

KOREAN EMBEDDING STUDY

- Chapter 2 -

20200218

Kwang-June Choi

자연어를 컴퓨터가 처리할 수 있는 숫자들의 나열, **벡터로 변환한 결과**

컴퓨터는 임베딩 값을 계산 및 처리함으로써 사람이 알아들을 수 있는 형태의 자연어로 출력

자연어의 통계적 패턴 정보를 임베딩에 넣어 의미를 함축함

임베딩 생성시 사용되는 통계 정보

1. 문장에 어떤 단어가 (많이) 쓰여 있는가?

-> *Bag of words*

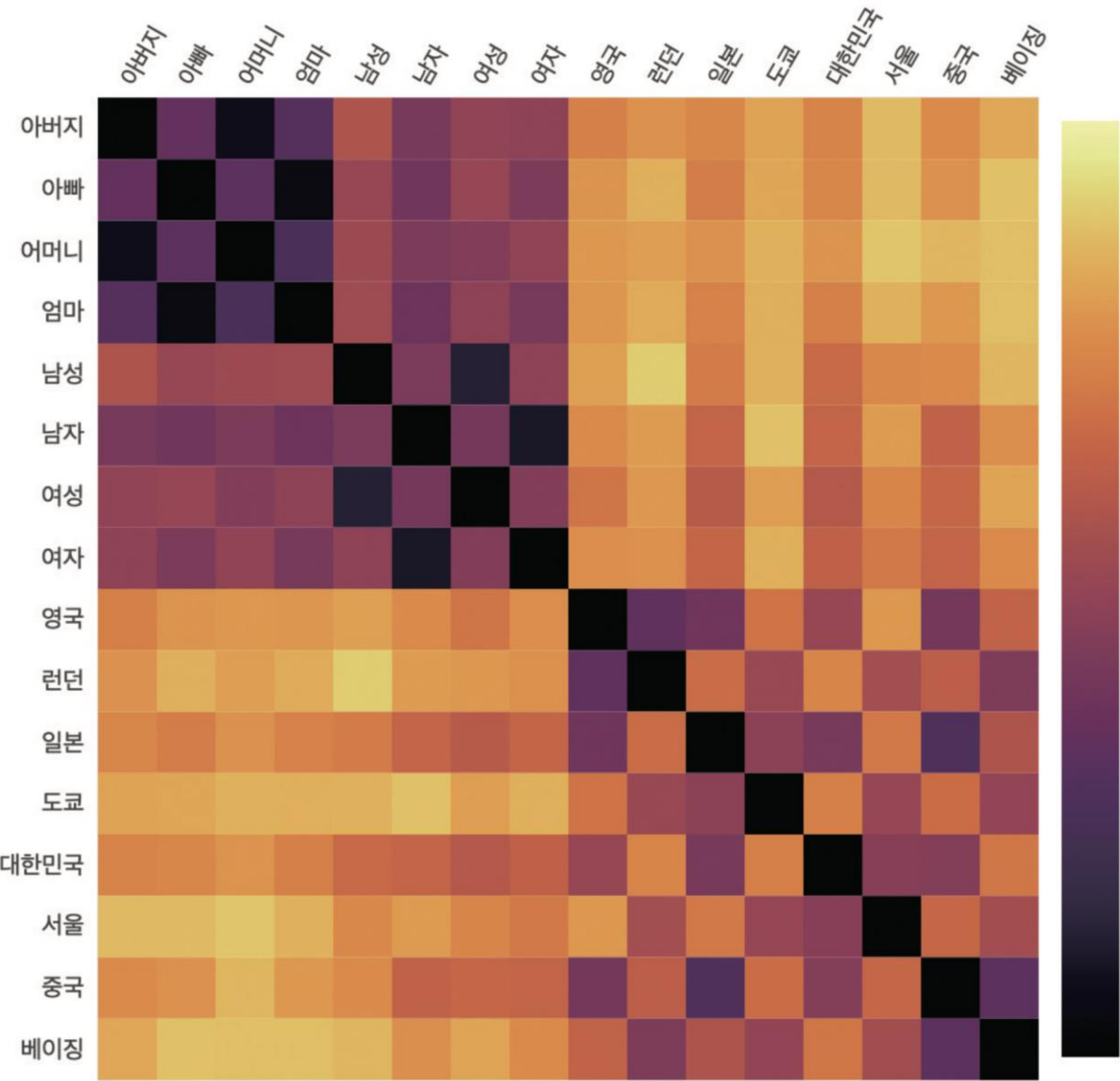
2. 문장에 단어가 어떤 순서로 등장하는가?

-> *Language Model*

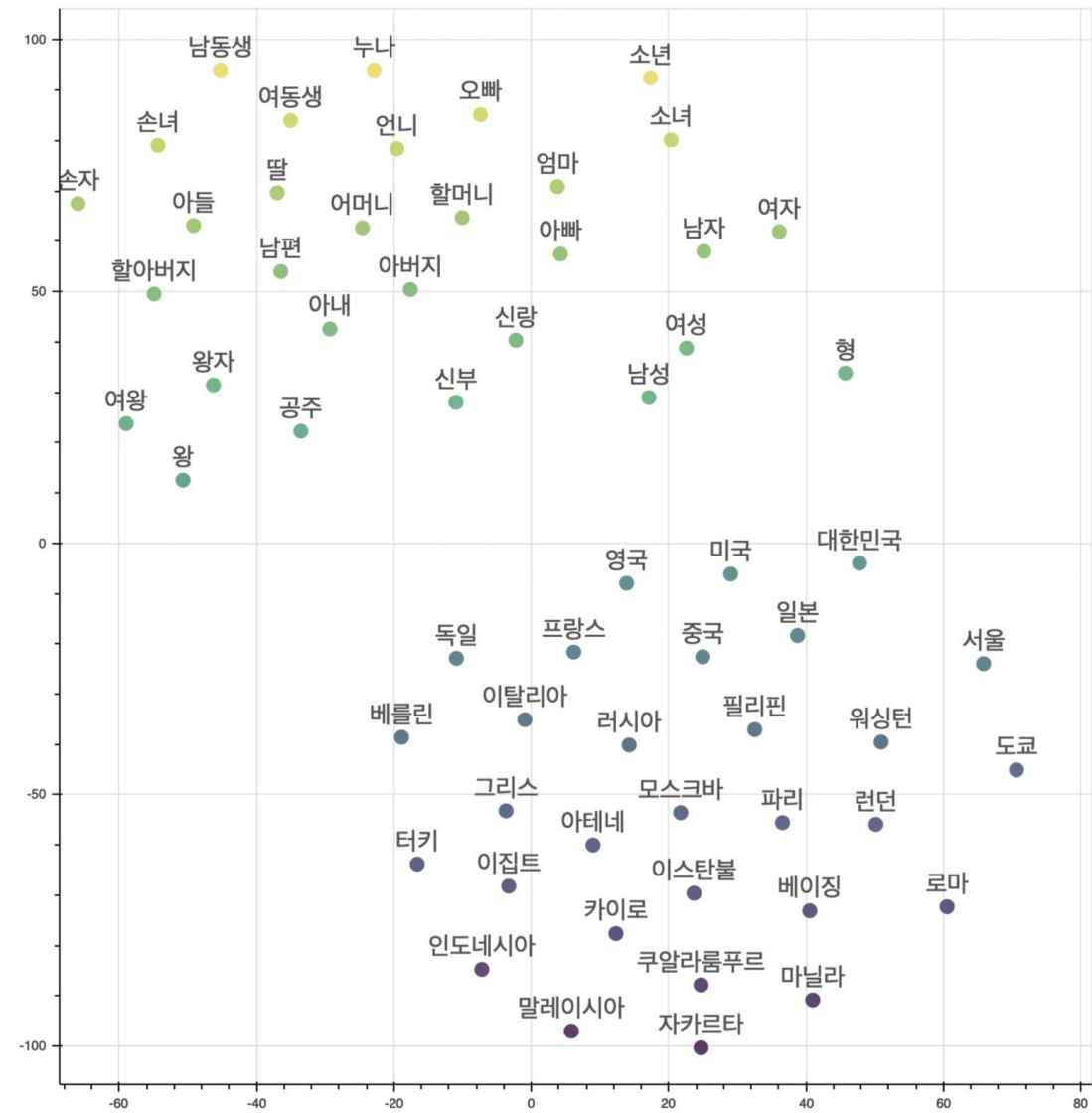
3. 문장에 어떤 단어가 함께 나타났는가?

-> *Distributional Hypothesis*

NLP Embedding

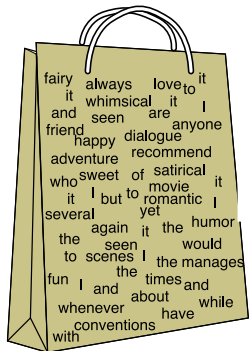


관련도 / 유사도 계산



시각화

Bag of Words



별 하나 에 추억 과
별 하나 에 사랑 과
별 하나 에 쓸쓸함 과
별 하나 에 동경 과
별 하나 에 시 와
별 하나 에 어머니 , 어머니

어머니 사랑
과 하나 별
시 하나 과 함 쓸쓸
하나 동경 추억 하나
과 별 하나 별
와 어머니

별	하나	에	추억	과	사랑	쓸쓸	함	동경	시	와	어머니	,
6	6	6	1	4	1	1	1	1	1	1	2	1

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

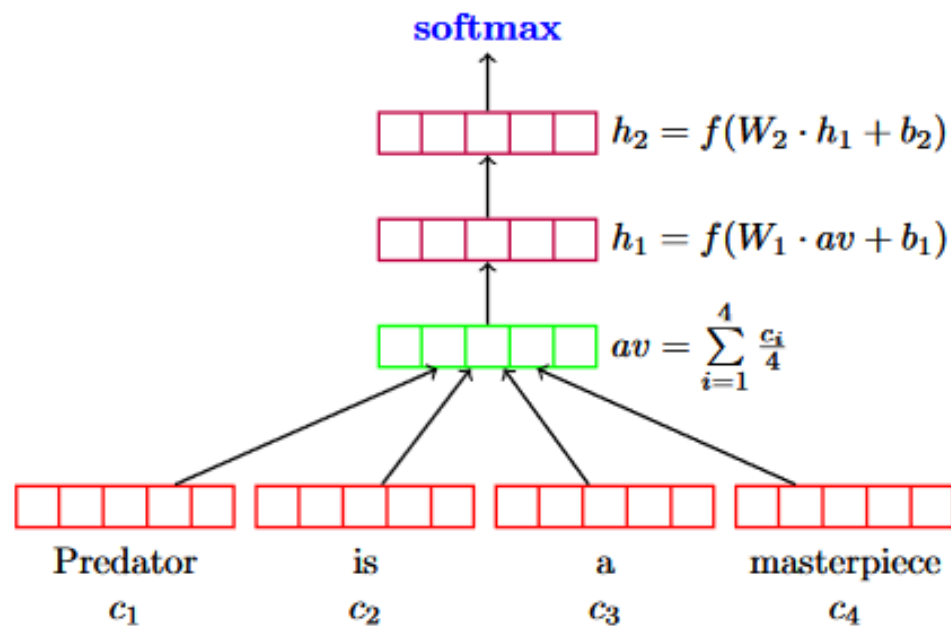
N = total number of documents

- 중복 원소를 허용한 집합
- 단어의 등장 순서와 상관없이 문서 내 등장 빈도를 임베딩으로 쓰는 기법

-> 저자가 생각한 주제가 문서 내 단어 사용에 녹아 있다는 가정

대표적 BoW기법, TF-IDF

- 조사와 같이 수없이 반복되는 단어 배제를 위한 기법
- 문서 단어 출현 빈도 x log(전체 문서의 수 / 단어가 나타난 문서의 수)
- 정보성이 없는 단어들의 가중치가 0으로 수렴하며 불필요한 정보 제거



Deep Averaging Network

- BoW의 Neural Network 버전
- 중복 집합에 속한 단어의 임베딩을 평균을 취해 생성

LM은 단어 시퀀스 또는 문장에 확률을 할당하는 모델

-> 가장 자연스러운 단어 시퀀스를 찾아내는 모델

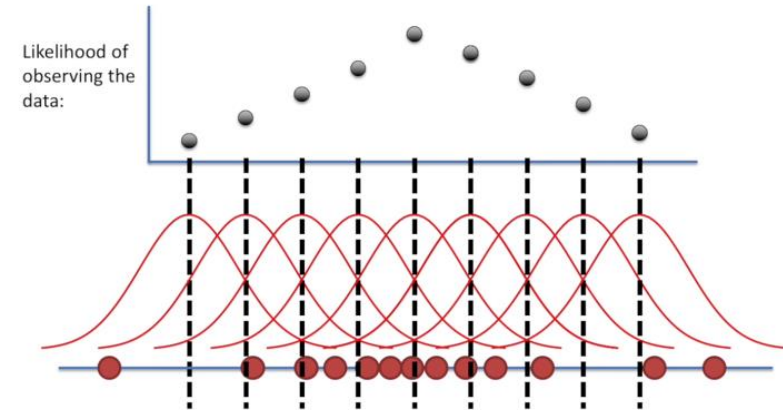
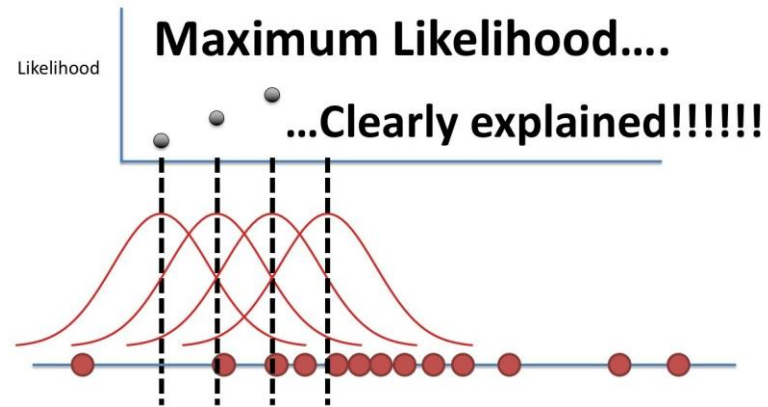
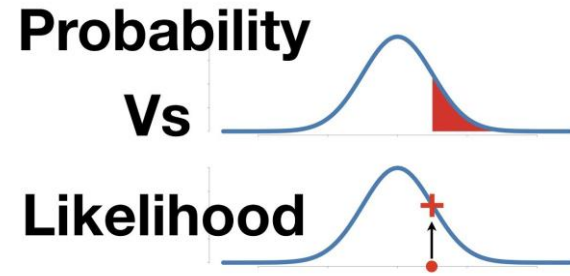
ex) $P(\text{나는 오늘 학교에 갔다}) > P(\text{나는 오늘 학교에 탔다})$

LM을 만드는 방법은 크게 두 가지

1. Statistical method (n-gram)
2. Neural Network method



Language Model



$$P(\text{명작이다}|\text{내, 마음, 속에, 영원히, 기억될, 최고의}) = \frac{\text{Freq}(\text{내, 마음, 속에, 영원히, 기억될, 최고의, 명작이다})}{\text{Freq}(\text{내, 마음, 속에, 영원히, 기억될, 최고의})}$$

ex) 2-gram

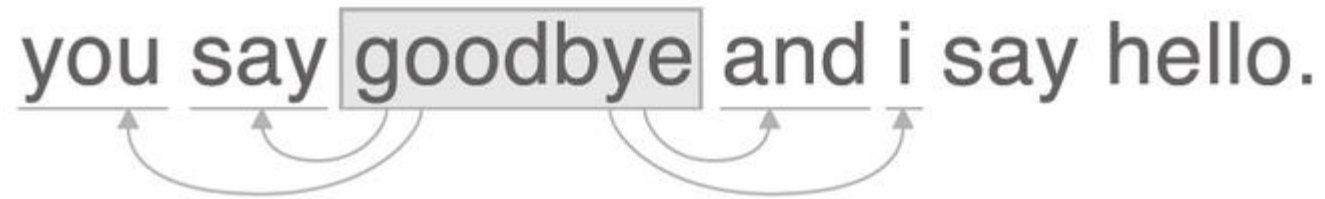
$$P(w_n|w_{n-1}) = \frac{\text{Freq}(w_{n-1}, w_n)}{\text{Freq}(w_{n-1})}$$

$$P(w_1^n) = P(w_1, w_2, \dots, w_n) = \prod_{k=1}^n P(w_k|w_{k-1})$$

분포 가정

- 자연어의 의미는 그 주변 문맥을 통해 유추 할 수 있음
- 어떤 단어의 pair는 비슷한 문맥 환경에서 자주 등장한다면 그 의미 또한 유사할 것

you say goodbye and i say hello.



(window_size = 2)

점별 상호 정보량(PMI, Pointwise Mutual Information)

- 두 확률변수 사이의 상관관계성을 계량화 하는 단위
- 말뭉치가 적더라도 제법 제대로 된 유사도 추출 가능

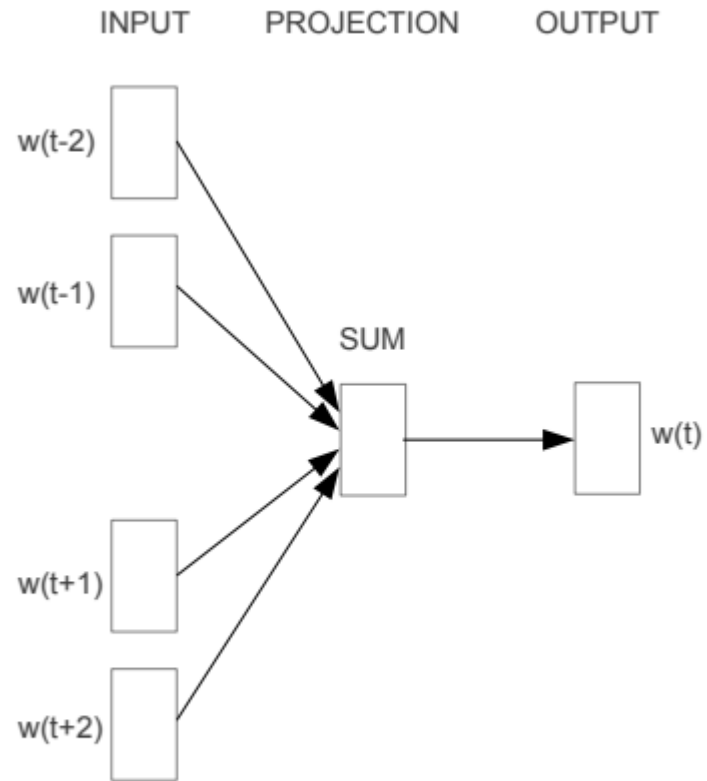
$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$



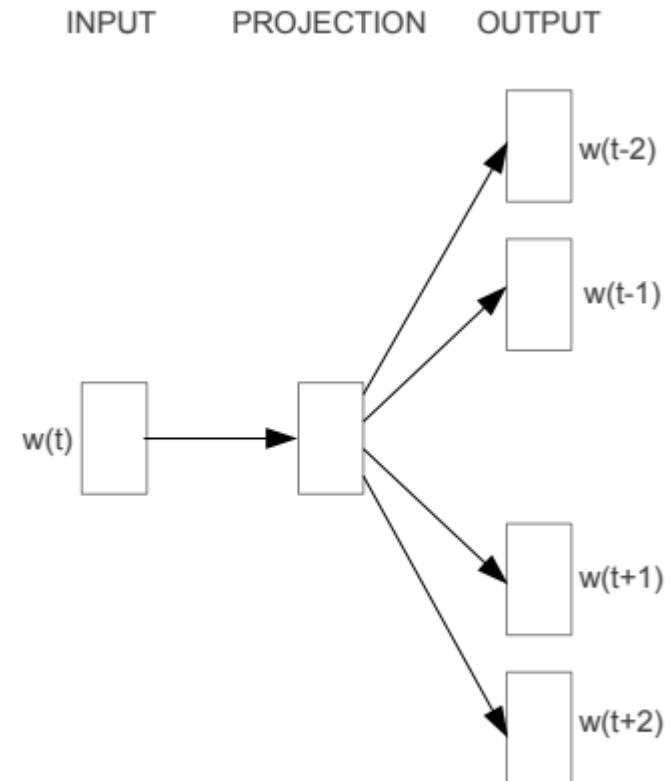
$$\text{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} = \log_2 \frac{\frac{C(x, y)}{N}}{\frac{C(x)}{N} \frac{C(y)}{N}} = \log_2 \frac{C(x, y) \cdot N}{C(x)C(y)}$$

Distributional Hypothesis

Word2Vec

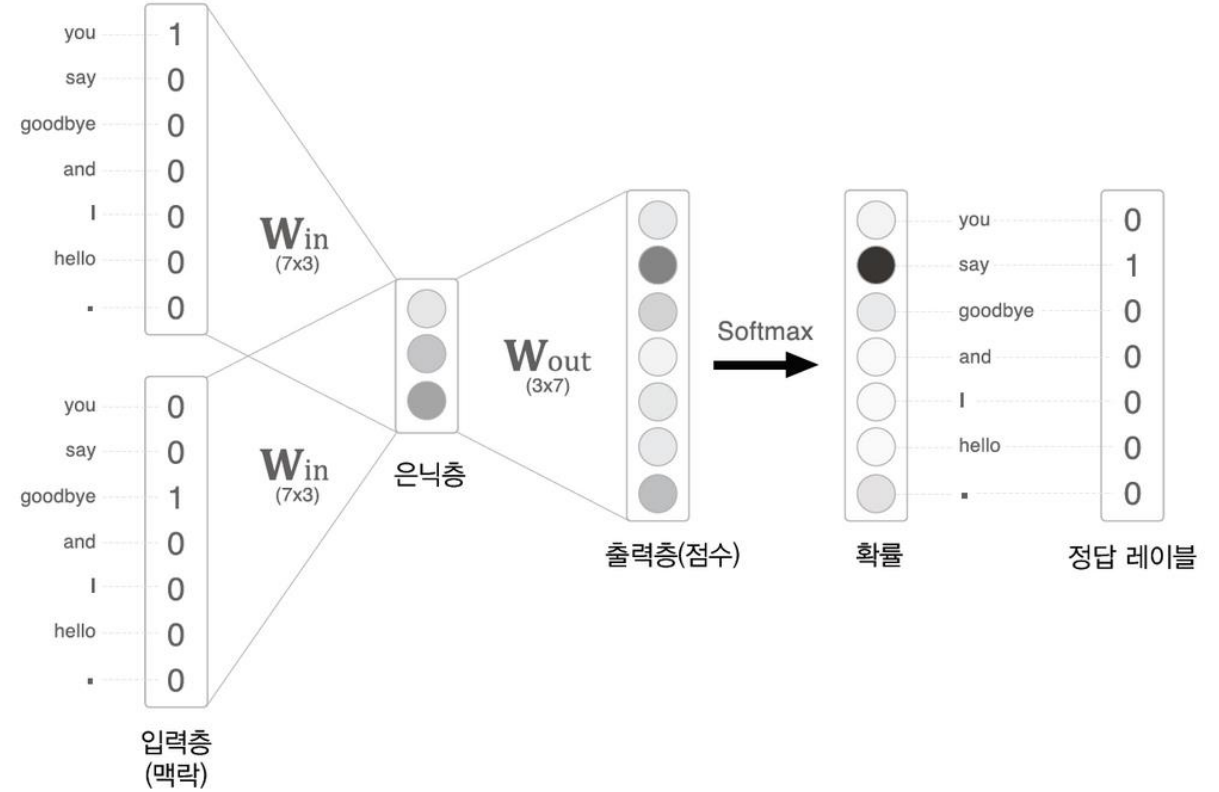
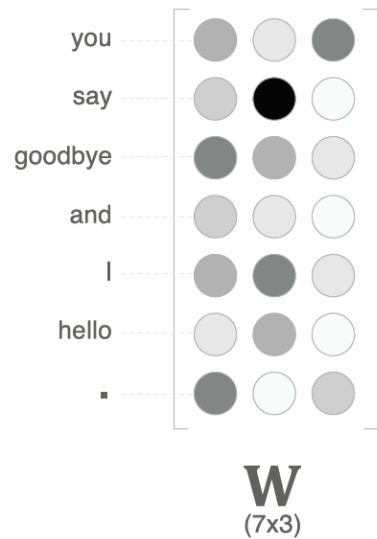
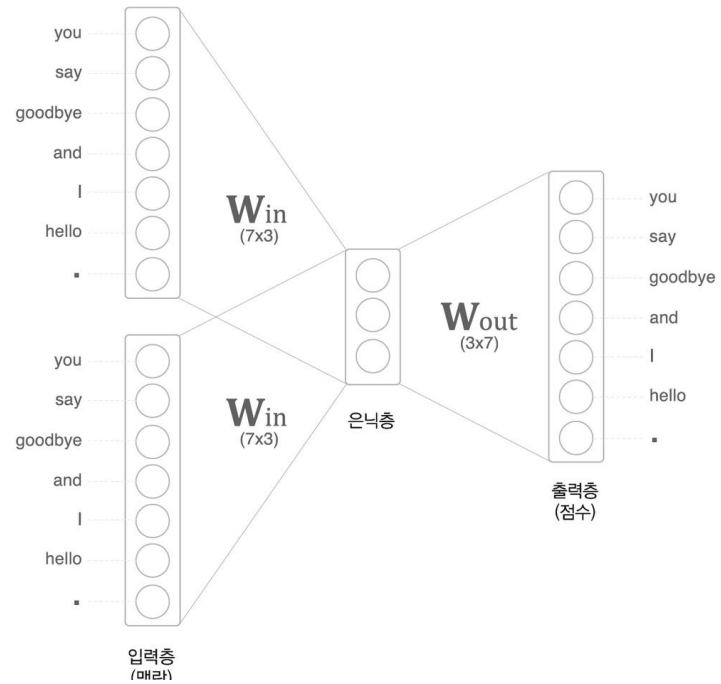


CBOW



Skip-gram

Distributional Hypothesis



벡터가 어떻게 의미를 갖게 되는가?

- Corpus의 statistical pattern을 서로 다른 각도에서 분석
- 각각의 가정은 상호보완적
- 어떠한 각도에서 해석하느냐에 따라 달라짐 (도메인? 목적?)

(예를 들어 Attention은 LM과 DH가 모두 고려됨)

Thank you