Neural Machine Translation by jointly learning to align and translate

임송현

기계번역의 종류

규칙 기반 기계번역

- 주어진 문장의 구조를 분석해 규칙을 세움
- 분류를 나누어 정해진 규칙에 따라 번역
- 매번 새로운 규칙을 찾아내고 적용해야 하는 한계

통계 기반 기계번역

- 언어쌍을 확장할 때 대부분의 알고리즘이나 시스템은 유지됨
- 여러 가지 모듈로 이루어져 매우 복잡

신경망 기계번역

- 인코더-디코더 형태의 구조
- 성능이 낮음

딥러닝을 이용한 기계번역

- 입력 문장(x)이 주어졌을 때 신경망 모형을 통해 **조건부 확률을 최대화**하는 번역 문장(y)를 찾는 것 $\arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$
- 인코더-디코더 형태의 구조
- RNN, LSTM, GRU를 네트워크로 사용
  ① 입력, 출력 길이 제한 x
  ② Sequential data 에 대한 **순서 학습** 가능 (문장 전체 -〉 의미 -〉 목표 언어로 번역)
  ③ **예외 상황에 강건** (사전 기반 알고리즘에 비해 예외 상황에 유연)

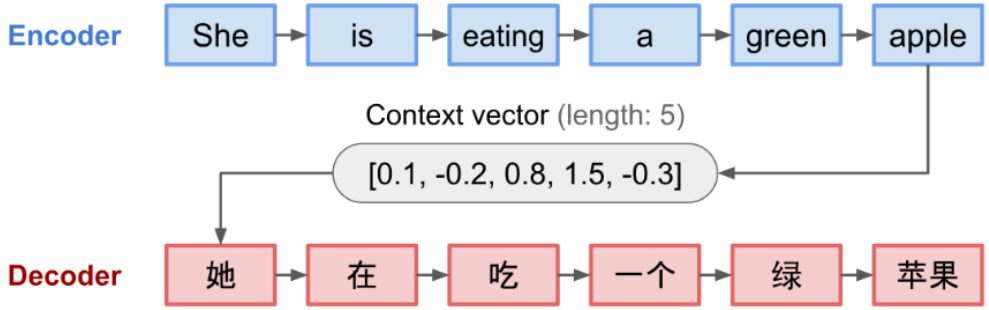
(https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html?fbclid=IwAR1QpSK0Eqf20b7YADsK9cOilhvBuF6lStybrN7rr815tbhf2bZEAaeB86U)

Seq2seq 모형

인코더 : 입력 시퀀스 처리 / 고정된 길이의 컨텍스트 벡터로 정보 압축

디코더 : 인코더 신경망의 마지막 상태를 초기 상태로 사용함

인코더에서 압축된 컨텍스트 벡터로 초기화하여 변역 결과를 내놓음



고정된 길이의 context 벡터의 사용

- 문장 앞 부분의 정보가 희석됨
- 길이가 긴 입력 시퀀스에 대해 long term dependency
- LSTM으로 일정 수준 개선되지만 완전한 해결은 못함 **-> 어텐션 매커니즘**

동기 및 아이디어

- 동기 소스 문장의 길이가 긴 문장의 기억을 돕기 위해 만들어짐
- 아이디어 '중요한 부분만 집중하게 만들자'

ex. 나는 자연어 처리 스터디에 참여했다.

- 하나의 문장 내의 가까운 문맥에서 단어의 관계 설명
- 다음 단어를 예측하는데 단어 사이의 상관관계 분석을 통해 focusing

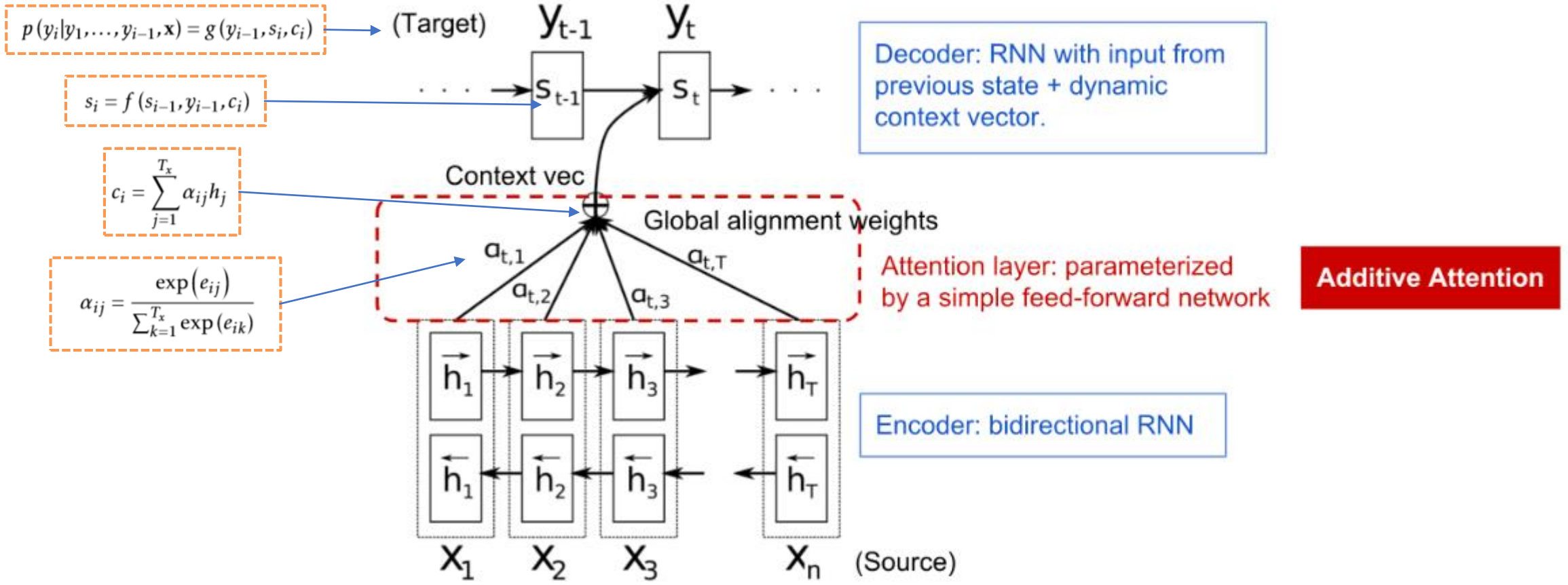
나는 자연어처리 스터디에 참여했다. high attention

나는 자연어처리 스터디에 참여했다. high attention

나는 자연어처리 스터디에 참여했다. low attention

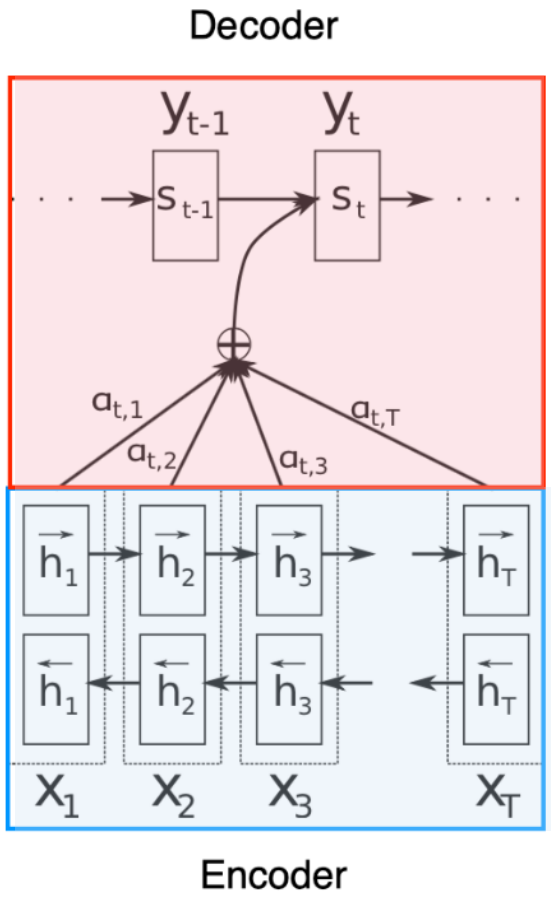
(https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html?fbclid=IwAR1QpSK0Eqf20b7YADsK9cOilhvBuF6lStybrN7rr815tbhf2bZEAaeB86U)

어텐션 매커니즘 구조



(https://arxiv.org/pdf/1409.0473.pdf)

어텐션 매커니즘 구조



인코더

- Bi-directional RNN 사용
- 입력 x -〉 은닉 상태 h

디코더 (y_i를 예측할 때)

$$e_{ij} = a(s_{i-1}, h_j) \qquad \text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

- e_{ij} : 은닉 상태 벡터 s_t와 인코더의 j 번째 열벡터가 얼마나 유사한지 나타내는 값
- alignment model은 다양한 형태로 변형가능

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

- 인코더의 j번째 열벡터에 대한 어텐션 확률의 가중합

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

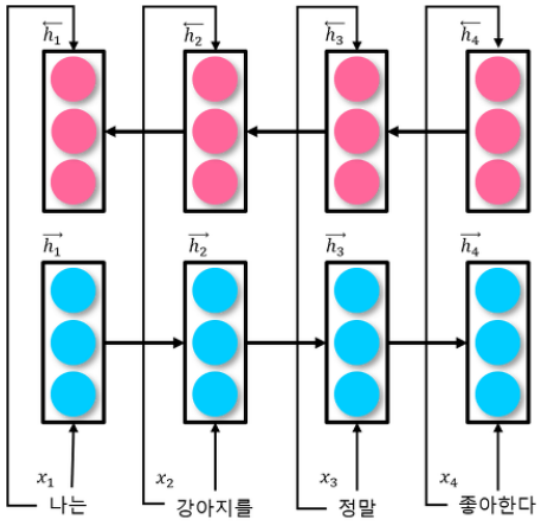
- 직전 시점의 은닉 상태 벡터 / 직전 시점의 디코더 출력 / 현재 컨텍스트 벡터

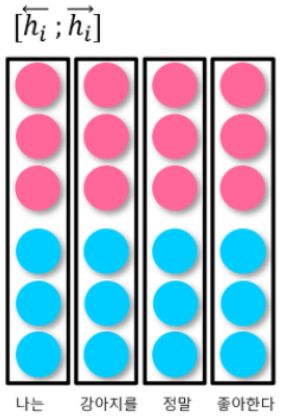
$$p(y_i | y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

예시 : 나는 강아지를 정말 좋아한다. ➡ I really like puppies.

인코더 (bi-directional model)

- 입력 시퀀스를 받아 은닉상태 벡터 h_j생성
- 생성된 벡터들을 차례로 쌓아 은닉 상태 벡터 행렬 F 생성

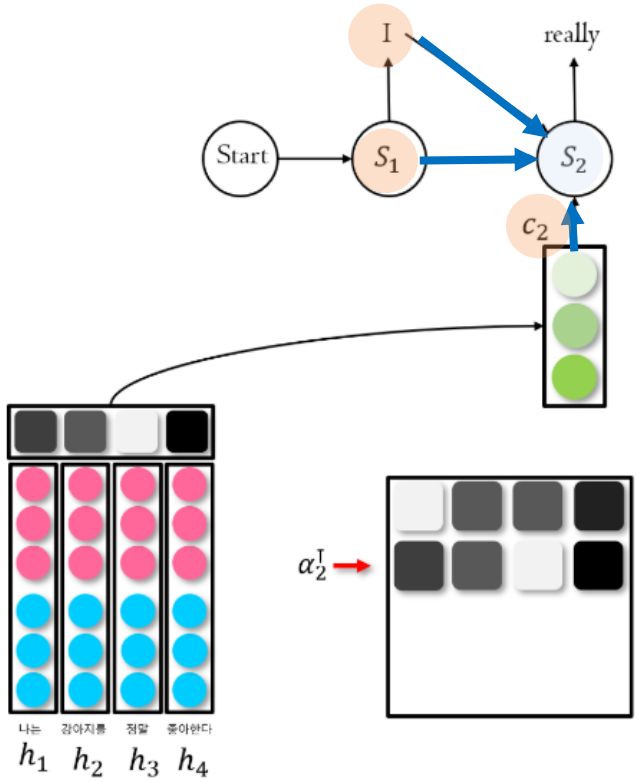


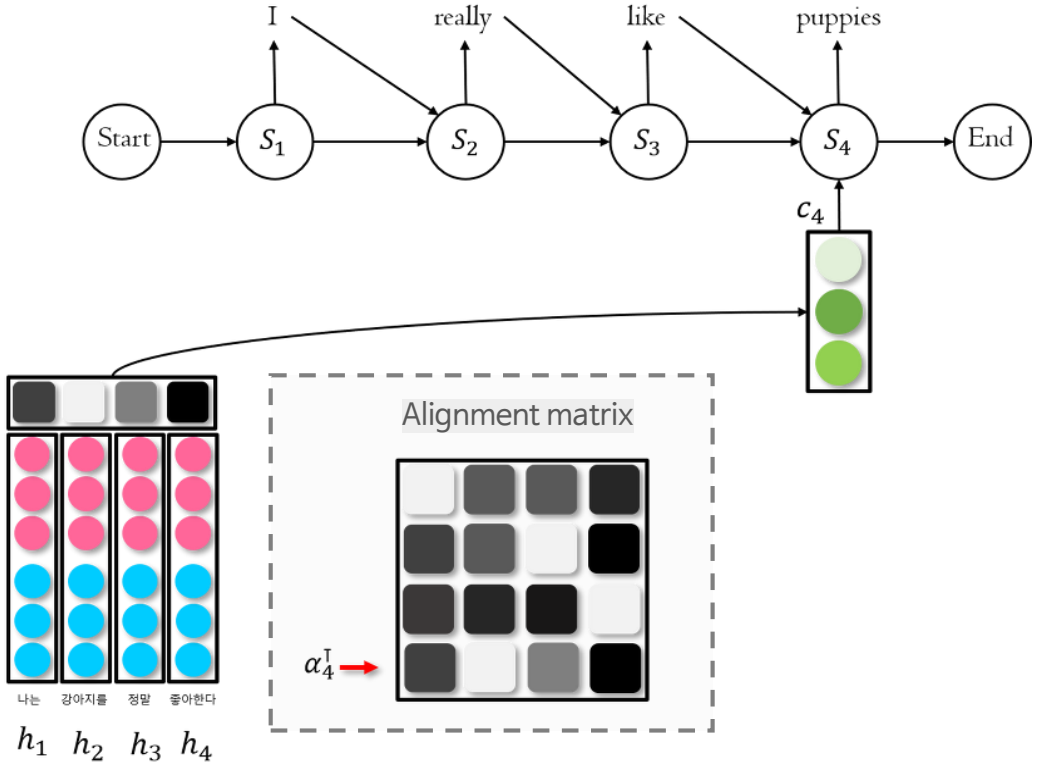


예시 : 나는 강아지를 정말 좋아한다. ➡ I really like puppies.

디코더

- 입력 : 인코더의 은닉 상태 벡터, 직전 시점의 은닉 벡터, 컨텍스트 벡터
- 출력 : 번역된 단어





(https://arxiv.org/pdf/1409.0473.pdf)

논문 결과 해석 English-to-French 번역

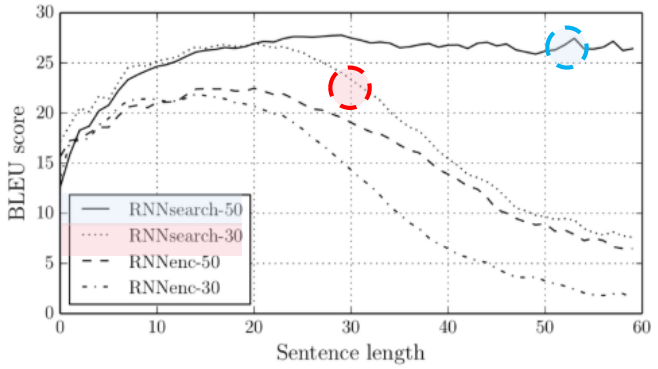
- 데이터 ACL WMT14 자료 : 850M -〉 데이터 선택[1] -〉 348M
- 평가 성능 측정 지표 BLEU ↑
- 모형 RNNsearch (어텐션) / RNNencdec (RNN 인코더-디코더)
- 세팅 인코더와 디코더 각각 1000개의 은닉 단위

  각 언어에서 자주 등장하는 단어 30,000개에 대한 shortlist를 사용

  포함되지 않은 단어는 특수 토큰([UNK])

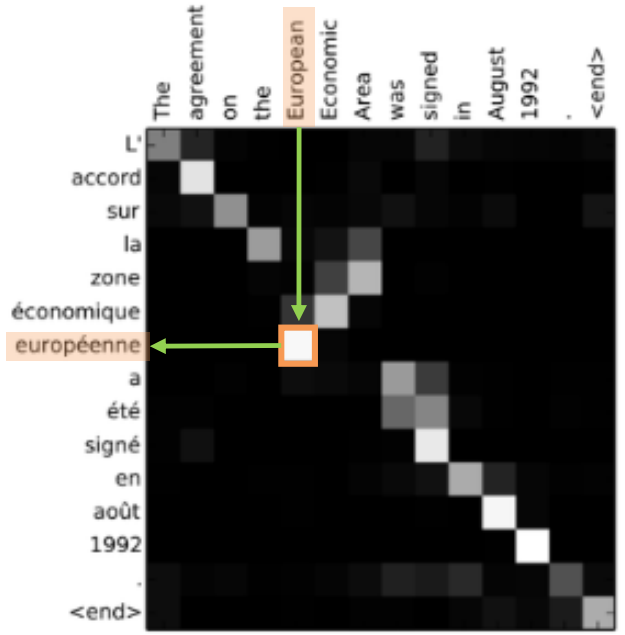
  학습 시 문장의 단어 길이 30 / 50

문장 길이에 따른 테스트 셋의 번역 결과에 대한 BLEU점수



훈련데이터의 BLEU 점수

Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63



테스트 문장에 대한 RNNsearch-50 Alignment matrix

1. Axelrodetal. (2011)

어텐션의 장점

- 행렬이기 때문에 네트워크로 학습가능
- 언어 모델에 적합한 특성을 지님

  언어별 문법, 번역 형태가 다른 것처럼 언어는 유연하기 때문에 예외 상황에 대해 스스로 학습

- 입력과 출력의 길이가 달라도 번역 가능

Reference

- https://arxiv.org/pdf/1409.0473.pdf
- https://curaai00.tistory.com/9
- https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/10/06/attention/
- https://nlpstudynote.tistory.com/18
- https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html?fbclid=IwAR1QpSK0Eqf20b7YADsK9cOilhvBuF6lStybrN7rr815tbhf2bZEAaeB86U
- 소문난 명강의 김기현의 자연어 처리 딥러닝 캠프