청년 AI Bigdata 교육

Machine translation

NLP lab 김영재
yj122198@postech.ac.kr

How?

토큰화

```
from transformers import BertTokenizer

tokenizer = BertTokenizer.from_pretrained("bert-base-cased")
```

```
tokenizer("Using a Transformer network is simple")
```

```
{'input_ids': [101, 7993, 170, 11303, 1200, 2443, 1110, 3014, 102],
 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0],
 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1]}
```

임베딩

- 희소 표현

  벡터의 유의미한 유사성을 표현할 수 없음

  ex) one-hot 방식

- 밀집 표현

  벡터의 차원을 단어 집합의 크기로 상정하지 않음.

  사용자가 설정한 값으로 모든 벡터표현의 차원을 맞춤

  ex) embedding 방식

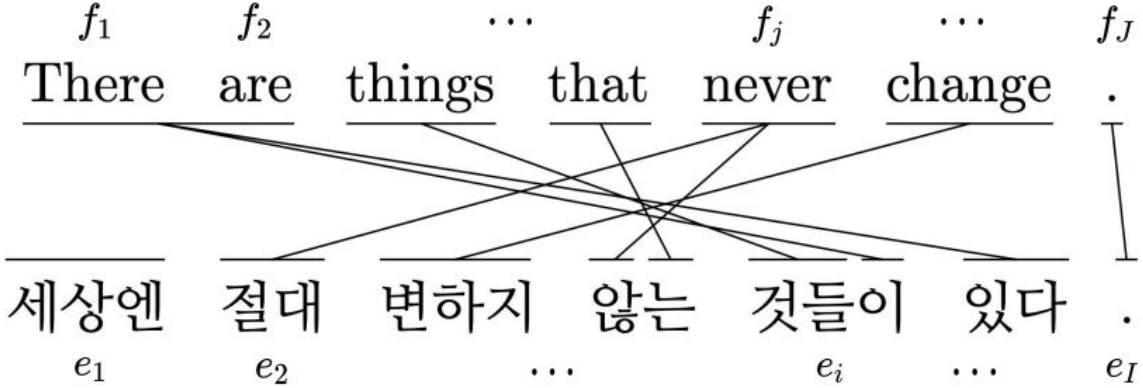
$$\textit{expect} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$$

https://www.youtube.com/watch?v=0RacJ0MQcDA

NMT(Neural Machine Translation)

Machine Translation

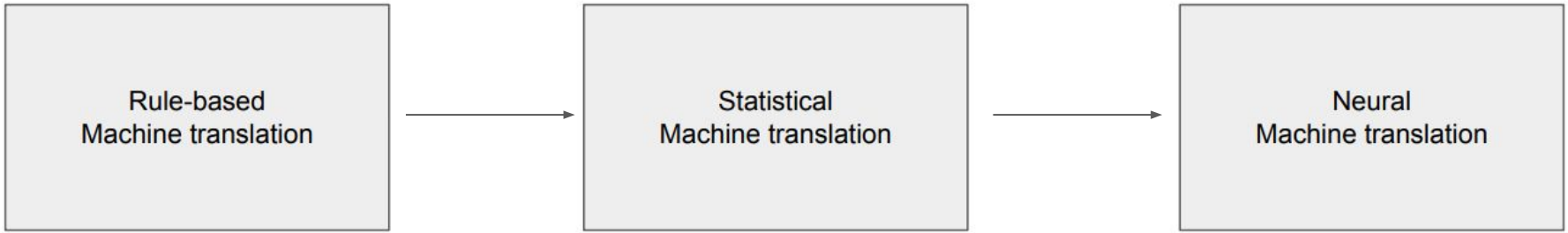
Input = **source** sentence



Output = **target** sentence

- Source length $J \neq$ Target length I

Machine Translation



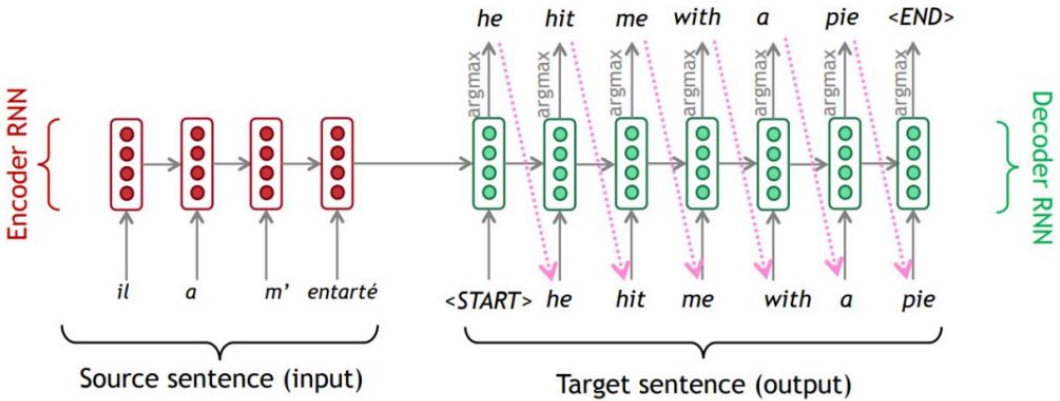
인공지능을 활용한 주요 통·번역 방식

순위	회사명
1 규칙 기반 기계 번역 (RBMT)	- 개별 단어 위주로 접근한 다음 규칙(문법)을 적용 - '모자를 쓰다'와 '편지를 쓰다'의 '쓰다'를 같은 뜻으로 인식 - 문법의 다양한 용례를 짚어내는 데는 효과적
2 통계 기반 기계 번역 (SMT)	- 방대한 분량의 기록을 활용해 언어 데이터 수집 - 다른 언어권 고유의 단어나 관용적 표현에는 약점 - '육회'를 '6회'와 같은 의미로 짚어 'Six times'로 번역
3 인공 신경망 기반 기계 번역(NMT)	- 사람 뇌 학습법 본떠 인공지능 스스로 빅데이터 학습 - 단어보다 전체 문장을 통째로 인식하면서 의미 파악 나서 - '육회'를 'Raw meat'로 알맞게 번역할 만큼 수준 높아져

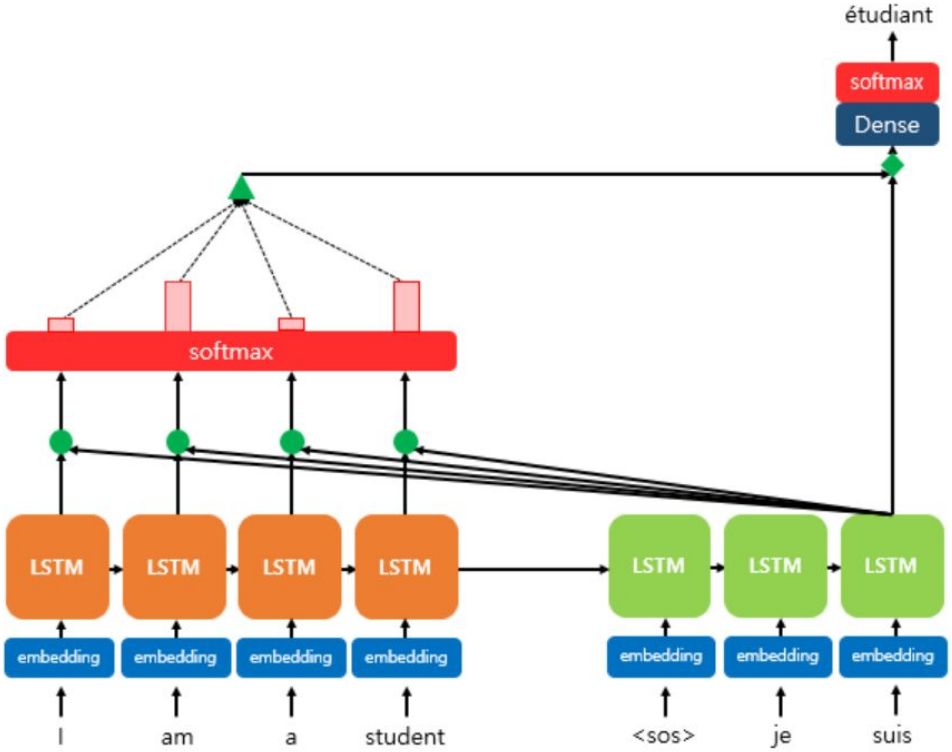
How?

Seq2Seq NMT via fixed-length representations

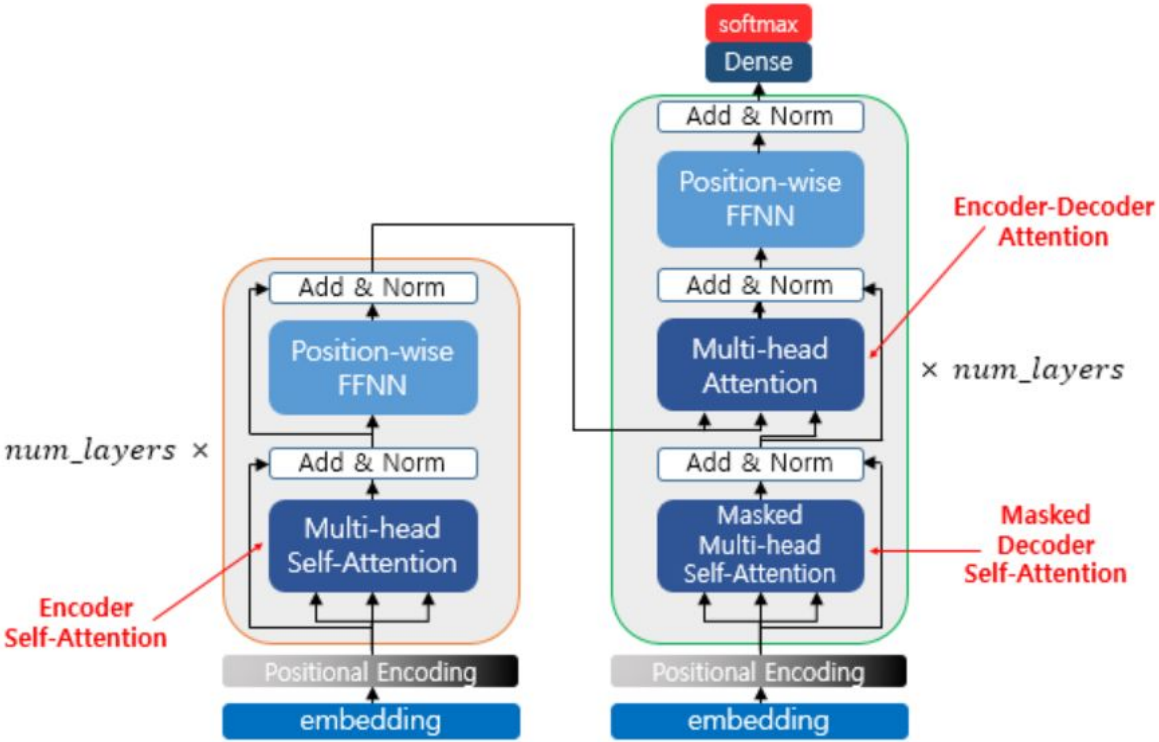
- Encoder RNN compresses input sequence into a fixed-length representation
- Decoder RNN produces output sequence from the representation
  - Each produced output token is fed into the next RNN's input



Attention


étudiant
softmax
Dense
softmax
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
embedding
embedding
embedding
embedding
embedding
embedding
embedding
I
am
a
student
<sos>
je
suis

Transformer



Evaluation

번역 평가는 왜 어려울까?

- 맞다 / 틀리다로 단순하게 구분하기 어렵다. (정답이 많다.)
- 비슷한 번역 결과는 비슷한 점수를 얻어야 한다.

Evaluation

번역 평가의 기준

- 정확도
  - 번역의 의미가 얼마나 정확한가?
  - 더해지거나 없어지거나 대체된 부분이 있는가?

- 유창성
  - 얼마나 문법적으로 자연스러운가? 번역체스럽지 않은가?
  - 얼마나 자연스럽게 읽히는가? 어순이 이상하지는 않은가?

Evaluation

번역 평가 방식

- Human evaluation
  - 주관적이다. 평가자의 기준에 따라서 점수가 바뀔 수 있다.
  - 신뢰성이 없다. 평가자별로 기준이 다르기에 서로의 점수에 동의하지 않을 수 있다.
  - 비싸다. 평가자를 고용하는 비용이 든다.
  - 평가를 원복할 수 없다.

- Auto evaluation
  - 문장이 얼마나 가까운지 평가하는 방식.
  - 빠르고 저렴하다.
  - 원복이 가능하다.

Evaluation

Auto evaluation - Precision and Recall

$$\text{Precision} = \text{Recall} = \text{F1} = 1.0$$

- reordering에 대한 penalty가 전혀 없음.

Evaluation

Auto evaluation - BLEU(Bilingual Evaluation Study)

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Evaluation

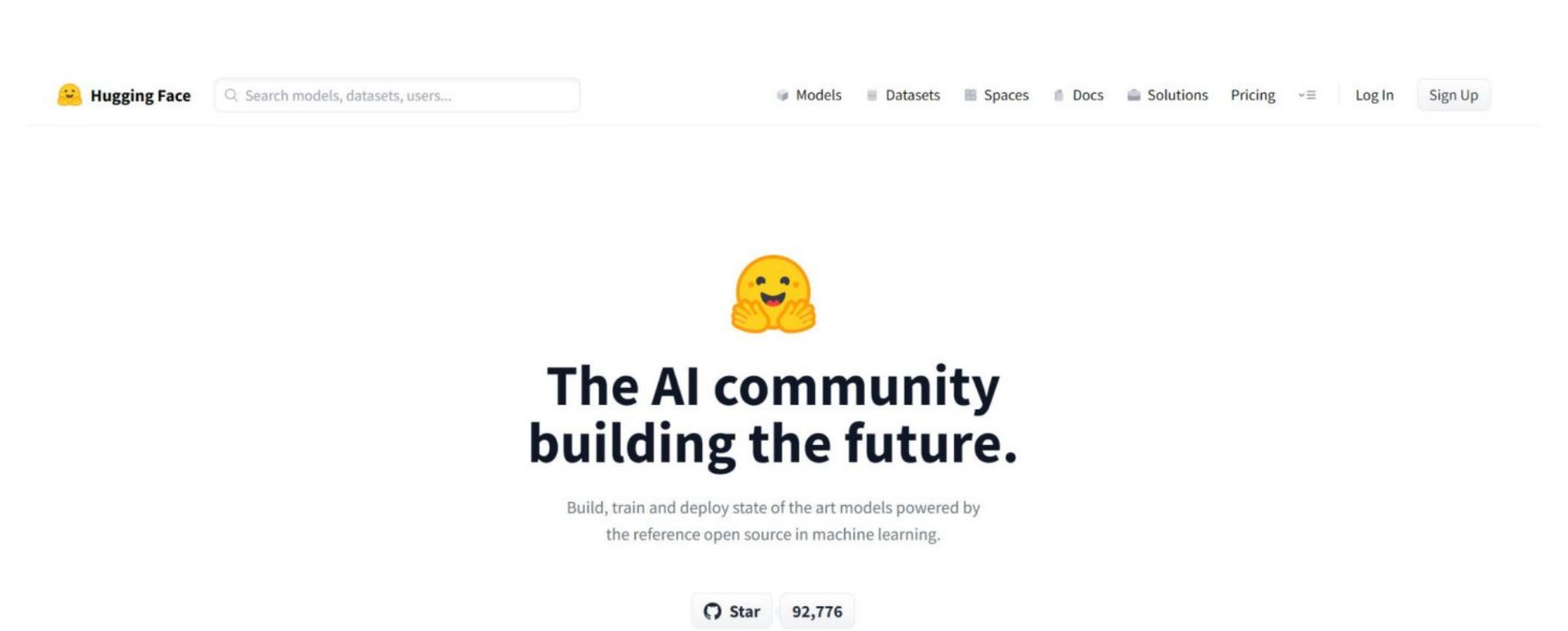
Auto evaluation - WER(Word Error Rate)

Minimum number of **edit**ing steps to transform output to reference

- Match: words match, no cost
- Substitution: replace one word with another
- Insertion: add a word
- Deletion: drop a word

Levenshtein distance, a.k.a. edit distance:

$$\mathrm{WER}\left(\hat{e}_1^{\hat{I}}, \dot{e}_1^{\dot{I}}\right) = \frac{\#\text{substitutions} + \#\text{insertions} + \#\text{deletions}}{\dot{I}}$$



https://wikidocs.net/166832