2022학년도 1학기 컴퓨터언어학

제6강 벡터의미론과 임베딩 (2)

박수지

서울대학교 인문대학 언어학과

2022년 3월 23일 수요일

오늘의 목표

1. Word2Vec의 skip-gram with negative sampling 방식으로 단어 벡터의 가중치를 구할 수 있다.
2. 단어 임베딩이 실제 의미 관계를 어떻게 포착하는지를 두 가지 측면에서 설명할 수 있다.

숙제

- 정규 과제 [숙제01]: 필수
- 심화 과제 [숙제01심화]: 선택

제출 시 파일명 엄수!

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. $\rightarrow$ 벡터의 차이가 크다.

단어의 벡터를 얻는 몇 가지 방법

- 공기행렬의 행벡터를 취한다. — TF-IDF
  - Long & sparse
- 로지스틱 회귀분석에서 학습시킨 가중치 벡터를 사용한다. — Word2Vec
  - Short & dense

로지스틱 회귀분석을 사용한다면…

- 무엇을 분류할 것인가?
- 어떤 데이터(입력 및 정답)를 사용할 것인가?

Skip-gram의 아이디어

- 무엇을 분류할 것인가?
  어떤 단어 c가 대상 단어 w의 이웃인지(+) 아닌지(−) 예측한다.
- 어떤 데이터(입력 및 정답)를 사용할 것인가?
  코퍼스에서 단어 c가 대상 단어 w의 문맥 단어로 관측되면 입력 w, c의 정답을 +로 간주한다.

... lemon, a [tablespoon of apricot jam, a] pinch ...
c1 c2 w c3 c4

Negative Sampling

문제 +와 −를 분류하려면 −에 해당하는 데이터도 있어야 한다.

해법 코퍼스에서 문맥 바깥의 단어(노이즈 단어)를 랜덤으로 추출하여 −로 분류한다.

positive examples +

w	c_{pos}
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

w	c_{neg}	w	c_{neg}
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

분류기

$w \in V$ 대상 단어
$\vec{w} \in \mathbb{R}^d$ w의 대상 임베딩 벡터

$c \in V$ 문맥 단어 또는 노이즈 단어
$\vec{c} \in \mathbb{R}^d$ c의 문맥 임베딩 벡터

주의!

어휘 집합 V에 속하는 단어는 대상 임베딩과 문맥 임베딩 두 가지 벡터로 표현된다.

확률 추정치 정의

단어쌍 (w, c)가 긍정$(+)$으로 분류될 확률

$$P(+|w, c) = \sigma(\vec{c} \cdot \vec{w}) = \frac{1}{1 + \exp(-\vec{c} \cdot \vec{w})}$$

분류기

$$
\begin{aligned}
P(-|w,c) = 1 - P(+|w,c) &= 1 - \sigma(\vec{c}\cdot\vec{w}) \\
&= 1 - \frac{1}{1+\exp(-\vec{c}\cdot\vec{w})} \\
&= \frac{1+\exp(-\vec{c}\cdot\vec{w})}{1+\exp(-\vec{c}\cdot\vec{w})} - \frac{1}{1+\exp(-\vec{c}\cdot\vec{w})} \\
&= \frac{\exp(-\vec{c}\cdot\vec{w})}{1+\exp(-\vec{c}\cdot\vec{w})} \\
&= \frac{\exp(-\vec{c}\cdot\vec{w})}{1+\exp(-\vec{c}\cdot\vec{w})} \times \frac{\exp(\vec{c}\cdot\vec{w})}{\exp(\vec{c}\cdot\vec{w})} \\
&= \frac{1}{\exp(\vec{c}\cdot\vec{w})+1} \\
&= \frac{1}{1+\exp(\vec{c}\cdot\vec{w})} = \sigma(-\vec{c}\cdot\vec{w})
\end{aligned}
$$

분류기

독립성 가정

윈도 $c_{1:L}$ 내의 모든 문맥 단어 $c_1, c_2, \cdots, c_L$ 은 서로 독립적으로 출현한다. (사실은 아님)

$$P(+|w, c_{1:L}) = P(+|w, c_1) \times P(+|w, c_2) \times \cdots \times P(+|w, c_L)$$
$$\log P(+|w, c_{1:L}) = \log P(+|w, c_1) + P(+|w, c_2) + \cdots + \log P(+|w, c_L)$$

주의!

- 일반적인 로지스틱 회귀분석에서 $\sigma(\vec{w} \cdot \vec{x} + b)$의 매개변수: $\vec{w}$, b
- Word2Vec 분류기에서 $\sigma(\vec{c} \cdot \vec{w})$의 매개변수: $\vec{c}$, $\vec{w}$

분류기

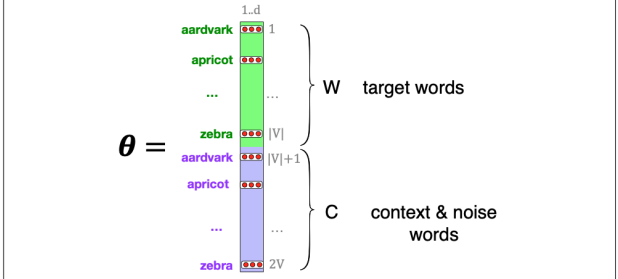


Figure 6.13 The embeddings learned by the skipgram model. The algorithm stores two embeddings for each word, the target embedding (sometimes called the input embedding) and the context embedding (sometimes called the output embedding). The parameter θ that the algorithm learns is thus a matrix of $2|V|$ vectors, each of dimension d, formed by concatenating two matrices, the target embeddings $\mathbf{W}$ and the context+noise embeddings $\mathbf{C}$.

매개변수 개수

$d \times 2|V|$

- d 한 벡터의 임베딩 차원
- $|V|$ 대상 임베딩 벡터 $\vec{w}$의 개수
- $|V|$ 문맥 임베딩 벡터 $\vec{c}$의 개수

Skip-gram 임베딩 학습

훈련 집합

대상 단어 w의 문맥 단어 c_{pos} 마다 k개의 노이즈 단어 c_{neg_i} 를 추출하는 경우

$\{((w, c_{pos}), +), ((w, c_{neg_1}), -), , \cdots, ((w, c_{neg_k}), -), \cdots\}$

모형 학습 목표

1. $P(+|w, c_{pos})$의 값을 크게 만든다.
2. $P(+|w, c_{neg_i})$의 값을 작게 만든다(= $P(-|w, c_{neg_i})$의 값을 크게 만든다).

Skip-gram 임베딩 학습

모형 학습 목표

1. $P(+|w, c_{pos})$의 값을 크게 만든다.
2. $P(+|w, c_{neg_i})$의 값을 작게 만든다(= $P(-|w, c_{neg_i})$의 값을 크게 만든다).

교차엔트로피 손실함수

$$
\begin{aligned}
L_{CE} &= -\log P(\texttt{class}|\texttt{data}) \\
&= -\log\left[P(+|w, c_{pos}) \times P(-|w, c_{neg_1}) \times \cdots \times P(-|w, c_{neg_k})\right] \\
&= -\left[\log P(+|w, c_{pos}) + \log P(-|w, c_{neg_1}) + \cdots + \log P(-|w, c_{neg_k})\right] \\
&= -\left[\log \sigma(\overrightarrow{c_{pos}} \cdot \vec{w}) + \log \sigma(-\overrightarrow{c_{neg_1}} \cdot \vec{w}) + \cdots + \log \sigma(-\overrightarrow{c_{neg_k}} \cdot \vec{w})\right]
\end{aligned}
$$

Skip-gram 임베딩 학습

교차엔트로피 손실함수

$$L_{CE} = -\left[\log \sigma(\overrightarrow{c_{pos}} \cdot \vec{w}) + \sum_{i=1}^{k} \log \sigma(-\overrightarrow{c_{neg_i}} \cdot \vec{w})\right]$$

학습 방법

각 매개변수에 대한 편도함수 $\frac{\partial L_{CE}}{\partial \overrightarrow{c_{pos}}}$, $\frac{\partial L_{CE}}{\partial \overrightarrow{c_{neg}}}$, $\frac{\partial L_{CE}}{\partial \vec{w}}$ 의 식을 사용하여 경사하강법을 적용한다.

Skip-gram 임베딩 학습

단어 임베딩 선택

사실 단어 i는 대상 임베딩 $\vec{w}_i$와 문맥 임베딩 $\vec{c}_i$ 두 가지 벡터로 학습된다.

문제 두 임베딩 중 어느 것을 단어 벡터로 선택하는가?

1. $\vec{w}_i + \vec{c}_i$ (두 벡터의 차원이 d로 같으므로 서로 더할 수 있음)
2. $\vec{w}_i$ ($\vec{c}_i$ 벡터는 버림)

Word2Vec 모형의 성능에 영향을 주는 초매개변수

- 단어 벡터의 차원 d
- Negative sample 개수 k
- Window size L

유사성과 연관성

Skip-gram 모형의 Window size와 단어의 의미 관계

"Hogwarts"와 가장 "가까운" 단어는? (Levy and Goldberg 2014)

Window size ± 2 유사한 단어 (2차 공기, paradigmatic association)

- "Sunnydale" (from Buffy the Vampire Slayer)
- "Evernight" (from a vampire series)

"Welcome to Hogwart/Sunnydale!"

Window size ± 5 연관된 단어 (1차 공기, syntagmatic association)

- "Dumbledore"
- "Malfoy"
- "half-blood"

"Malfoy's first year at Hogwarts School..."

단어 유추: 평행사변형 모형

$$\overrightarrow{\text{tree}} - \overrightarrow{\text{apple}} = \overrightarrow{\text{vine}} - \overrightarrow{\text{grape}} \quad \Rightarrow \quad \overrightarrow{\text{vine}} = \overrightarrow{\text{tree}} - \overrightarrow{\text{apple}} + \overrightarrow{\text{grape}}$$

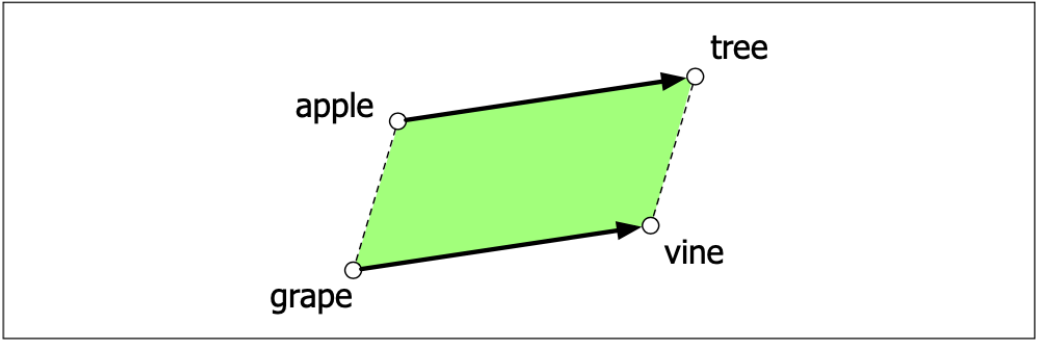


Figure 6.15 The parallelogram model for analogy problems (Rumelhart and Abrahamson, 1973): the location of $\overrightarrow{\text{vine}}$ can be found by subtracting $\overrightarrow{\text{apple}}$ from $\overrightarrow{\text{tree}}$ and adding $\overrightarrow{\text{grape}}$.

단어 유추: 평행사변형 모형

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

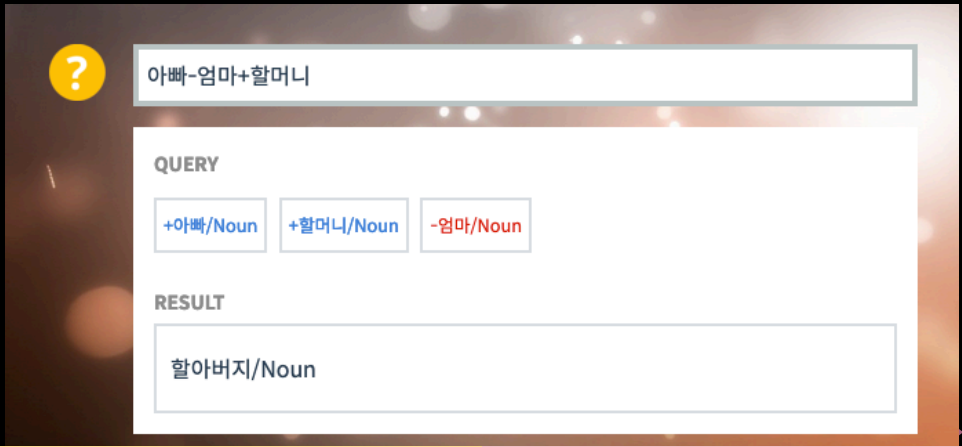
예시: 단어쌍 사이의 공통된 의미 관계

- 아빠 : 엄마 = 할아버지 : 할머니 (= 남자 : 여자)
- $\overrightarrow{\text{아빠}} - \overrightarrow{\text{엄마}} = \overrightarrow{\text{할아버지}} - \overrightarrow{\text{할머니}}$
- $\overrightarrow{\text{할아버지}} = \overrightarrow{\text{아빠}} - \overrightarrow{\text{엄마}} + \overrightarrow{\text{할머니}}$

단어 유추: 평행사변형 모형

아빠 : 엄마 = 할아버지 : 할머니 관계가 성립한다는 것을 실제로 확인해 보자!

- Korean Word2Vec `https://word2vec.kr`



단어 유추: 평행사변형 모형

한국어 Word2Vec 예시

$$\text{한국:서울=X:도쿄} \Rightarrow \overrightarrow{\text{한국}} - \overrightarrow{\text{서울}} + \overrightarrow{\text{도쿄}} = \text{X}$$

$$\text{태극기:한국=X:미국} \Rightarrow \overrightarrow{\text{태극기}} - \overrightarrow{\text{한국}} + \overrightarrow{\text{미국}} = \text{X}$$

$$\text{긴:짧은=X:작은} \Rightarrow \overrightarrow{\text{긴}} - \overrightarrow{\text{짧은}} + \overrightarrow{\text{작은}} = \text{X}$$

$$\text{순욱:조조=X:유비} \Rightarrow \overrightarrow{\text{순욱}} - \overrightarrow{\text{조조}} + \overrightarrow{\text{유비}} = \text{X}$$

$$\text{문재인:한국=X:프랑스} \Rightarrow \overrightarrow{\text{문재인}} - \overrightarrow{\text{한국}} + \overrightarrow{\text{프랑스}} = \text{X} \cdots \text{앗!}$$

문제

현실을 반영하는 데이터의 한계

father : doctor = mother : X $\Rightarrow$ X = nurse??

편향의 문제

- 인종·성 등에 관한 편견을 단순히 반영할 뿐만 아니라 확대하고 재생산한다.
- 실제로 유색 인종의 이력서가 부정적으로 평가되는 등 사회적 문제를 일으킨다.

신경망 모형을 위한 준비

1. 로지스틱 회귀: 입력 벡터를 선형변환한 후 분류함수를 적용하여 출력값을 얻는다.
2. 단어 임베딩: 단어의 의미를 밀집벡터로 표상한다.

다음 주에 할 일

- SLP3 7장 읽기
- (선택)《밑바닥부터 시작하는 딥러닝 1》2장 읽기