밑바닥부터 시작하는 딥러닝2 CHAPTER 2

자연어와 단어의 분산 표현

자연어 처리란

자연어(natural language): 사람들이 평소에 쓰는 말

자연어 처리(natural language processing): 컴퓨터가 사람 말을 이해하도록 하는 것

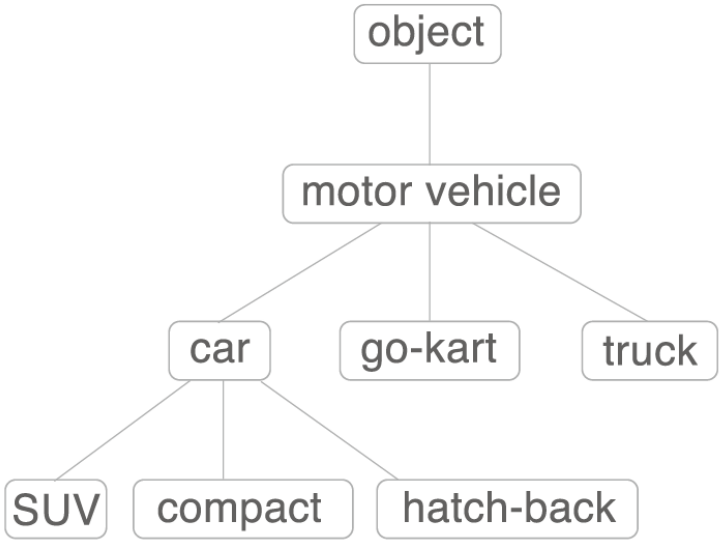
자연어 처리 예시: 검색 엔진, 번역, 질의응답 시스템, 문장 자동 요약, 감정분석 등등

단어의 의미를 이해시키기 위한 방법으로는 시소러스를 활용한 기법, 통계 기반 기법, 추론 기반 기법

시소러스

유의어 집합 + 그래프

그림 2-2 단어들을 의미의 상 · 하위 관계에 기초해 그래프로 표현한다(문헌 [14]를 참고하여 그림).



WordNet

시대 변화에 대응하기 어려움, 엄청난 인적 비용, 단어의 미묘한 차이 구분이 어려움 등의 문제점

통계 기반 기법

말뭉치(corpus): 자연어 처리 연구나 애플리케이션을 염두에 두고 수집된 텍스트 데이터, 문장들은 자연어의 특성

말뭉치에서 자동적, 효율적으로 단어의 의미들을 추출

말뭉치에 전처리 과정 필요: 예시에서는 말뭉치를 단어로 분할, 단어들을 단어 id 목록으로 변환

단어의 분산 표현(distributional representation): 단어의 의미를 벡터로 표현, 최종적으로 밀집벡터(대부분의 원소가 0이 아닌 실수인 벡터)로 표현 <- 아 이래서 차원 감소가 필요한 이유 중 하나?!

분포 가설(단어의 의미는 주변 단어에 의해 형성된다)에 기반 -> '맥락' 중요

동시 발생 행렬(co-occurrence matrix): 어떤 단어에 대해서 윈도우 크기에 따른 주변에 동시 발생한 단어들이 어떤 것들이 있고 몇 번 등장했는지를 집계하여 행렬로 나타낸 것, 각 행은 해당 단어를 표현한 벡터가 된다

단어 벡터 간 유사도: 대표적으로 코사인 유사도두 벡터 x, y가 가리키는 방향이 얼마나 비슷한지 -> 1이면 완전히 같다 / -1이면 완전히 반대

$$\text{similarity}(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{x_1y_1+\cdots+x_ny_n}{\sqrt{x_1^2+\cdots+x_n^2}\sqrt{y_1^2+\cdots+y_n^2}}$$

통계 기반 기법 개선하기

점별 상호정보량(poinwise mutual information)

단지 발생 횟수만으로 관련성을 판단하기에는 관사같이 고빈도 단어들같은 예외가 있음

PMI 척도: 값이 높을수록 관련성도 높다

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

두 단어의 동시발생 횟수가 0일 경우 에러 발생 -> PPMI(positive pmi)

$$\text{PPMI}(x,y) = \max(0,\text{PMI}(x,y))$$

동시발생행력 -> ppmi 행렬: 더 나은 척도 하지만 말뭉치의 어휘 수가 증가함-> 각 단어 벡터 차원 수도 증가라는 문제점

원소 대부분이 0 -> 중요하지 않다, 노이즈에 약함, 견고하지 못함

차원 감소(dimensionality reduction)

중요한 정보는 최대한 유지하면서 벡터의 차원 줄이기

데이터를 최대한 넓게 분포시키는 축 찾기

희소 벡터 (원소 대부분이 0) -> 밀집벡터(원소 대부분이 0이 아닌 값)

특잇값분해(singular value decomposition)

그림 2-10 SVD에 의한 차원 감소

벡터 공간에서 위치가 가깝다는 것은 곧 의미가 가깝다

PTB(penn treebank) 데이터셋 평가 : 고속 SVD 추천(sklearn 모듈)