Natural Language Processing

언어 모델: Seq2Seq

강사 양석환

Sequence To Sequence 모델

번역

- 번역
  - 어떤 언어로 된 글을 다른 언어의 글로 옮기는 것
  - 인류가 언어를 사용하기 시작한 이래로 계속되어온 큰 관심사의 하나
  - 자연어 처리 분야에서의 종합예술이라고 칭할 정도로 다양한 기술이 통합됨
  - 번역의 궁극적인 목표
    - 어떤 언어 f의 문장이 주어졌을 때, 가능한 e 언어의 번역 문장 중에서 최대 확률을 가지는 $\hat{e}$를 찾아내는 것

$$\hat{e} = argmax\ P_{f \to e}(e|f)$$

번역이 어려운 이유

- 번역이 어려운 이유
  - 인간의 언어(자연어)는 컴퓨터 프로그래밍 언어처럼 명확하지 않다.(모호성)
  - 자연어는 그 활용에 있어서의 효율을 극대화하는 쪽으로 흘러간다.
    - 우리는 정보나 단어를 생략하고, 문장을 짧게 만들며, 동일한 단어와 어절을 상황에 따라서 다른 의미로 사용(경제성, 효율성)한다.
    - 특히 한국어는 어순이 불규칙하고, 주어가 생략되는 등(그래도 다 이해한다) 그 효율성의 추구가 극대화된 언어의 하나이다.
  - 언어는 문화를 내포하고 있으므로 수천 년간 쌓여온 사람의 의식, 철학 등이 녹아 들어가 있어서 그러한 문화의 차이가 번역을 더욱 어렵게 만든다.

번역이 어려운 이유

- **영화 <그린 랜턴> 대사의 오번역 예**

원문	오번역
In brightest day, in blackest night.	일기가 좋은 날, 진흙같은 어두운 밤.
No evil shall escape my sight.	아니다 이 악마야, 내 앞에서 사라지지
Let those who worship evil's might.	누가 사악한 수도악마를 숭배하는지 볼까
Beware my power, Green Lantern's light!!!	나의 능력을 조심해라, 그린 랜턴 빛!!!

올바른(???) 번역 예

- 가장 밝은 낮에도, 가장 어두운 밤에도,
- 나의 시야에서 벗어날 악은 결코 없으니.
- 악의 힘을 숭배하는 자들이여,
- 나의 힘을 경계하라, 그린 랜턴의 빛을!!!

https://blog.naver.com/blaze_terran/222073387132

기계번역의 역사

- 기계번역의 역사
  - 규칙 기반 기계 번역
  - 통계 기반 기계 번역
  - 딥러닝 이전의 신경망 기계 번역
  - 딥러닝 이후의 신경망 기계 번역

규칙 기반 기계 번역

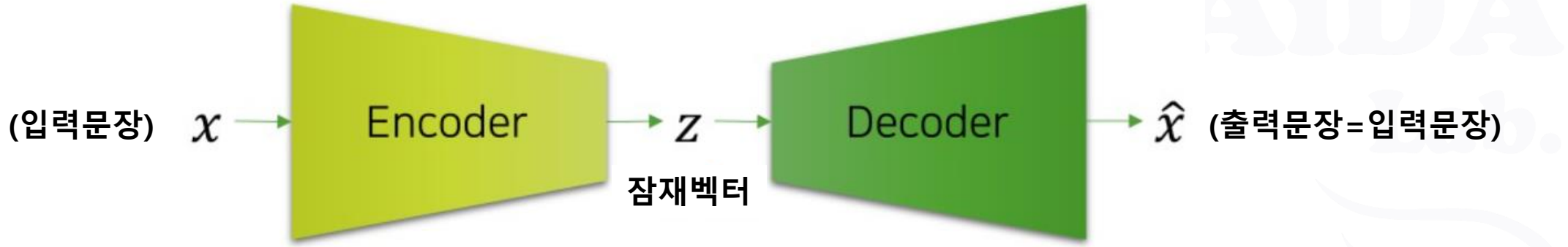
- 규칙 기반 기계 번역(Rule-Based Machine Translation, RBMT)
  - 가장 전통적인 번역 방식
  - 주어진 문장의 구조를 분석하고, 그 분석에 따라 규칙을 세운 후, 분류를 나누어 정해진 규칙에 따라 번역
  - 사람의 경우는 일반화 능력이 뛰어나므로 몇 가지 규칙으로 번역을 수행할 수 있지만... 컴퓨터에서는 매우 어려움
  - 규칙이 잘 만들어진다면 통계 기반 기계 번역보다 자연스러운 표현이 가능
  - 그러나 규칙을 사람이 일일이 만들어야 함 → 자원과 시간 등 소요 비용 높음

통계 기반 기계 번역

- 통계 기반 기계 번역(Statistical Machine Translation, SMT)
  - 신경망 기계 번역 이전에 세상을 지배하던 번역 방식
  - 대량의 양방향 코퍼스에서 통계를 얻어내어 번역 시스템 구성
  - 구글의 초기 번역 시스템에 채용되면서 유명해짐
  - 많은 모델로 구성되므로 매우 복잡함
  - 통계 기반 방식이므로 언어쌍의 확장 시, 대부분의 알고리즘, 시스템이 유지되므로 기존의 규칙 기반 기계 번역에 비해 비용적으로 유리했음

딥러닝 이전의 신경망 기계 번역

- 딥러닝 이전의 신경망 기계 번역(Neural Machine Translation, NMT)
  - 신경망 모델이 외면 받던 도중에도 신경망을 이용하여 기계 번역을 해결하려는 시도가 있었음
  - 현재의 언어모델에 사용된 인코더-디코더(Encoder-Decoder) 형태를 가지고 있었으나 컴퓨터의 성능 부족, 데이터의 부족 등의 이유로 제대로 된 성능을 발휘하지 못함



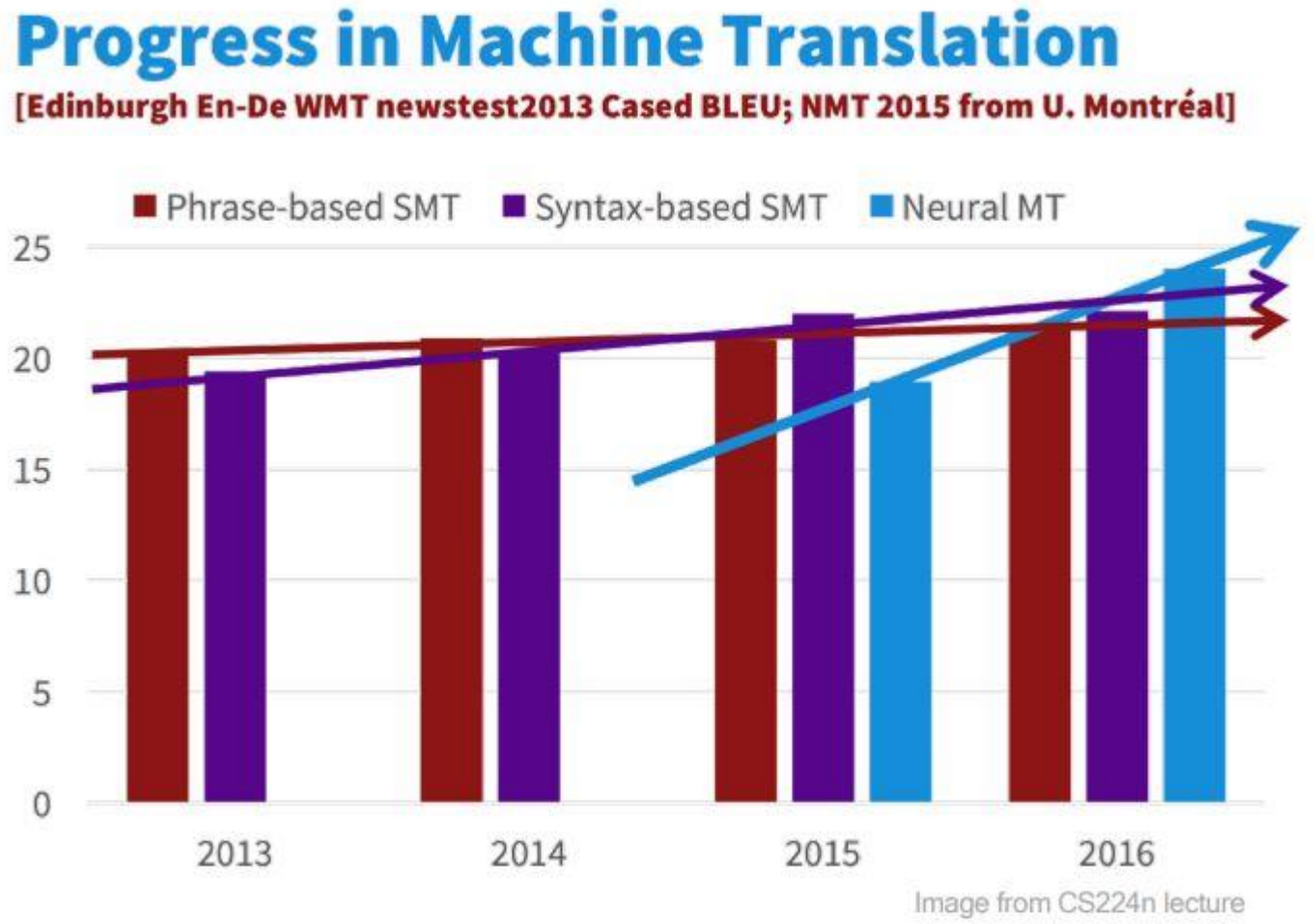
딥러닝 이후의 신경망 기계 번역

- **딥러닝 이후의 신경망 기계 번역**
  - **기존 통계 기반 기계 번역 방식을 순식간에 앞질러버림**
  - **구글 번역기를 포함하여 대부분의 상용 번역기가 딥러닝 기술로 대체됨**
  - **신경망 기반 기계번역의 장점**

항목	내용
end-to-end 모델	기존 통계기반 기계번역 시스템은 수많은 모듈로 구성되어 매우 복잡하고 훈련이 어려웠지만, 딥러닝 기반 방식은 단 하나의 모델로 번역을 수행하여 성능을 극대화 시킴
더 나은 언어 모델	신경망 기반의 모델이므로 기존의 n-gram 방식보다 강력하며, 희소성 문제의 해결 및 자연스러운 번역 결과와 문장생성이 가능해짐
훌륭한 문장 임베딩	신경망의 특징에 따라 문장의 차원축소, 임베딩 능력이 뛰어남. 단어 단위보다 문장단위에서 심각하게 발생하는 노이즈, 희소성 문제에 대한 대처능력이 뛰어남

딥러닝 이후의 신경망 기계 번역

- 통계 기반 기계번역 vs 신경망 기계 번역



Seq2Seq 모델

- Seq2Seq (Sequence to sequence) 모델
  - 한 도메인(예, 영어 문장)에서 다른 도메인(예, 한국어 문장)으로 시퀀스(Sequence)를 변환하는 모델
  - 시퀀스를 다루는 모델이므로 당연히 순차데이터를 대상으로 함. 주로 시계열 데이터에 대하여 많이 활용됨
  - 대표적인 시퀀스 모델인 RNN 모델을 기반으로 하고 있음
  - 2개의 RNN 모델을 이용하여 인코더와 디코더를 구현하였고, 이 때문에 Encoder-Decoder 모델이라고도 부름

Seq2Seq 모델

- **시퀀스에 대한 예측을 목표로 하고 있으며**
  - 자연어 모델링: 각 타임 스텝에서 주어진 시퀀스를 기반으로 다음 단어 예측
  - 품사 태깅: 단어의 문법 품사 예측
  - 개체명 인식: 단어가 사람, 위치, 제품, 회사 같은 개체명에 속하는지 예측

등에 많이 활용되고 있음

Seq2Seq 모델의 목적

- **Seq2Seq 모델의 목적은**
  - **모델 구조를 이용하여 MLE를 수행, 주어진 데이터를 가장 잘 설명하는 파라미터 θ를 찾는 것**
    - MLE(Maximum Likelihood Estimation, 최대 우도법)
      - **주어진 데이터를 토대로 확률 변수의 모수를 구하는 방법**
      - **모수가 주어졌을 때, 원하는 값들이 나올 가능도 함수를 최대로 만드는 모수를 선택하는 방법**
        - 가능도 함수(Likelihood Function): “그럴듯한”의 의미로 모수 θ가 주어졌을 때 주어진 표본 x가 얻어질 확률을 말함
          - 가능도 함수가 크다 → 해당 모수가 θ일때 해당 표본 x가 수집될 확률이 높다.
          - $L(\theta_1|x) > L(\theta_2|x)$라면 θ_1이 θ_2보다 모수일 확률이 높음

Seq2Seq 모델의 목적

- **seq2seq 모델의 목적을 수식으로 나타내면**

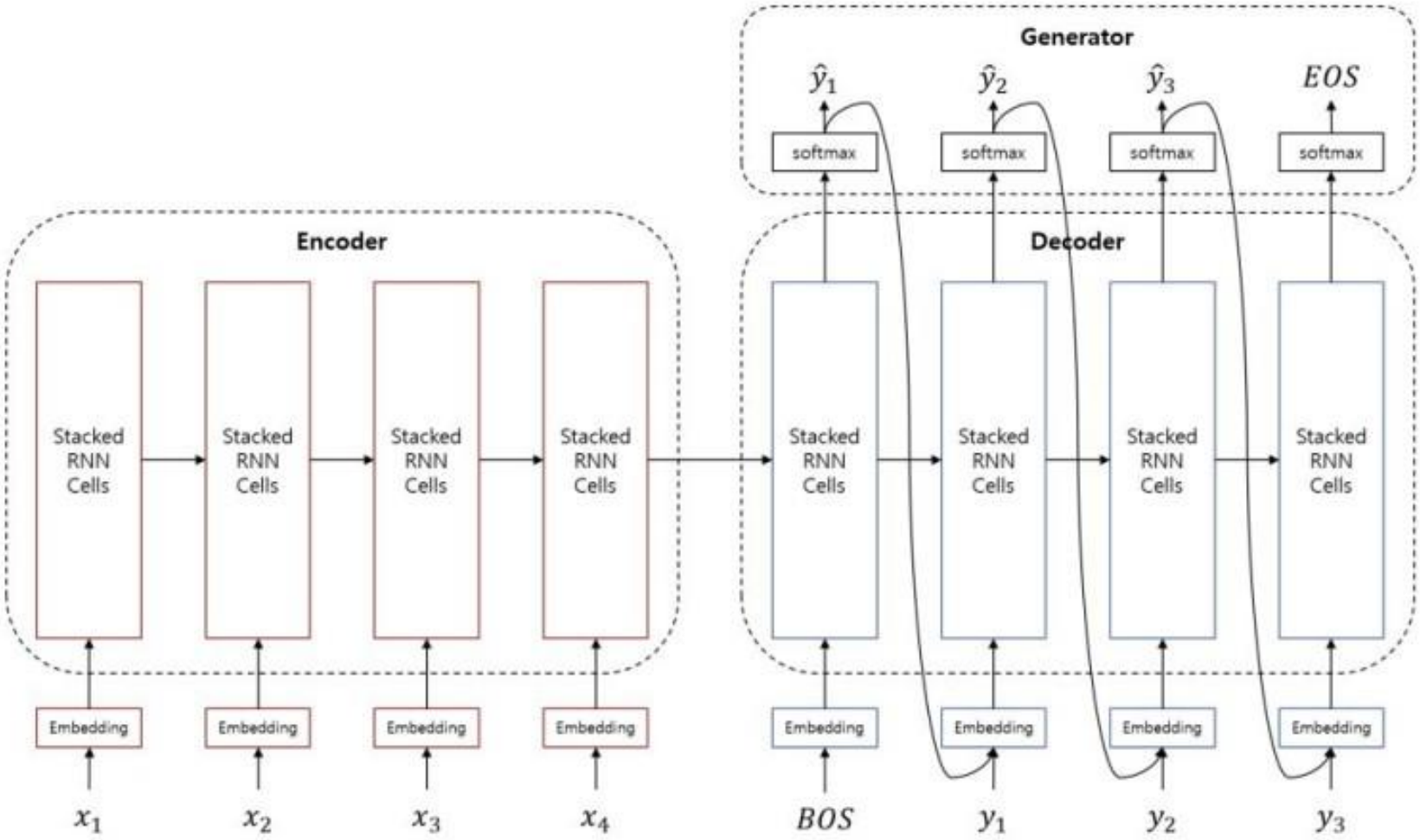
$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})\ \boldsymbol{where}\ \ \boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots \boldsymbol{x_n}\}, \boldsymbol{Y} = \{\boldsymbol{x_1}, \boldsymbol{y_2}, \cdots \boldsymbol{y_m}\}$$

  - $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$**를 최대로 하는 모델 파라미터를 찾는 작업**
  - **파라미터에 대한 학습이 완료되면 사후 확률을 최대로 하는 Y 도출**

$$\hat{\boldsymbol{Y}} = \underset{Y \in y}{\operatorname{argmax}}\, \boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta})$$

Seq2Seq 모델의 구성

- 이 계산을 위해 seq2seq는 크게 3개의 서브모듈(인코더, 디코더, 생성자)로 구성됨

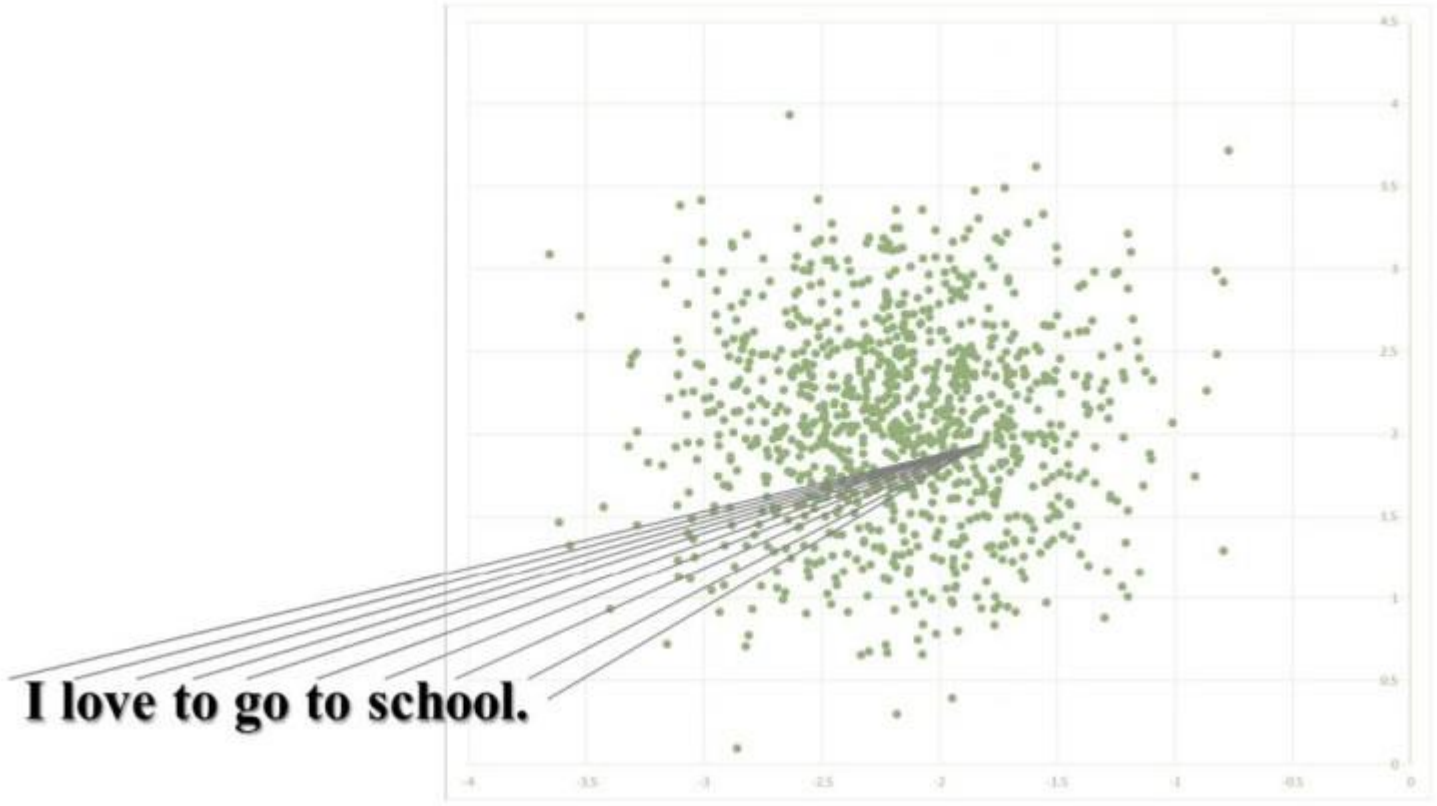


Seq2Seq 모델: 인코더

- 인코더
  - 주어진 소스 문장인 여러 개의 벡터를 입력으로 받아 문장을 함축하는 문장 임베딩 벡터 생성 ($P(z|X)$의 모델링)
  - RNN 분류 모델과 거의 같음
  - $P(z|X)$를 모델링하고, 주어진 문장을 매니폴드를 따라 차원 축소하여 해당 도메인의 잠재 공간(Latent Space)에 있는 어떤 하나의 점에 투영하는 작업

Seq2Seq 모델: 인코더

- 인코더는 문장을 하나의 벡터로 압축하여 표현함



Seq2Seq 모델: 인코더

- 기존의 텍스트 분류 문제에서는 모든 정보(특징, Feature)가 필요하지는 않음

  → 벡터 변환 시, 많은 정보를 간직할 필요는 없음

  - 예: 나는.. 과 같은 중립적인 단어는 감성 분류에 불필요
- 그러나, 기계 번역을 위한 문장 임베딩 벡터에서는 최대한 많은 정보를 요구
- 인코더를 수식으로 나타내면,

$$h_t^{src} = RNN_{enc}(emb_{src}(x_1), (h_{t-1}^{src})$$

$$H^{src} = [h_1^{src}; h_2^{src}; \cdots; h_n^{src}]$$

- emb_{src} : 인코더의 임베딩 레이어
- [;] : 이어 붙이는 작업
- 수식은 time-step별로 RNN 모델을 통과시켰음을 나타낸 것

- 인코더를 실제 구현할 때는 전체 time-step을 병렬로 한 번에 처리함

$$H^{src} = RNN_{enc}(emb_{src}(X), h_0^{src})$$

Seq2Seq 모델: 디코더

- 디코더
  - 인코더와 반대의 역할을 수행
  - 앞에서 살펴 본 seq2seq의 수식을 time-step에 관해 출어서 나타내면

$$P_\theta(Y|X) = \prod_{t=1}^{m} P_\theta(y_t|X, y_{<t})$$

$$\log P_\theta(Y|X) = \sum_{t=1}^{m} \log P_\theta(y_t|X, y_{<t})$$

  - 조건부 확률 변수에 X 가 추가됨

Seq2Seq 모델: 디코더

- seq2seq 식에서 조건부 확률에 X가 추가된 것은
- 인코더의 결과인 문장 임베딩 벡터와 이전 time-step까지 번역하여 생성한 단어들에 기반하여 현재 time-step의 단어를 생성함을 의미
- 수식으로 표현하면

$$h_t^{tgt} = RNN_{dec}(emb_{tgt}(y_{t-1}), (h_{t-1}^{tgt}))$$

$$where \quad h_0^{tgt} = h_n^{src} \quad and \quad y_0 = BOS$$

BOS : Beginning Of Sentence
EOS : End Of Sentence

디코더 자체만으로도 신경망 언어모델에 속함 → 디코더 입력의 초깃값으로 y_0 에 BOS 토큰을 입력으로 줌

Seq2Seq 모델: 생성자

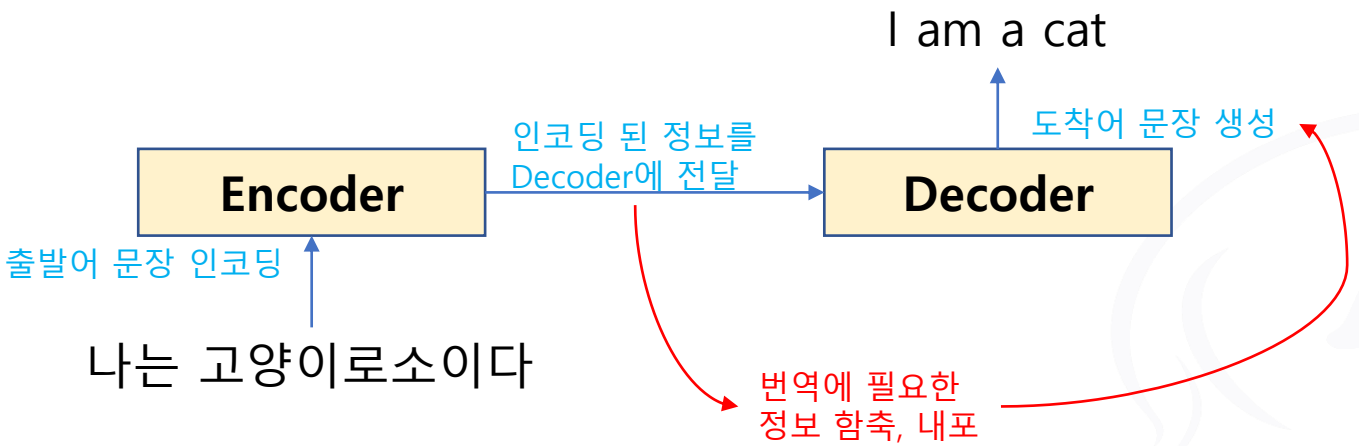
- 생성자(Generator)
  - 디코더에서 각 time-step 별로 결과 벡터 h_t^{tgt}를 받아 softmax를 계산하여 각 타겟 언어의 단어(어휘)별 확률 값을 반환하는 작업 수행
  - 생성자의 결과 값은 각 단어가 나타난 확률인 이산 확률 분포가 됨
  - 문장의 길이가 $|Y| = m$일때 맨 마지막에 반환되는 단어 y_m은 EOS 토큰이 됨
  - EOS 토큰: 디코더 계산의 종료를 나타냄

$$\widehat{y_t} = softmax\left(linear_{hs \to |V_{tgt}|}\left(h_t^{tgt}\right)\right) \quad and\ \widehat{y_m} = EOS$$

where hs is hidden size of RNN, and $|V_{tgt}|$ is size of output vocabulary

Seq2Seq 모델: 번역 수행 예시

- **Encoder와 Decoder가 번역을 수행하는 예**

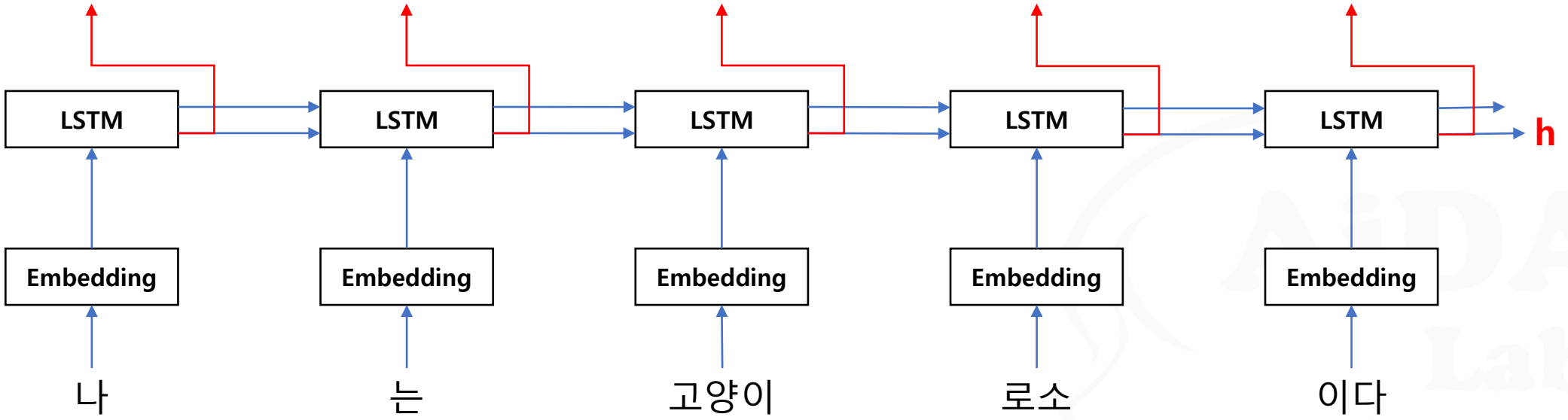


인코딩(부호화) : 정보를 어떤 규칙에 따라 변환하는 것

디코딩(복호화) : 인코딩된 정보를 원래의 정보로 되돌리는 것

Seq2Seq 모델: Encoder를 구성하는 계층

- **Encoder를 구성하는 계층**

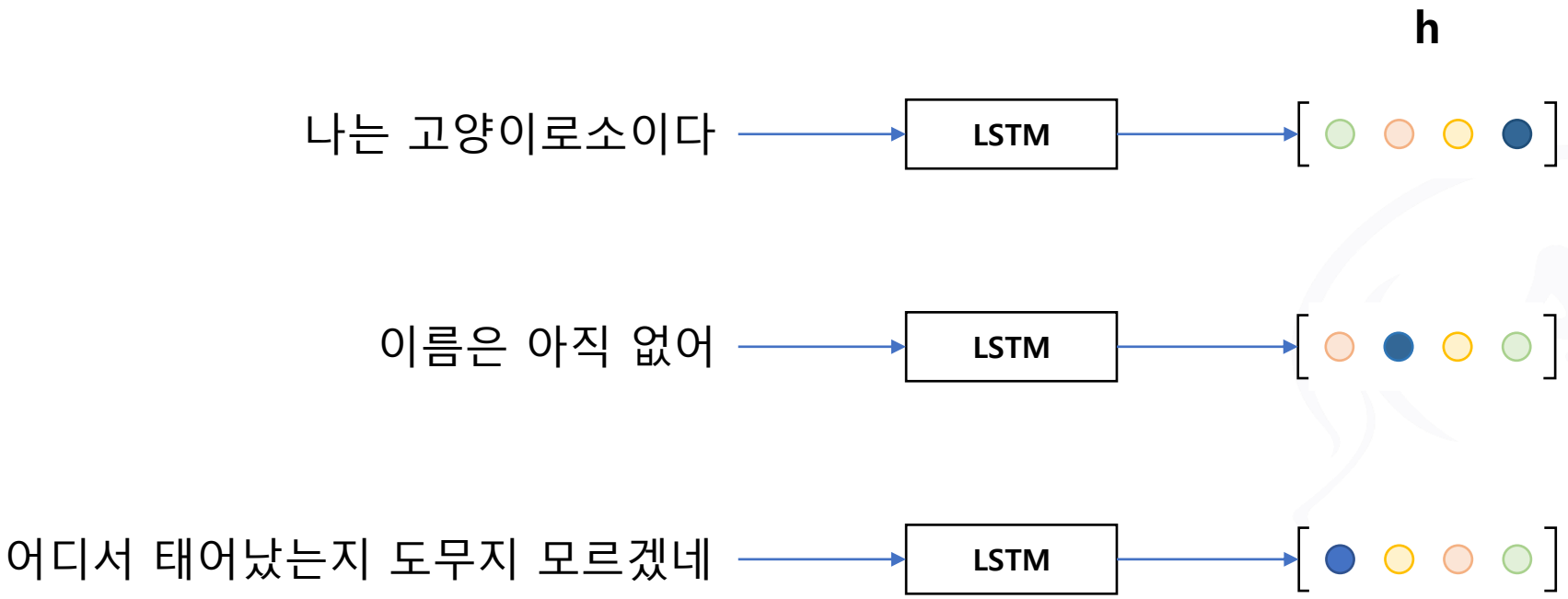


RNN(LSTM)을 이용해서
시계열 데이터를 h라는 은닉상태 벡터로 변환

- 출력벡터 **h**는 LSTM 계층의 마지막 은닉상태
→ 입력 문장을 번역하는데 필요한 정보가 인코딩 됨
- h는 고정길이 벡터
→ 인코딩 작업 = 임의 길이의 문장을 고정길이벡터로 변환하는 작업

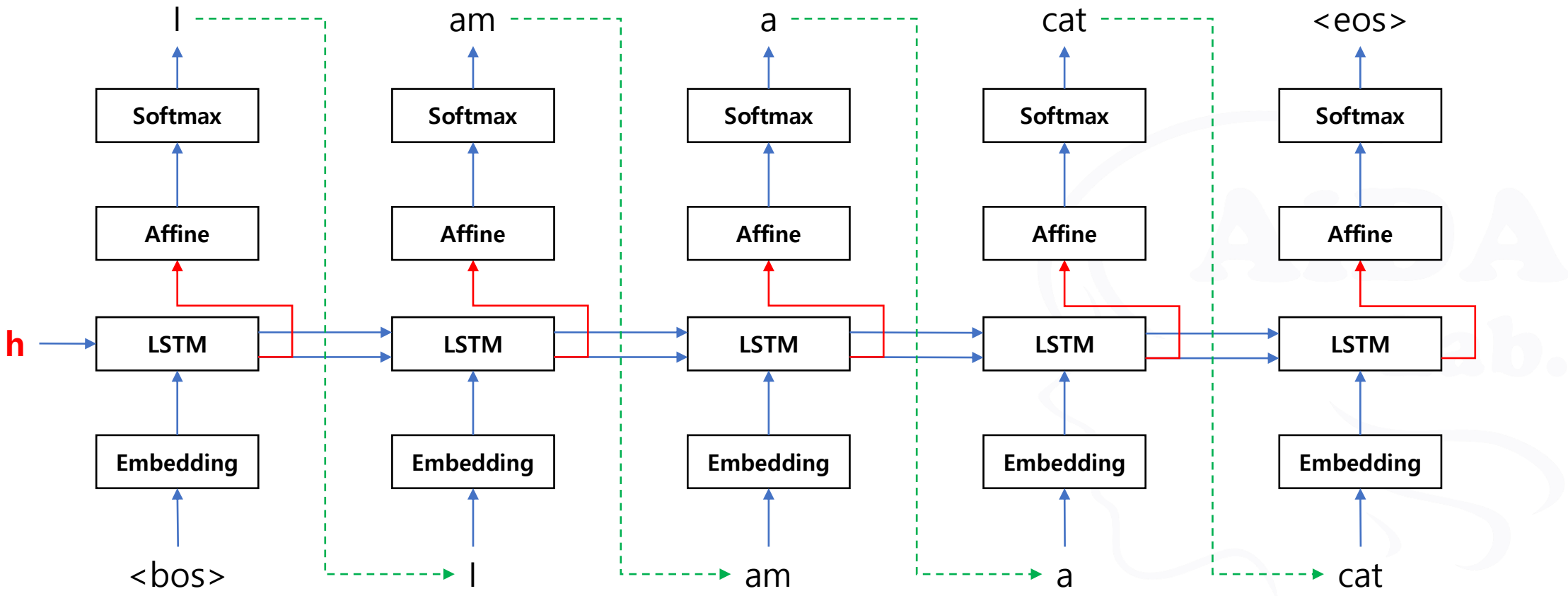
Seq2Seq 모델: Encoder를 구성하는 계층

- **Encoder는 문장을 고정 길이 벡터로 인코딩한다.**



Seq2Seq 모델: Decoder를 구성하는 계층

- **Decoder를 구성하는 계층**



Seq2Seq 모델 의 활용 분야

- **seq2seq의 활용 분야**
  - **특정 도메인의 시계열 데이터 또는 시퀀스 데이터 입력을**
  - **다른 도메인의 시계열 또는 시퀀스 데이터로 출력하는데 탁월한 능력을 보임**

활용분야	입력과 출력
기계 번역	특정 언어 문장을 입력으로 받아 다른 언어의 문장으로 출력
챗봇	사용자의 문장 입력을 받아 대답을 출력
문서 요약	긴 문장을 입력으로 받아 같은 언어의 요약된 문장으로 출력
기타 자연어 처리	사용자의 문장 입력을 받아 프로그래밍 코드로 출력 등
음성 인식	사용자의 음성을 입력으로 받아 해당 언어의 문자열(문장)로 출력
독순술	입술 움직임의 동영상을 입력으로 받아 해당 언어의 문장으로 출력
이미지 캡셔닝	변형된 seq2seq를 사용하여 이미지를 입력으로 받아 그림을 설명하는 문장을 출력

Seq2Seq 모델: 한계점

- 한계점
  - seq2seq는 오토인코더(AutoEncoder)의 일종으로 특히 시계열 또는 시퀀스 데이터에 강점이 있는 모델이라고 볼 수 있음
  - 장기 기억력 문제
    - 신경망 모델은 차원축소를 통한 데이터 압축에 탁월한 성능을 보이지만
    - 정보를 무한하게 압축할 수는 없음 → 압축 가능한 정보량의 한계
      → 문장(또는 time-step)이 길어질수록 압축 성능 하락
    - LSTM, GRU 등을 사용하여 RNN보다는 높은 성능을 낼 수 있지만 역시 한계가 존재함

Seq2Seq 모델: 한계점

- **구조 정보의 부재**
  - 현재의 딥러닝 자연어 처리 추세에서는
  - 문장을 이해할 때 구조 정보를 사용하기보다는 단순히 시퀀스 데이터로 다루는 경향
  - 아직까지는 성공적으로 사용되지만 다음 단계로 나아가려면 구조 정보가 필요할 것으로 추정되고 있음
- **챗봇 또는 QA봇**
  - 시퀀스 데이터 입력을 다른 도메인의 시계열/시퀀스 데이터로 출력하는데 뛰어남
  - 많은 사람이 seq2seq를 학습 시키면 더욱 뛰어난 성능을 보일 것으로 기대
  - 그러나 지속적인 데이터, 정보 추가가 필요하지 않은 번역, 요약과 달리 대화의 경우 지속적인 관리가 요구되지만 해당하는 구조적 기능이 부족함 → 발전된 구조 요구