💛

10장

10.1 임베딩

입력받은 사람의 언어(자연어)를 컴퓨터가 이해할 수 있는 형태의 벡터로 변환하는 것.

희소 표현 기반 임베딩, 횟수 기반 임베딩, 예측 기반 임베딩…

10.1.1 희소 표현 기반 임베딩

sparse: 대부분의 값이 0으로 채워져 있음→ 원-핫 인코딩

- 원-핫 인코딩의 단점
- 1) 단어끼리의 유사성을 고려하지 못한다.
- 2) 차원의 저주에 빠질 수 있다.
- 3) 이를 대체하여 word2vec, glove, fasttext 등이 대체로 할 수 있다.

10.1.2 횟수 기반 임베딩

- 카운터 벡터, TF-IDF
- 카운터 벡터(CounterVectorizer())
- TF-IDF: 특정 문서 내에서 단어의 출현 빈도가 높거나 전체 문서에서 특정 단어가 포함된 문서가 적을수록 TF-IDF값이 크다.→ 단어의 중요도

10.1.3 예측 기반 임베딩

- 특정 문맥에서 어떤 단어가 나올지 예측하면서 단어를 벡터로 만드는 방식
- 워드투벡터: 신경망 알고리즘으로, 텍스트의 각 단어마다 벡터를 출력한다. → 의미론적으로 유사한 벡터는 가깝게 출력(코사인 유사도)
- 윈도우의 크기에 따라 쓸고가면서? 각 단어에 대한 가중치도 함께 출력한다.

CBOW(continuous bag of words)
- 단어를 여러 개 나열한 후, 이와 관련된 단어를 추정하는 방식. 문장에서 다음에 등장할 단

어를 예측하는 것.

- 각 문맥 단어를 은닉층으로 투사하는 가중치 행렬은 모든 단어에서 동일함.

skip-gram

- 특정한 단어에서 문맥이 될 수 있는 단어를 예측한다.
- 중심 단어에서 주변 단어를 예측하는 방식을 사용.
- 입력 단어 주변의 단어 k개를 문맥을 보고 예측 모형을 생성(k: 윈도우 크기)

fastText

- 워드투벡터의 단점을 보완하고자 개발
- 기존(워드투벡터): 분산 표현(단어의 분산 분포가 유사한 단어들에 비슷한 벡터 값을 할당하여 표현), 사전에 없는 단어에 대해서는 벡터 값 할당 못한다. 자주 사용되지 못하는 단어에 불안정하다.
- 변경(word representation): 노이즈에 강하고 새로운 단어에 대해 형태적 유사성을 고려한다.

** 사전에 없는 단어에 벡터 값 부여하기

- n개의 값에 따라 n개로 단어의 분리 수준을 결정한다.
- 모든 단어를 n그램에 대해 임베딩 한 후, 분리된 단어와 유사도를 계산하여 사전에 없는 단어도 의미를 유추할 수 있다.

10.1.4 횟수/예측 기반 임베딩

글로브(GloVe)

- 단어에 대한 글로벌 동시 발생 확률 정보를 포함한다.
- 단어에 대한 통계 정보+skip-gram

10.2 트랜스포머 어텐션

언어 번역에서 사용되므로 인코더+디코더 네트워크를 이용한다.

- 모든 벡터를 디코더로 보내는데, 기울기 소멸 문제를 해결하기 위함이다. → 행렬 크기가 매우 커지는 단점이 발생. (softmax 함수를 이용→ 디코더에 전달)
- 디코더에 갑자기 많은 정보가 들어오므로 더 집중할(attention) 벡터를 소프트맥스 함수로 점수를 매겨 계산한다.

- 인코더: 셀프 어텐션+feed forward neural network
- 디코더: 3개의 층으로 이루어짐. 인코더-디코더 어텐션 층

10.2.1 seq2se1

번역에 초점을 둔 모델이다.

10.2.2 BERT

기존의 단방향 자연어 처리 모델의 단점을 보완한 양방향 자연어 처리 모델이다. 트랜스포머를 이용하여 구현된 pretrained model이다.

10.3 한국어 임베딩