

Winter School

# BERT 구동해보기

# 강의 목표

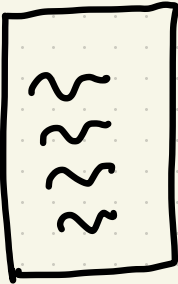
1. 단어 임베딩의 종류를 알고, 실험해본다.
2. BERT의 탄생 배경을 이해한다.
3. BERT code 돌려본다.

# 목차

- 단어 단위 임베딩 : Word-embedding
  - King-man + woman = ? : Word2Vec
  - Assume라 Assumption이 그렇게 다른가요? : FastText
  - 주변 단어만 고려한다고요? : GloVe
- 문장 단위 임베딩 : Language Model
  - BERT의 탄생 : RNN부터 Transformer까지!
  - Attention is all you need
  - BERT의 구조와 쓰임.
- BERT 구동해보기!

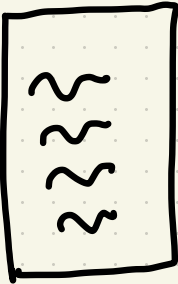
# 자연어 처리 (Natural Language Processing, NLP)

## - Encoding

input =   
text

# 자연어 처리 (Natural Language Processing, NLP)

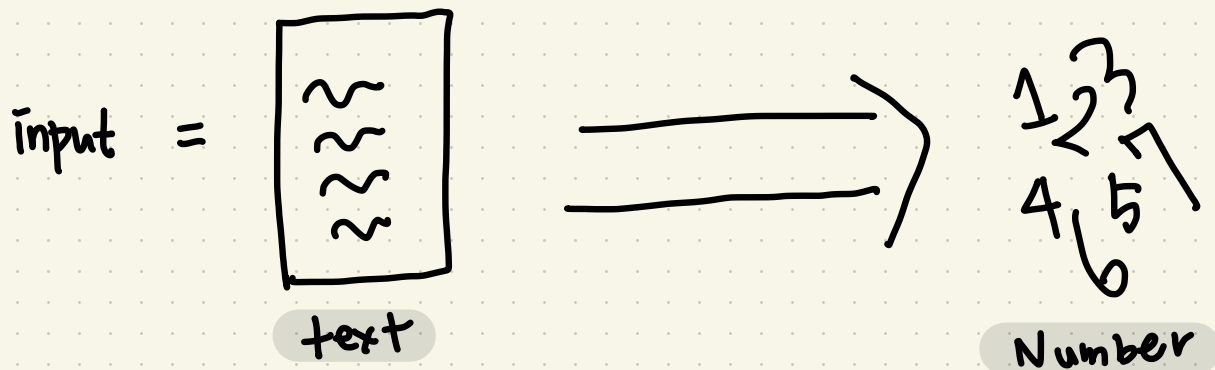
## - Encoding

input =   
text

↑  
컴퓨터가 이해할 수 없음.

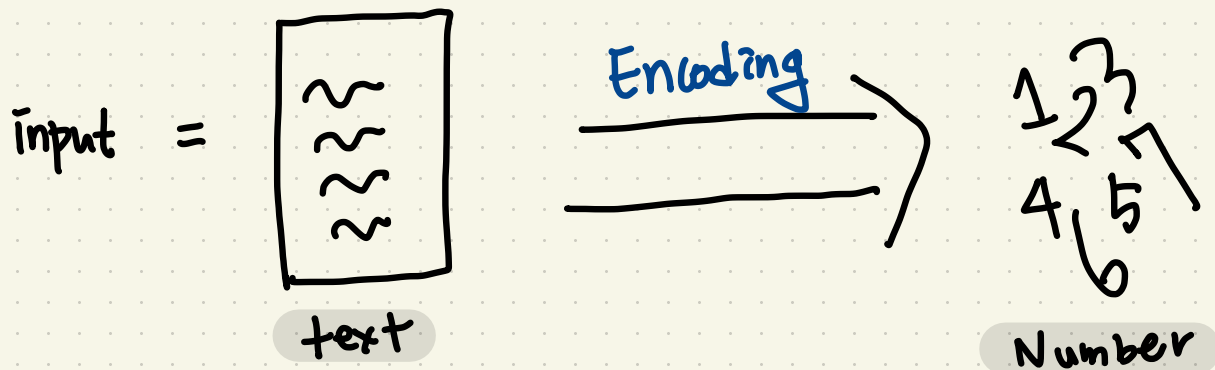
# 자연어 처리 (Natural Language Processing, NLP)

## - Encoding



# 자연어 처리 (Natural Language Processing, NLP)

## - Encoding



# 자연어 처리 (Natural Language Processing, NLP)

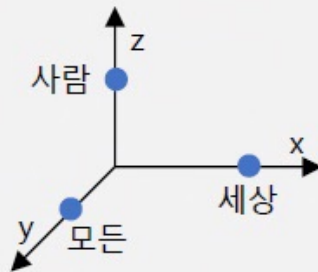
## One-hot encoding

- 자연어를 좌표평면이 나타낸 가장 쉬운 방법
- Sparse representation

세상 모든 사람

단어	Vector
세상	[1, 0, 0]
모든	[0, 1, 0]
사람	[0, 0, 1]

n개의 단어는 n차원의 벡터로 표현

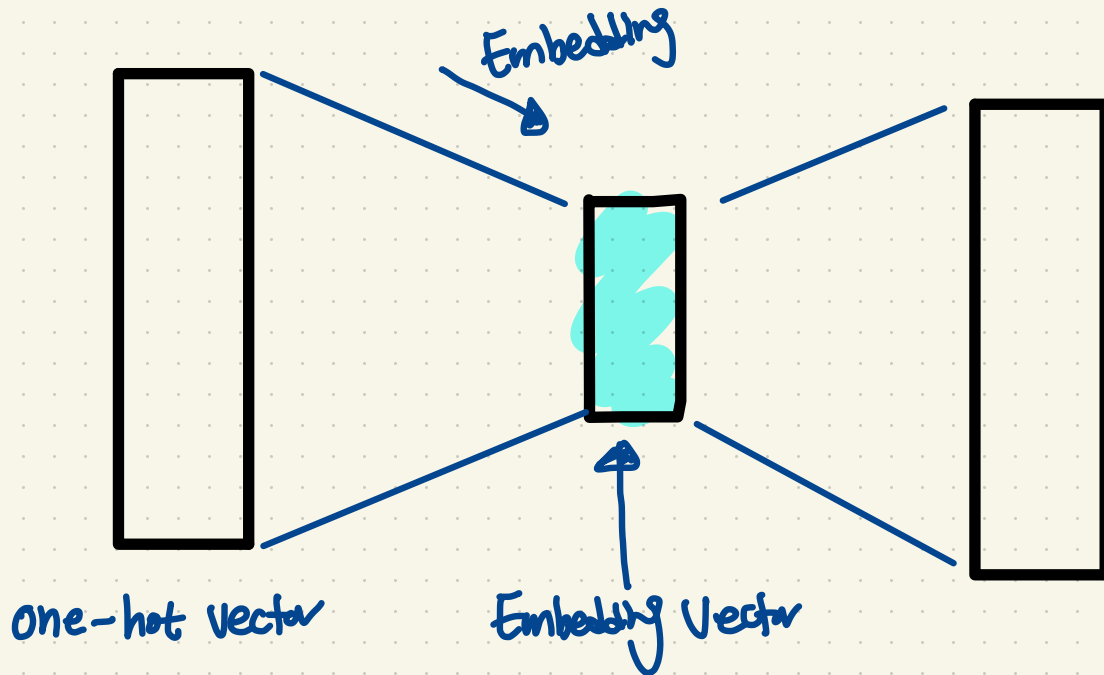


단어의 의미 정보 표현 불가능



# 자연어 처리 (Natural Language Processing, NLP)

## Word Embedding



무엇 정보를 이용해 단어를 표현?

## 무엇 정보를 이용해 단어를 표현?

- 주변 단어  $\Rightarrow$  Word2Vec
- 주변 단어 + n-gram  $\Rightarrow$  FastText
- 주변 단어 + 분포  $\Rightarrow$  GloVe

Word2Vec

# 자연어 처리 (Natural Language Processing, NLP)

## Word Embedding - Word2Vec

- 한 단어의 주변 단어들 통계, 그 단어의 의미를 파악

جرو

كلب



컴퓨터에게 자연어는 '기호'일 뿐

# 자연어 처리 (Natural Language Processing, NLP)

## Word Embedding - Word2Vec

- target word
- context word
- window size

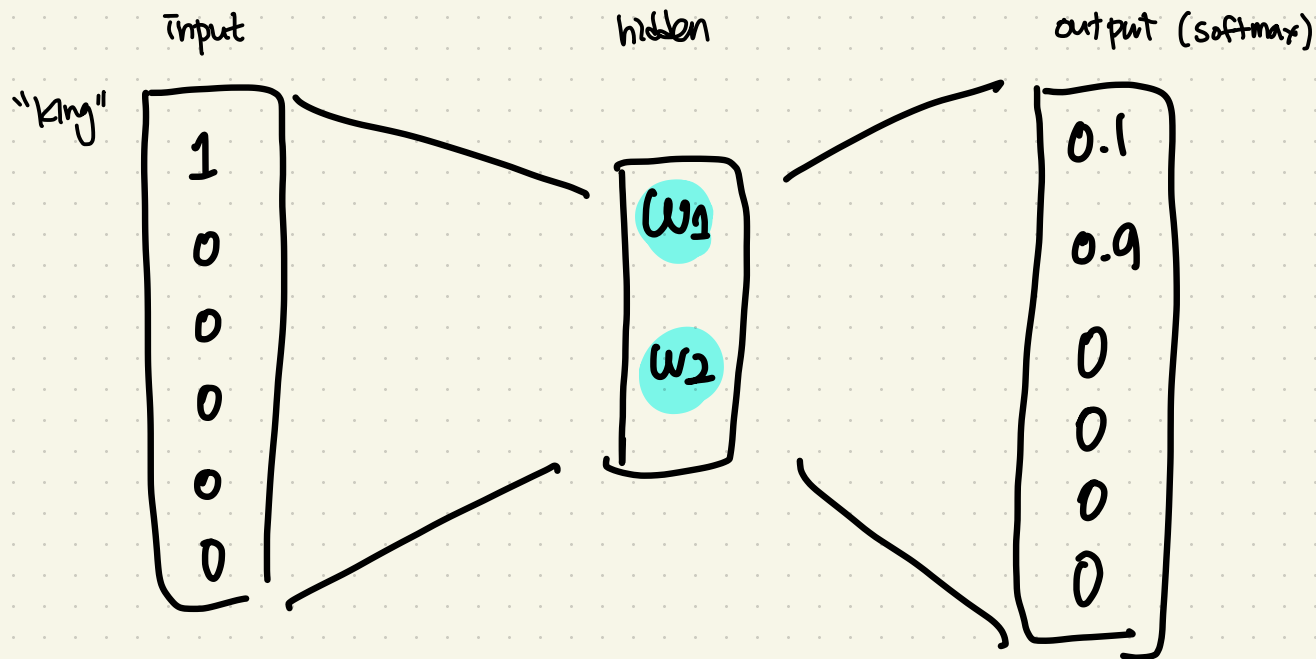
جرو (개) 가 멍멍! 하고 짖었다.

كلب (강아지) 가 멍멍! 하고 짖었다.

⇒ 각 단어의 의미는 모르겠지만, 주변 단어가 비슷하니 의미도 비슷할 것이다.

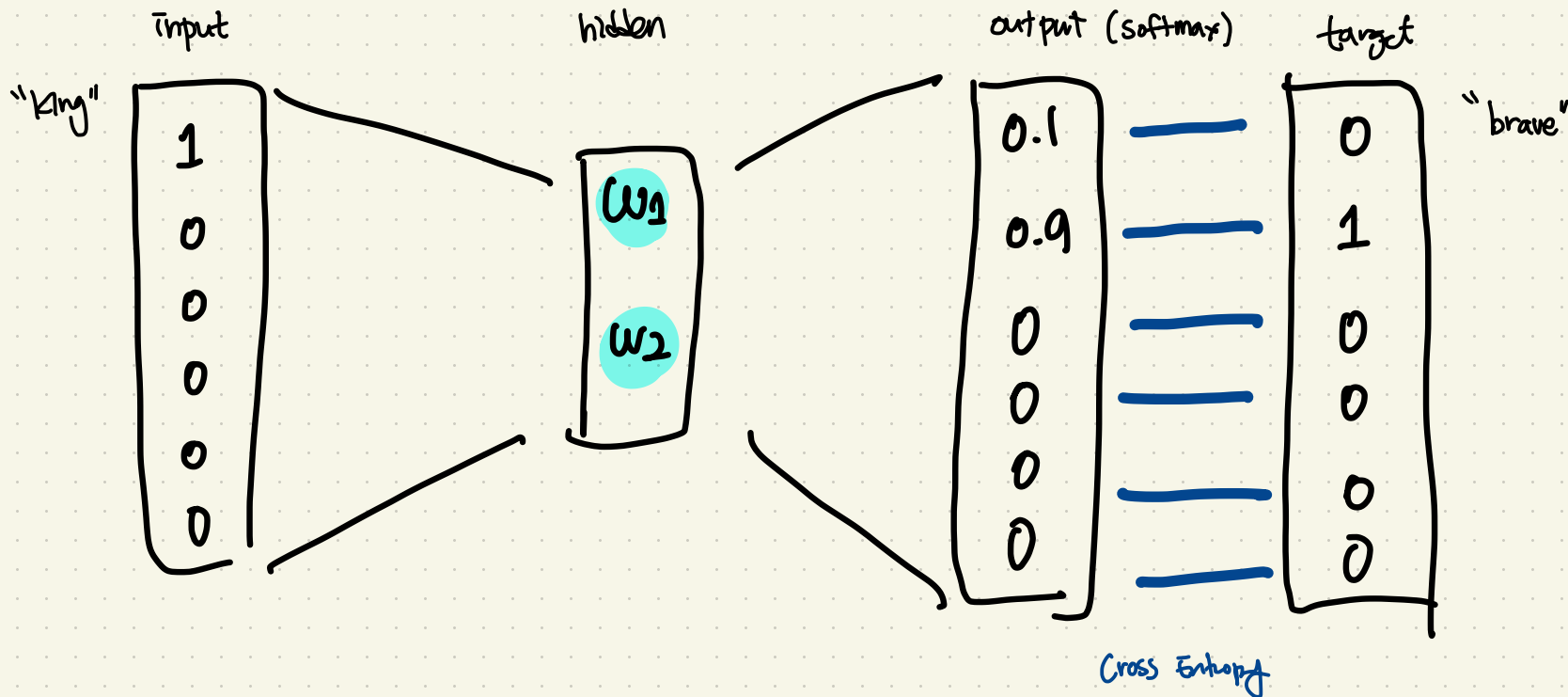
# 자연어 처리 (Natural Language Processing, NLP)

## Word2Vec - 신경망 구조



# 자연어 처리 (Natural Language Processing, NLP)

## Word2Vec - 신경망 구조

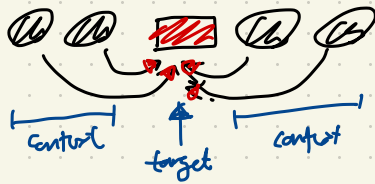




# 자연어 처리 (Natural Language Processing, NLP)

## Word2Vec - 종류

**CBoW** : Context words로 target word 예측하기

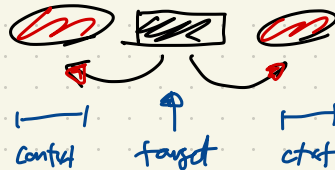


[학습 데이터 쌍]

{ Context words : target word }

→ 1개의 pair

**Skip-gram** : target word로 Context word 예측하기



[학습 데이터 쌍]

{ Context word 1 : target }

{ Context word 2 : target }

{ Context word 3 : target }

⋮

window-size  
\*  
2

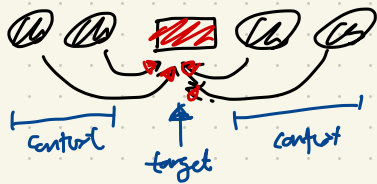
→ 여러개의 pair

→ 같은 코스로 더 많은 학습 데이터 확보 = 임베딩 품질 ↑

# 자연어 처리 (Natural Language Processing, NLP)

## Word2Vec - 종류

**CBoW** : Context words로 target word 예측하기

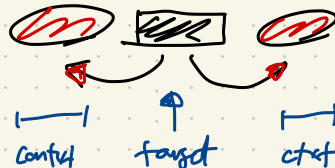


[학습 데이터 쌍]

{ Context words : target word }

→ 1개의 pair

**Skip-gram** : target word로 Context word 예측하기



[학습 데이터 쌍]

{ Context word 1 : target }

{ Context word 2 : target }

{ Context word 3 : target }

⋮

window-size  
\*  
2

→ 여러개의 pair

→ 같은 context로 더 많은 target word 예측 = 임베딩 품질 ↑

# 자연어 처리 (Natural Language Processing, NLP)

## Word2Vec - Data generation

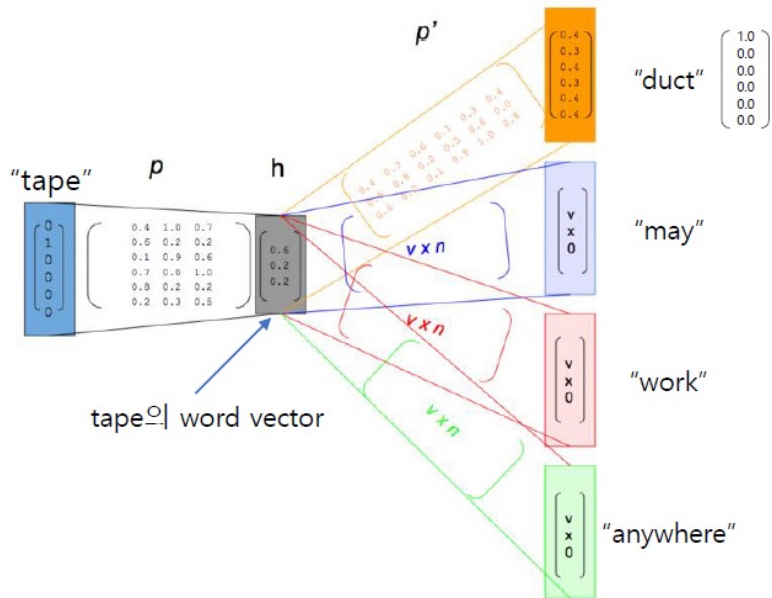
[ 코드를 통해 살펴볼게요. ]

# 자연어 처리 (Natural Language Processing, NLP)

## word2vec - skip-gram

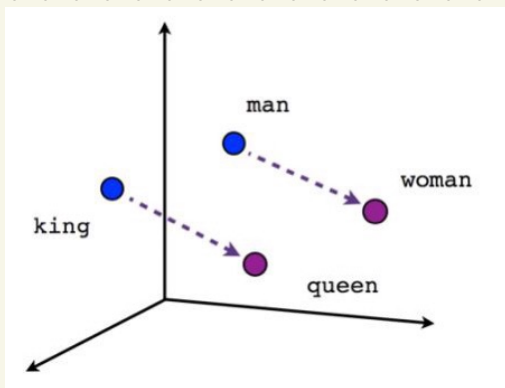
"Duct tape may works anywhere"

Word	One-hot-vector
"duct"	[1, 0, 0, 0, 0]
"tape"	[0, 1, 0, 0, 0]
"may"	[0, 0, 1, 0, 0]
"work"	[0, 0, 0, 1, 0]
"anywhere"	[0, 0, 0, 0, 1]



# 자연어 처리 (Natural Language Processing, NLP)

## Word embedding - word2vec



- 중심 단어 (target word) 의 주변 단어 (context word) 들을 이용해, 중심 단어를 표현하는 방식의 학습
- 단어의 의미는 분포와 연상
- CBow와 skip gram 방식 존재

- [장점]
- 단어만 유선 특징
  - 단어 간 관계 파악
  - 벡터 연산을 통한 추론

- [단점]
- 단어의 subword information 무시
  - Out of Vocabulary (OOV)

FastText

Assume vs Assumption

- Word2Vec에따라 단어의 다양한 용인 표현들이 서로 관련된 어휘로 란기

### 동사 원형: 모르다

모르네  
모르데  
모르지  
모르더라  
모르리라  
모르는구나  
모르잖아  
모르려나  
모르니  
모르고  
모르나  
모르면  
모르면서  
모르거든  
모르는데  
모르지만  
모르더라도  
모르다가도  
모르지조차

모르기까지  
모르기를  
모르기는  
모르기도  
모르기만  
모르는  
모르던  
모른다  
모른다면  
모른다만  
모른다치고  
모르겠다  
모르겠네  
모르겠지  
모르겠더라  
모르겠구나  
모르겠니  
모르겠고

모르겠으나  
모르겠으면  
모르겠으면서  
모르겠거나  
모르겠거든  
모르겠는데  
모르겠지만  
모르겠더라도  
모르겠다가도  
모르겠던  
모르겠다면  
모르겠다면  
모를까  
모를지  
모를지도  
모를수록  
몰라  
몰라도  
몰라서

몰라야  
몰라요  
몰라라  
몰랐다  
몰랐네  
몰랐지  
몰랐더라  
몰랐으리라  
몰랐구나  
몰랐잖아  
몰랐으려나  
몰랐으니  
몰랐거나  
몰랐거든  
몰랐는데  
몰랐지만  
몰랐더라도  
몰랐다가도  
몰랐던

몰랐다면  
몰랐지만  
몰랐을  
몰랐을까  
몰랐을지  
몰랐을지도  
몰랐어  
몰랐어도  
몰랐어야  
몰랐어요  
몰랐더라면  
몰랐더라도  
몰랐겠다  
몰랐겠네  
몰랐겠지  
몰랐겠더라  
몰랐겠구나  
몰랐겠니  
몰랐겠고

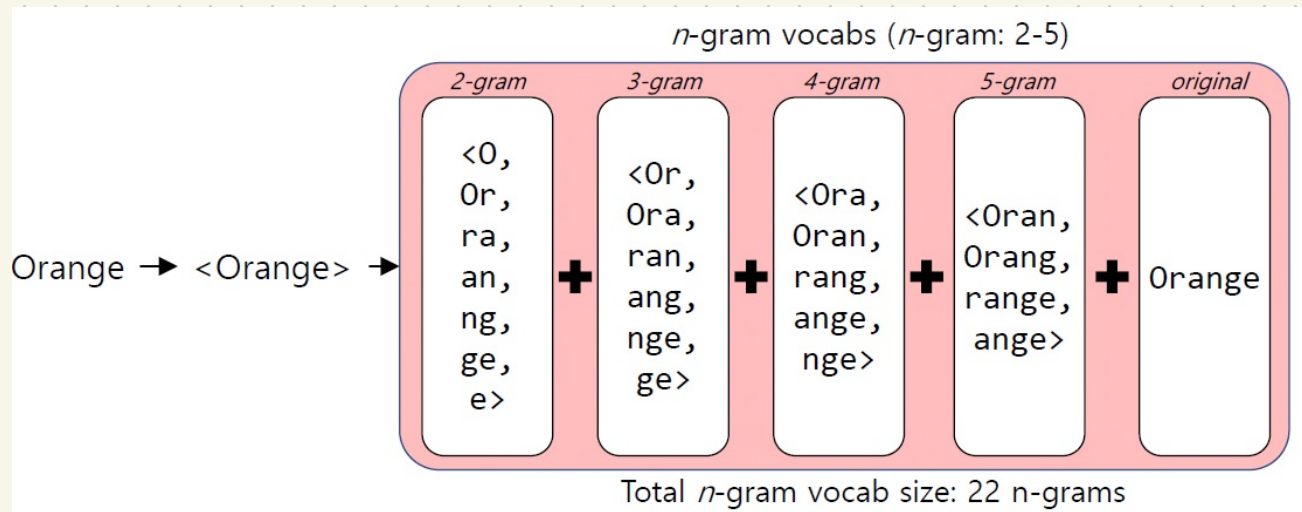
몰랐겠으나  
몰랐겠으면  
몰랐겠으면서  
몰랐겠거나  
몰랐겠거든  
몰랐겠는데  
몰랐겠지만  
몰랐겠더라도  
몰랐겠다가도  
몰랐겠던  
몰랐겠다면  
몰랐겠다면  
몰랐겠어  
몰랐겠어도  
몰랐겠어서  
몰랐겠더라  
몰랐겠어나  
몰랐겠더라면  
몰랐겠더라도



# 자연어 처리 (Natural Language Processing, NLP)

## Word embedding - FastText

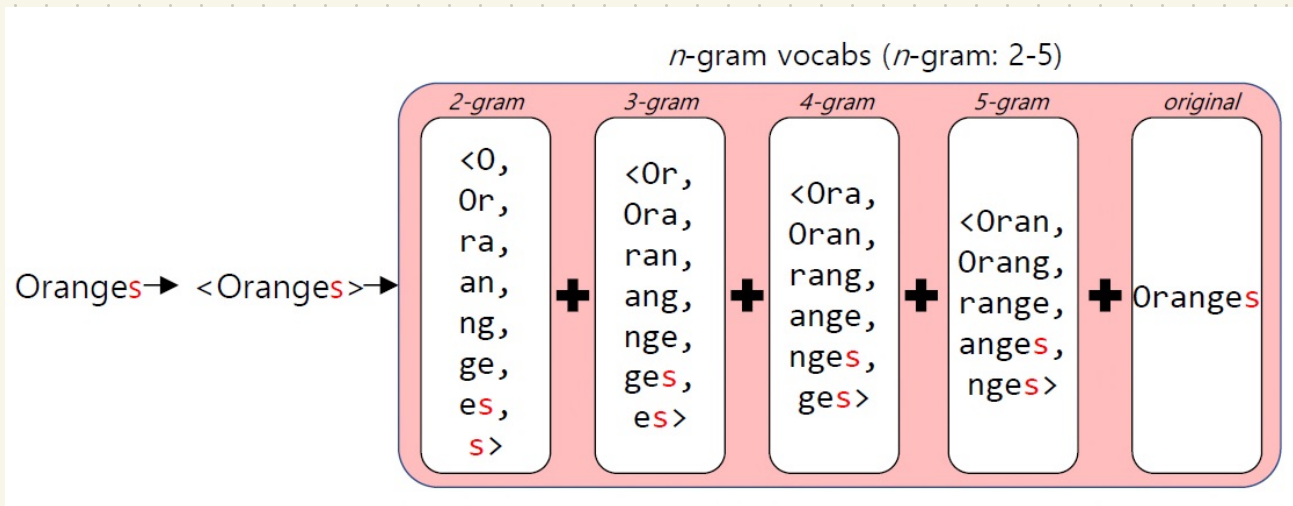
- Word2Vec과 유사한 방식의 학습
- [차이] 단어를  $n$ -gram으로 나누어 학습함.
- 이때,  $n$ -gram으로 나누어진 단어는 사전에 들어가지 않고, 별도의  $n$ -gram Vector 생성



# 자연어 처리 (Natural Language Processing, NLP)

## Word embedding - FastText

- Word2Vec과 유사한 방식의 학습
- [차이] 단어를 n-gram으로 나누어 학습함.
- 이때, n-gram의 시작원 단어는 사전에 들어가지 않고, 별도의 n-gram Vector 생성



주변 단어만 보도 충분할까?

주변 단어만 보도 충분할까?

- 전체 코퍼스의 통계 정보를 고려해볼까

GloVe

# 자연어 처리 (Natural Language Processing, NLP)

## Word Embedding - GloVe

- Word2Vec이 전체 Corpus의 정보를 담지 못한다는 문제 보완
- 두 단어의 유사도에 통계 정보가 반영
- [목적 함수] 두 벡터의 내적 =  $\log(\text{종시 등장 확률})$ 
  - 유사도
  - Corpus 전체의 통계 정보

$$\sum_{i,j=1}^{|V|} f(A_{ij}) \cdot (U_i \cdot V_j + b_i + b_j - \log A_{ij})^2$$

[  $\frac{1}{2G}$  Time ]