

# 2021학년도 2학기 언어와 컴퓨터

## 제23강 벡터 의미론 (1)

박수지

서울대학교 인문대학 언어학과

2021년 12월 1일 수요일

## 오늘의 목표

- 1 분포 가설이 무엇인지 설명할 수 있다.
- 2 두 단어 벡터 사이의 코사인 유사도를 계산할 수 있다.
- 3 TF-IDF 방식으로 벡터의 가중치를 구할 수 있다.
- 4 Word2Vec에서 skip-gram with negative sampling 방식으로 벡터의 가중치를 구할 수 있다.

## 벡터 의미론의 두 가지 직관

- 1 분포주의 직관(distributionalist intuition)  
유사한 문맥에서 나타나는 단어는 유사한 의미를 가진다.
- 2 벡터 직관(vector intuition)  
단어를 벡터공간상의 점으로 표현한다.

## Word embedding

단어를 다차원 의미 공간에 표상하는 것

## 문제

두 벡터가 얼마나 유사한지를 어떻게 측정하는가?

# 코사인 유사도(cosine similarity)

두 벡터 사이의 각이 작을수록 가깝다.

## 정의

벡터  $v = [v_1, v_2, \dots, v_N]$ 와  $w = [w_1, w_2, \dots, w_N]$  사이의 각이  $\theta$ 일 때 ( $0 \leq \theta \leq \pi$ )

$$\text{cos-sim}(v, w) = \cos \theta = \frac{v \cdot w}{\|v\| \|w\|} = \frac{v_1 w_1 + v_2 w_2 + \dots + v_N w_N}{\sqrt{v_1^2 + v_2^2 + \dots + v_N^2} \sqrt{w_1^2 + w_2^2 + \dots + w_N^2}}$$

$$\cos \theta = +1 \Rightarrow \theta = 0 \Rightarrow \text{같은 방향}$$

$$\cos \theta = -1 \Rightarrow \theta = \pi \Rightarrow \text{반대 방향}$$

## 단어에 벡터의 값을 부여하는 다양한 기법

- TF-IDF: long & sparse
- Word2Vec: short & dense
- ...

# TF-IDF(Term Frequency - Inverse Document Frequency)

## 개념

$$\text{TF-IDF}(t, d) = \text{tf}_{t,d} \times \text{idf}_t$$

$$\text{tf}_{t,d} \log_{10} [[\text{단어 } t \text{가 문서 } d \text{에 출현한 횟수}] + 1]$$

- 문서에 자주 나타날수록 중요한 단어(주제어)다.

- 문서에 한 번도 나오지 않은  $\text{tf}_{t,d}$ 의 값은 0이다.
- $\text{tf}$ 값이 높을수록 문서 내에서 중요하다.

$$\text{idf}_t \log_{10} \frac{[\text{전체 문서의 개수}]}{[\text{단어 } t \text{를 포함하는 문서의 개수}]}$$

- 이 문서 저 문서에 다 나오는 단어(예: the)은 정보량이 적다.

- 모든 문서에 나오면  $\text{idf}_t$  값이 0이 된다.
- $\text{idf}$ 값이 높을수록 일반적으로 중요하다.

100개 문서에 1번씩 나오기 vs. 10개 문서에 10번씩 나오기

# TF-IDF(Term Frequency - Inverse Document Frequency)

## 특징

- 차원 수가 크다. (코퍼스 내 전체 문서의 개수와 같음)
- 대부분의 값이 0이다. (한 문서에 들어가는 어휘는 전체의 일부에 불과함)

## 희소벡터(sparse vector)의 문제

- 1 기계학습의 특성값으로 사용하기 어렵다.
- 2 계산해야 할 매개변수의 수가 많다.
- 3 동의어를 포착하기 어렵다.
  - car와 automobile이 같은 문서에 출현하는 일은 적다. → 벡터의 차이가 크다.

# Word2Vec

## Skip-gram with negative sampling

### 특징

- 차원 수가 작다. (50-1000 정도)
- 0이 아닌 값들로 이루어져 있다.

### 주요 기법

- Skip-gram
- Continuous Bag-of-words



## Skip-gram with negative sampling

- 1 대상 단어 및 그와 이웃하는 문맥 단어를 긍정적인 사례로 다룬다.
- 2 임의의 다른 단어를 추출하여 부정적인 사례로 삼는다.
- 3 위의 두 사례를 구별하는 분류기를 로지스틱 회귀분석으로 훈련시킨다.
- 4 훈련된 매개변수 가중치 값들을 단어 벡터로 삼는다.

대상 단어(t)와 문맥 단어(c)의 예시(window size  $L = 2$ ):

... lemon, a [tablespoon of apricot jam, a] pinch ...

c1 c2 t c3 c4

# Word2Vec

## Skip-gram with negative sampling

### Negative sampling

**문제** 분류기를 위해서는 부정적인 사례도 있어야 한다.

**해결** 문맥 바깥의 단어를 추출해서 부정적인 사례로 삼는다.

대상 단어(t)의 문맥 단어(c) 하나마다  $k = 2$ 개의 노이즈 단어(n)를 추출하는 예시:

#### positive examples +

t	c
apricot	tablespoon
apricot	of

#### negative examples -

t	c	t	c
apricot	aardvark	apricot	seven
apricot	my	apricot	forever

# Word2Vec

## Skip-gram with negative sampling

단어를  $d$ 차원으로 표상하고 window size를  $L = 2$ , 부정 샘플 개수를  $k = 2$ 로 놓았을 때

$$\mathbf{t} = [t_1, t_2, \dots, t_d], \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{bmatrix}, \quad \mathbf{n}_1 = \begin{bmatrix} n_{11} \\ n_{12} \\ \vdots \\ n_{1d} \end{bmatrix}, \quad \mathbf{n}_2 = \begin{bmatrix} n_{21} \\ n_{22} \\ \vdots \\ n_{2d} \end{bmatrix}$$

### 목표

- 1  $P(+|\mathbf{t}, \mathbf{c}) \cdots (\mathbf{t}, \mathbf{c})$ 가 이웃일 확률  $\cdots \mathbf{t} \cdot \mathbf{c}$ 를 높게!
- 2  $P(-|\mathbf{t}, \mathbf{n}_i) \cdots (\mathbf{t}, \mathbf{n}_i)$ 가 이웃이 아닐 확률  $\cdots \mathbf{t} \cdot \mathbf{n}_i$ 를 낮게!

# Word2Vec

## Skip-gram with negative sampling

$$P(+|t, c) = \frac{1}{1 + e^{-t \cdot c}}$$

$$\begin{aligned} P(-|t, n_i) &= 1 - P(+|t, n_i) \\ &= \frac{1 + e^{-t \cdot n_i}}{1 + e^{-t \cdot n_i}} - \frac{1}{1 + e^{-t \cdot n_i}} \\ &= \frac{e^{-t \cdot n_i}}{1 + e^{-t \cdot n_i}} \\ &= \frac{e^{-t \cdot n_i}}{1 + e^{-t \cdot n_i}} \times \frac{e^{t \cdot n_i}}{e^{t \cdot n_i}} \\ &= \frac{1}{1 + e^{t \cdot n_i}} \end{aligned}$$

# Word2Vec

## Skip-gram with negative sampling

대상  $t$ , 문맥  $c$ , 노이즈  $n_i$  ( $i = 1, 2$ )가 주어졌을 때

$$\begin{aligned}\text{손실함수 } L_{CE} &= - \left[ \log P(+|t, c) + \sum_{i=1}^2 \log P(-|t, n_i) \right] \\ &= - \left[ \log \frac{1}{1 + e^{-t \cdot c}} + \sum_{i=1}^2 \log \frac{1}{1 + e^{t \cdot n_i}} \right]\end{aligned}$$

### 비교

로지스틱 회귀분석 분류기 가중치  $a$ 는 변수, 데이터  $x$ 는 상수

Word2Vec (Skip-gram) 대상 벡터  $t$ 와 문맥  $c$  모두 변수

# Word2Vec

## Skip-gram with negative sampling

**대상 단어** 어휘  $V$ 의  $i$ 번째 단어

**문맥 단어** 어휘  $V$ 의  $j$ 번째 단어

$$T \times C = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1d} \\ \dots & \dots & \dots & \dots \\ t_{i1} & t_{i2} & \dots & t_{id} \\ \dots & \dots & \dots & \dots \\ t_{|V|1} & t_{|V|2} & \dots & t_{|V|d} \end{bmatrix} \cdot \begin{bmatrix} c_{11} & \vdots & c_{1j} & \vdots & c_{1|V|} \\ c_{21} & \vdots & c_{2j} & \vdots & c_{2|V|} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{d1} & \vdots & c_{dj} & \vdots & c_{d|V|} \end{bmatrix}$$

## 임베딩 선택 방법

- 1  $t_i = [t_{i1}, \dots, t_{id}]$
- 2  $t_i + c_i = [t_{i1} + c_{i1}, \dots, t_{id} + c_{id}]$
- 3  $t_i \oplus c_i = [t_{i1}, \dots, t_{id}, c_{i1}, \dots, c_{id}]$

## 모형 매개변수

- $d$  (벡터의 차원 수)
- $L$  (window size — 이웃의 기준)

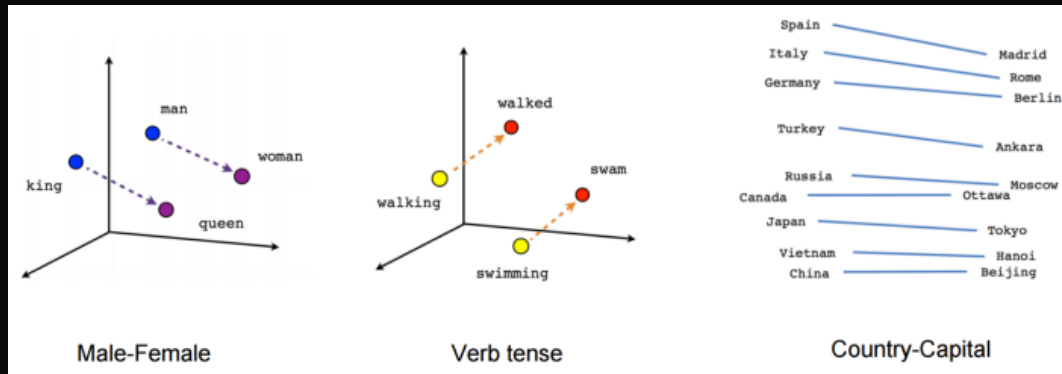
## 한계

개별 문맥에 따라 변하는 단어의 의미는 포착할 수 없다.

# Word2Vec

## 단어 유추

의미를 수치(벡터)로 표현하기 → 단어 사이의 의미 관계를 계산할 수 있다.



<https://towardsdatascience.com/>



# Word2Vec

## 단어 유추

### 예시

단어쌍 사이의 공통된 의미 관계

- 아빠 : 엄마 = 할아버지 : 할머니 (= 남자 : 여자)
- [벡터공간에서] 아빠 - 엄마 = 할아버지 - 할머니

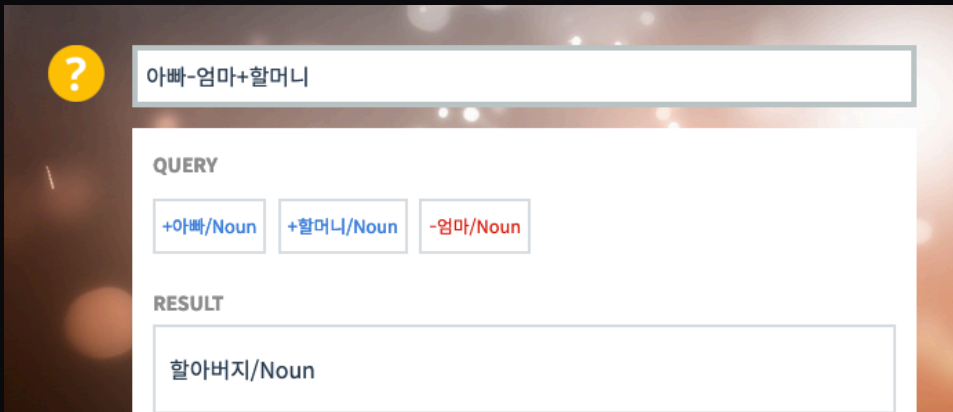
$$A : B = X : D \quad \Rightarrow \quad X = A - B + D$$

# Word2Vec

## 단어 유추

아빠 : 엄마 = 할아버지 : 할머니 관계가 성립한다는 것을 어떻게 확인할 수 있는가?

- Korean Word2Vec <https://word2vec.kr>



The screenshot shows the Korean Word2Vec web interface. At the top left, there is a yellow circle with a question mark. Below it, a search bar contains the text "아빠-엄마+할머니". Under the search bar, the word "QUERY" is displayed. Below "QUERY", there are three buttons: "+아빠/Noun" (blue text), "+할머니/Noun" (blue text), and "-엄마/Noun" (red text). Below these buttons, the word "RESULT" is displayed. Below "RESULT", there is a box containing the text "할아버지/Noun".

# Word2Vec

단어 유추

## 다양한 의미 관계

- 나라-국기
- 주군-책사
- 반의 관계
- 동사 활용

## 데이터의 한계

- 대통령-나라

# 유사한/연관된 단어 찾기

## 단어의 의미 관계

**연관성** 두 단어가 서로 이웃인 경우 (syntagmatic association)

- 예: 연필-공책

**유사성** 두 단어의 이웃이 비슷한 경우 (paradigmatic association)

- 예: 연필-볼펜

실습 코드 [https://colab.research.google.com/drive/1bBg\\_CcZduJiFS7DVTMQxg0iPgMyqodZT?usp=sharing](https://colab.research.google.com/drive/1bBg_CcZduJiFS7DVTMQxg0iPgMyqodZT?usp=sharing)