삼성전기 AI전문가 양성과정 - 프로젝트 실습 (비영상)

자연어처리를 위한 Machine Translation

현청천

2022.02.28

What is Machine Translation

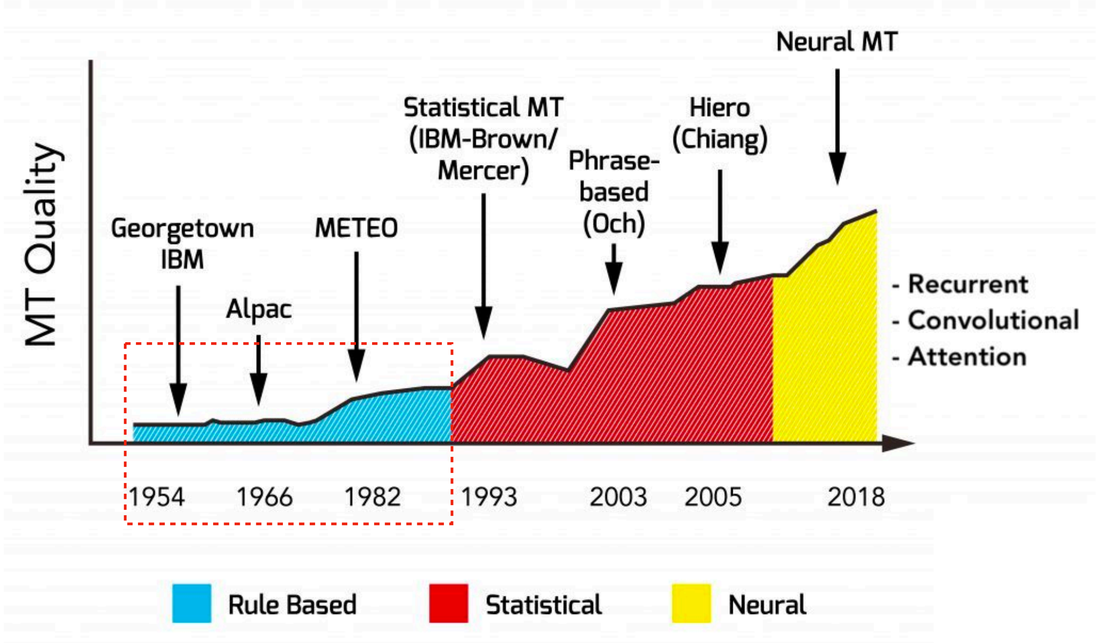
인간이 사용하는 자연 언어를 컴퓨터를 사용하여 다른 언어로 번역하는 것

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

교육은 세상을 바꿀 수 있는 가장 강력한 무기이다.

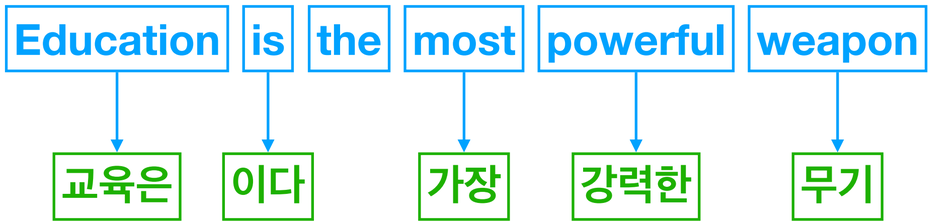
출처: https://ko.wikipedia.org/wiki/%EA%B8%B0%EA%B3%84_%EB%B2%88%EC%97%AD

What is Machine Translation (history)



Rule-based Machine Translation

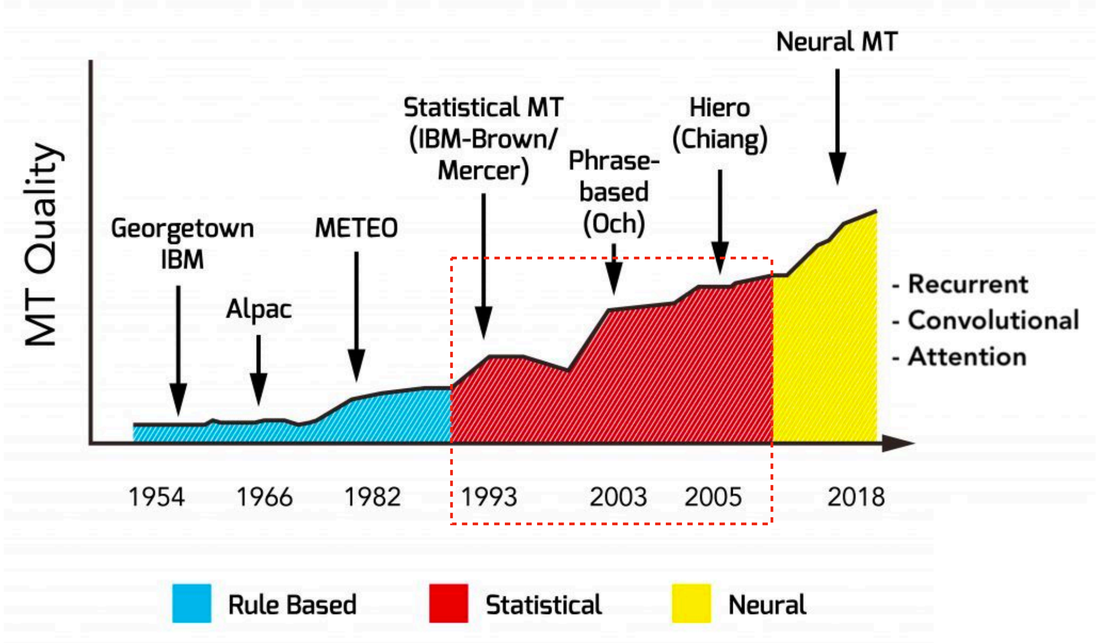
- Bilingual dictionary
- Linguistic rules for each language



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
- $argmax_y P(y|x)$

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

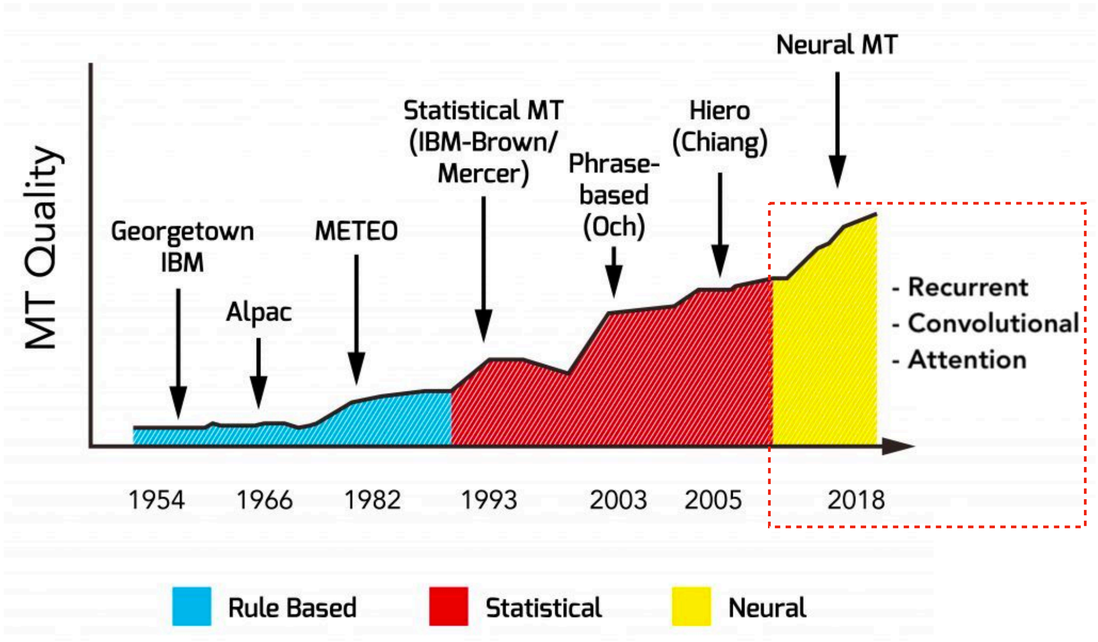
$$\text{'교육은'} = argmax_y P(y|x)$$

$$\text{'가장'} = argmax_y P(y|x, \text{'교육은'})$$

$$\text{'강역한'} = argmax_y P(y|x, \text{'교육은'}, \text{'가장'})$$

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)

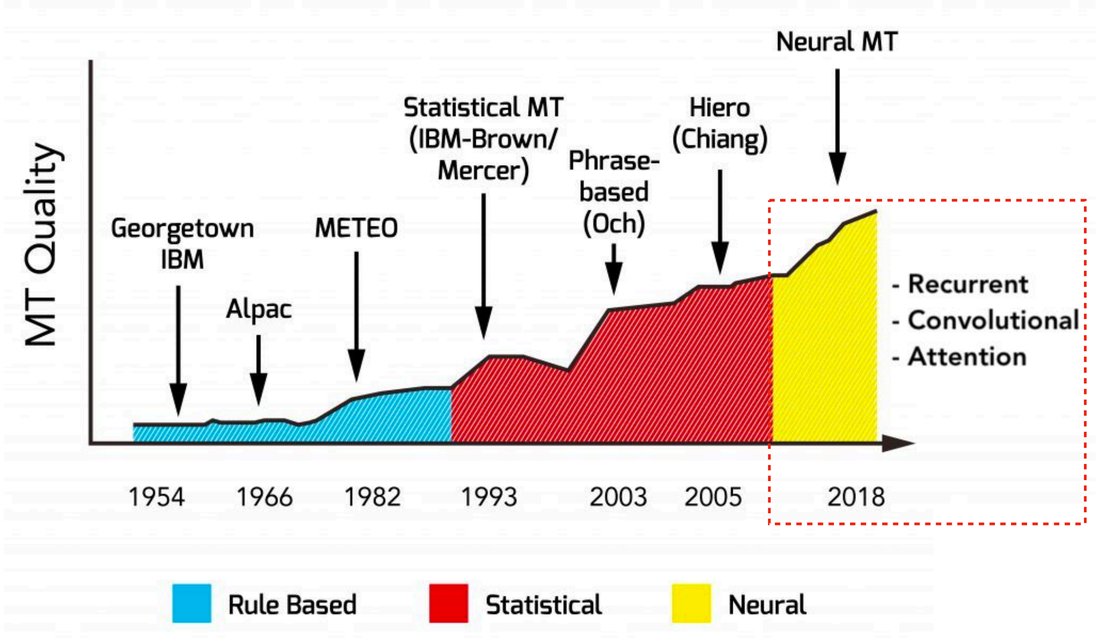


Neural Machine Translation

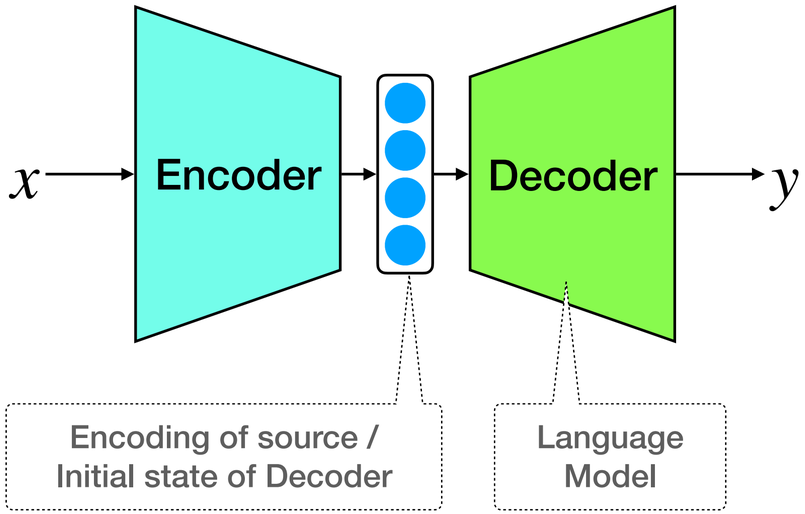
- 2041년 sequence to sequence 모델 등장
- 데이터에서 Neural Network 학습
- $P(y \mid x; \theta)$
  - 어순 오류 감소
  - 어휘 오류 감소
  - 문법 오류 감소

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



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

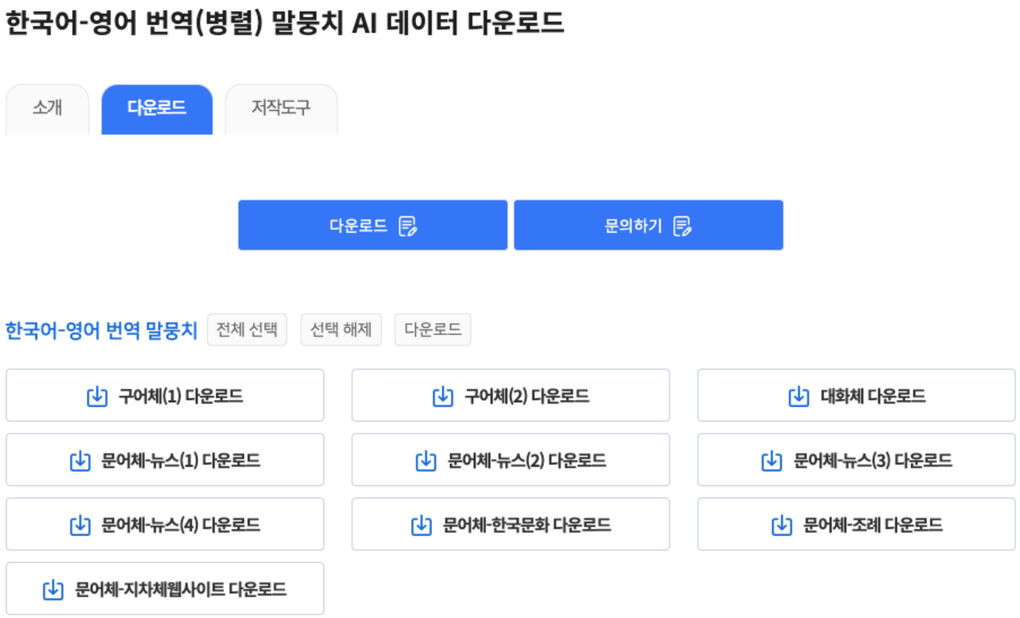
- WMT Dataset

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

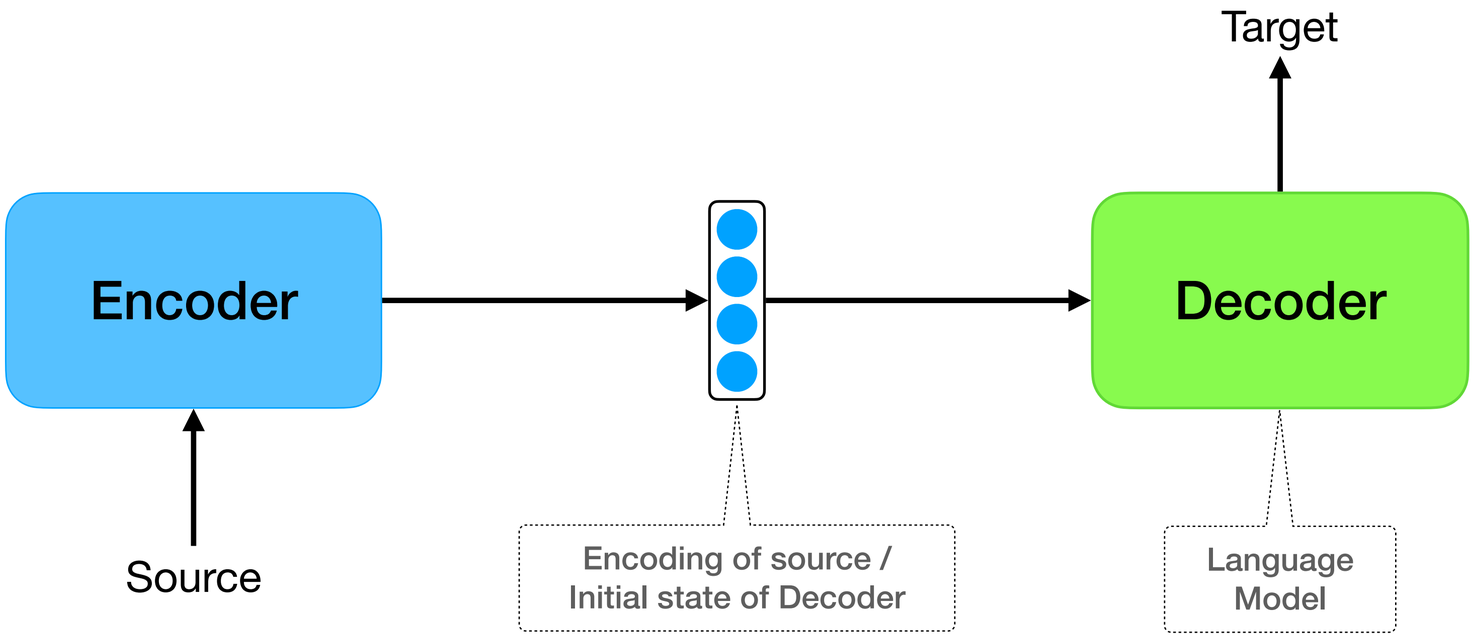
Machine Translation DataSet

- WMT Dataset
- AI-Hub 한국어-영어 병렬 말뭉치

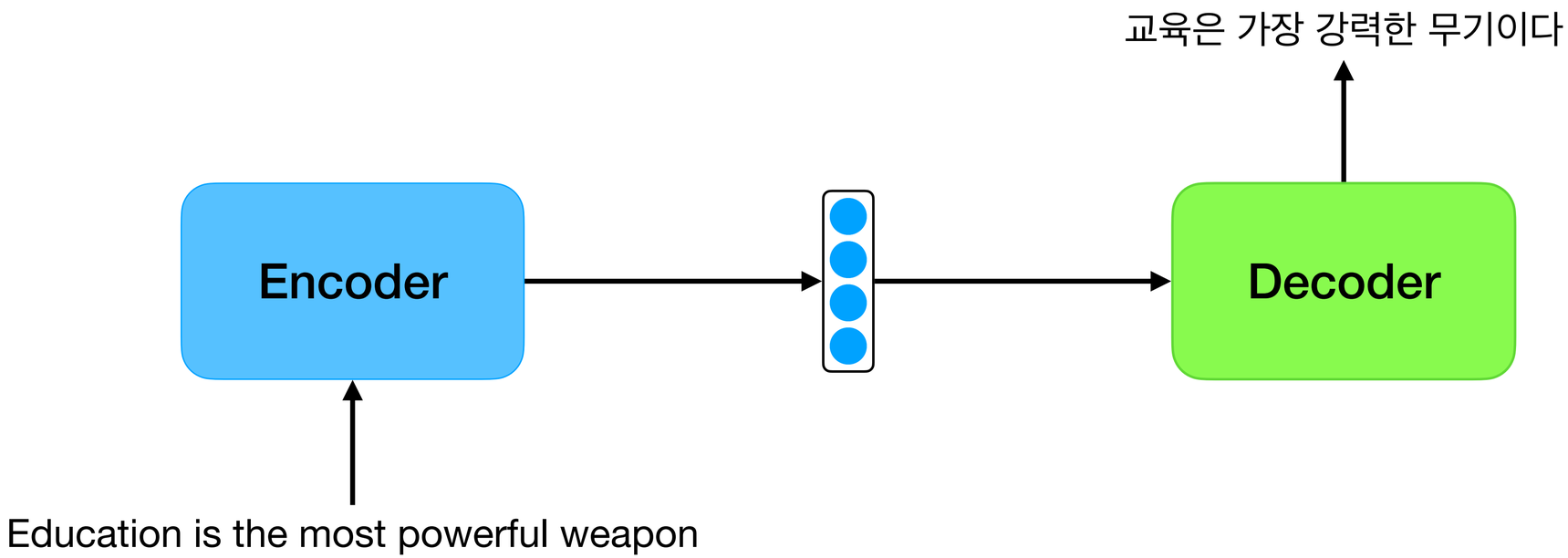


- AI-Hub 한국어-영어 데이터
- 회원 가입 및 별도의 서류제출 후 다운로드
- https://aihub.or.kr/aidata/87/download

Machine Translation Model

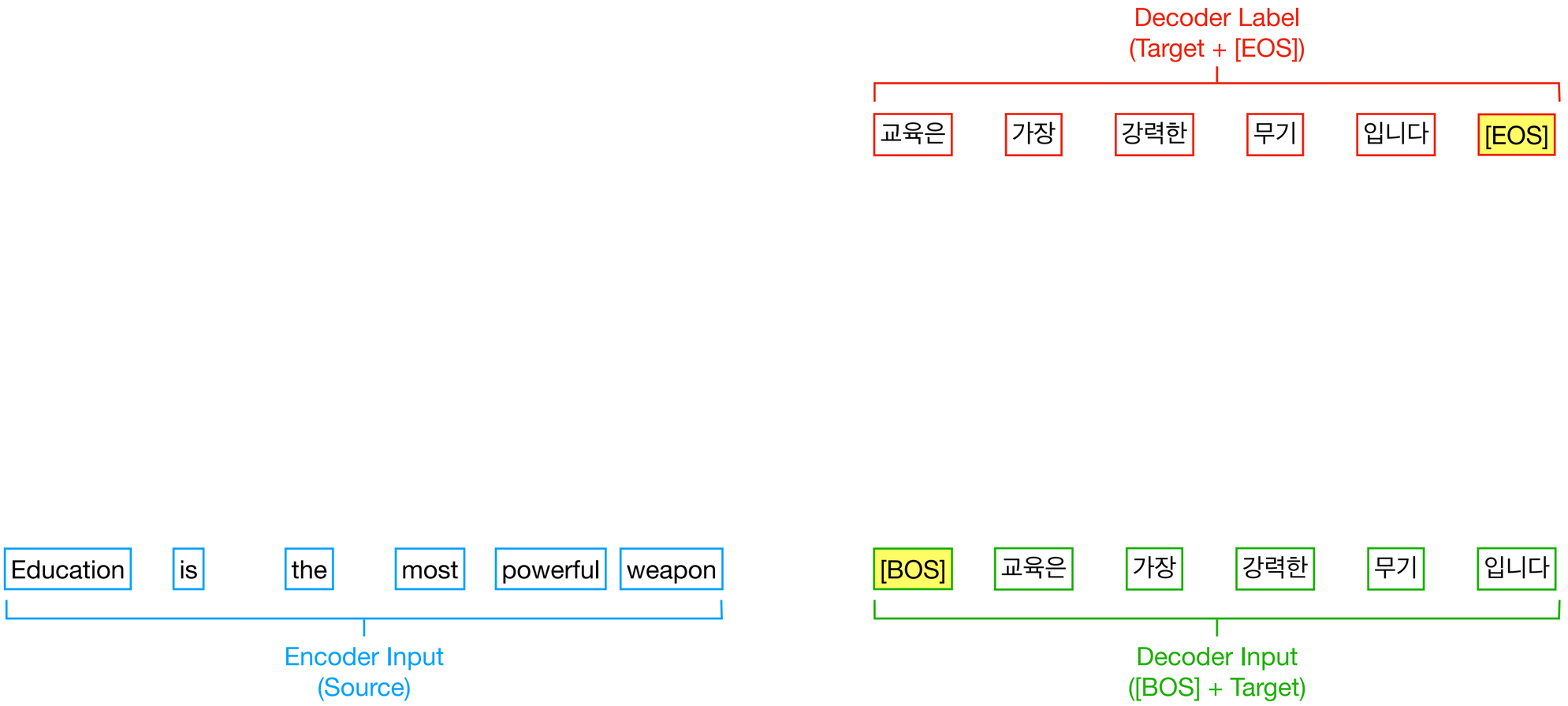


Machine Translation Model

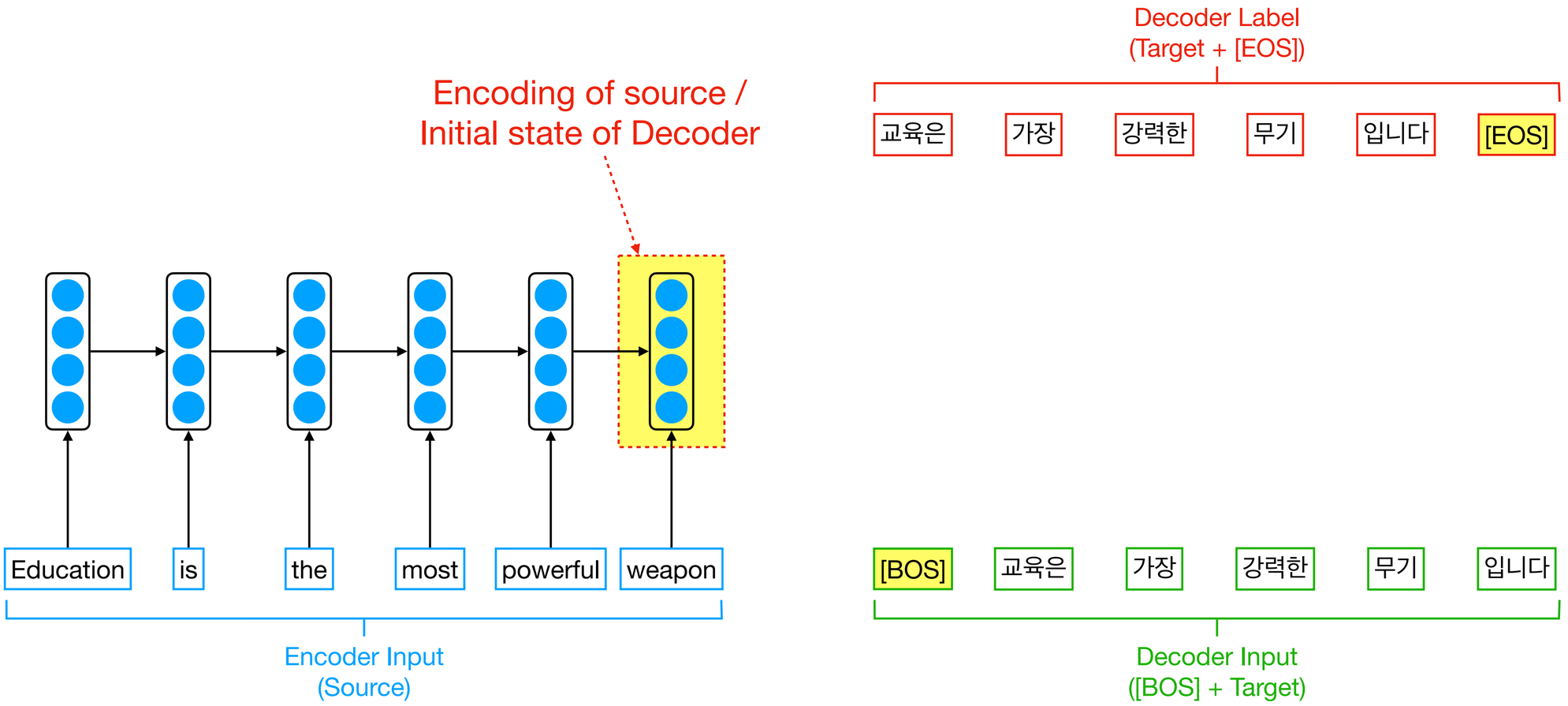


Encoder Decoder Architecture / Sequence to Sequence

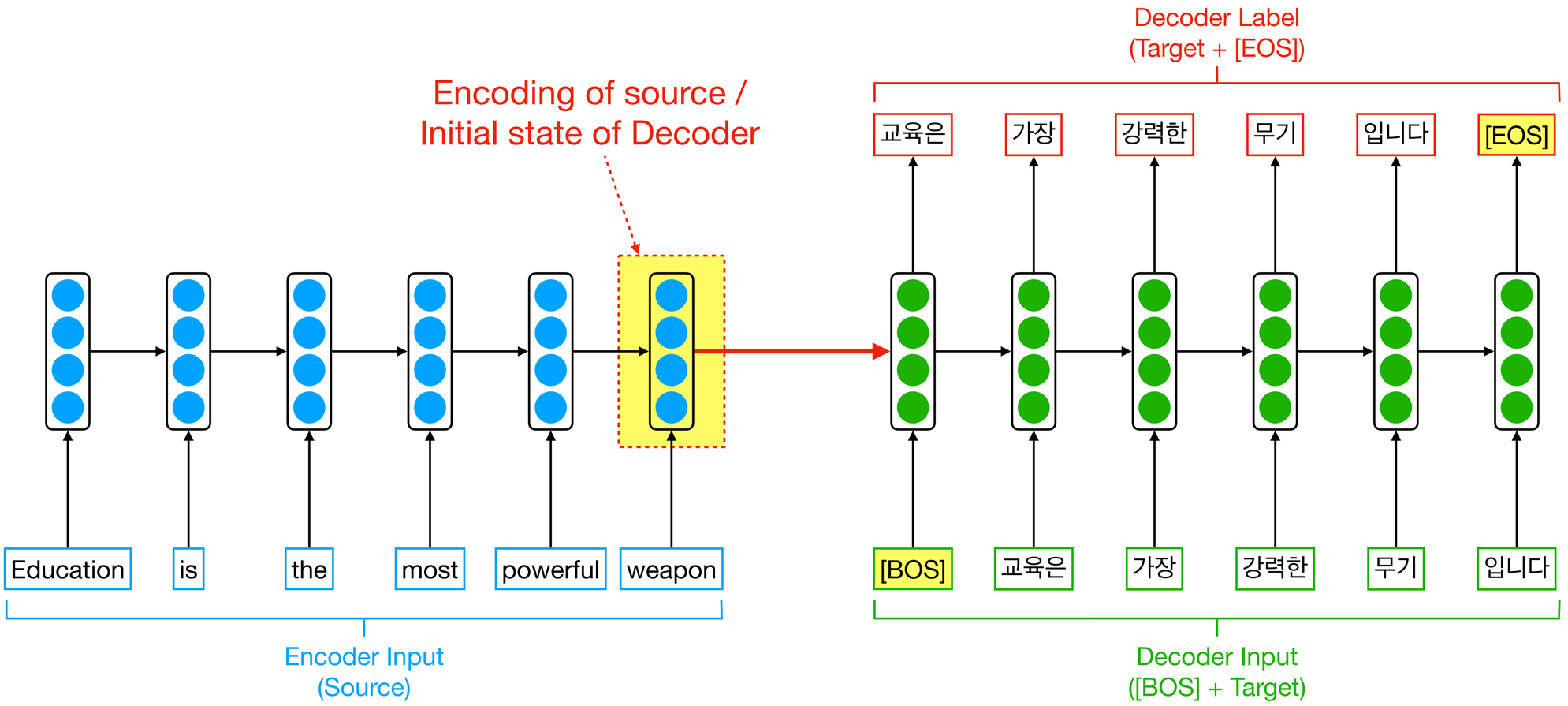
Machine Translation Model (Training)



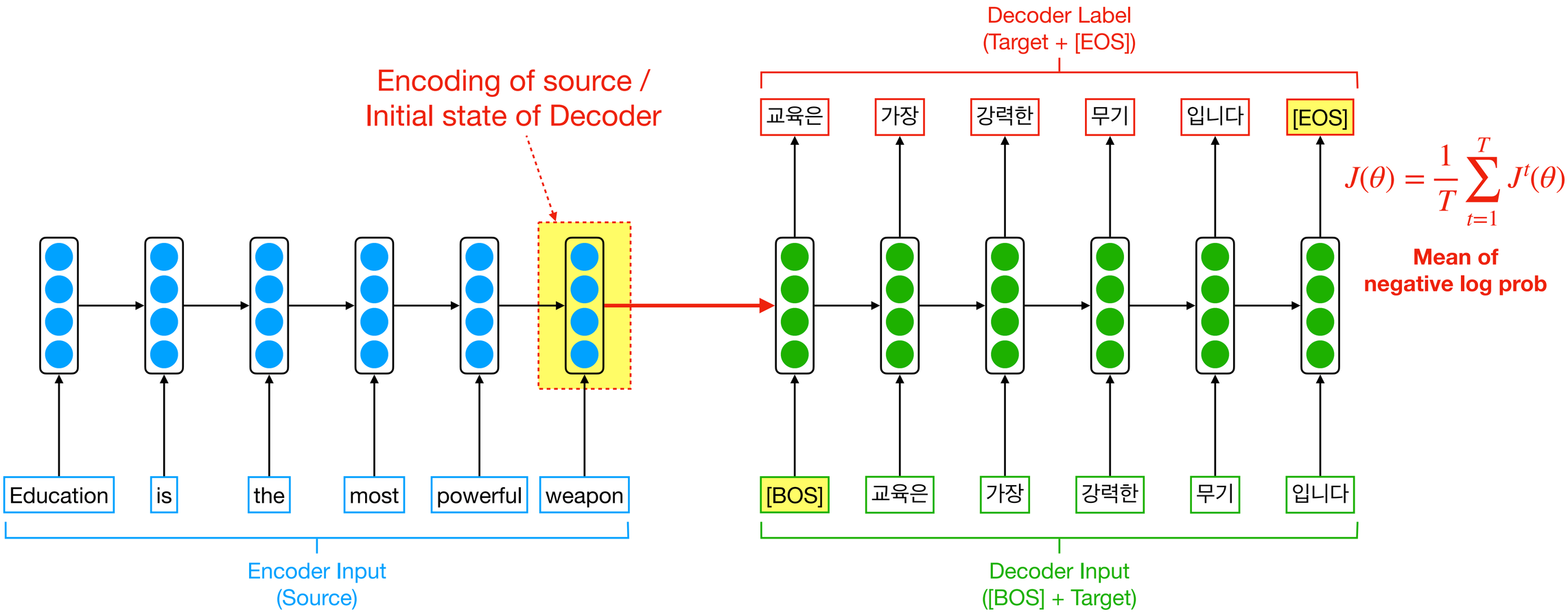
Machine Translation Model (Training)



Machine Translation Model (Training)

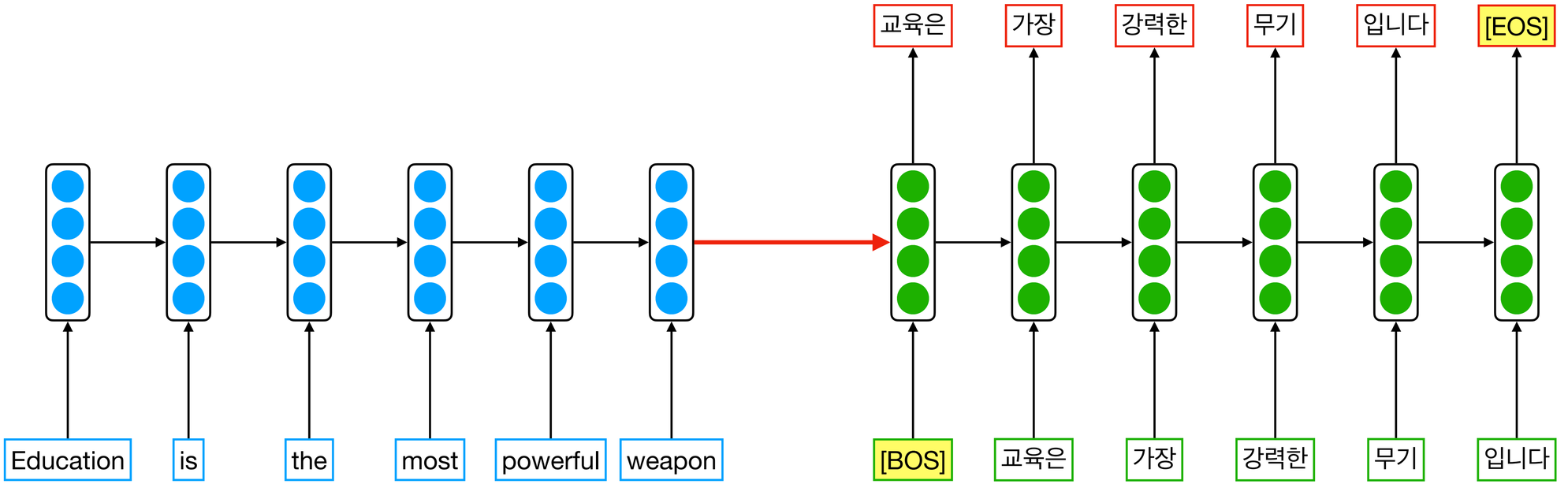


Machine Translation Model (Training)



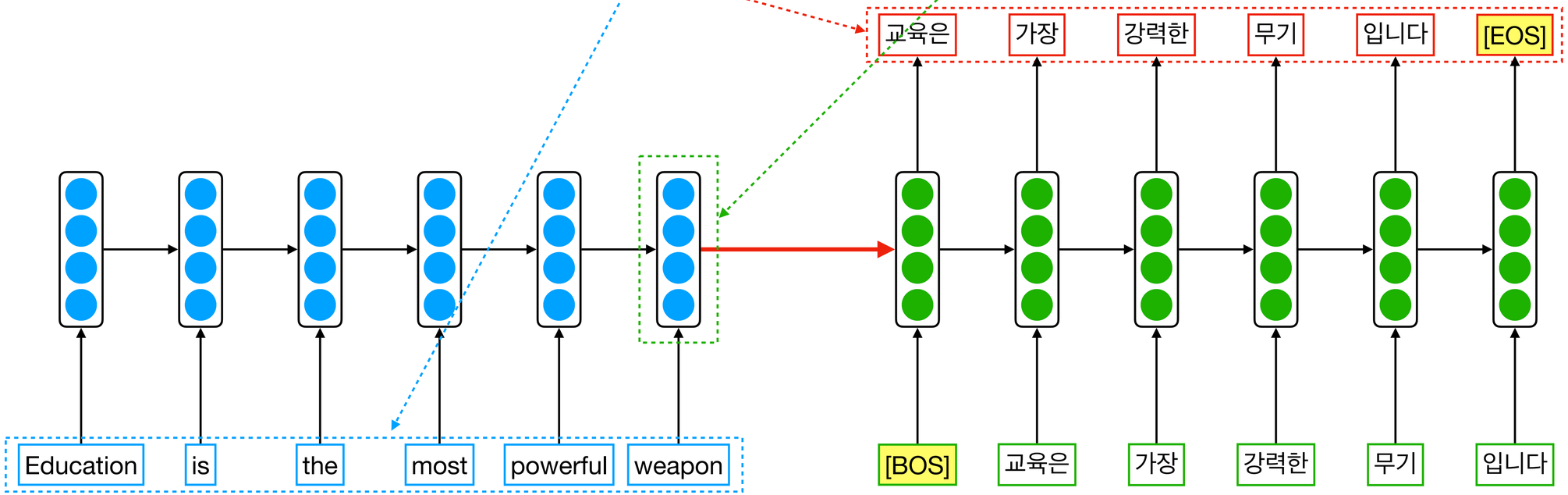
Machine Translation Model (Training)

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

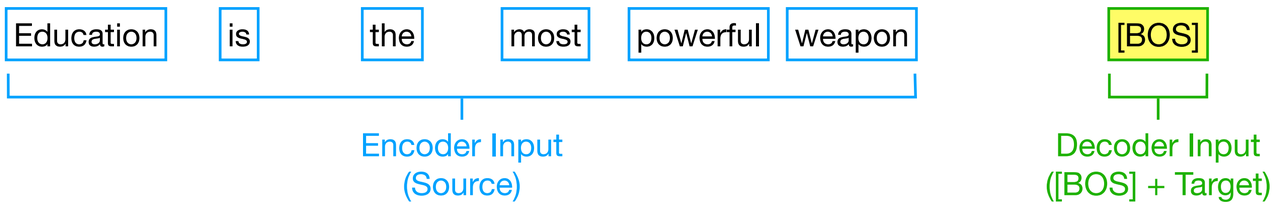


Machine Translation Model (Training)

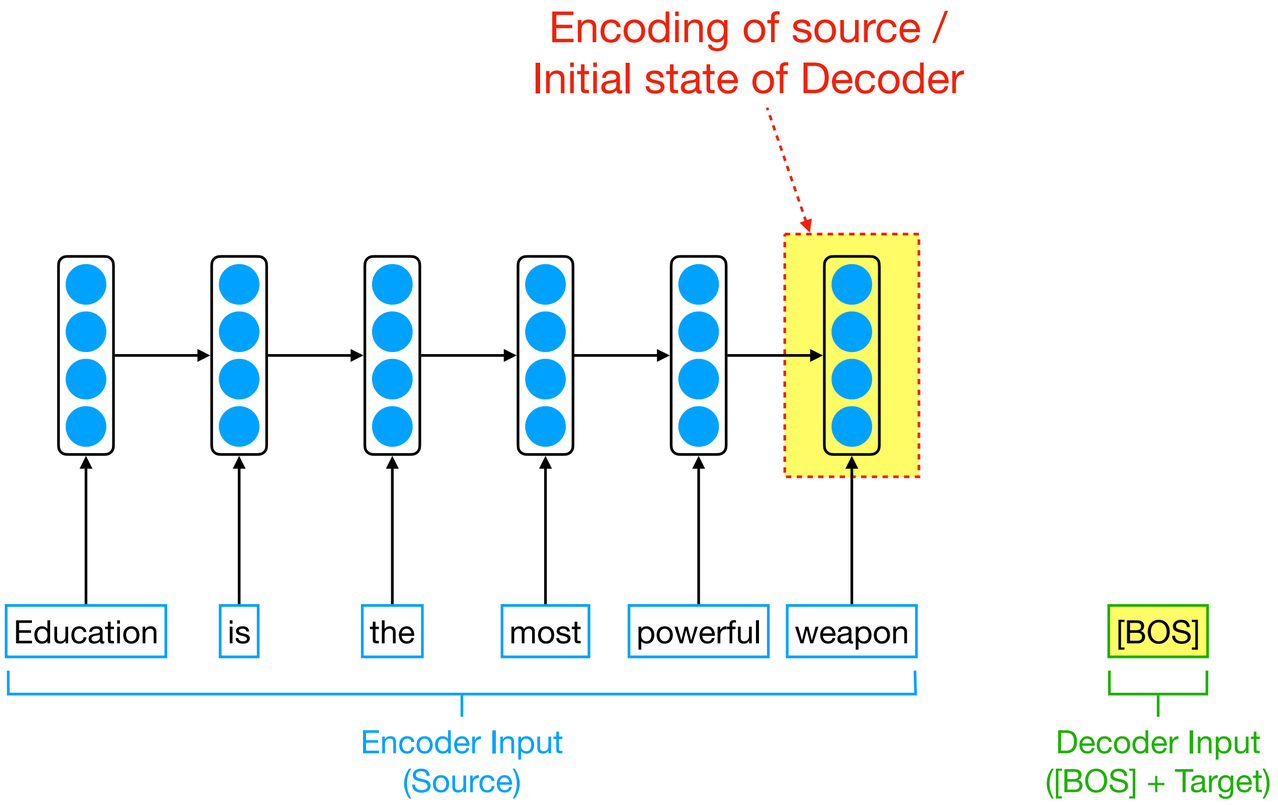
$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_m) = \prod_{t=1}^{n} p(y_t \mid v, y_1, \ldots, y_{t-1})$$



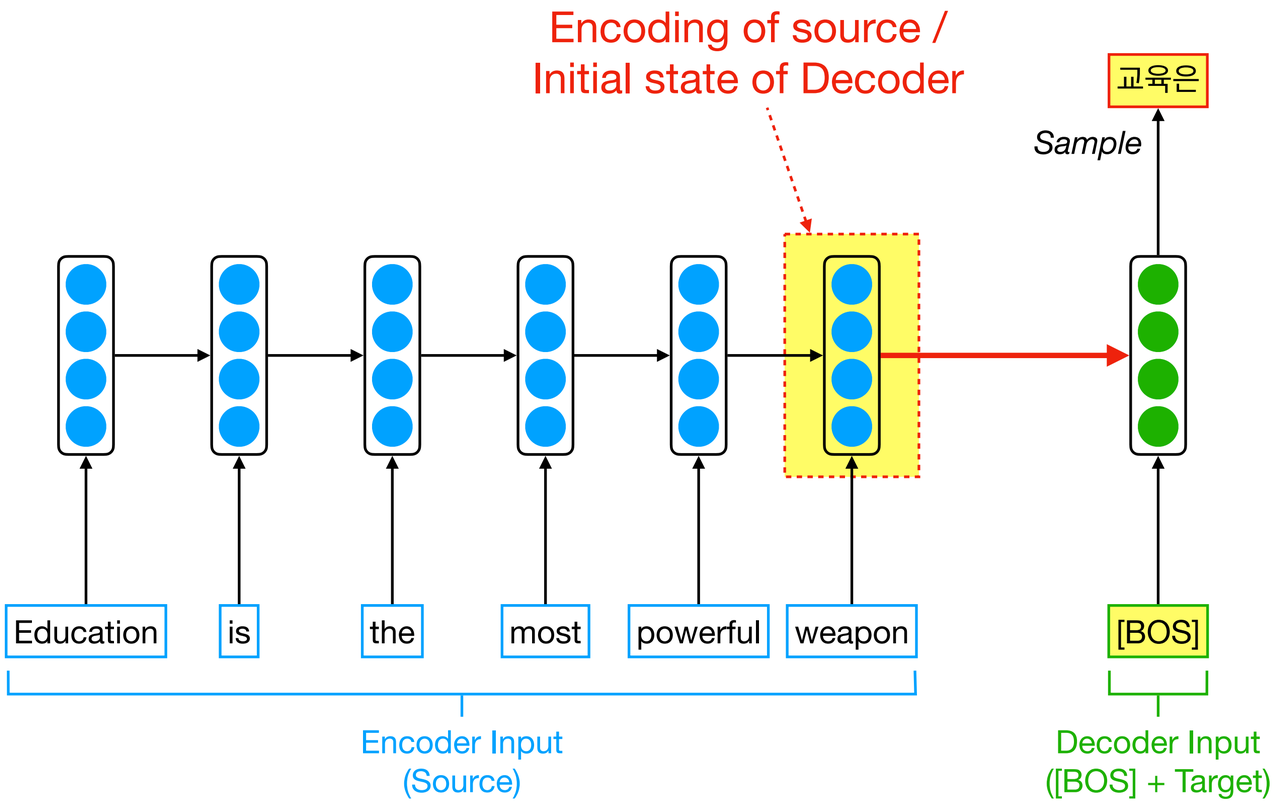
Machine Translation Model (Inference)



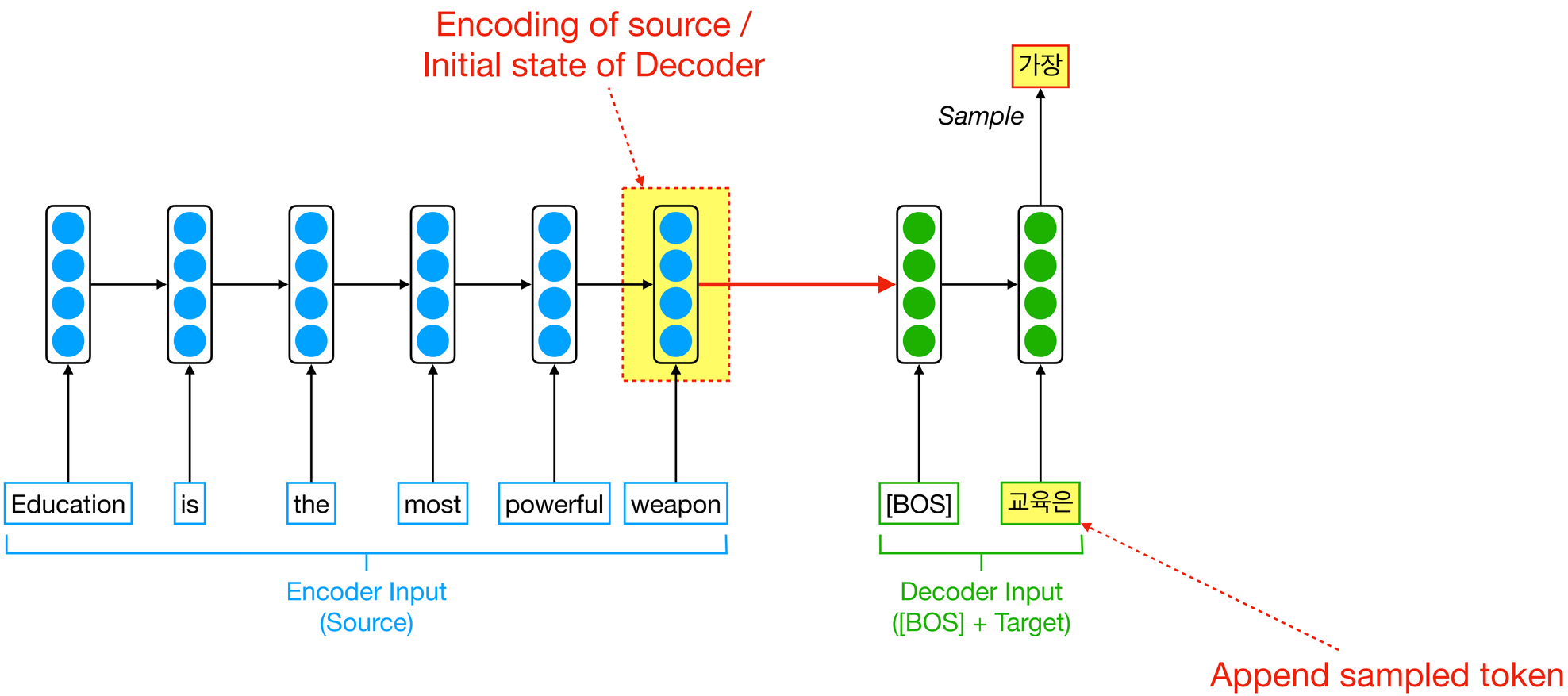
Machine Translation Model (Inference)



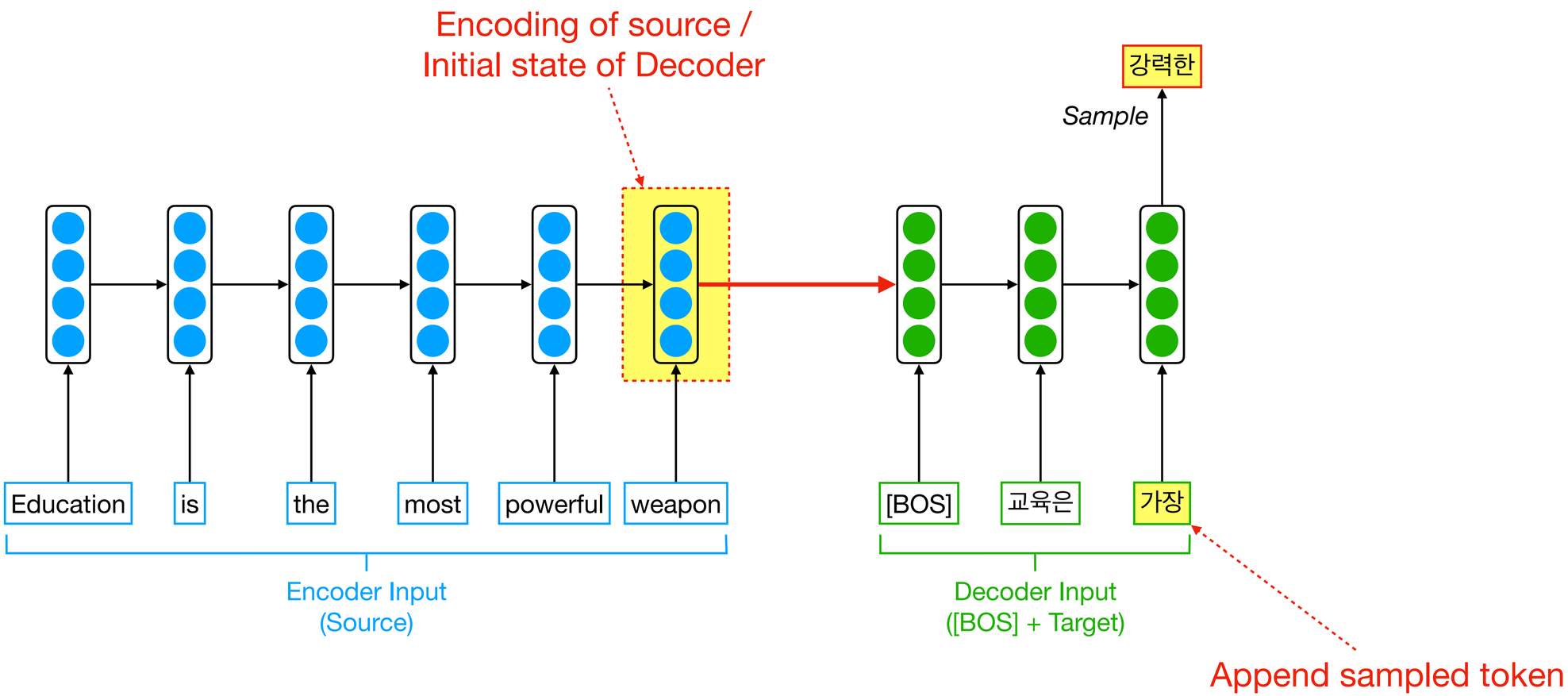
Machine Translation Model (Inference)



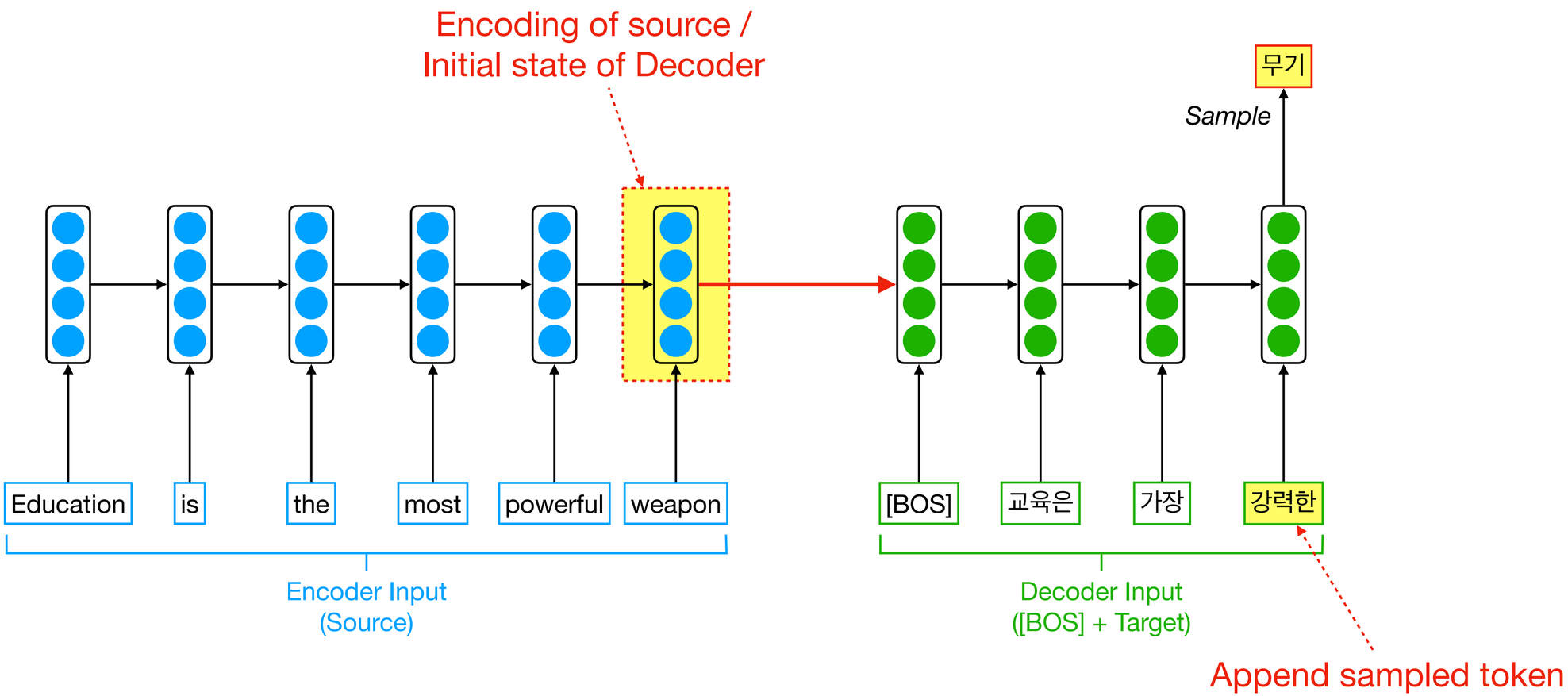
Machine Translation Model (Inference)



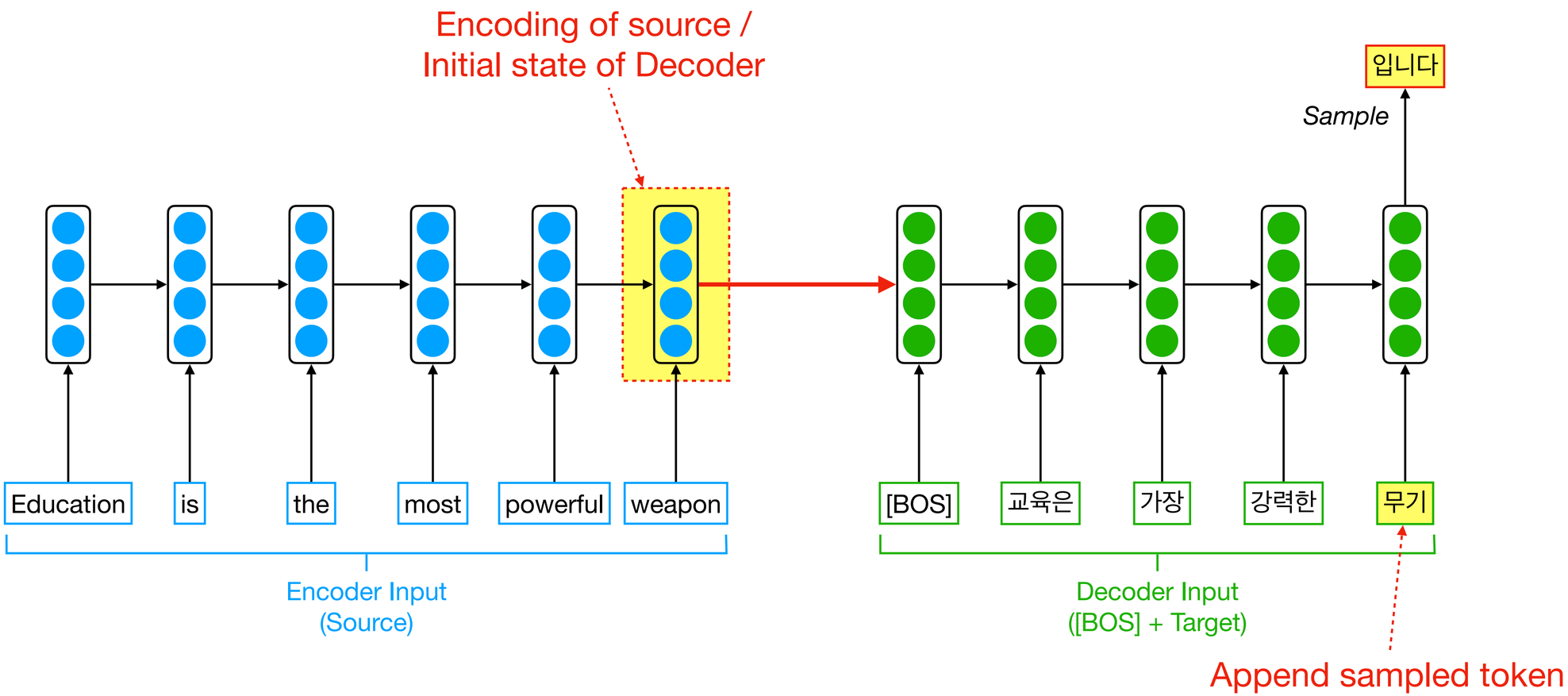
Machine Translation Model (Inference)



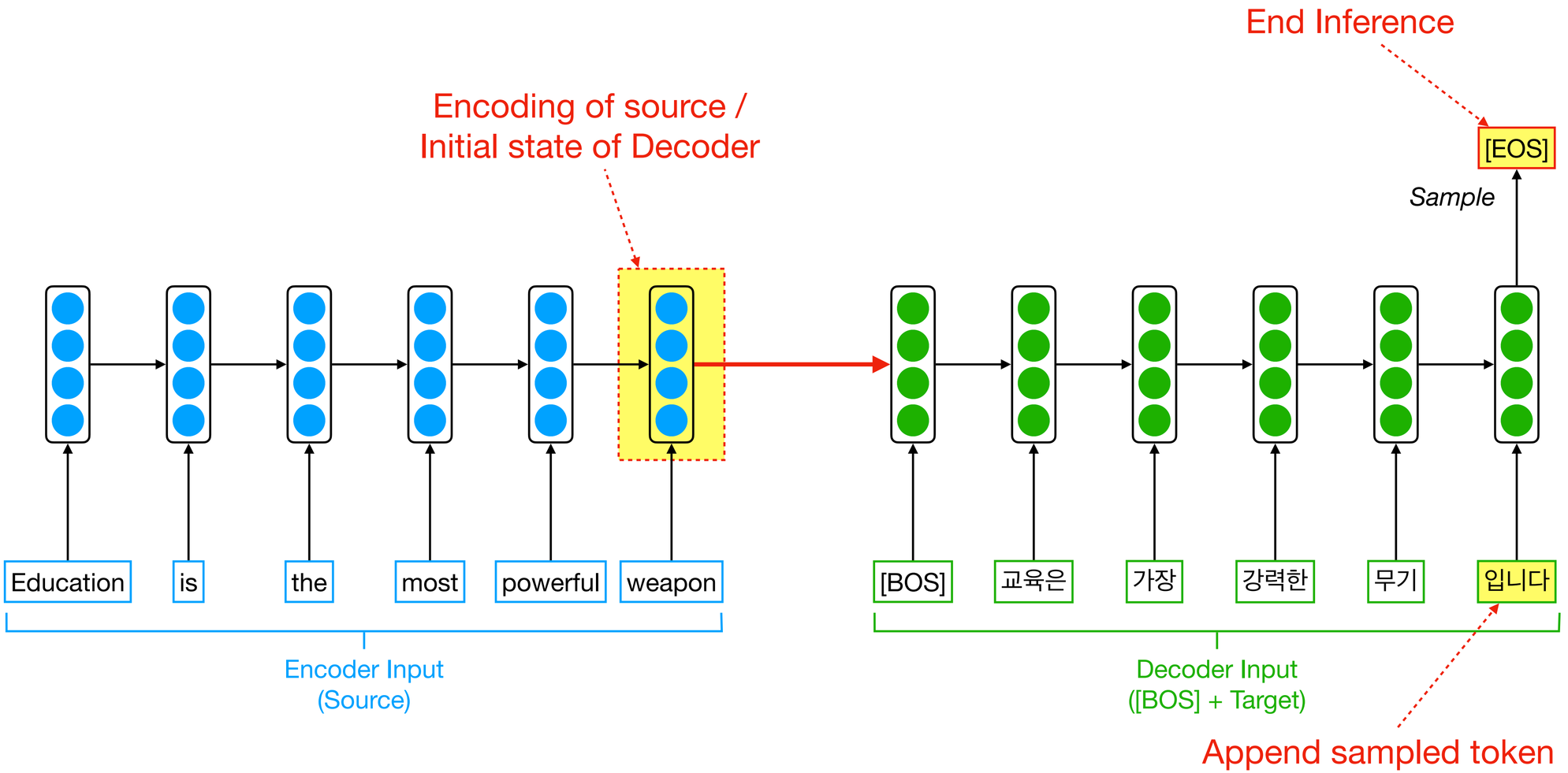
Machine Translation Model (Inference)



Machine Translation Model (Inference)



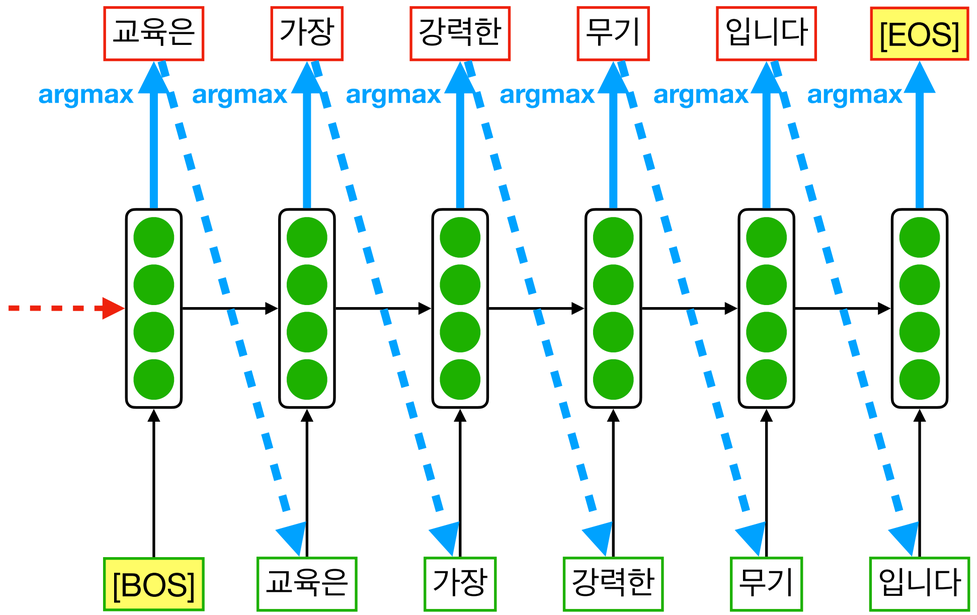
Machine Translation Model (Inference)



Machine Translation Model (Versatile)

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

Machine Translation Model (Decoding)



Greedy Search

Machine Translation Model (Decoding)

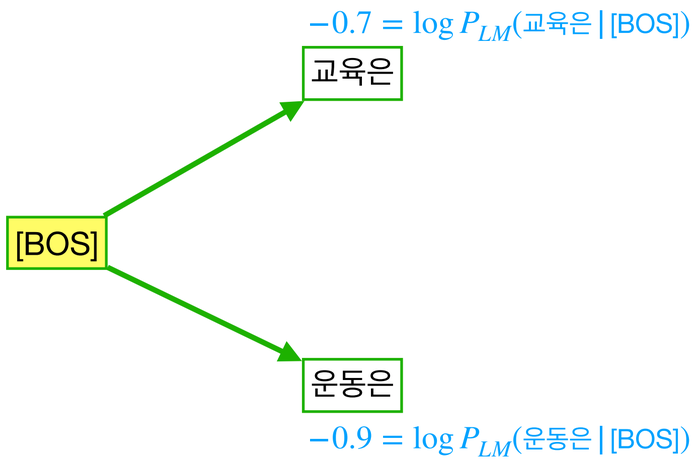
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$

[BOS]

Beam Search (k=2)

Machine Translation Model (Decoding)

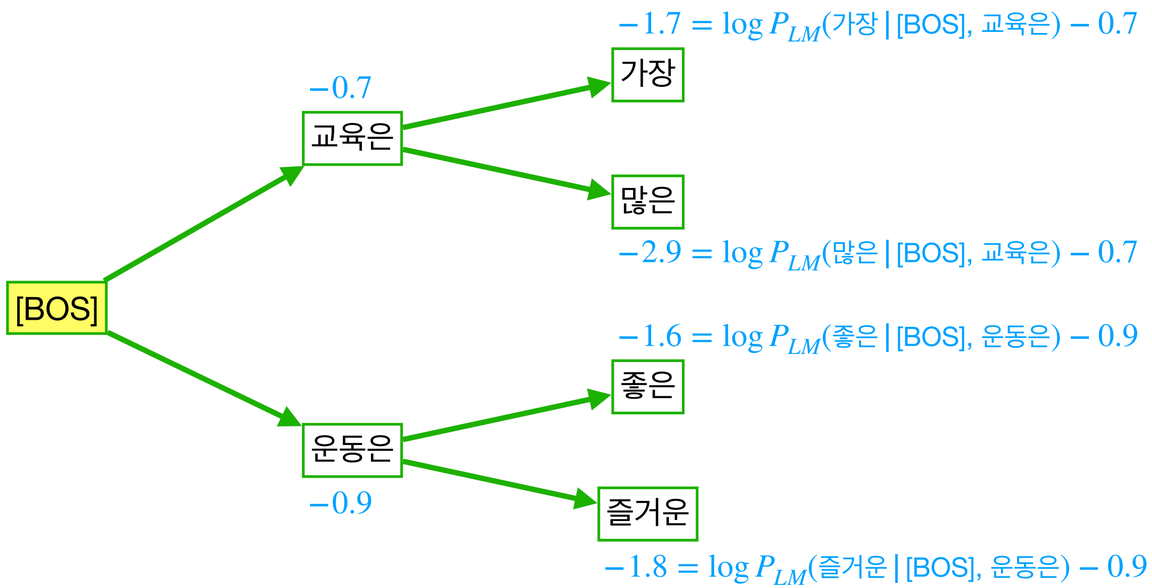
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i \mid y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

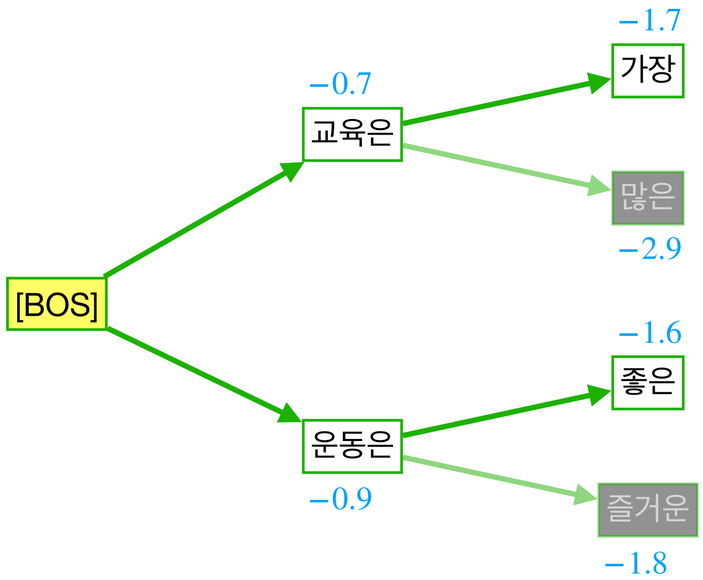
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i \mid y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

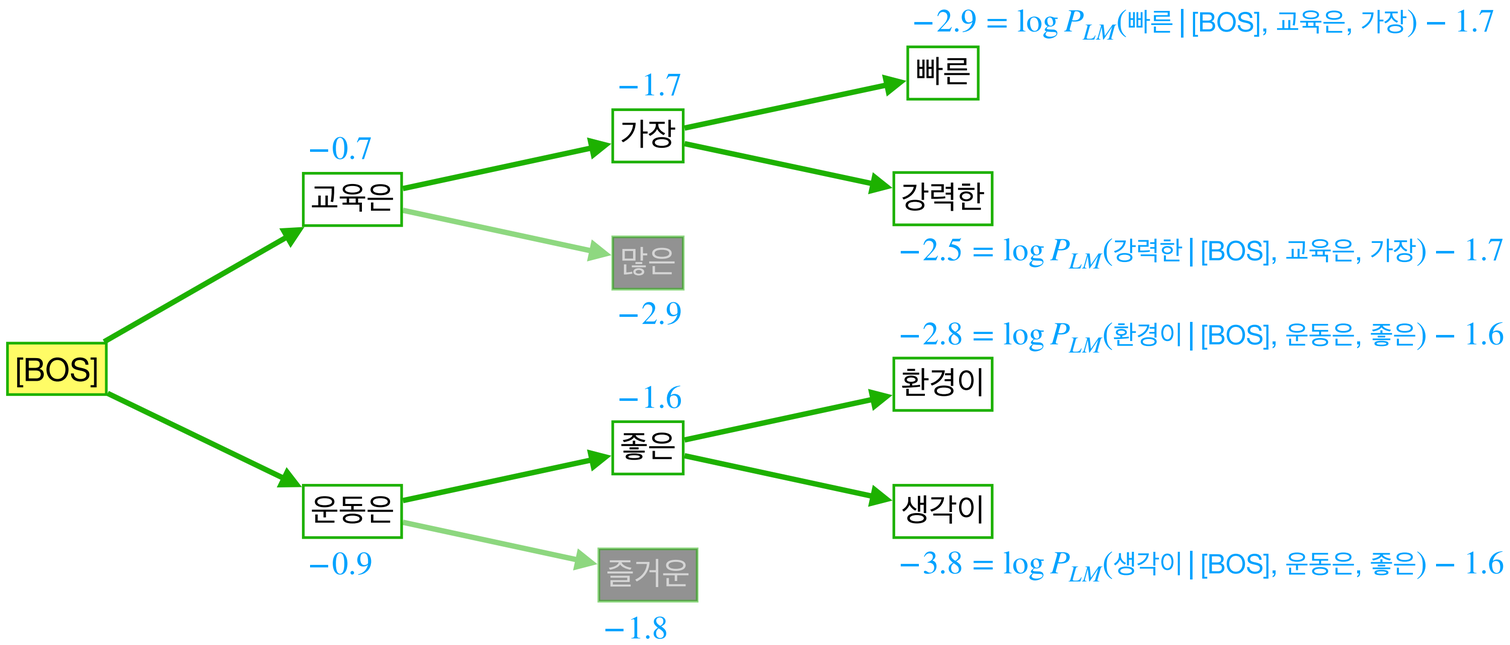
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

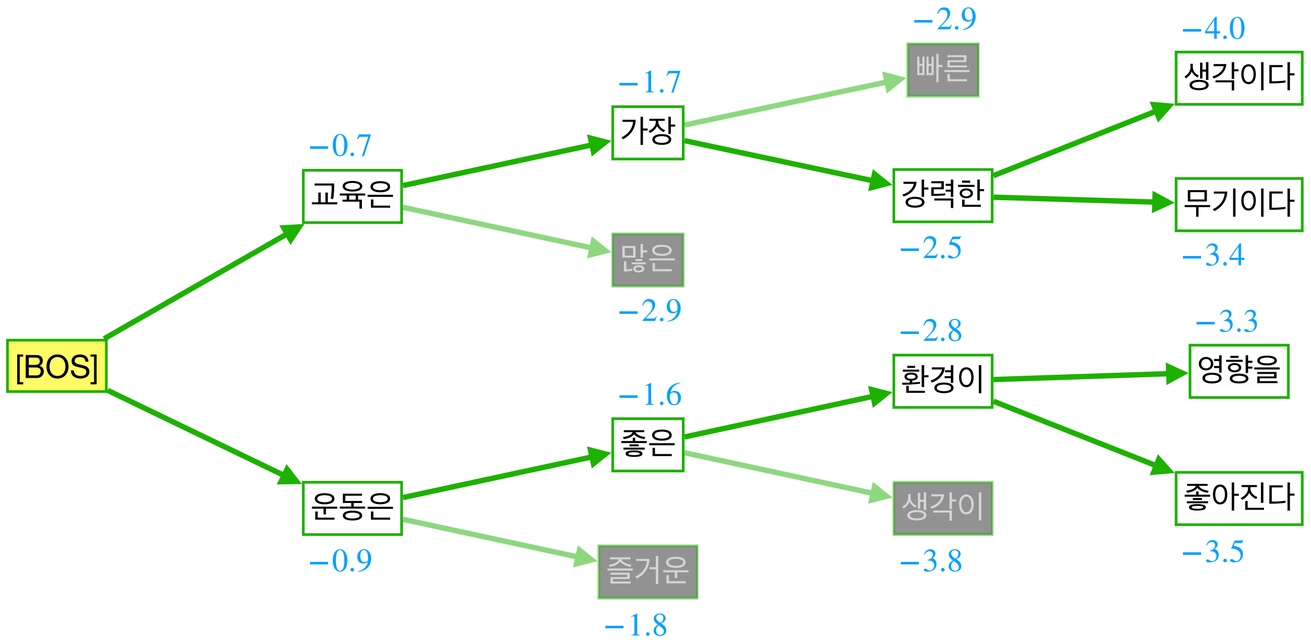
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i \mid y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

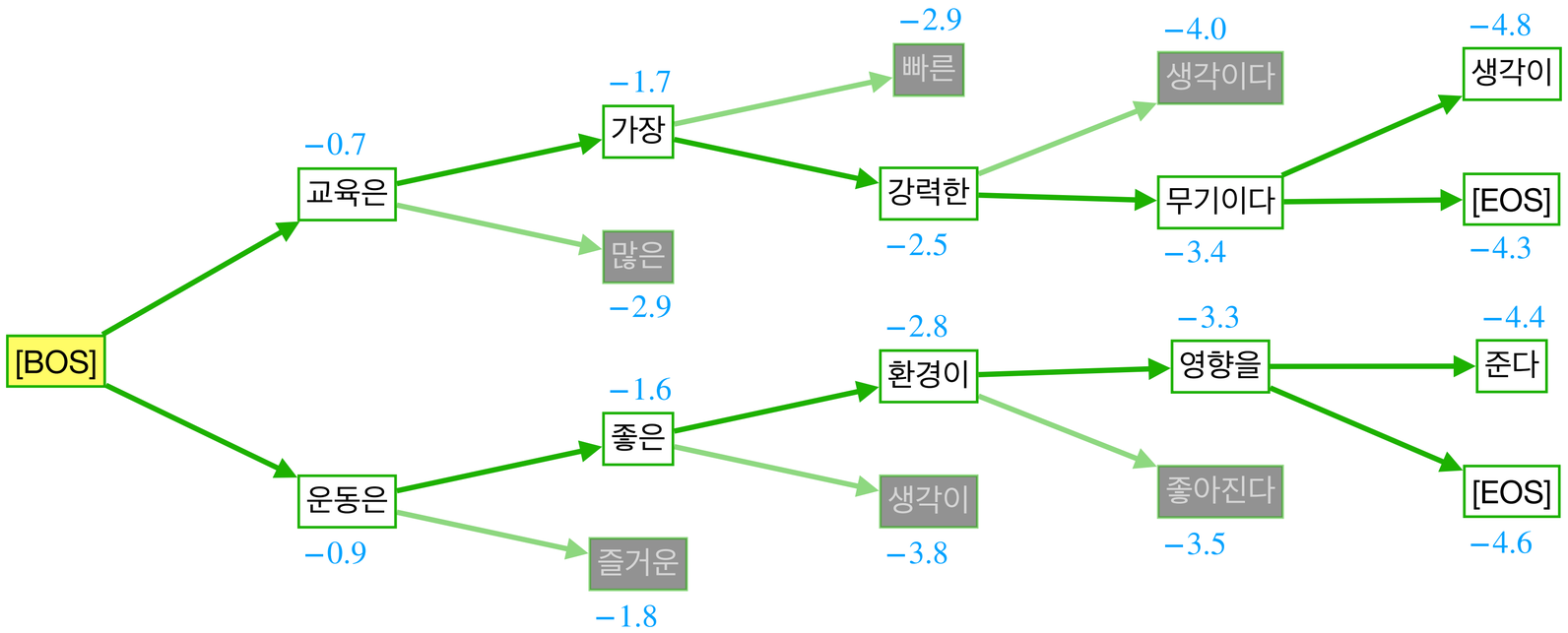
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

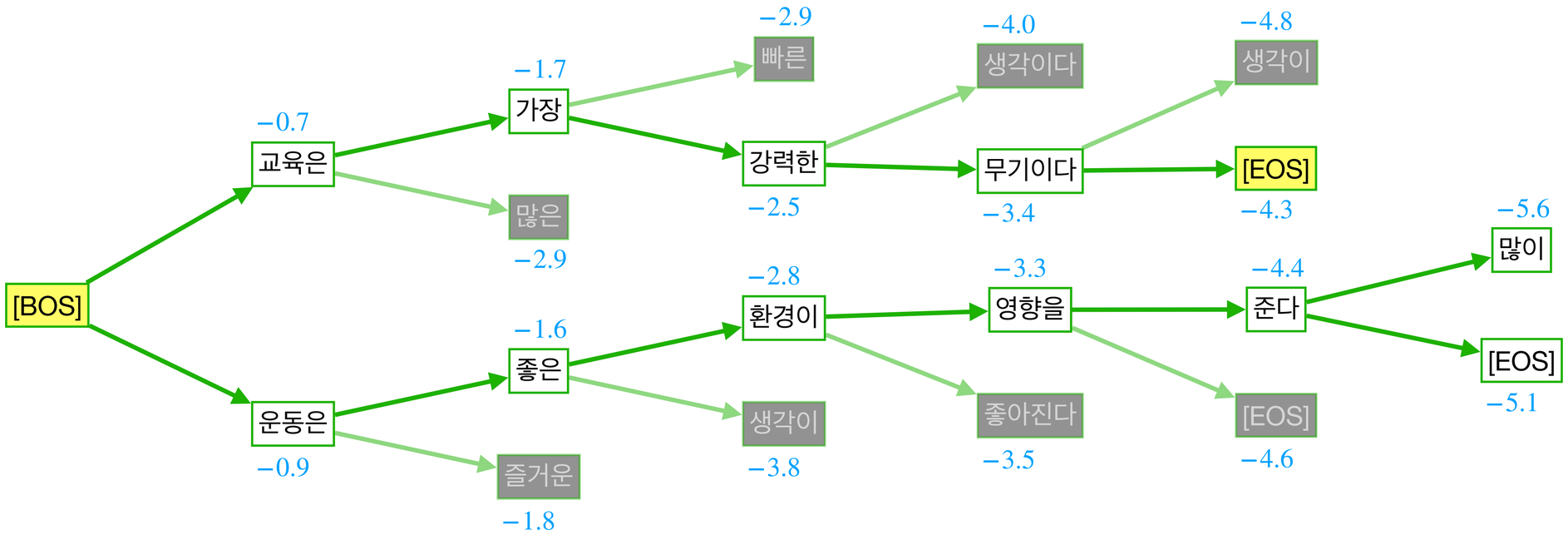
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

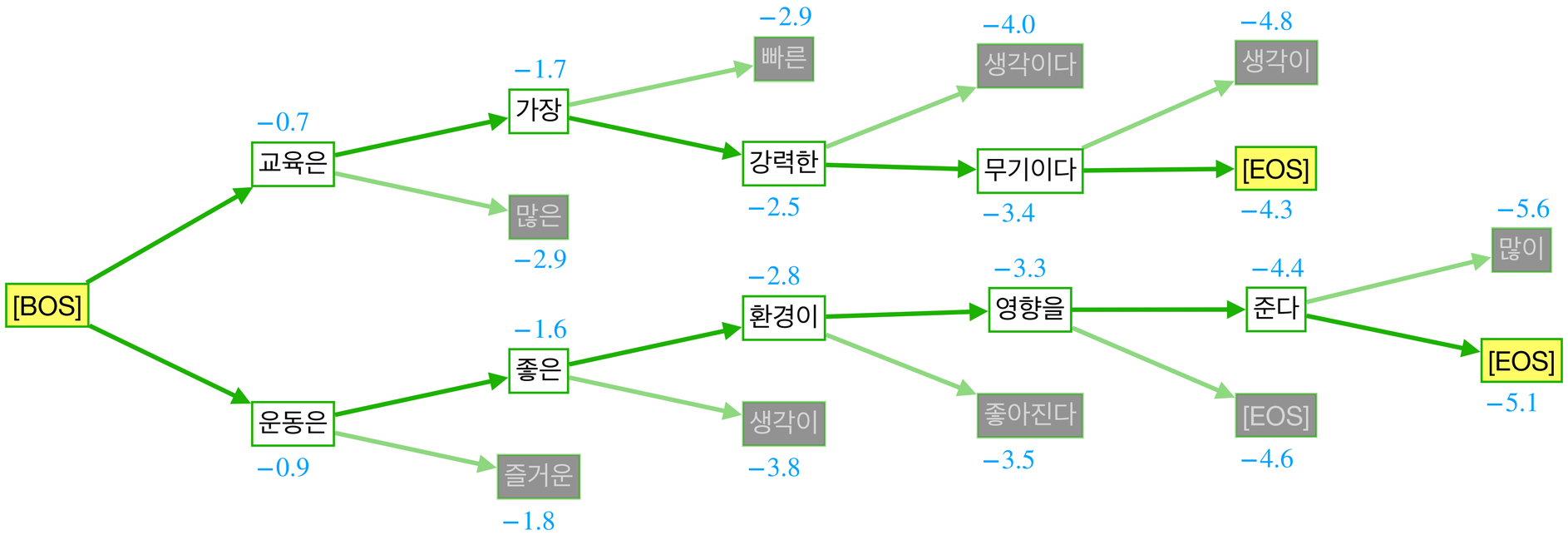
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

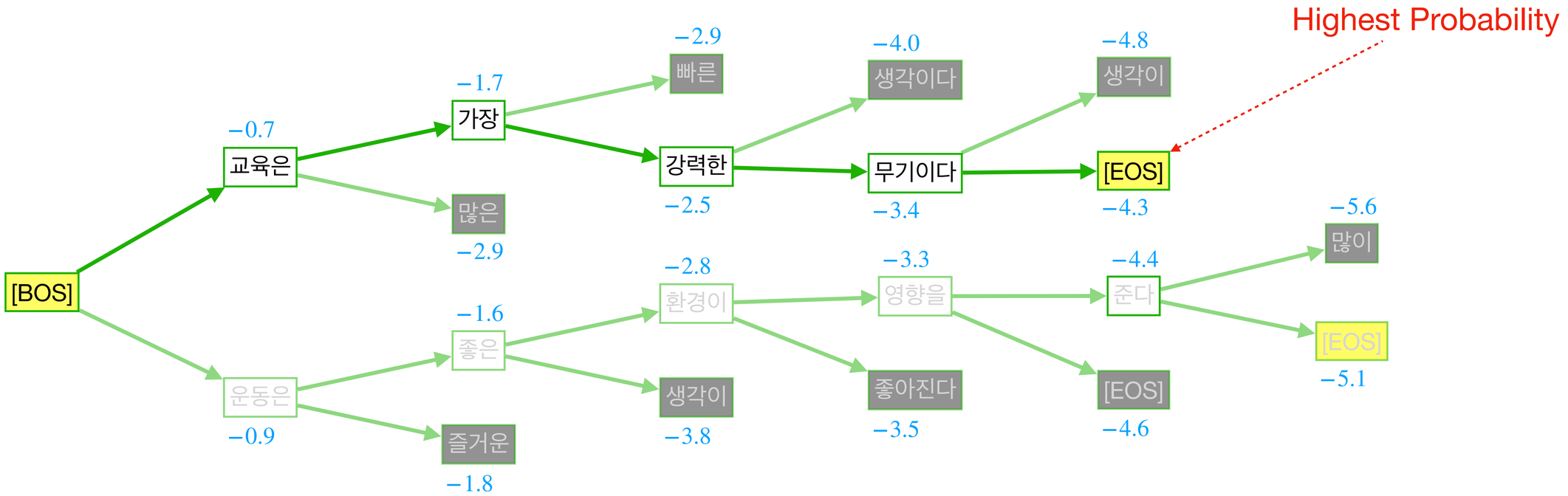
$$score(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{LM}(y_i | y_1, \ldots, y_{i-1})$$



Beam Search (k=2)

Machine Translation Model (Decoding)

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



Beam Search (k=2)

Machine Translation Model (Decoding)

Length Penalty

- Longer hypotheses have lower score

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

- Normalize by length

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

Beam Search (k=2)

Machine Translation Model (Metric)

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

How do you evaluate???

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.

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

- Machine Translation된 결과와 사람이 Translation한 결과를 비교하여 품질을 평가
  - n-gram precision
  - penalty for too-short system translations

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

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

N-gram precision

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

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

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.

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

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

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.

Reference 2: There is a cat on the mat.

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

$$Count_{clip} = \min(Count, Max\_Ref\_Count)$$

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

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

Machine Translation Model (Metric)

BLEU(Bilingual Evaluate Understudy) Score

N-gram precision

Candidate의 n-gram의 $Count_{clip}$개수

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

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

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.

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

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.

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

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

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

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

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n)$$

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

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

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

Reference: 나는 어제 집에 가서 잠을 설쳤다

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

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

감사합니다.