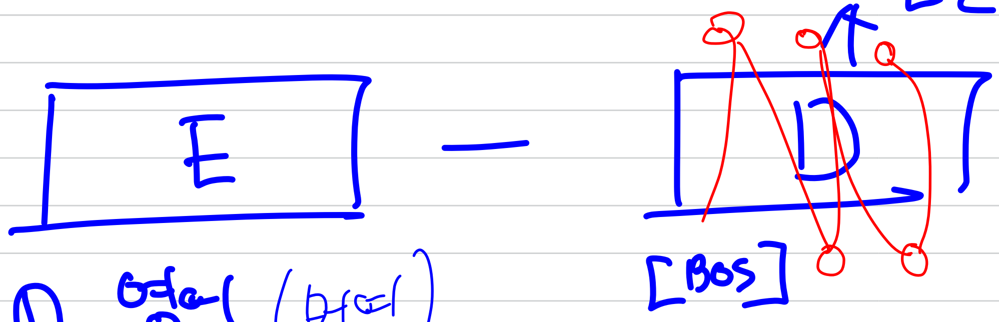
$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_m) = \boxed{\prod_{t=1}^{n} p(y_t \mid v, y_1, \ldots, y_{t-1})}$$



Machine Translation Model (Training)

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_m) = \prod_{t=1}^{n} p(y_t \mid v, y_1, \ldots, y_{t-1})$$





① 한글 ② Answer ③ Summery

④ 긍정 or 부정

① 영어 (번역)

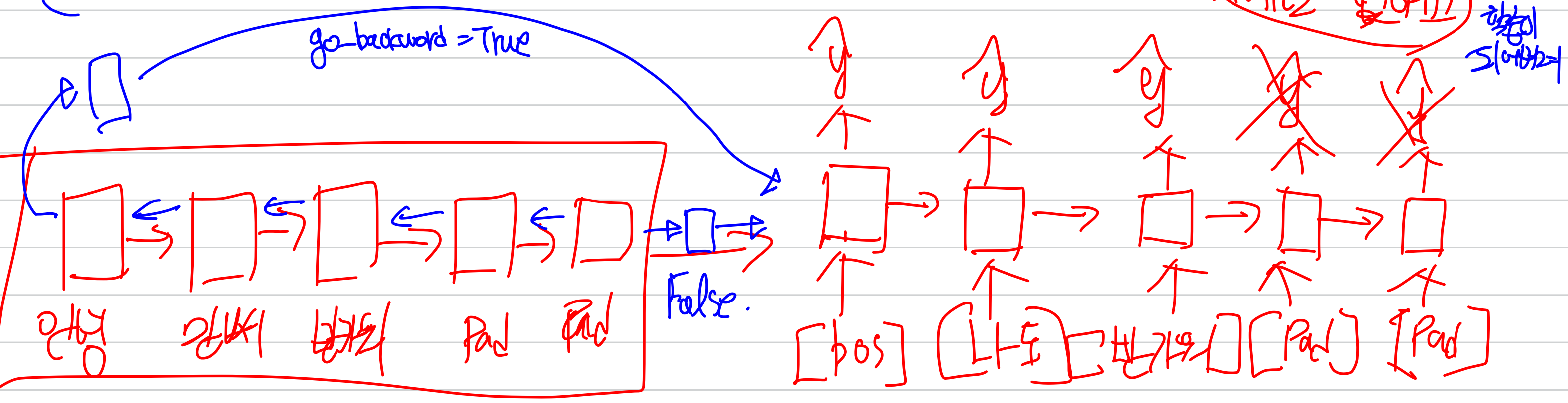
② Question (챗봇)

③ 긴 문장 (요약)

④ 문장

model.predict (en-input, dec-input)
[q-id] [q-id] batch-size=1 이니까

② 챗봇

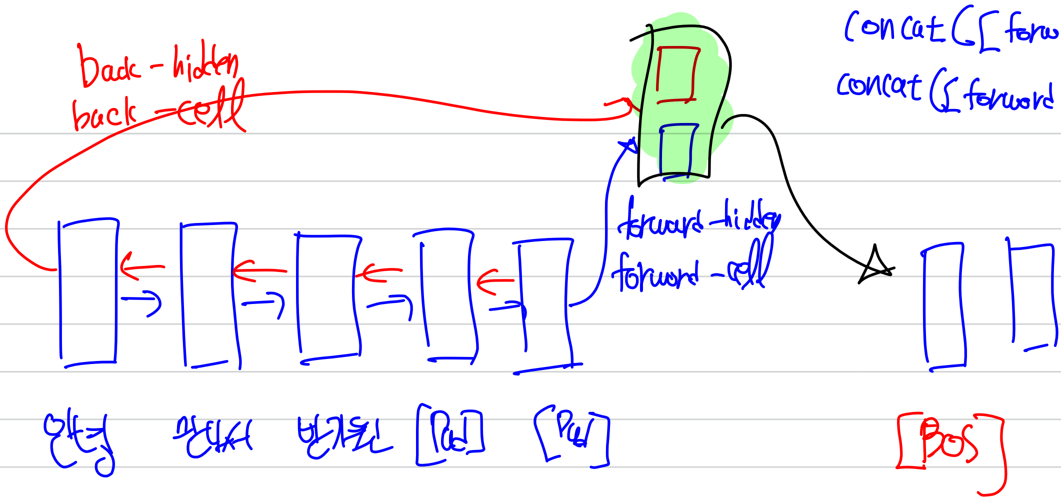


☆ 하다보니 토큰의 대부분이 Pad라 편향되어 학습이 되어버려

① ←← 인코더에서는 가능한가

① 디코더에서는 ←← 순서안된다

- 원래라는 Pad가 뒤에 있을테니 go-backword가 좋고
- 앞에서 Pad를 생성한다면 go-backword=False 가 좋고



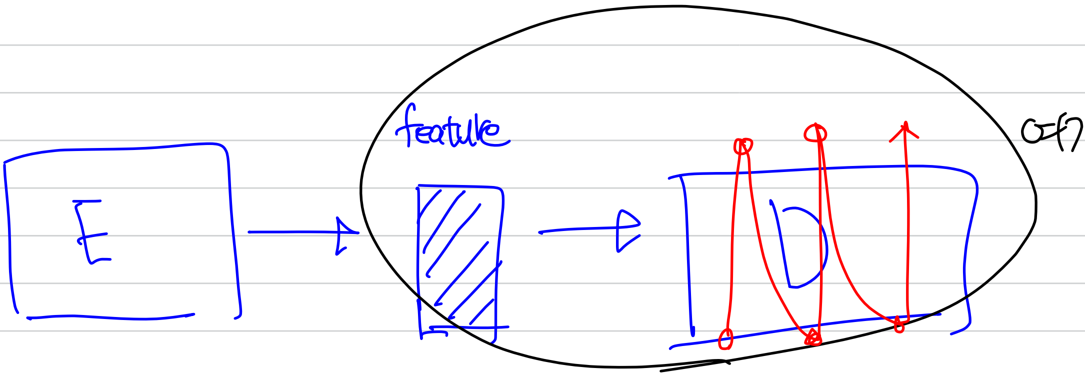
concat([forward-hidden, back-hidden])
concat([forward-cell, back-cell])

○ shape [bs, d-model*2], [x1], [x1], [x1], [x1]

○ concat 된거 아예 출력해버려

○ encoder는 양방향 가능

○ decoder는 단방향 만. (forward_)
- go-back하면 답보고 하는 꼴이 되잖아.



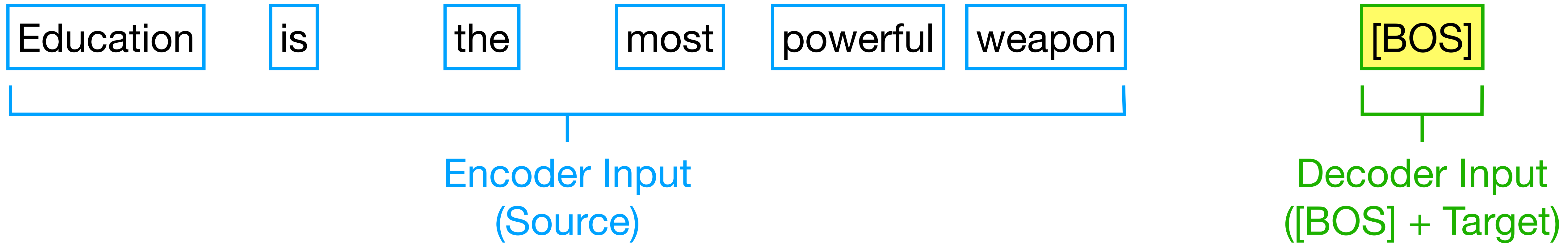
여기서 반복.

○ 문장이 긴 경우마다 E에서 feature를 추출하기

○ but) example에서는 짧은 문장, 간단한 것이니

D의 input이 새로 생성될때마다.
E를 다시, 매번 feature 추출을 함!

Machine Translation Model (Inference)



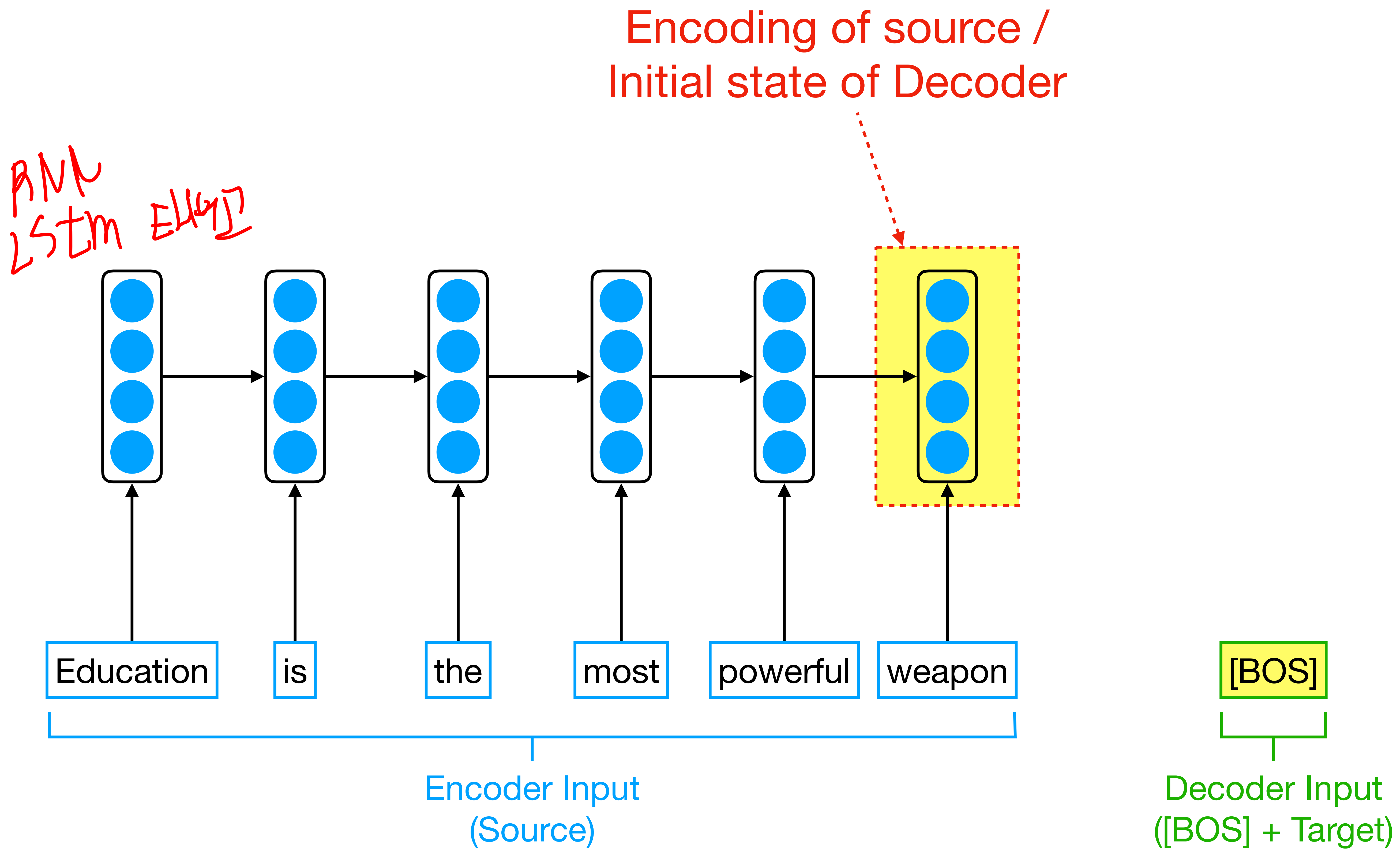
Education is the most powerful weapon

Encoder Input
(Source)

[BOS]

Decoder Input
([BOS] + Target)

Machine Translation Model (Inference)



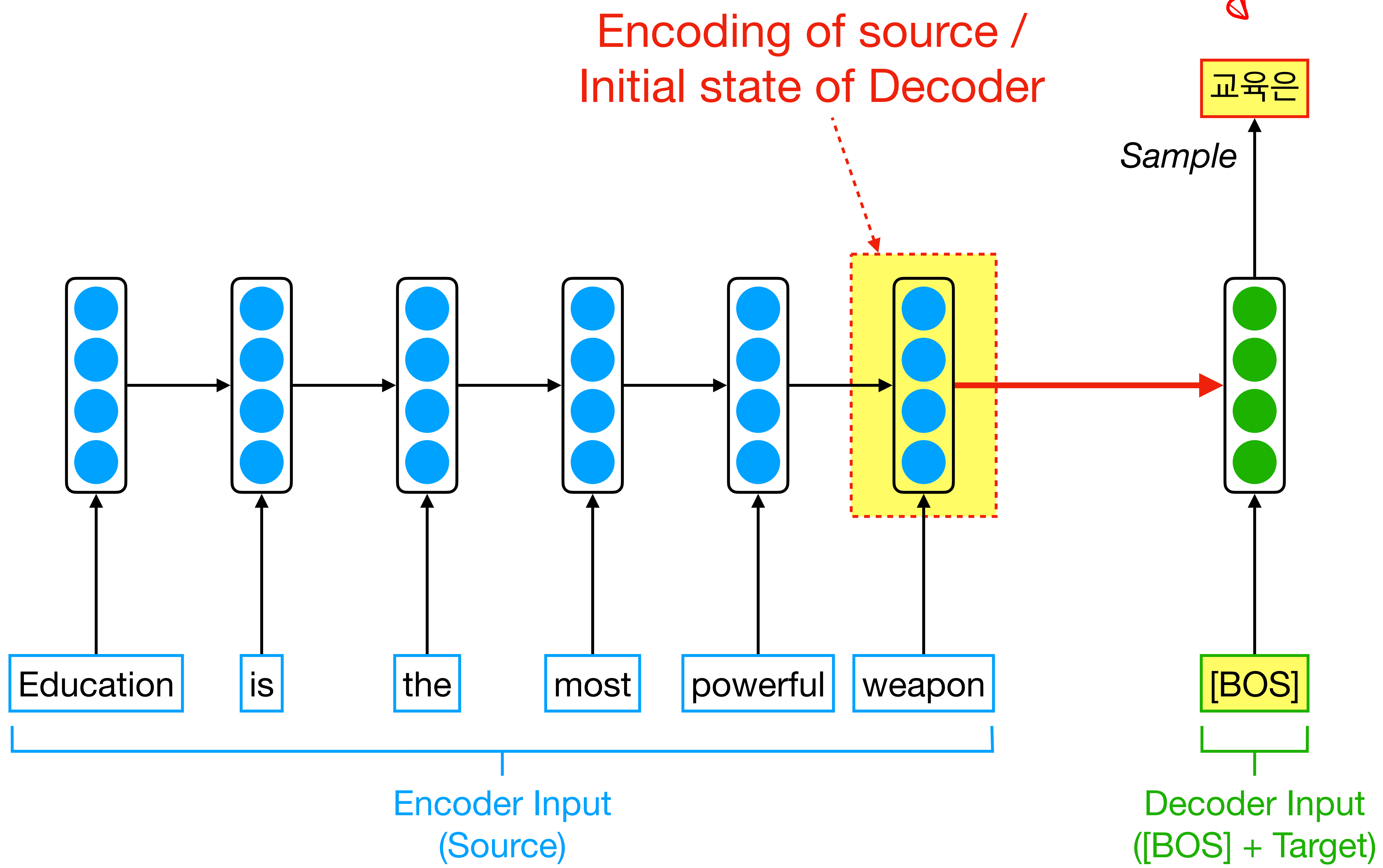
Machine Translation Model (Inference)

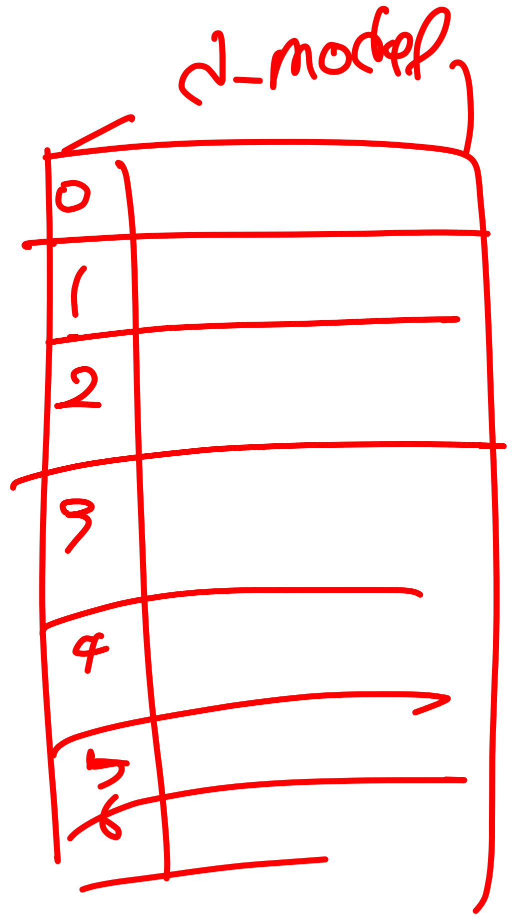
greedy

random(P)

(b,s) 1, [3,5,4]

(bs=1) n-seq



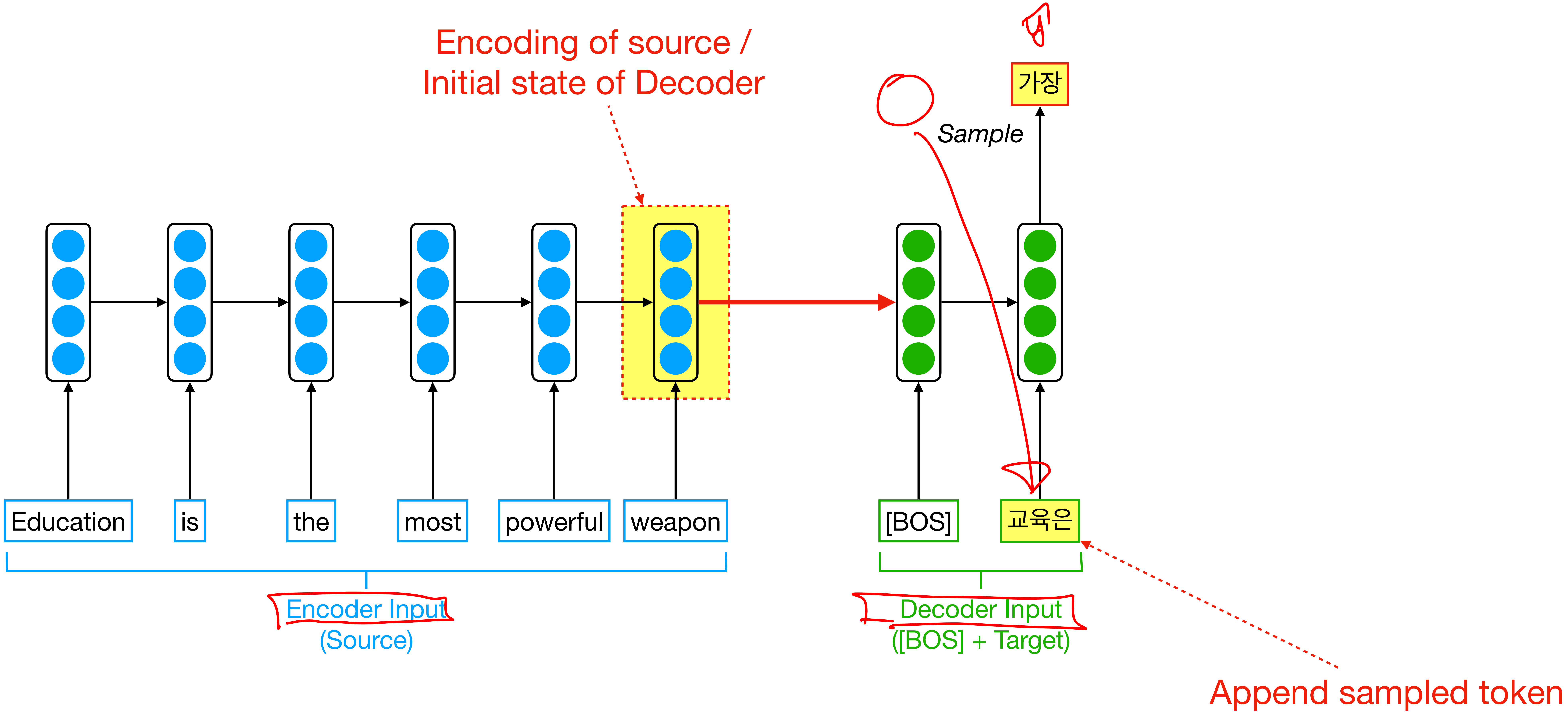


개너기!//

[[3,0,0,0] [5,0,0,0] [4,0,0,0]]

[1, 3, d_model]

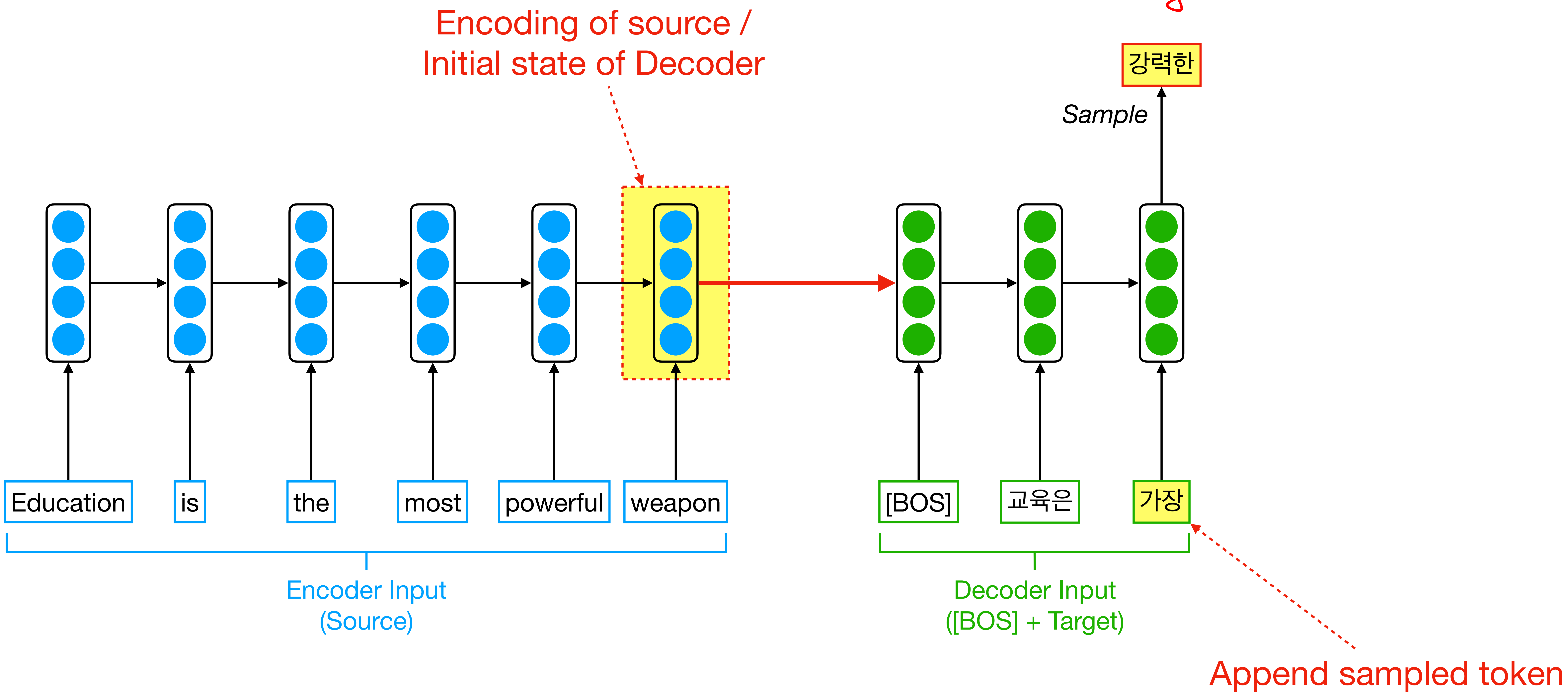
Machine Translation Model (Inference)



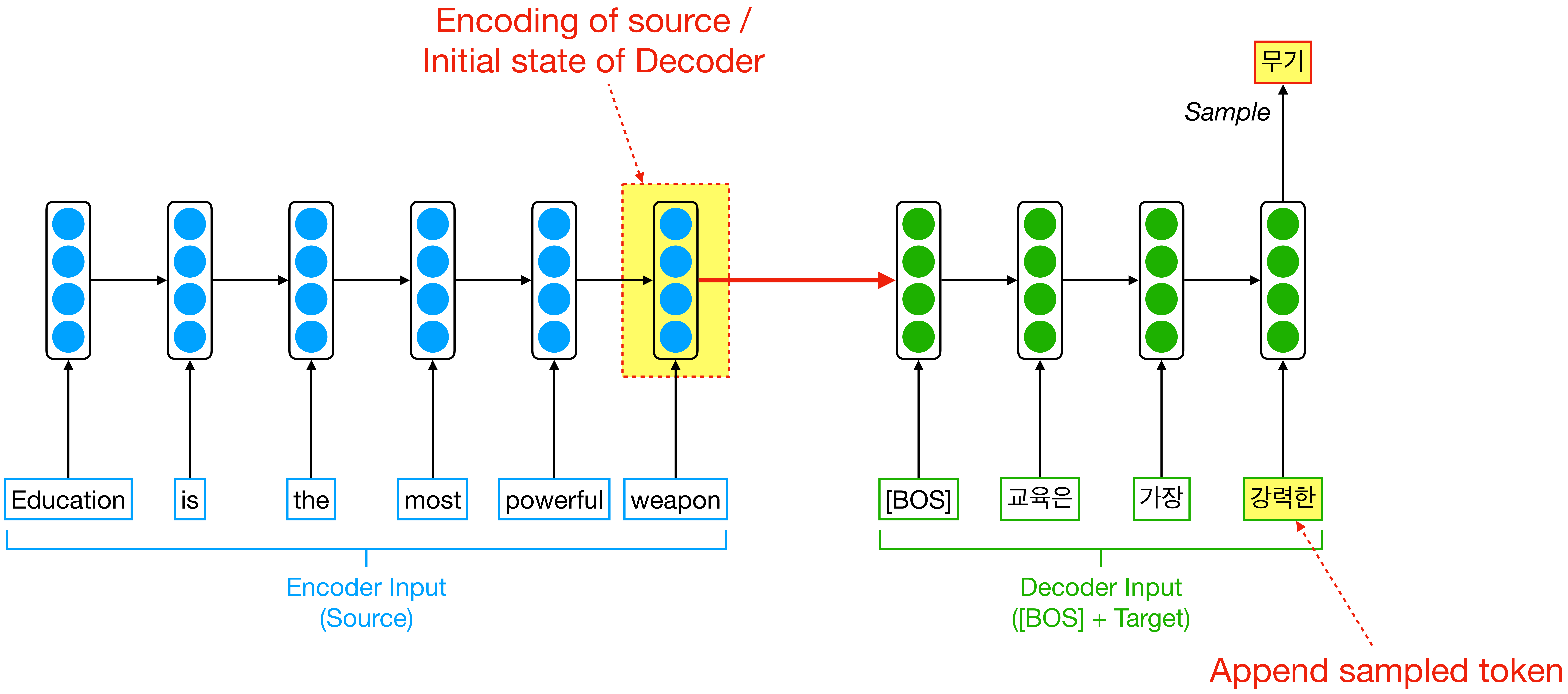
Encoding of source / Initial state of Decoder

Append sampled token

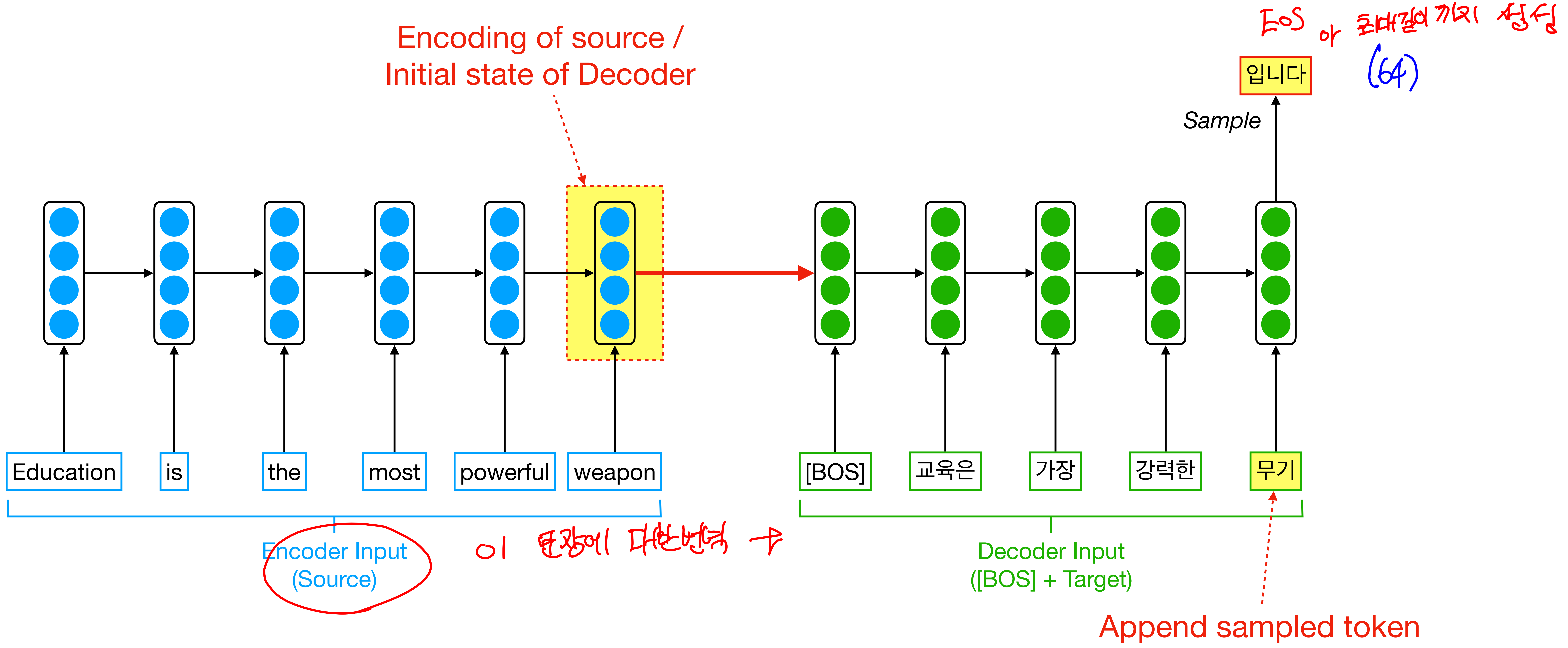
Machine Translation Model (Inference)



Machine Translation Model (Inference)



Machine Translation Model (Inference)



Encoding of source /
Initial state of Decoder

Encoder Input
(Source)

Decoder Input
([BOS] + Target)

Append sampled token

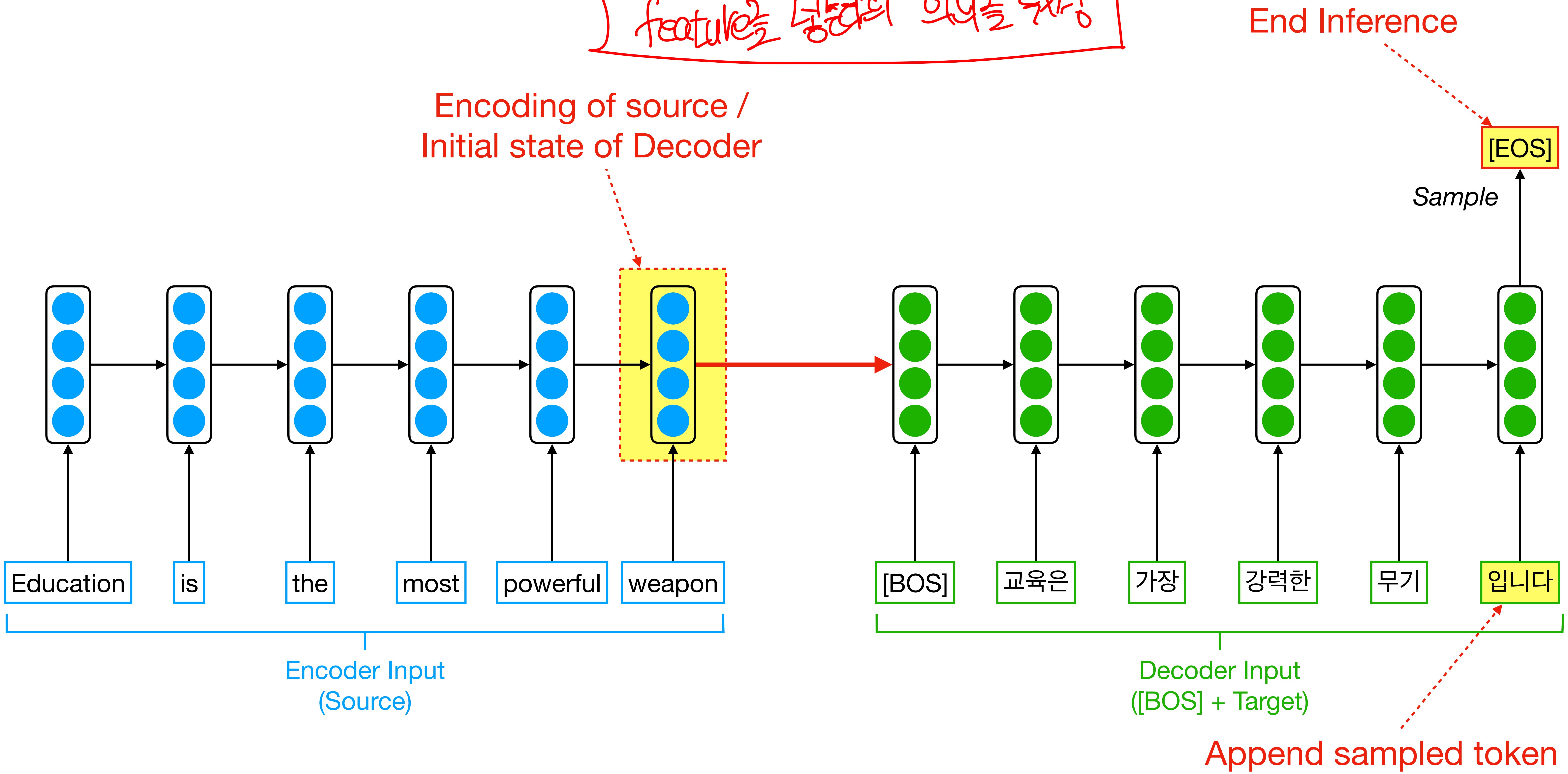
EOS 아 최대길이까지 생성
(64)

이 문장에 대한 번역 →

Machine Translation Model (Inference)

CNN, Lstm, maxpooling 등!

feature를 넣는데의 의의를 두자!



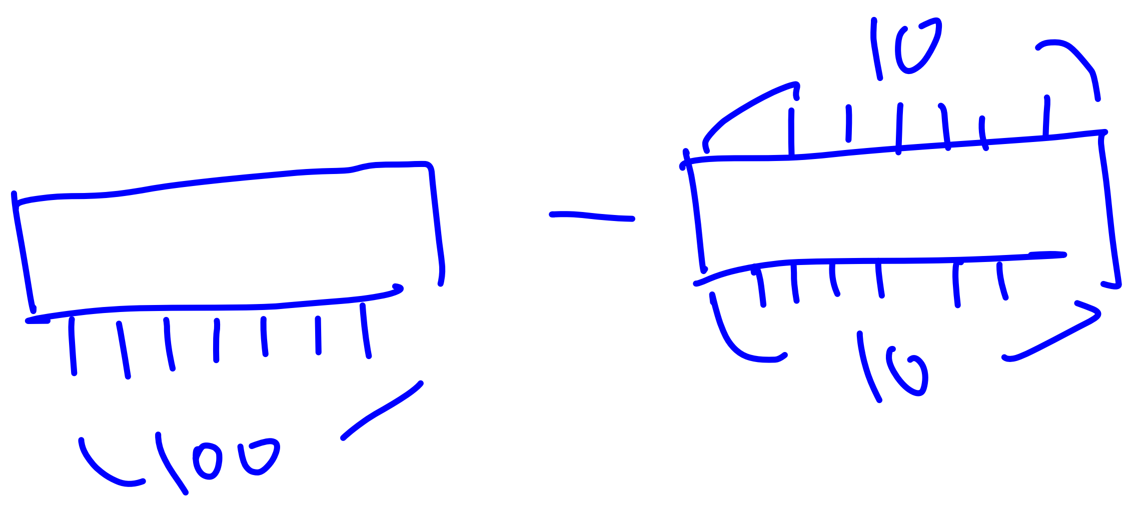
Machine Translation Model (Versatile)

이순신이 누구야

→ 그사람 고향은

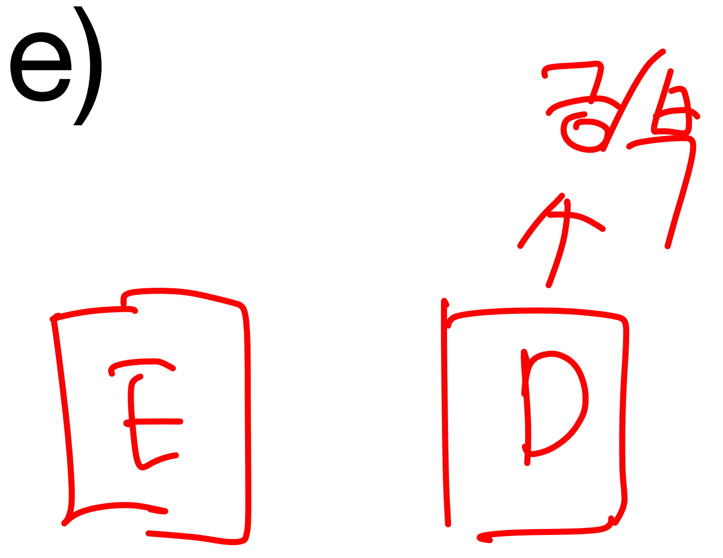
multi turn

- Summarization (long text → short text) - 요약문
- 잘채 → Dialog (user utterance → agent utterance) - 채팅
- Parsing (text → parsed sequence) → 검색
- Code generation (text → program code)
- etc.



- sequence size 다르게 가능
-

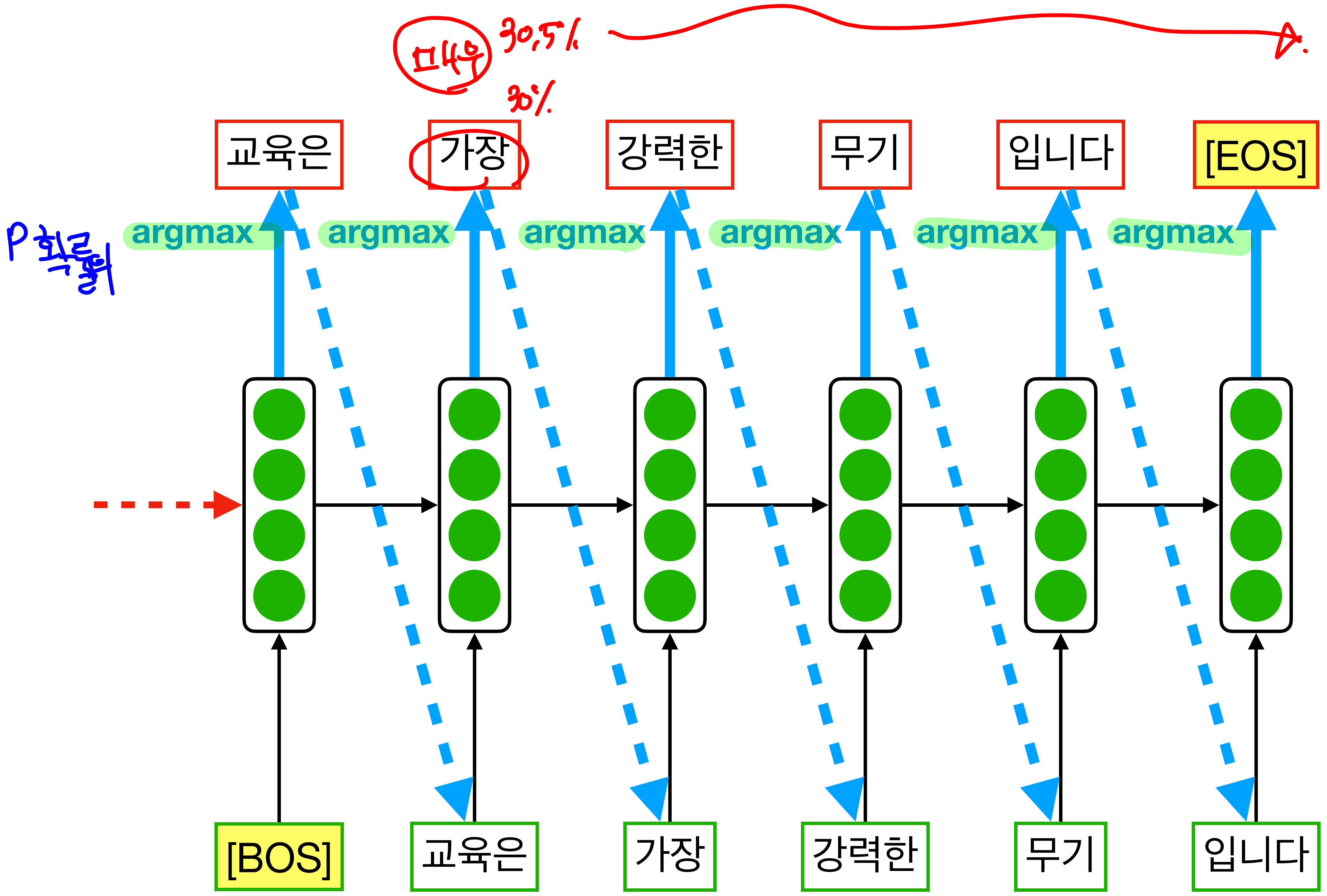
나는 개념 →



Machine Translation Model (Decoding)

argmax()3//

① greedy search의 문제점..



Greedy Search

Top 1만//

Machine Translation Model (Decoding)

$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \boxed{\log} P_{LM}(y_i \mid y_1, \ldots, y_{i-1})$$

$\pi \cdot P_{L/m}$

[BOS]

(k=5)까지 함

~

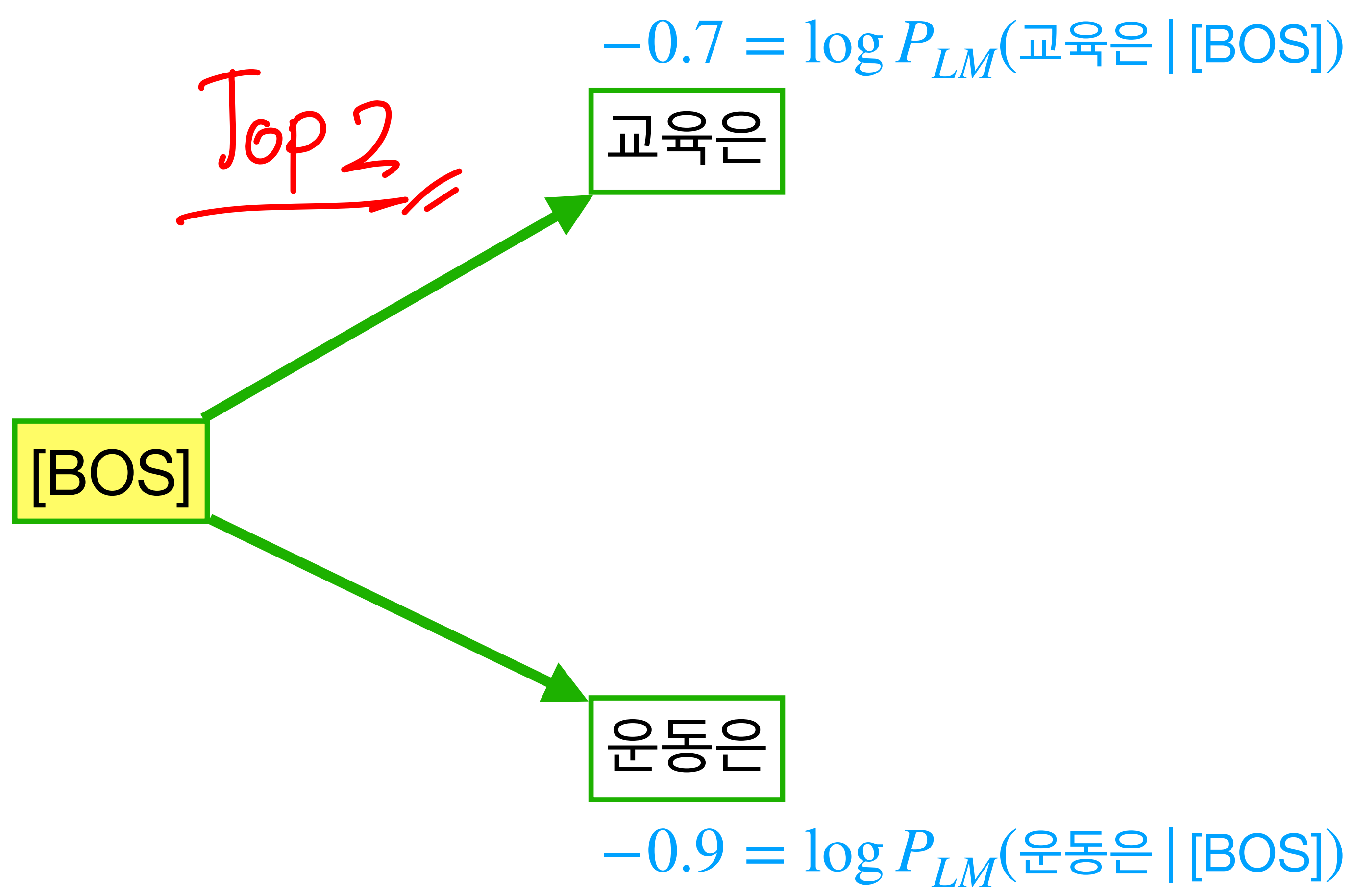
○ 문장을 2개로 뽑고
그중에 좋은 걸로 하지~

Beam Search (k=2)

Machine Translation Model (Decoding)

$\prod_{i=1}^{t} P \Rightarrow$ P를 곱하기

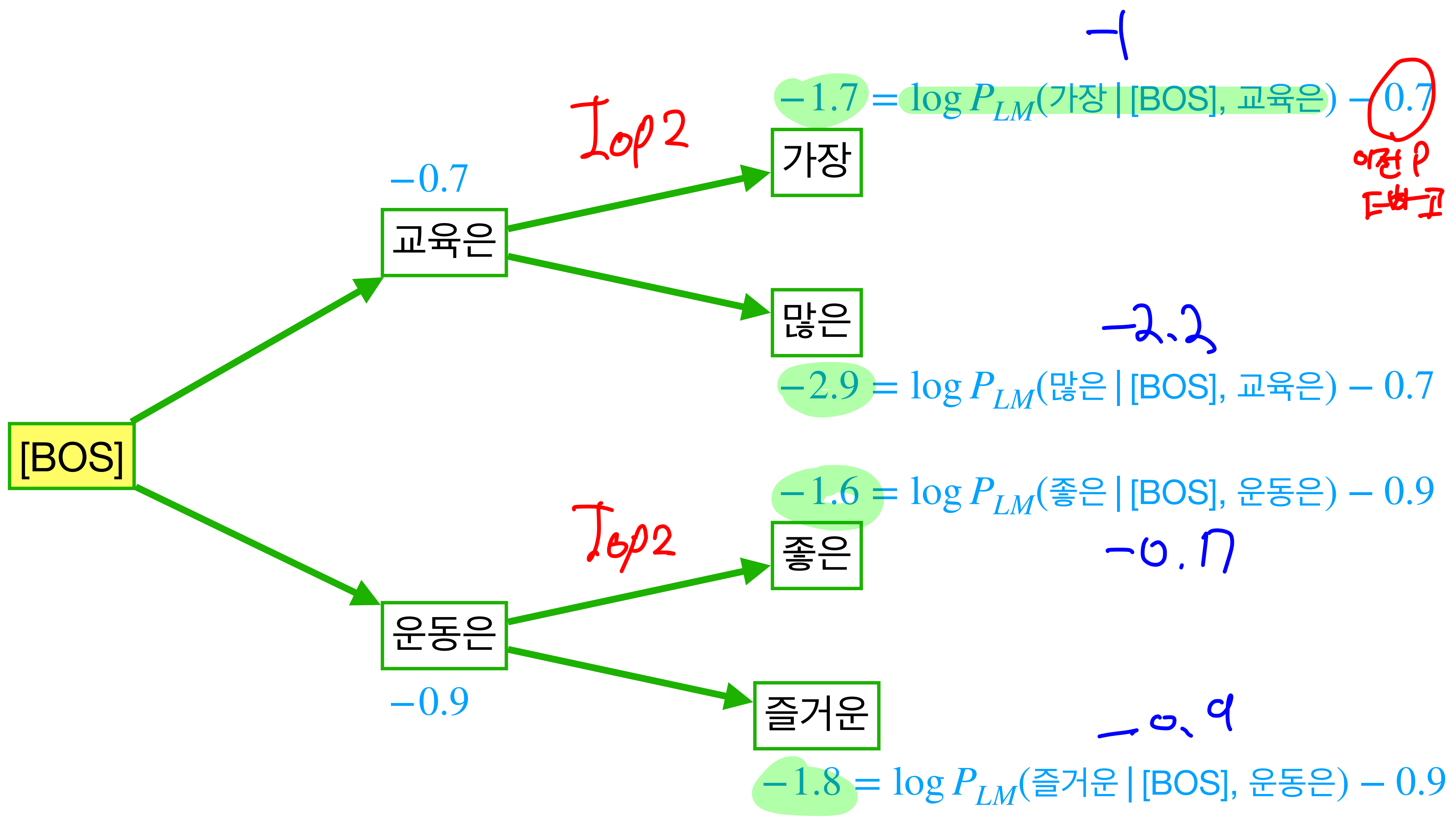
$$score(y_1, \dots, y_t) = \underbrace{\sum_{i=1}^{t}}_{\text{더하기}} \log P_{LM}(y_i \mid y_1, \dots, y_{i-1})$$


$-0.7 = \log P_{LM}$(교육은 | [BOS])
Top 2
교육은
[BOS]
운동은
$-0.9 = \log P_{LM}$(운동은 | [BOS])

Beam Search (k=2)

Machine Translation Model (Decoding)

$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i \mid y_1, \ldots, y_{i-1})$$



$-1.7 = \log P_{LM}(\text{가장} \mid \text{[BOS], 교육은}) - 0.7$

$-2.9 = \log P_{LM}(\text{많은} \mid \text{[BOS], 교육은}) - 0.7$

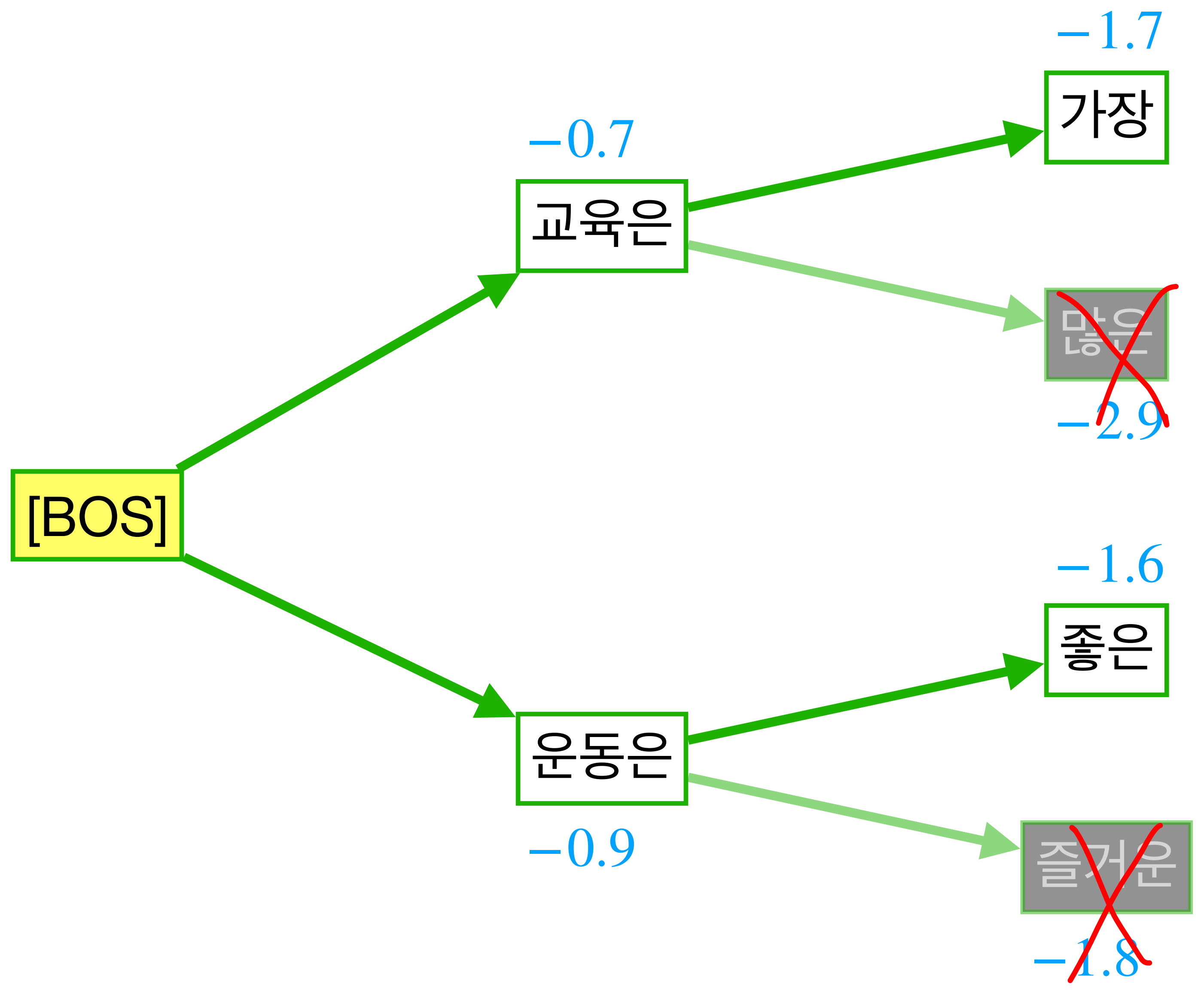
$-1.6 = \log P_{LM}(\text{좋은} \mid \text{[BOS], 운동은}) - 0.9$

$-1.8 = \log P_{LM}(\text{즐거운} \mid \text{[BOS], 운동은}) - 0.9$

Beam Search (k=2)

Machine Translation Model (Decoding)

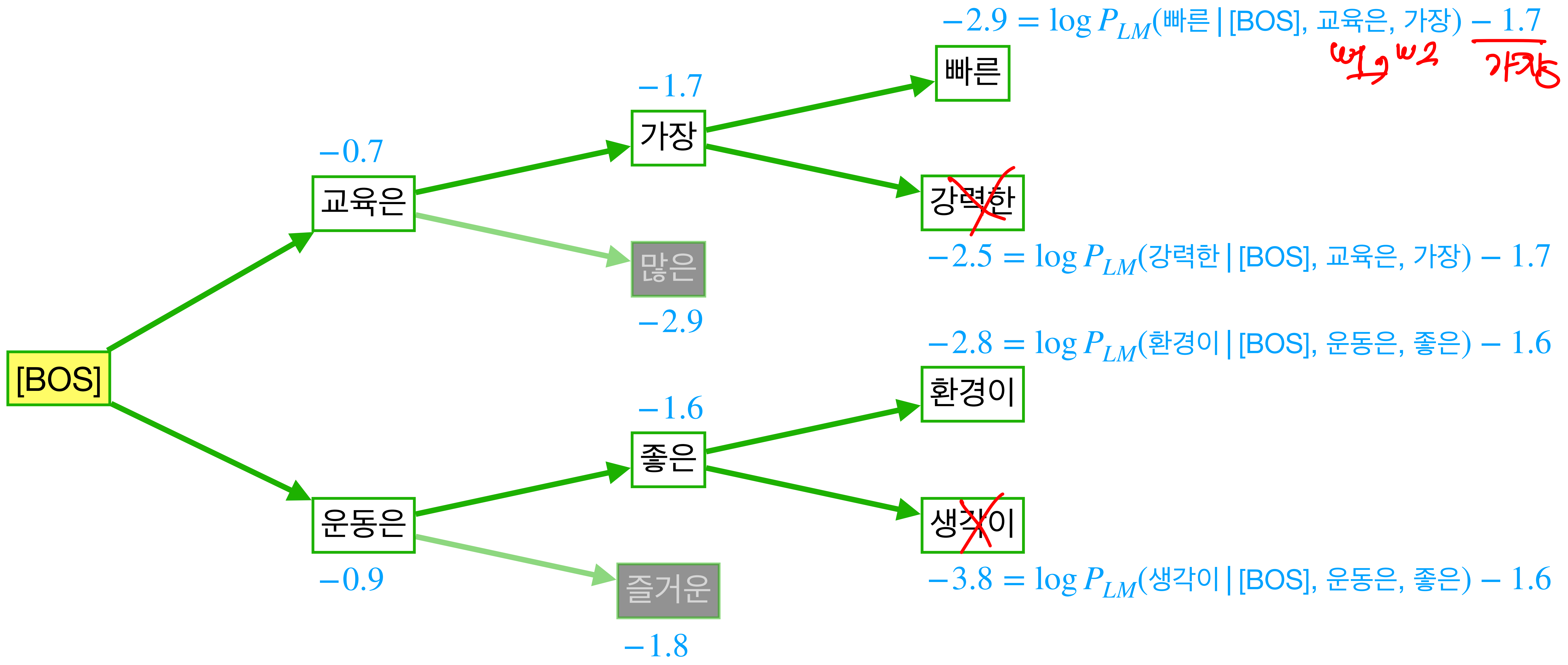
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



$-2.9 = \log P_{LM}$(빠른 | [BOS], 교육은, 가장) $\underline{-\ 1.7}$

$-2.5 = \log P_{LM}$(강력한 | [BOS], 교육은, 가장) $-\ 1.7$

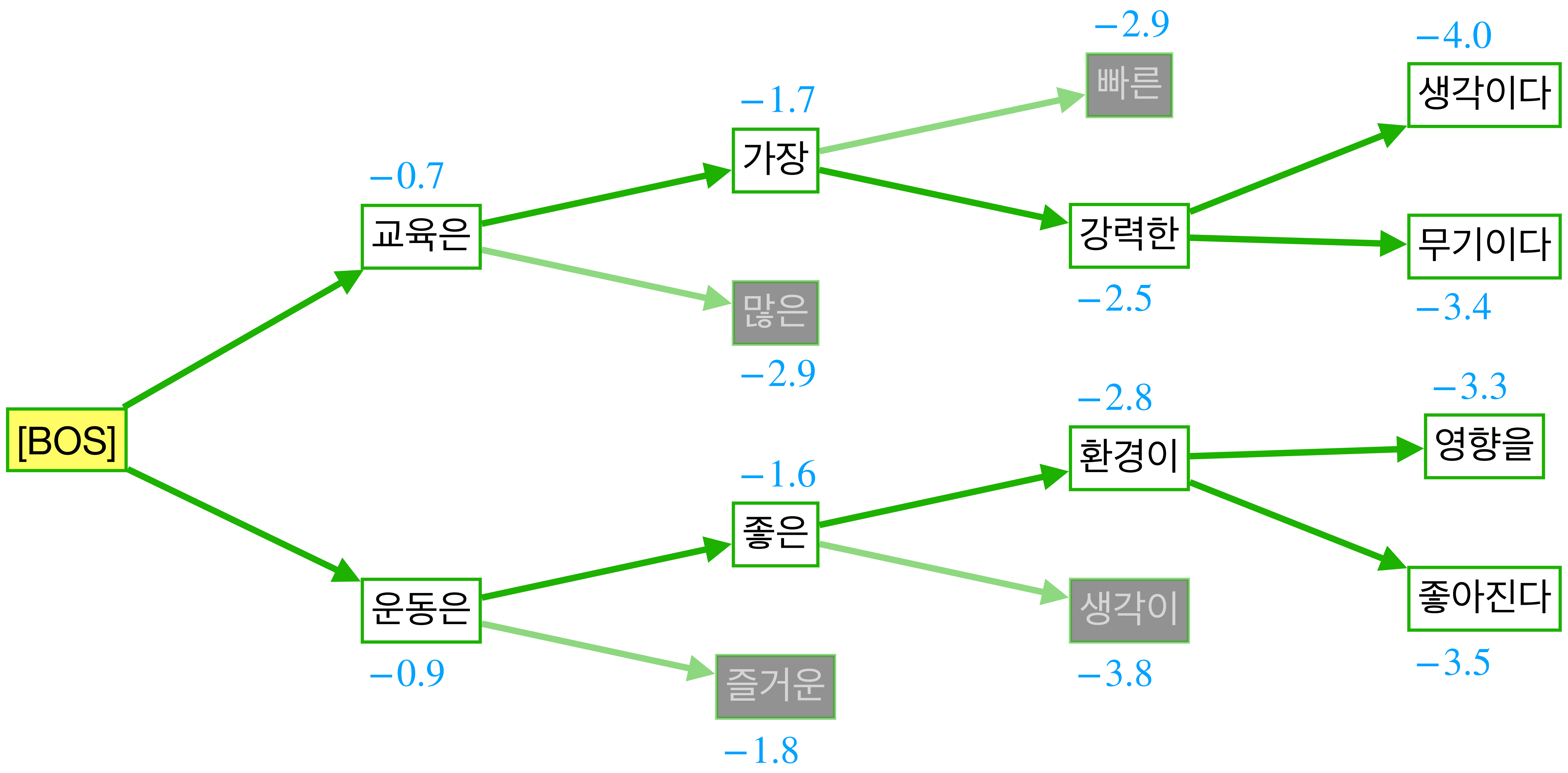
$-2.8 = \log P_{LM}$(환경이 | [BOS], 운동은, 좋은) $-\ 1.6$

$-3.8 = \log P_{LM}$(생각이 | [BOS], 운동은, 좋은) $-\ 1.6$

Beam Search (k=2)

Machine Translation Model (Decoding)

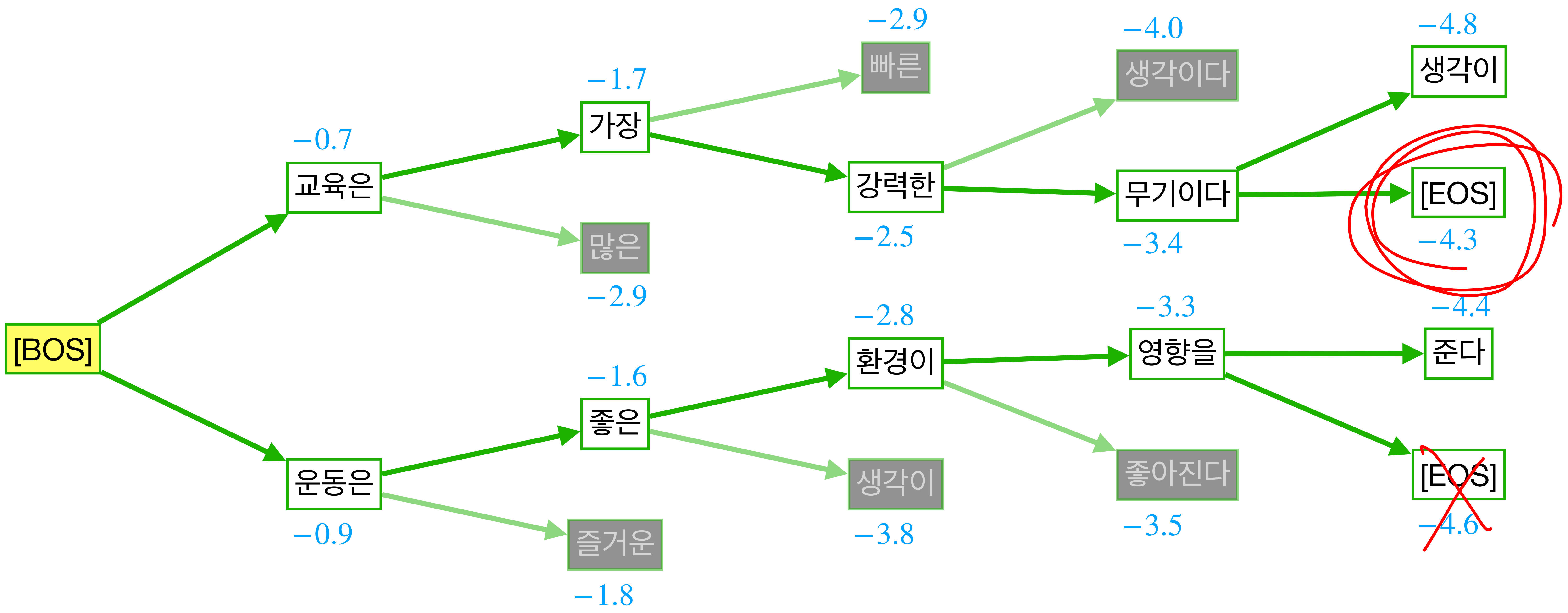
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

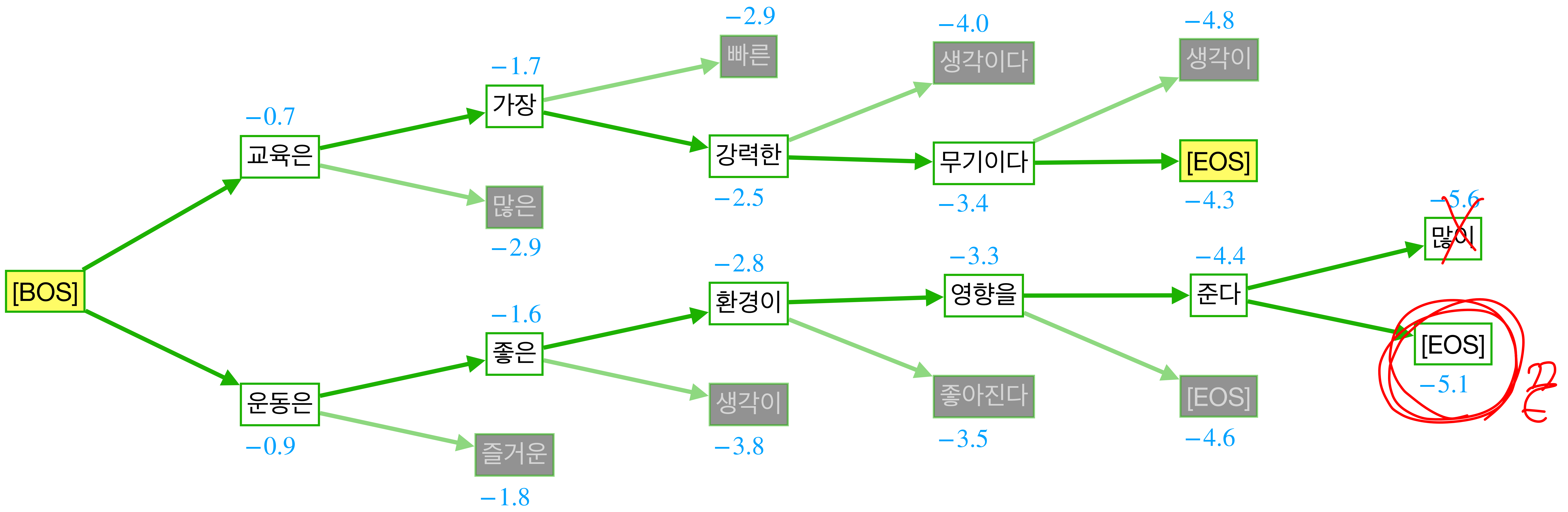
$$score(y_1, ..., y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, ..., y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

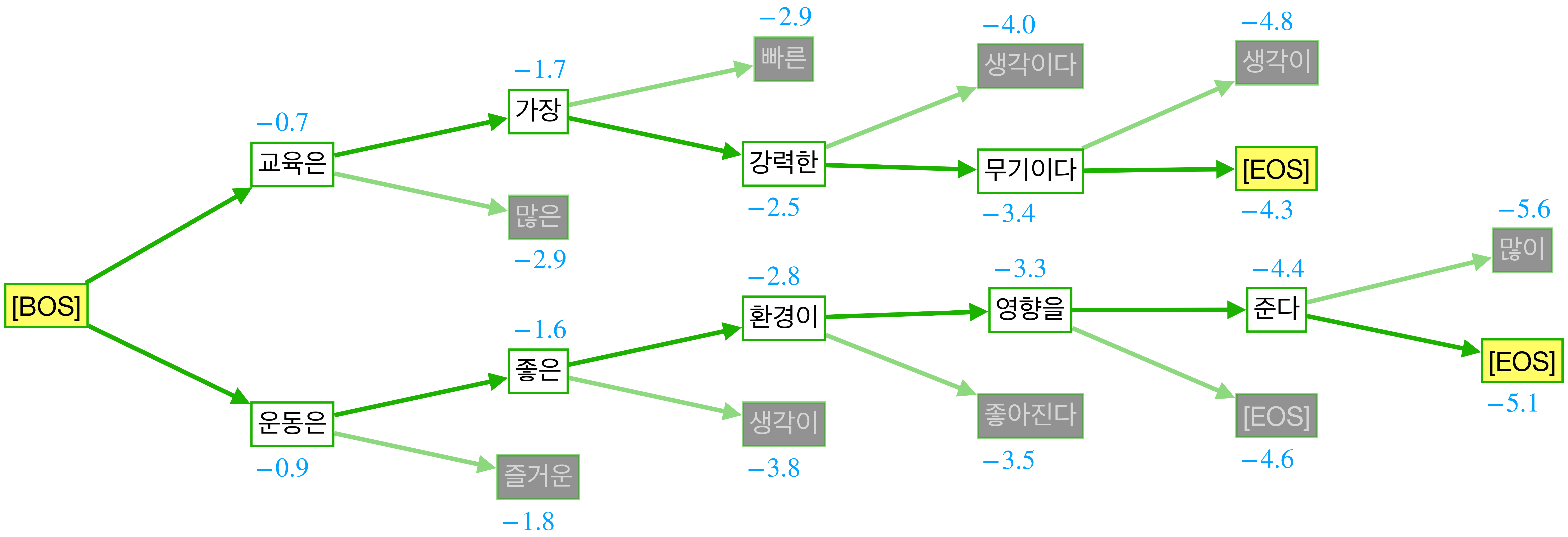
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i \mid y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



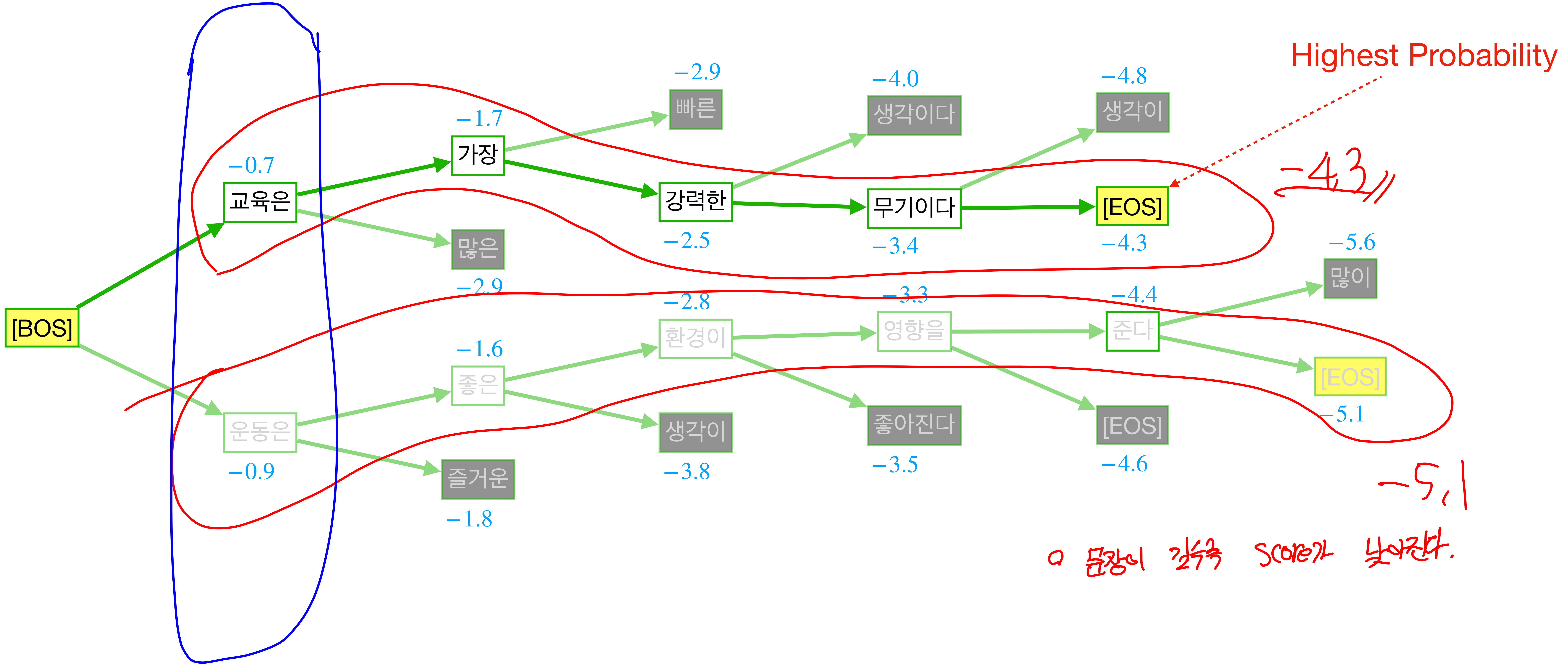
Beam Search (k=2)

Machine Translation Model (Decoding)

0 1 이때 -∞. 에까지 나와!!

32000개를 다 볼순 없으니까

$$score(y_1, \dots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i \mid y_1, \dots, y_{i-1})$$



○ 문장이 길수록 score가 낮아진다.

Beam Search (k=2)

Machine Translation Model (Decoding)

Length Penalty

- Longer hypotheses have lower score 길어질수록 더해지니까 score가 더 낮아지는거지 -

$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{n-1})$$

⇒ 그걸 짧은게 무조건 좋게봐

normalize 해주자!

- Normalize by length

$$score(y_1, \ldots, y_t) = \boxed{\frac{1}{t}} \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{n-1})$$

단어의 길이만큼 평균 해주면 해결 하는거지!

Beam Search (k=2)

Machine Translation Model (Metric)

Education is the most powerful weapon we can use to change the world.

↓

교육은 세상을 바꿀 수 있는 가장 강력한 무기이다. ✓

세상을 바꿀 수 있는 가장 강력한 무기는 교육이다. ✓

가장 강력한 무기인 교육을 통해 세상을 바꿀 수 있다. ✓

=> Q 어떻게 평가할건데?

How do you evaluate???

뭐가 좋은지 알수없잖아!

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

candidate　　　　　　　　　　　　　reference

- Machine Translation된 결과와 사람이 Translation한 결과를 비교하여 품질을 평가

True

  - n-gram precision) n-gram 단위로.
  - penalty for too-short system translations) → 짧은 문장에는 페널티 주겠다.

$$\frac{TP}{TP+FP} \Rightarrow \frac{\text{실제 정답.}}{\text{정답이라고 예측한것중에}}$$

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

N-gram precision

18개 17개 맞춤

생성값

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party. len ⇒ 18

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct. len ⇒ 14

정답값

14개중 8개

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

○ 번역 문장이다 3개를 만들어 놓은거임.

$$\text{Candidate 1 Unigram Precision} = \frac{17}{18}$$

$$\text{Candidate 2 Unigram Precision} = \frac{8}{14}$$

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

N-gram precision

Candidate: the the the the the the the.

Reference 1: The cat is on the mat. =2개

Reference 2: There is a cat on the mat. =1개

max => 2

$$Count_{clip} = \min(\overset{\text{7개}}{Count}, \overset{\text{2개}}{Max\_Ref\_Count})$$

2 = (7, 2)

$$\text{Candidate Unigram Precision} = \frac{7}{7}$$

$$\text{Candidate Modified Unigram Precision} = \frac{2}{7}$$

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

N-gram precision

1
2
3
4
:
5

$$p_n = \frac{\overbrace{\sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}^{\text{Candidate의 n-gram의 } Count_{clip}\text{개수}}}{\underbrace{\sum_{n\text{-}gram' \in C} Count(n\text{-}gram')}_{\text{Candidate의 n-gram 개수}}}$$

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

Penalty for too-short system translations

Candidate: of the 짧게 예측.

○사실상 ○점짜리 번역이지 => 근데 bleu가 만점//

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

$$\text{Candidate Unigram Precision} = \frac{2}{2}$$

당연히 Precision이 높아지지. Panelty를 주자!

짧은 문장일수록 n-gram precision이 높아지는 경향이 있음

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

Penalty for too-short system translations

Candidate: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party. len => 16

candidate > reference 중 가장 작은 len

brevity penalty

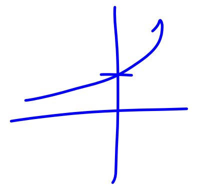
$$BP = \begin{cases} 1, & \text{if } c > \overset{16}{r} \\ \underline{exp(1 - \frac{r}{c})}, & \text{if } c \le r \end{cases}$$

= 음수이니까.

ex) $exp(1 - \frac{16}{4})$

$= exp(-3)$

< 1



BP = $exp(1 - \frac{r}{c})$는 $0 < x < 1$ 이지 => 페널티

짧은 문장일수록 n-gram precision이 높아지는 경향이 있음

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

gram	w	w
1	0.25	0.125
2	0.25	0.125
3	0.25	0.25
4	0.25	0.5

gram

○ 보통 $n=4$, $W_n = \frac{1}{N}$ $= \frac{1}{4}$ $= 0.25$

$$p_n = \frac{\sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram' \in C} Count(n\text{-}gram')}$$

$$\boxed{BP = \begin{cases} 1, & \text{if } c > r \\ exp(1 - \frac{r}{c}), & \text{if } c \leq r \end{cases}}$$

Sum

zip

$$BLEU = \underbrace{BP}_{=\text{blue Penelty}} \cdot exp(\boxed{\sum_{n=1}^{N}} \boxed{w_n} \boxed{\log p_n})$$

log 씌운

$$\log BLEU = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

Baseline: $N = 4, \quad w_n = \frac{1}{N}$

(0.25 0.25 0.25 0.25)

unit bi thr half

@2021 cchyun@gmail.com., Ltd All Rights Reserved.

참고: https://www.aclweb.org/anthology/P02-1040.pdf

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

Candidate: 나는 어제 집에 가서 잠을 잤다

↕ ⓧ 번역 잘안했지?

Reference: 나는 어제 집에 가서 잠을 설쳤다

Metric도
왜 자꾸 count 기반을 할까요
=

- BLEU는 유용한 지표지만 완벽하지 않음
- 그럼에도 → BLEU가 높으면 번역의 품질이 좋을 가능성이 높음
- 통계적인 지표

평가하는 방법이
마땅치 않다가.

Metric은
연구하는 분야가 따로
있음.

참고: https://www.aclweb.org/anthology/P02-1040.pdf

감사합니다.