302: Language Models

언어모델의 기능 방식과 흐름

1. Embedding Vector로 단어의 의미를 표현하는 방식
2. 언어모델의 종류들과 장, 단점. 그리고 그것을 보완한 다른 모델들
    1. RNN
    2. LSTM
    3. Seq2Seq
    4. Attention
    5. Transformer

Word Embedding

구분	One-Hot Vector	임베딩 벡터
차원	고차원(단어 집합의 크기)	저차원
다른 표현	희소 벡터의 일종	밀집 벡터의 일종
표현 방법	수동(index)	훈련 데이터로부터 학습함
값의 타입	1과 0	실수

단어간 유사성을 표현할 수 없는 One-Hot Vector의 한계 극복

단어 사이의 “유사도”를 벡터화한다.

Word Embedding

비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다

Embedding

구축하다, 기반을 두다.

Embedding Vector

단어의 의미 = 단어를 표현한 벡터

단어 사이의 유사도를 다차원 공간에 벡터화 하겠다.

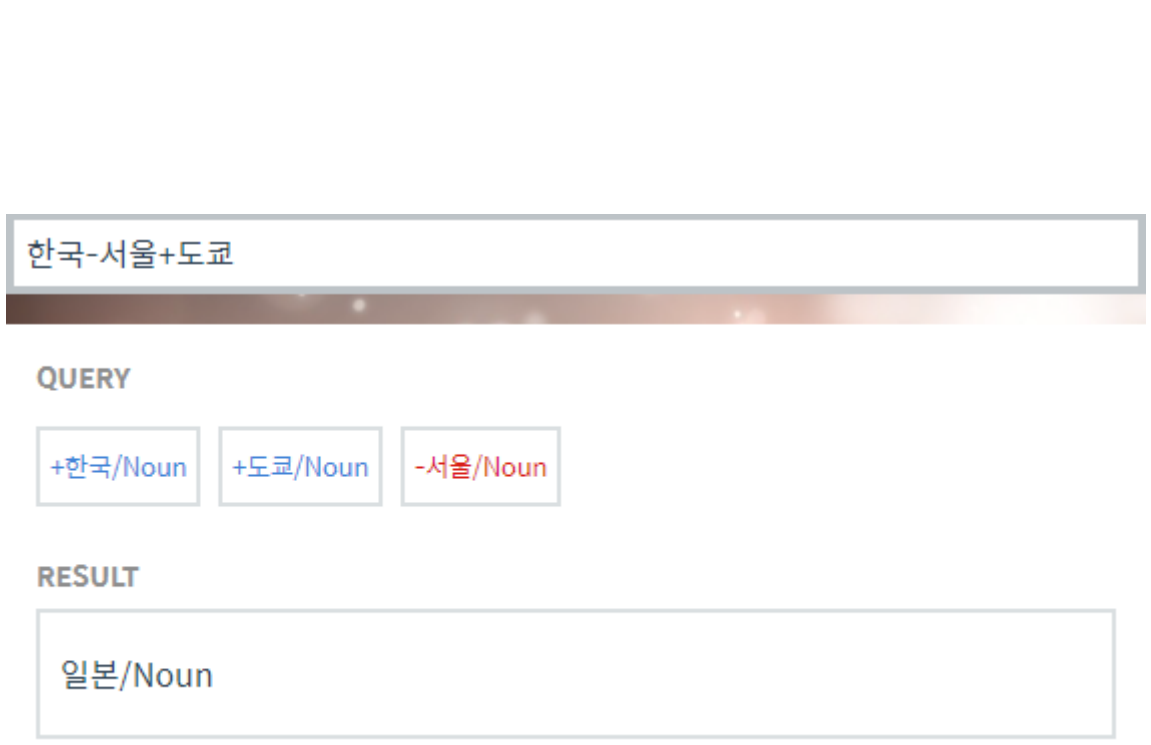
(밀집벡터 Dense Vector 생성)

Word Embedding

유사도와 의미가 무슨 관계인가?

나는	나의	깜찍한	강아지를	사랑한다.
		사랑스러운		
		귀여운		

Word Embedding: Word2Vec



http://w.elnn.kr/search/

ABOUT

이곳은 단어의 효율적인 의미 추정 기법(Word2Vec 알고리즘)을 우리말에 적용해 본 실험 공간입니다.

Word2Vec 알고리즘은 인공 신경망을 생성해 각각의 한국어 형태소를 1,000차원의 벡터 스페이스 상에 하나씩 매핑시킵니다. 그러면 비슷한 맥락을 갖는 단어들은 가까운 벡터를 지니게 되며, 벡터끼리 시맨틱 연산도 수행할 수 있습니다. 이는 분산 시맨틱스 가정에 기초하고 있습니다.

CORPUS

실험을 위해 한국어 위키백과와 나무위키에서 제공하는 자료를 사용했습니다. 주어진 자료를 특수문자 제거, 띄어쓰기 정정, 형태소 분석 등의 방법으로 처리한 결과, 약 45만 종류, 4.2억 개의 단어로 구성된 말뭉치를 생성할 수 있었습니다.

EXAMPLES

- 한국 - 서울 + 파리 = ?
- 컴퓨터공학 - 자연과학 + 인문학 = ?
- 사랑 + 이별 = ?

CONTACT

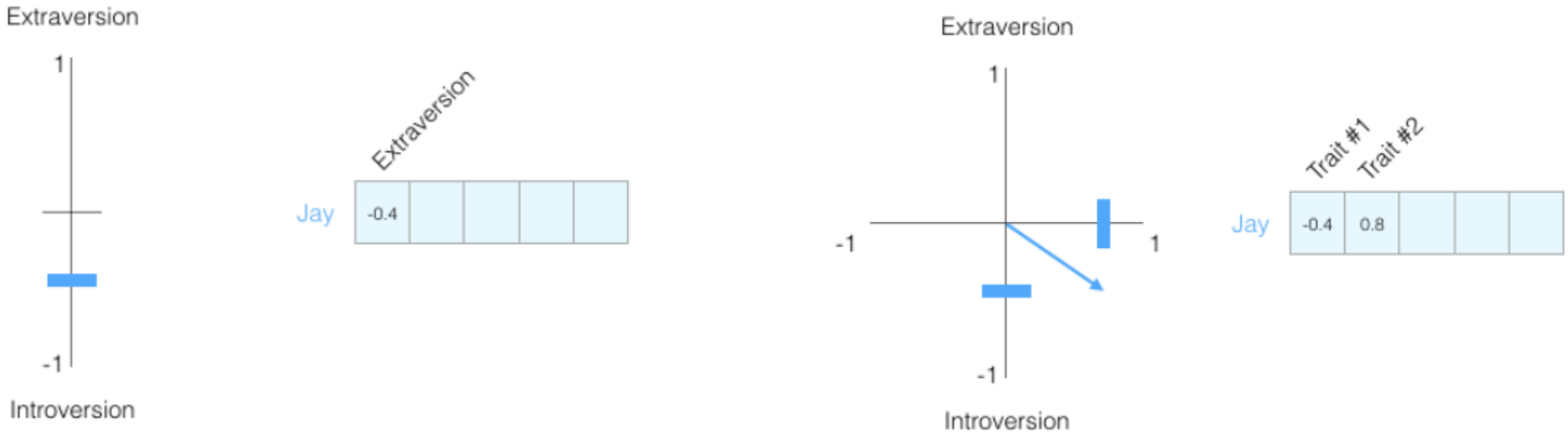
문의사항이 있으면 이메일 elnn@elnn.kr 로 연락주세요!

HISTORY

- 2014-02: 웹 서비스 시작
- 2015-04: 디자인 개선 및 DB 업데이트

Word Embedding: Word2Vec

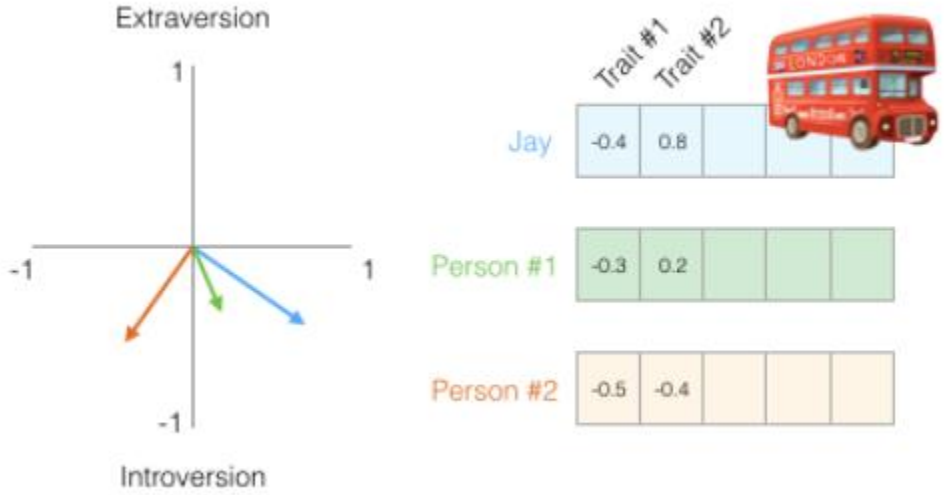
사람의 성격을 벡터로 표현한다고 하자.



https://databreak.netlify.app/2019-04-25-Illustrated_word2vec/

Word Embedding: Word2Vec

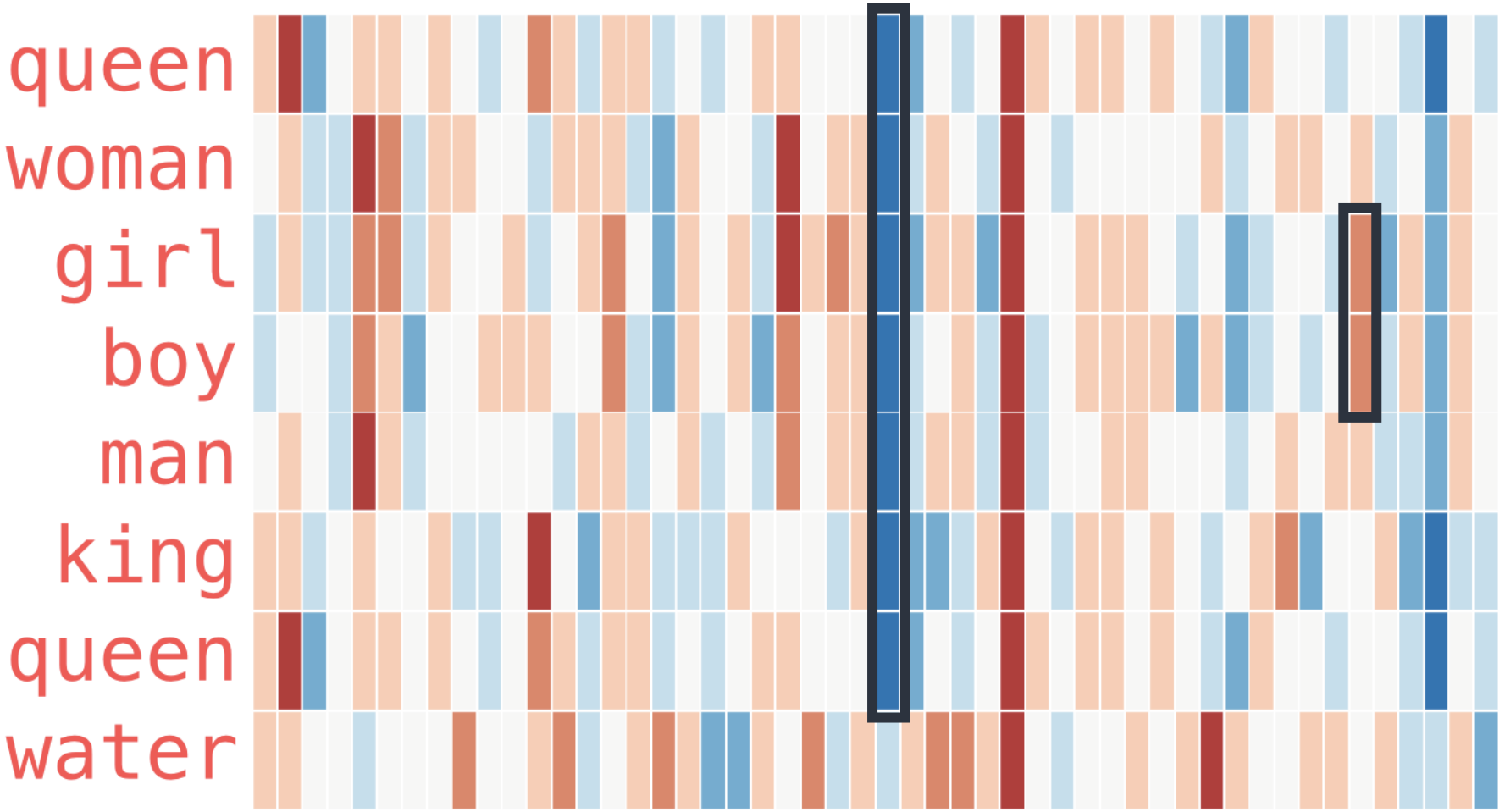
사람의 성격을 벡터로 표현한다고 하자.




```
cosine_similarity( Jay [-0.4, 0.8] , Person #1 [-0.3, 0.2] ) = 0.87
cosine_similarity( Jay [-0.4, 0.8] , Person #2 [-0.5, -0.4] ) = -0.20
```

https://databreak.netlify.app/2019-04-25-Illustrated_word2vec/

Word Embedding: Word2Vec



https://databreak.netlify.app/2019-04-25-Illustrated_word2vec/

Word Embedding: Word2Vec

CBOW (Continuous Bag of Words)
주변 단어 (맥락) 으로 부터 중간 단어를 예측

아버지가 _ 에 들어가신다.

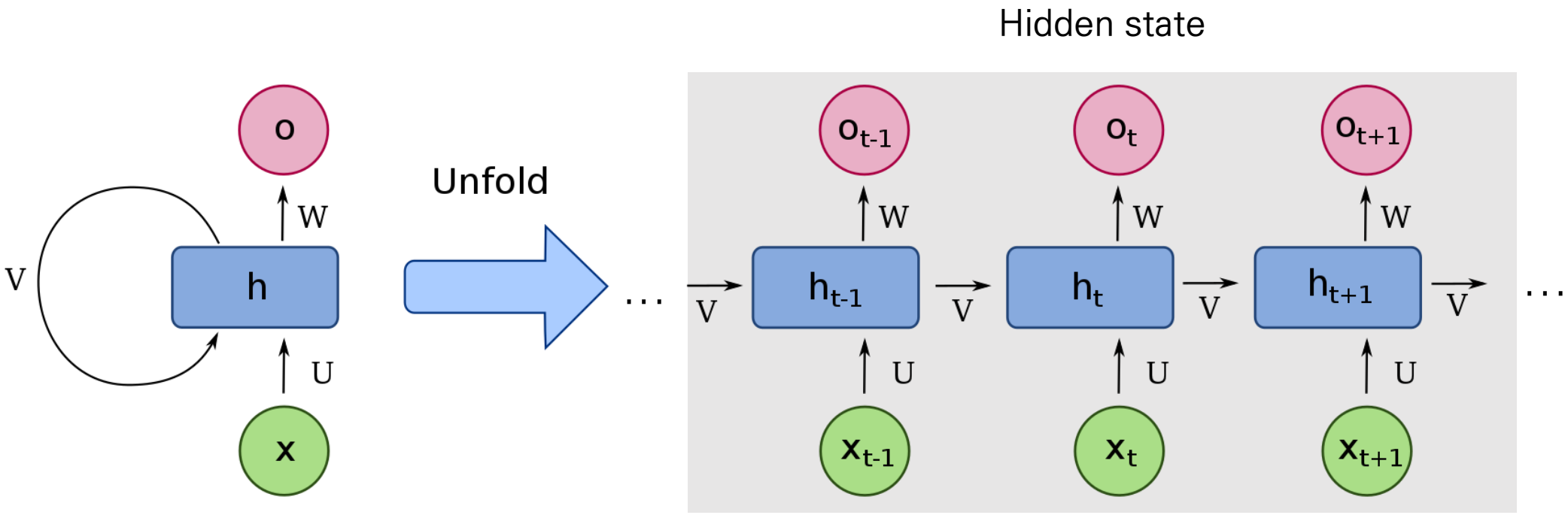
Skip-gram
중간 단어로부터 주변 단어 (맥락) 를 예측

___ 가 방에 ____

Word Embedding: Word2Vec

언어모델: RNN

Recurrent Neural Network, 순환 신경망



언어모델: RNN

We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

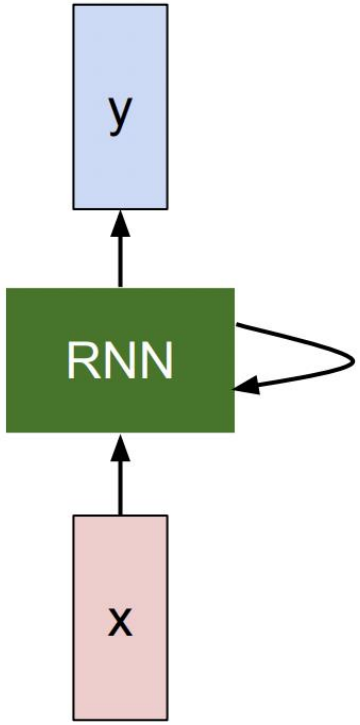
$$h_t = f_W(h_{t-1}, x_t)$$

new state

some function with parameters W

old state

input vector at some time step



언어모델: RNN

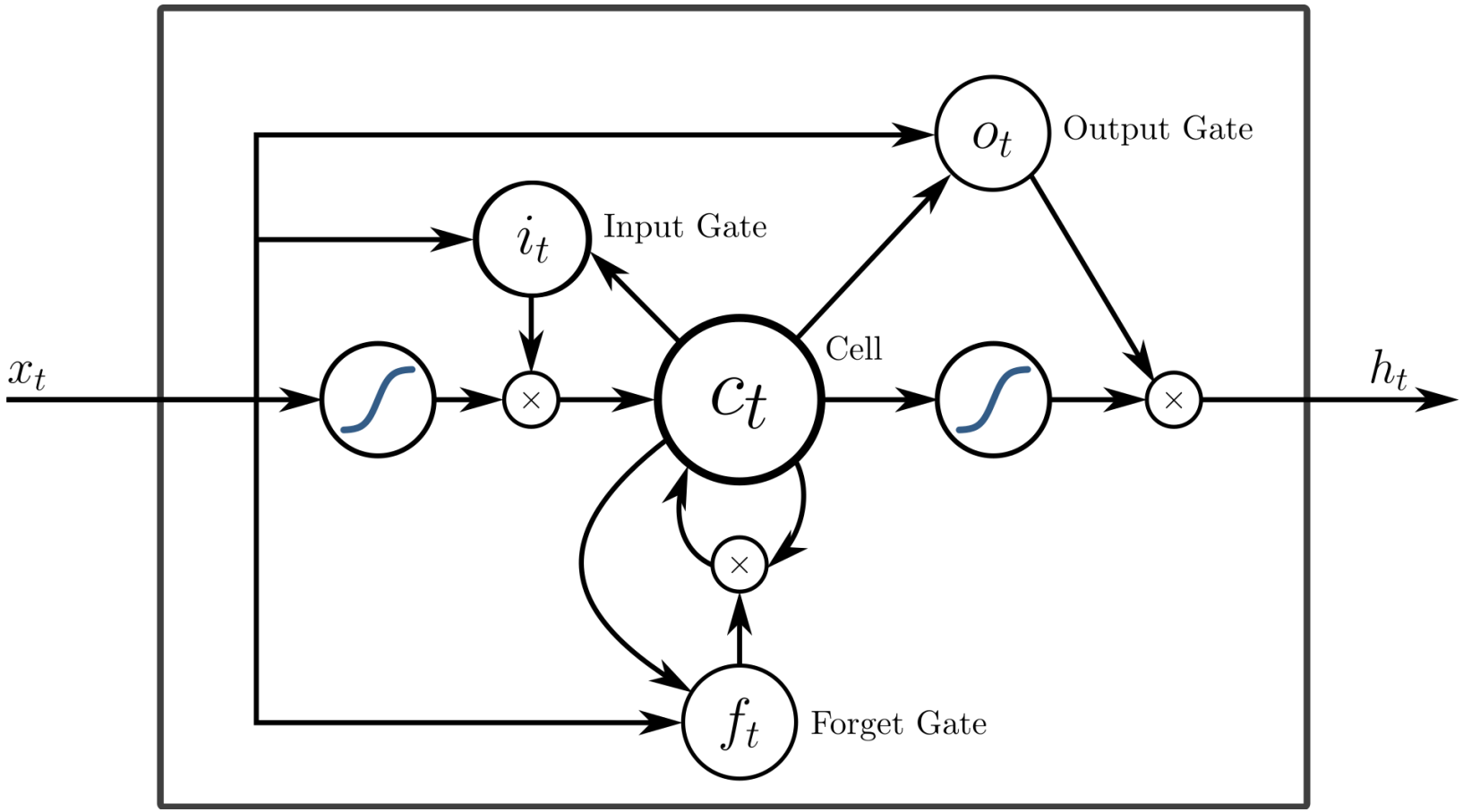
언어모델링에서

1. 모델에 기억이라는 개념이 생겨, 순서와 시간을 학습할 수 있다.
2. 입력과 출력값의 길이가 자유롭다.

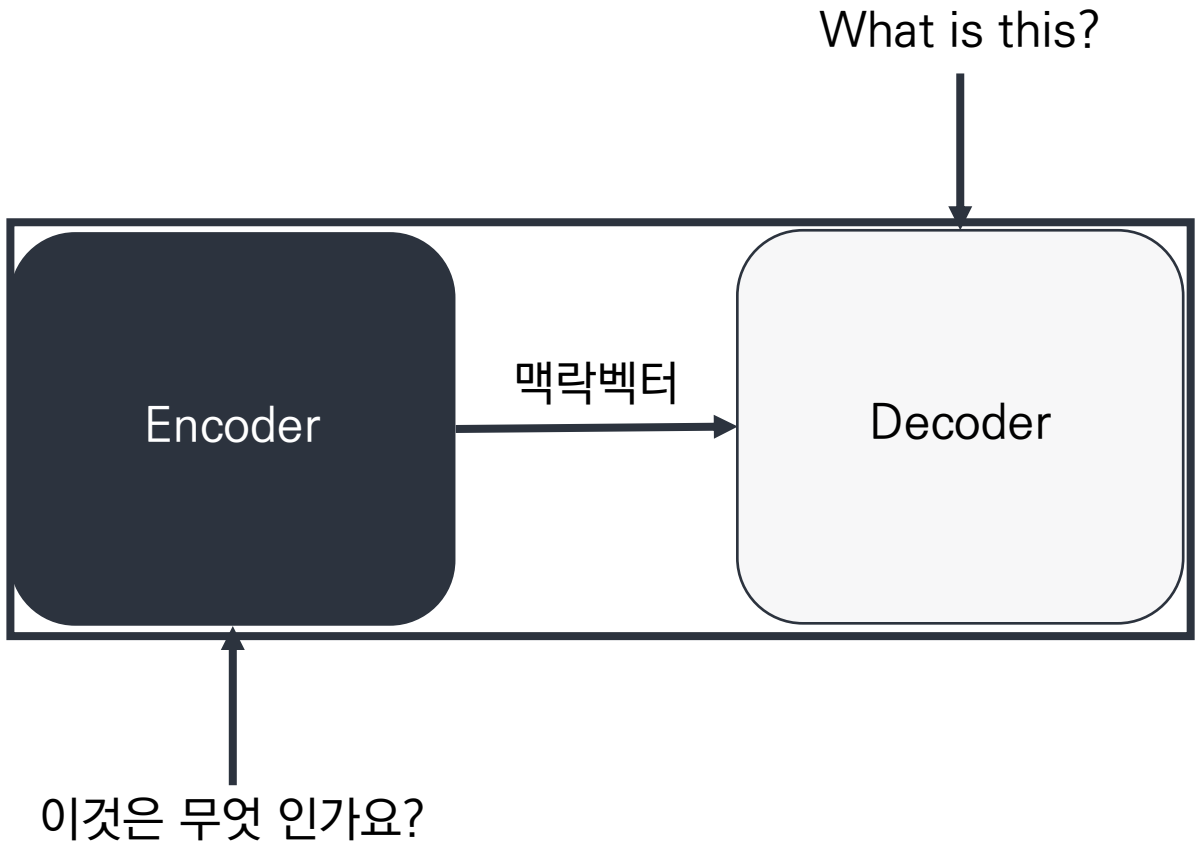
한계

1. 멀리 떨어진 데이터는 잘 학습하지 못한다. : 단기기억 상실증, LSTM으로 해결
2. 병렬 연산이 불가능하므로 느리다. : t-1 값이 필요하기 때문
3. 기울기 소실과 폭주 : RNN의 고질적 문제

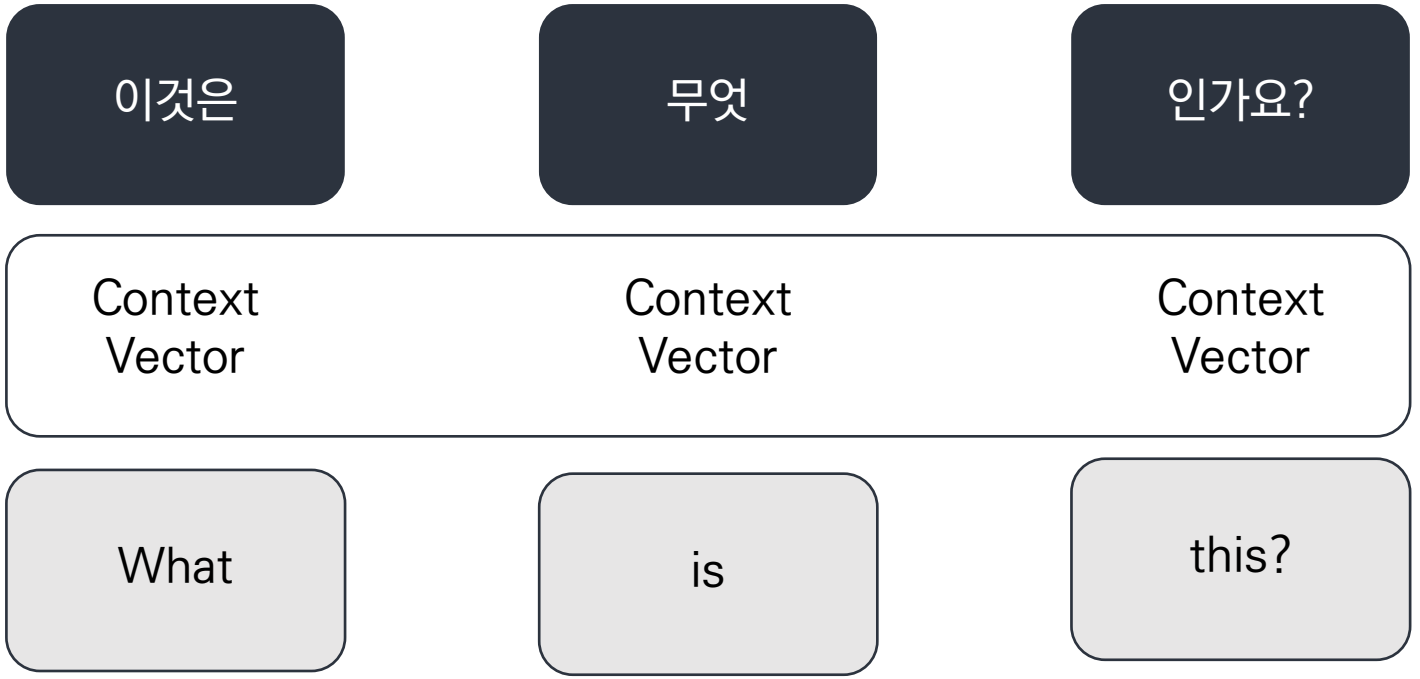
언어모델: LSTM



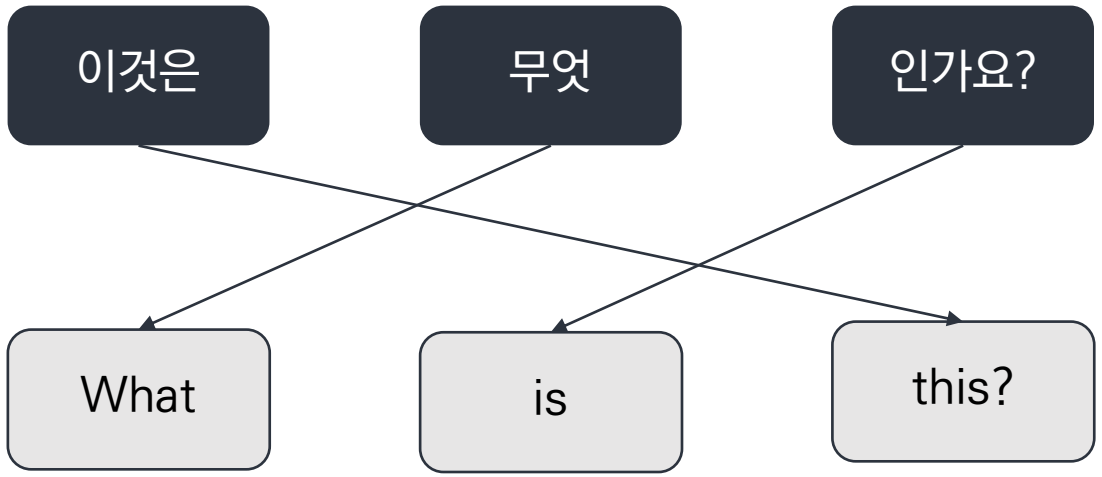
언어모델: Seq2Seq

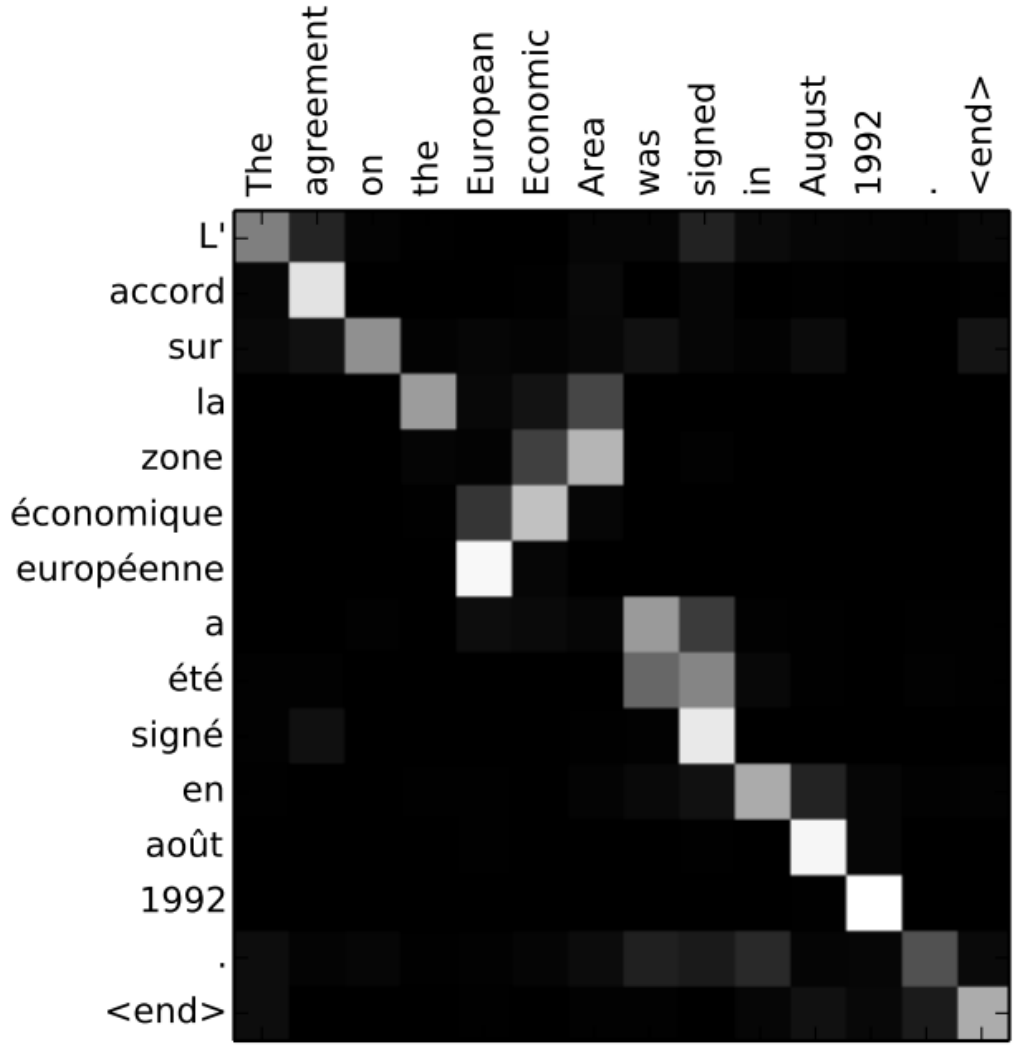


언어모델: Seq2Seq



언어모델: Attention





Bahdanau et al., 2015