NMT Tutorial

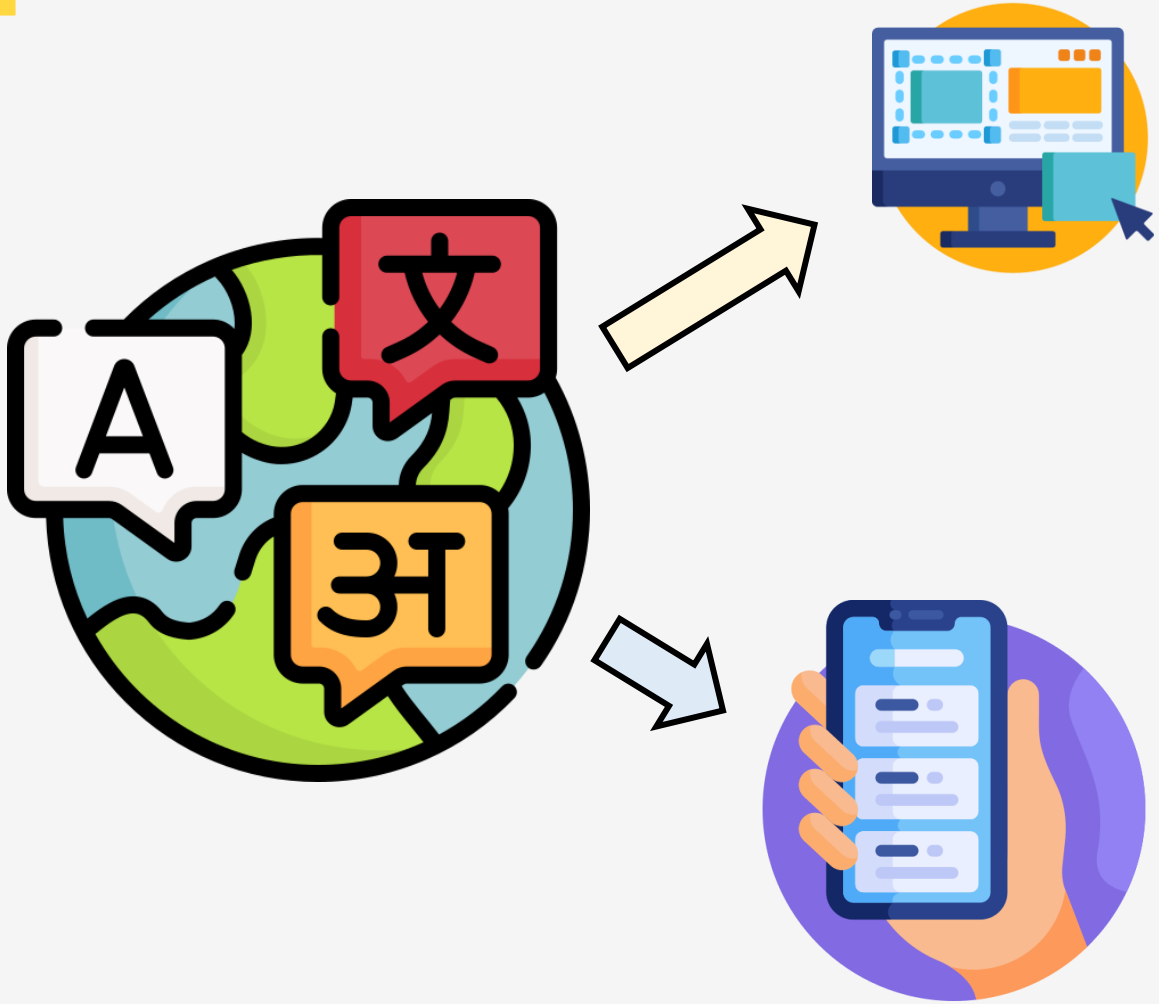
Neural Machine Translation Tutorial

발표자: 이세영

2023. 07. 29

Contents

1. NMT 모델이란?



- NMT: Neural Machine Translation
- 주로 인공지능 Machine 을 활용하여 translation 수행

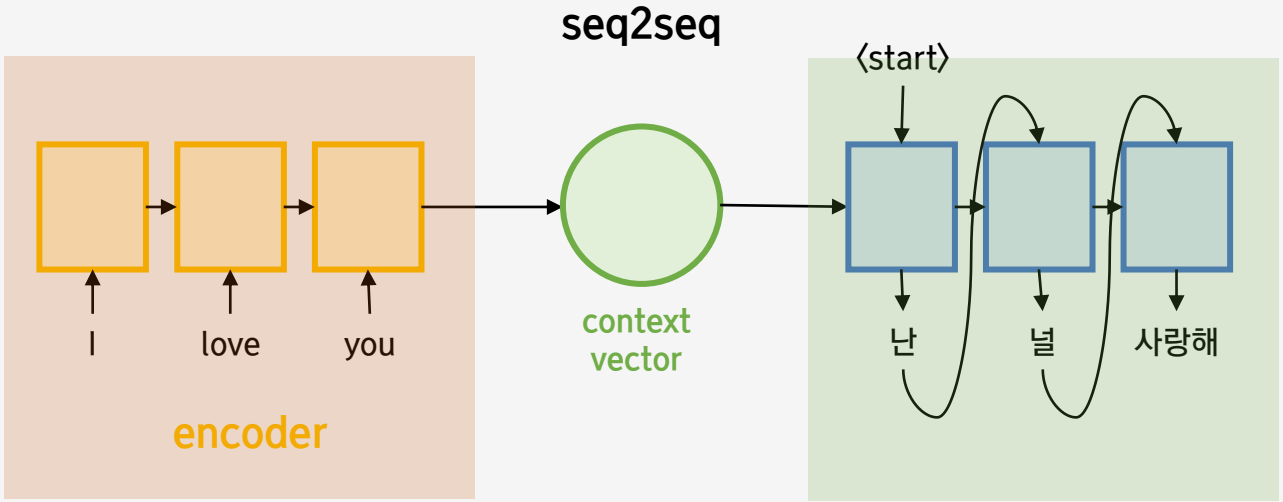
장점

- 실시간 서비스 가능
- 대량으로 서비스 가능

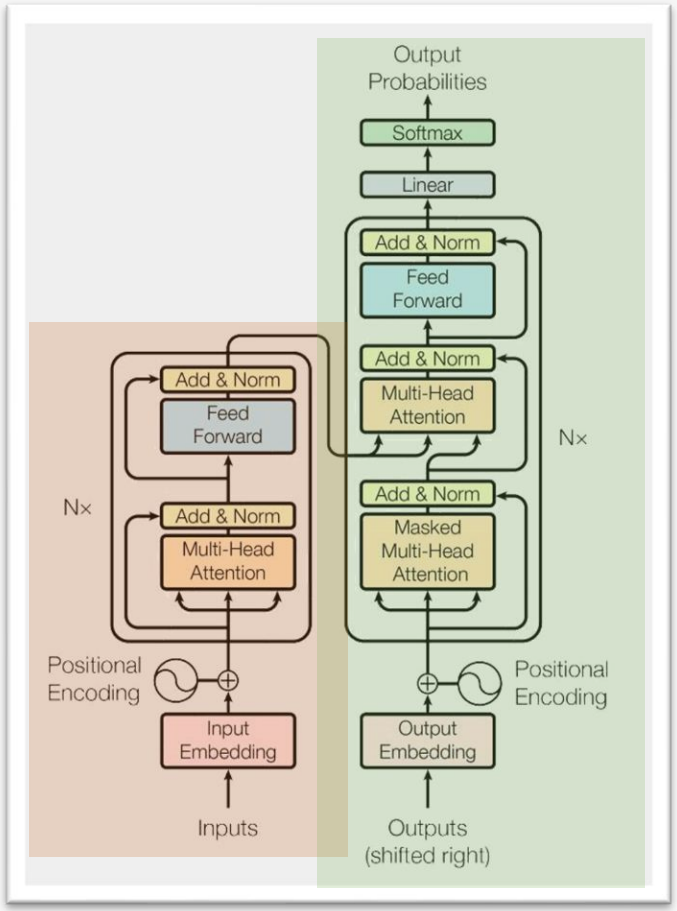
단점

- context 이해가 부족할 수 있음
- 대량의 고품질 데이터 필요
- 대량의 고품질 데이터로 "잘" 학습된 모델 필요
- 많은 리소스 필요 (GPU, Memory 등)

2. seq2seq model



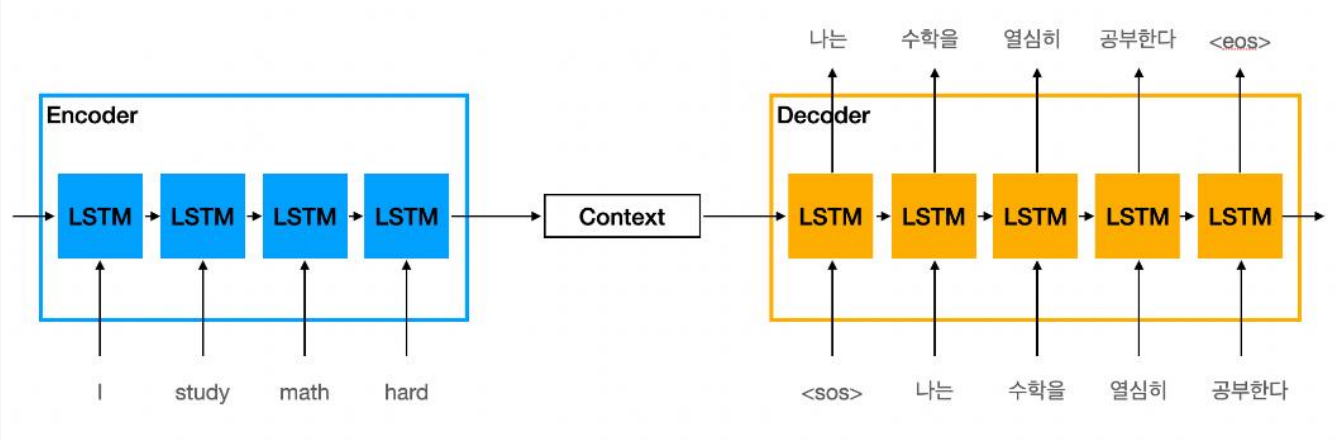
- Encoder – Decoder 구조
- Encoder에서는 source language 학습
- Decoder에서는 target language 학습



transformers

2. seq2seq model

- Encoder-Decoder
- Encoder: 입력 sequence의 정보를 압축해 context vector를 생성
    - Encoder RNN 모델의 마지막 시점 hidden state vector가 context vector로 사용되며 context vector는 Decoder에 있는 RNN 모델의 첫번째 시점 hidden state vector가 된다.
- Decoder: context vector를 활용해 출력 sequence 생성
    - Decoder의 RNN 모델은 sequence의 시작을 알리는 스페셜 토큰인 를 첫번째 시점의 입력으로 받는다. 토큰과 context 벡터를 바탕으로 첫번째 시점의 출력값을 생성한다.



출처: https://nkw011.github.io/nlp/seqtoseq/

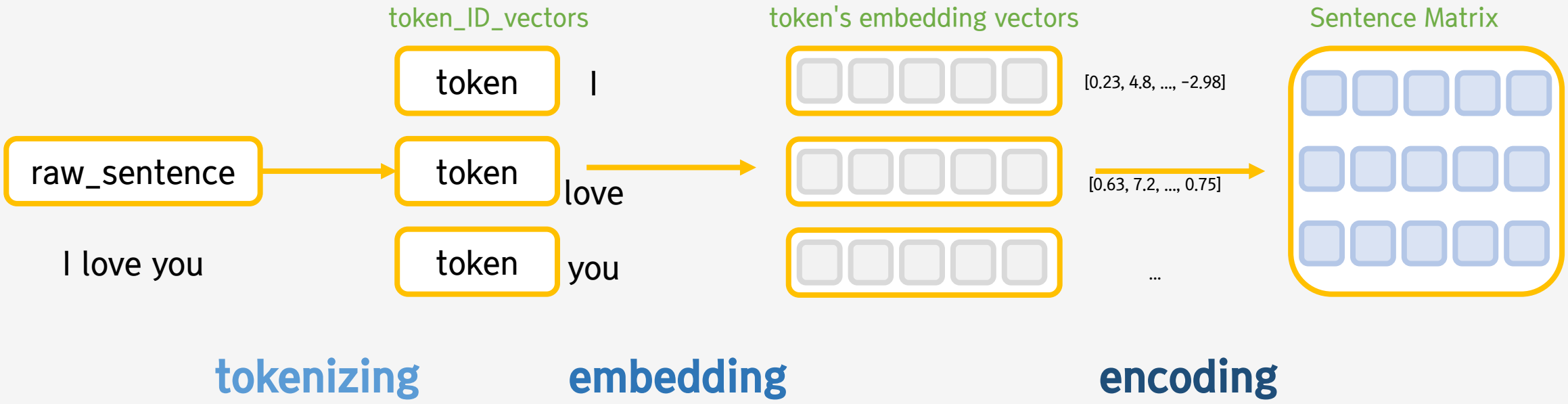
3. Encoder

Embedding: 토크나이징된 단어 토큰들을 벡터들로 변환하는 과정

Encoding: Embedding된 벡터들을 Sentence Matrix로 변환하는 과정

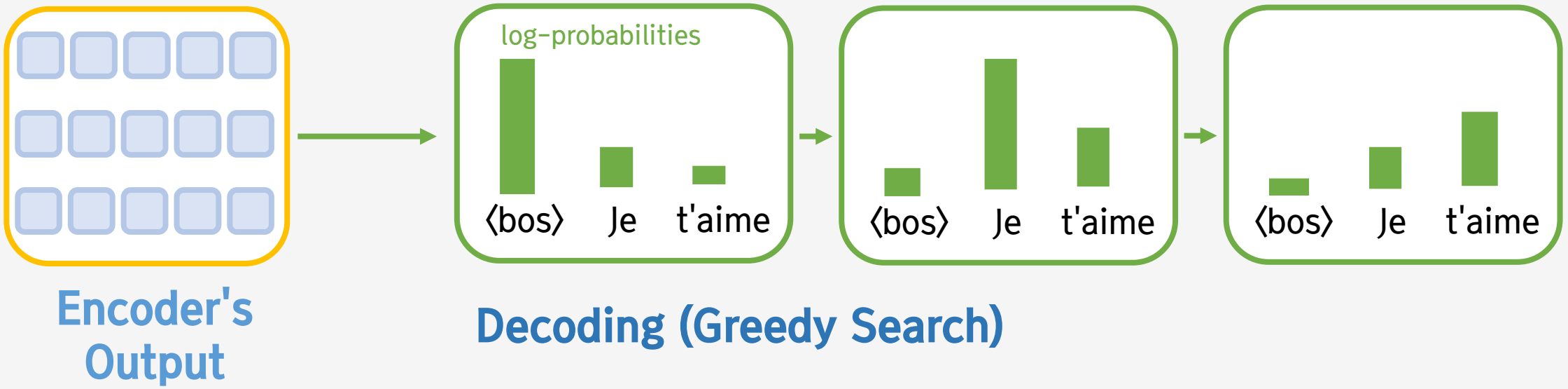
출처 https://beausty23.tistory.com/223

- source sentence를 학습하는 부분
- raw sentence -> tokenizing -> embedding -> encoding



4. Decoder

- target sentence 를 학습하는 부분
- encoder's output 을 통해 ⟨eos⟩토큰 혹은 max_length 까지 토큰을 생성

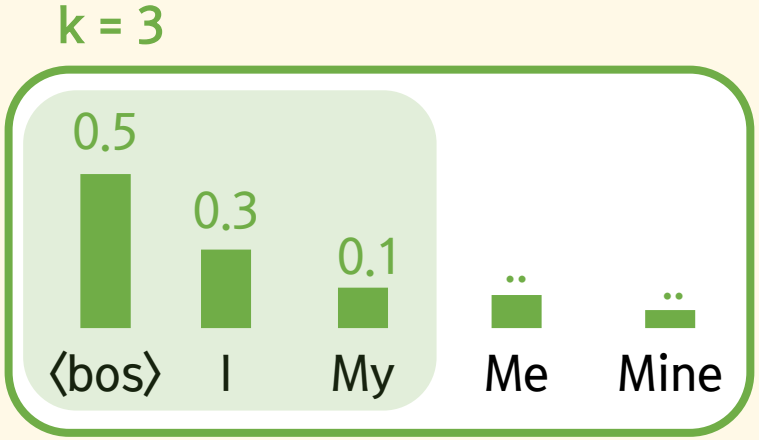


5. Decode strategy

- 크게 sampling(확률론적)과 search(결정론적) 방법이 있음

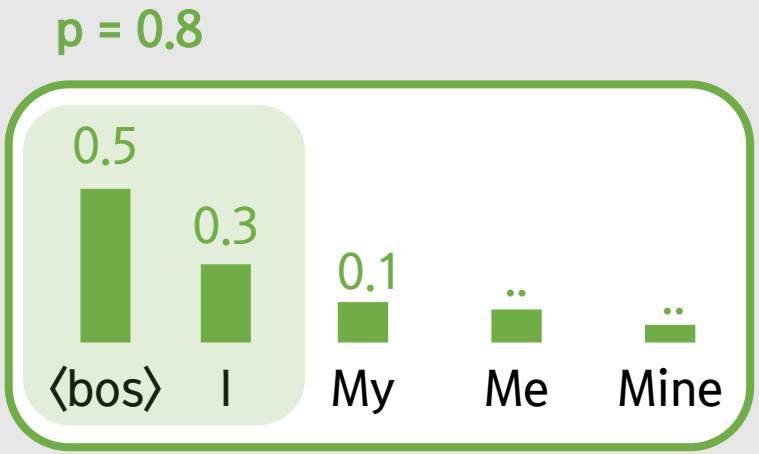
top-k sampling

- 후보군 중에서 상위 k개 만큼의 후보에서 next token 을 샘플링



top-p sampling

- 후보군 중에서 p확률만큼의 후보에서 next token 을 샘플링

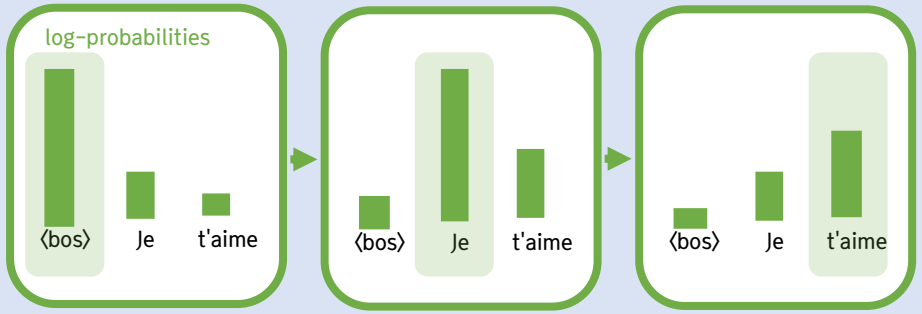


5. Decode strategy

- 크게 sampling(확률론적)과 search(결정론적) 방법이 있음

Greedy search

- 후보군 중에서 확률이 가장 높은 1개의 token 으로만 next token 선택



Beam Search

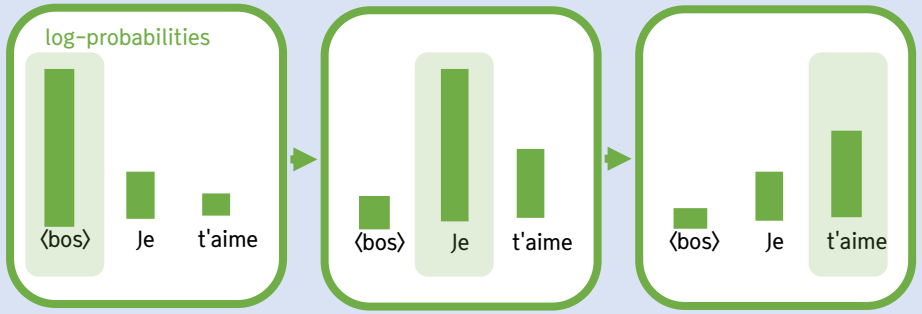
- n 개의 beam을 생성하여 search 수행

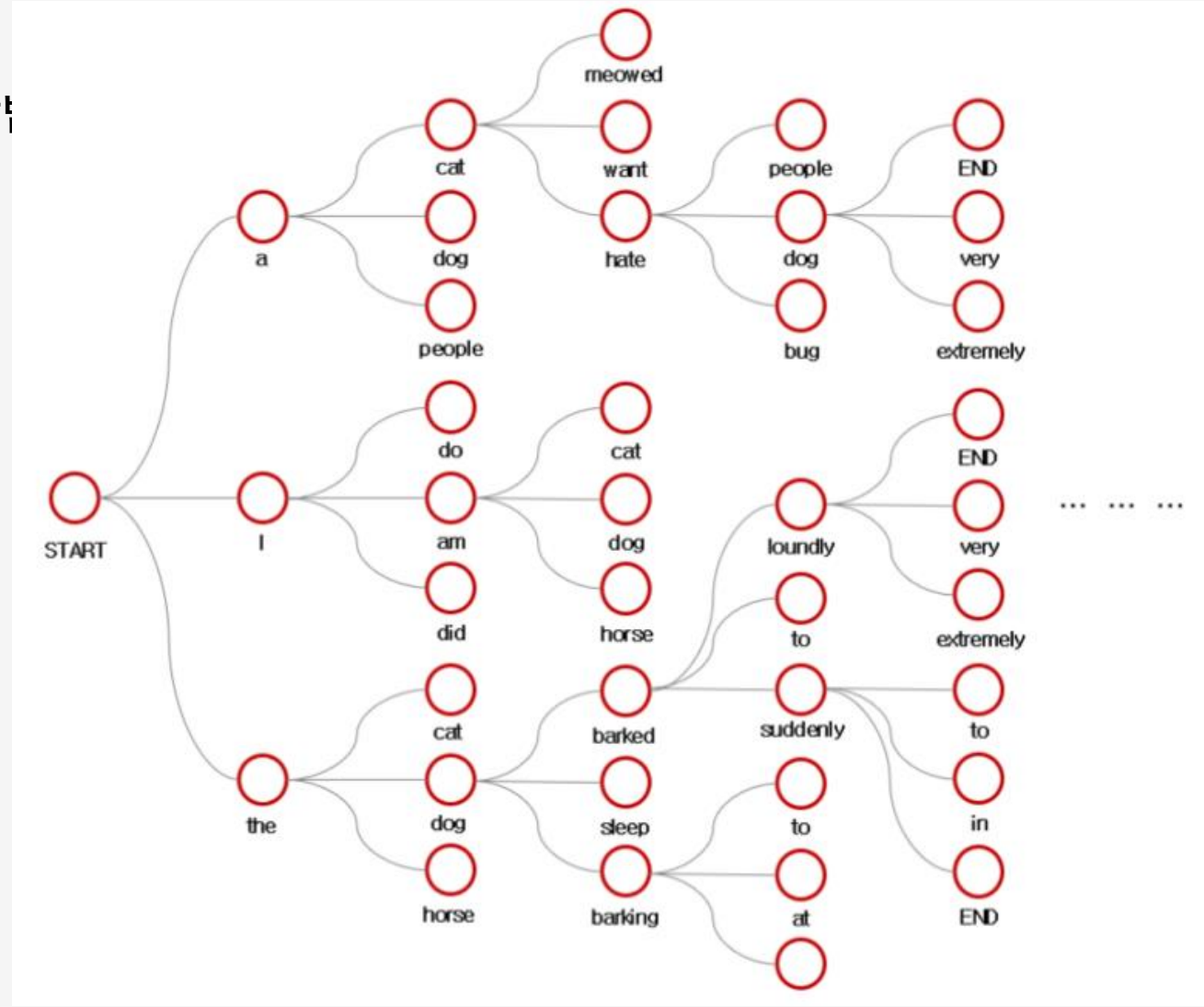
5. Decode strategy

- 크게 sampling(확률론적)과 search(결정론적) 방

Greedy search

- 후보군 중에서 확률이 가장 높은 1개의 token 으로만 next token 선택

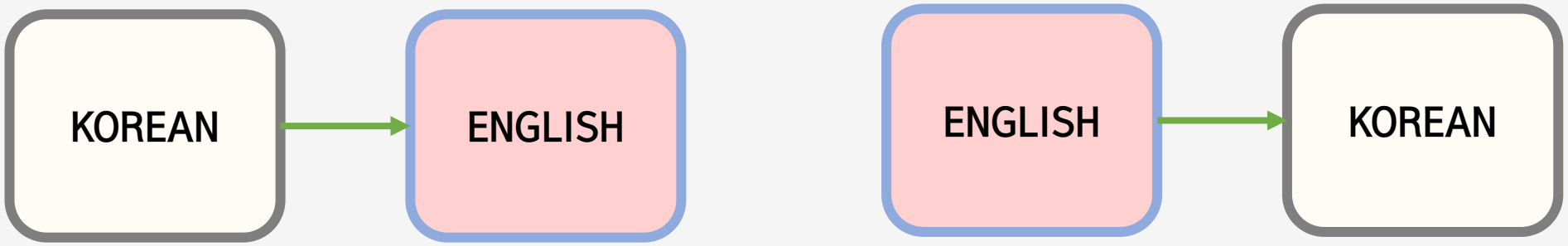




6. Technique of improving performance of MT models.

- Back-translation Paper Blog
- I have 10K Korean-English sentence Pairs, and I have 20K Korean Sentences

10K Korean-English sentence Pairs **(Training)**

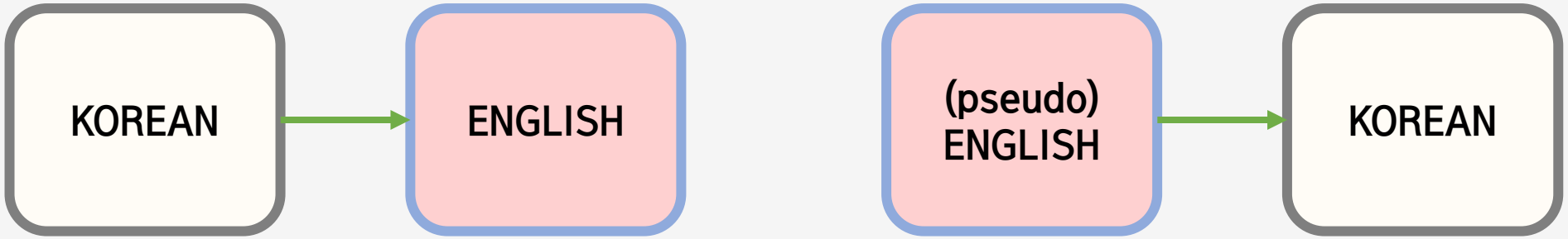


20K Korean Sentences

(Generation)

pseudo corpus가 너무 많아도 좋지 않음.
2~3배가 정도가 적당하다고 알려져 있음.

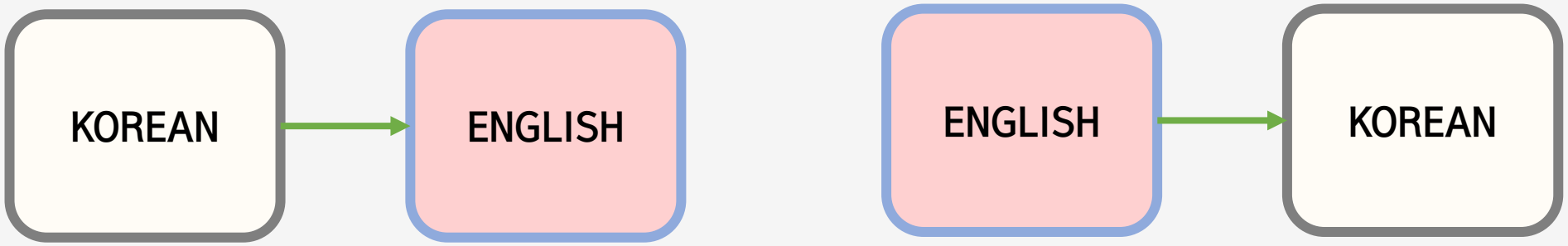
Training



6. Technique of improving performance of MT models.

- Back-translation Paper Blog
- I have 10K Korean-English sentence Pairs, and I have 20K Korean Sentences

10K Korean-English sentence Pairs **(Training)**

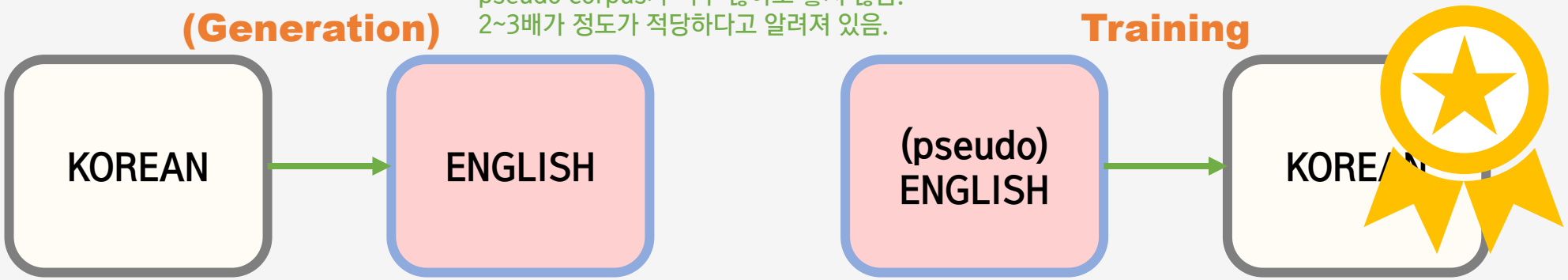


20K Korean Sentences

(Generation)

pseudo corpus가 너무 많아도 좋지 않음.
2~3배가 정도가 적당하다고 알려져 있음.

Training



References

- Embedding vs Encoding link
- top-p, top-k sampling link
- beamsearch written by sooftware link
- Back-translation written by kh-kim link
- Improving Neural Machine Translation Models with Monolingual Data link
- OpenNMT-py github

Thank You