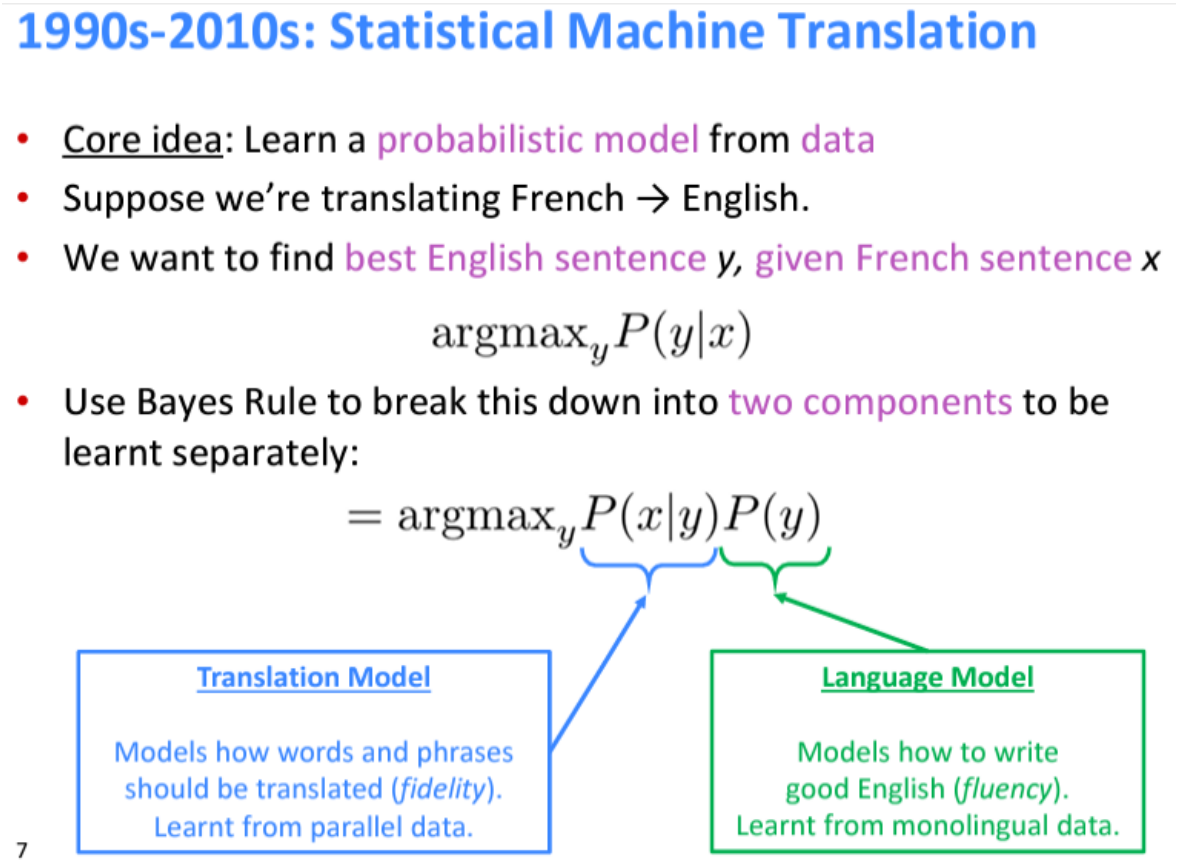
CS224n Lecture 8

Machine Translation

- Stastical Machine Translation(SMT)



alignment

- alignment: 서로 다른 언어로 번역된 문장들의 쌍 간에 '단어들의 일치'를 의미한다.

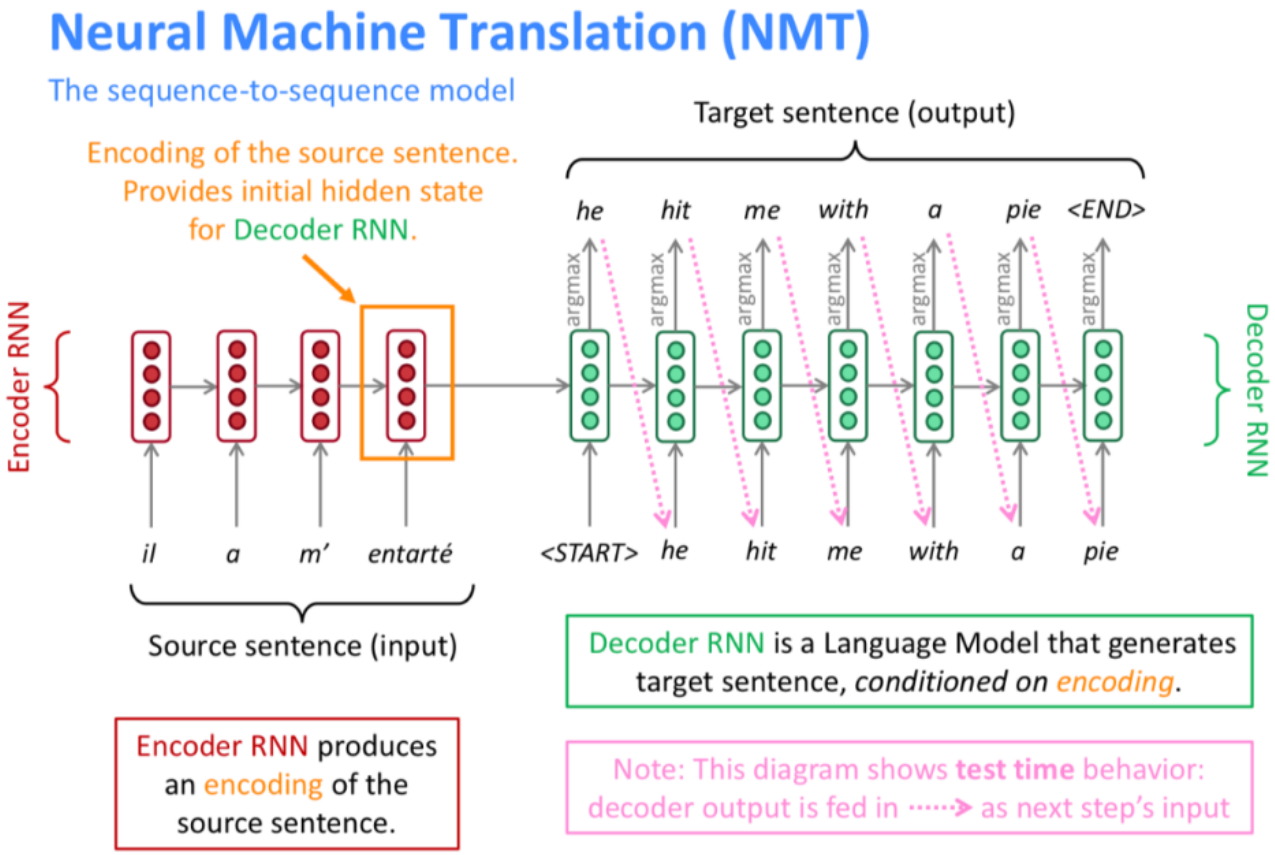
- 실제로 일치하는 counterpart 가 없거나 "one to many", "many to many", "many to one" 등 다양한 경우가 발생하는데, 이 때 계산 비용이 많이 발생하게 된다.

Neural Machine Translation(NMT)

: 하나의 neural network 로 machine translation 을 수행하는 것으로,

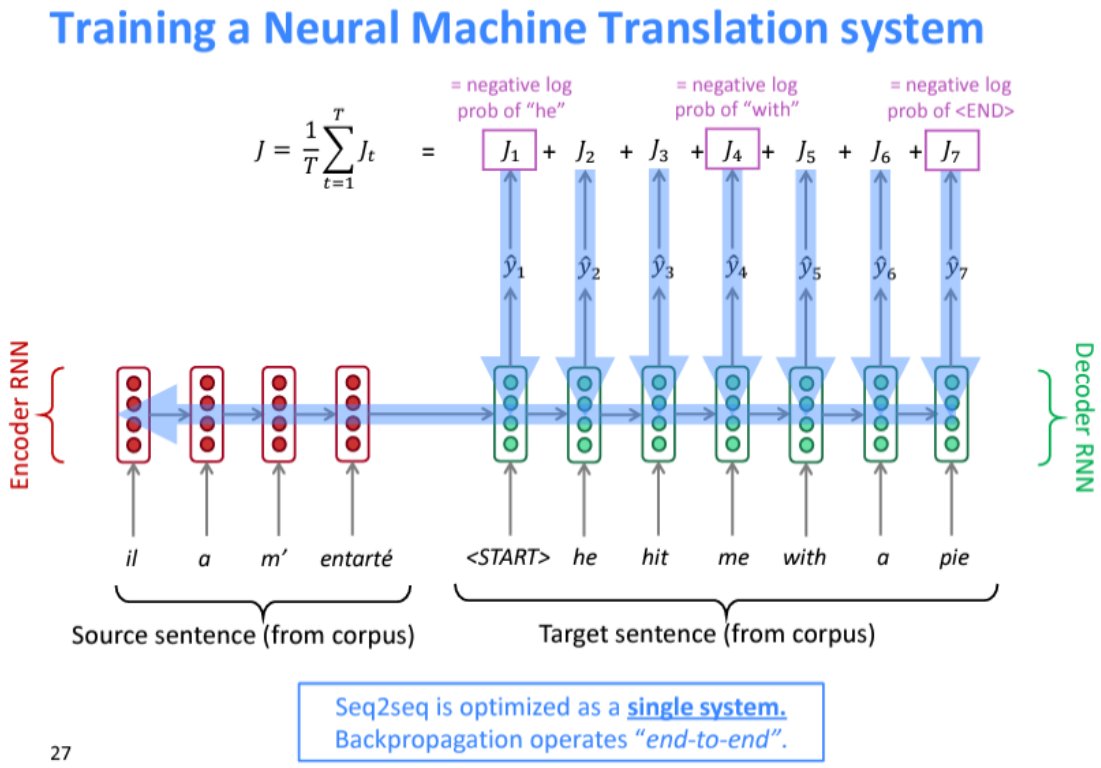
neural network architecture 은 sequence-to-sequence(seq2seq)라 한다.

Translation, Conversation, Summarization 과 같이 sequential output 에 의존적인 문제를 다루는 seq2seq 모델들이 있다.



Seq2Seq

- encoder: 모델의 input sequence 를 출력값으로 받고, 고정된 사이즈 'context vector'로 encode 한다.
- decoder: context vecotr 를 'seed'로 사용하며, seed 는 output sequence 를 생성한다.



* Training a Neural Machine Translation system

- Greedy decoding

- Beam search decoding

- NMT 의 장점: SMT 에 비해 더 나은 성능(fluency, 'context' & 'phrase similarities' usage), 최적화된 end-to-end 방식 등

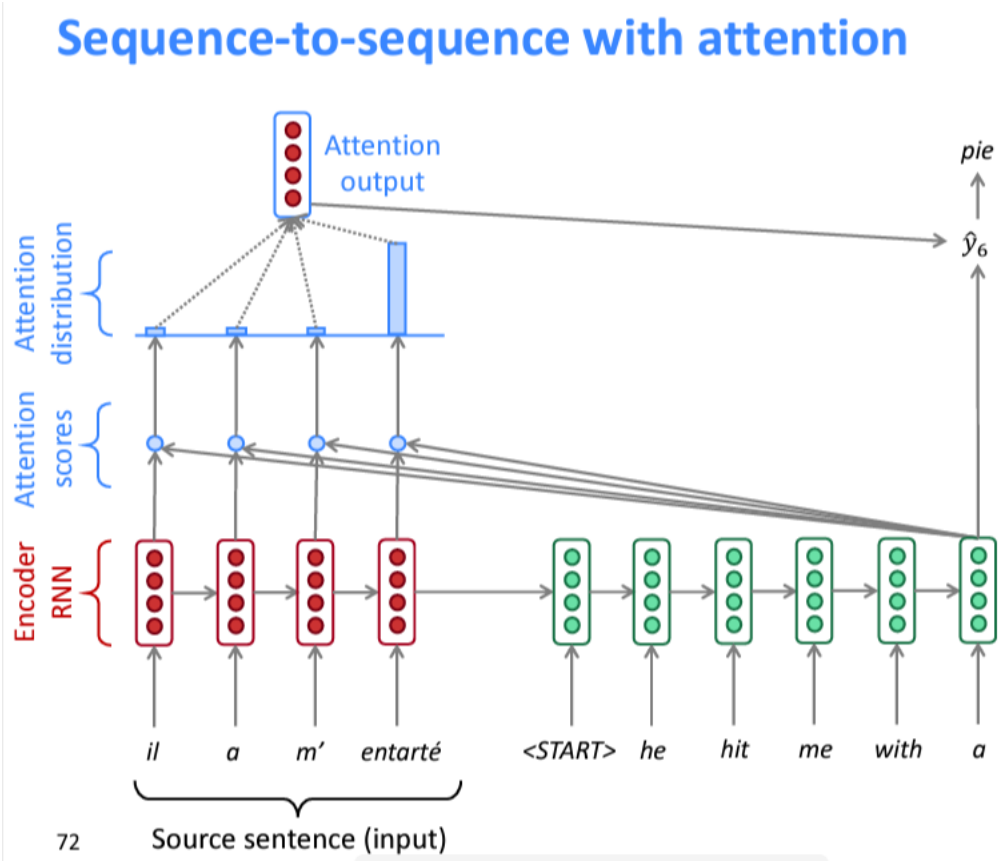
(그러나 디버그하기가 비교적 어렵다거나 control 이 어렵다는 단점도 가지고 있다.)

BLEU

: Bilingual Evaluation Understudy 로, 사람이 번역한 것과 기계가 번역한 것을 우선 n-gram similarity 로 스코어를 매긴다. 그리고 너무 함축하여 번역한 문장에 대해서는 패널티를 준다. 그렇게 사람이 직접 번역한 것과 기계가 번역한 것의 유사도를 판단한다.

Attention

: seq2seq 의 문제점은 encoder 에서 decoder 로 넘어갈 때, 하나의 hidden state 만을 가지기 때문에 information bottleneck 이 될 수 있다는 점이다. 그래서 이 점을 decoder 의 각 step 을 encoder 로 직접 연결하자는 점이다.



일단 가장 중요한 NMT 성능이 크게 향상되었다. 그리고 bottleneck 문제도 해결하였다. direct connection 이 생기니 vanishing gradient problem 도 많이 해결되었다. NMT 의 단점이로 평가받던 디버그 하기 어렵다던 문제도 어느정도 풀렸다. attention 을 시각화할 경우 alignment 처럼 나온다.