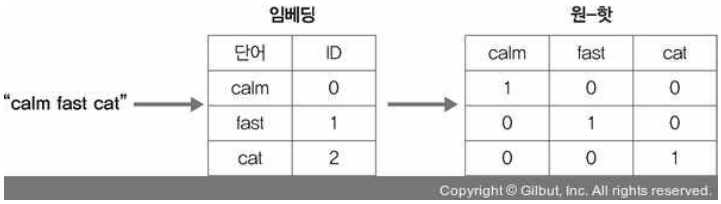
preview 15 nlp 임베딩

임베딩 : 자연어 -> 숫자 벡터
임베딩 역할 : • 단어 및 문장 간 관련성 계산 • 의미적 혹은 문법적 정보의 함축

희소 표현 기반 임베딩
• 원핫인코딩



단점 :
▫ 두 단어는 서로 직교하게 되므로, 단어끼리의 관계성 없이 서로 독립적이게 된다.
▫ 차원의 저주 (존나 커짐)

횟수기반 임베딩
• 카운터 벡터 (토크나이징과 벡터화가 동시에)
문서 집합에서 단어를 토큰으로 생성하고 각 단어의 출현 빈도수를 이용하여 인코딩해서 벡터를 만드는 방법
CountVectorizer()

• TF-IDF
정보 검색론(Information Retrieval, IR)에서 가중치를 구할 때 사용되는 알고리즘
TF(Term Frequency)(단어 빈도)

$$tf_{t,d} = \begin{cases} 1 + \log count(t,d) & count(t,d) > 0 \text{일 때} \\ 0 & \text{그 외} \end{cases}$$

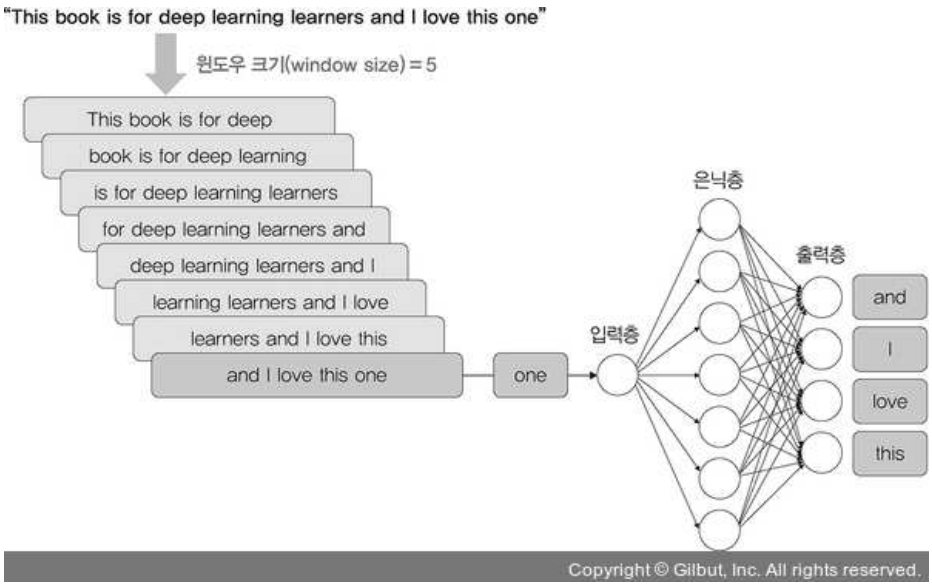
IDF(Inverse Document Frequency)(역문서 빈도)

$$idf_t = \log(\frac{N}{df_t}) = \log(\frac{\text{전체 문서 개수}}{\text{특정 단어 } t\text{가 포함된 문서 개수}})$$

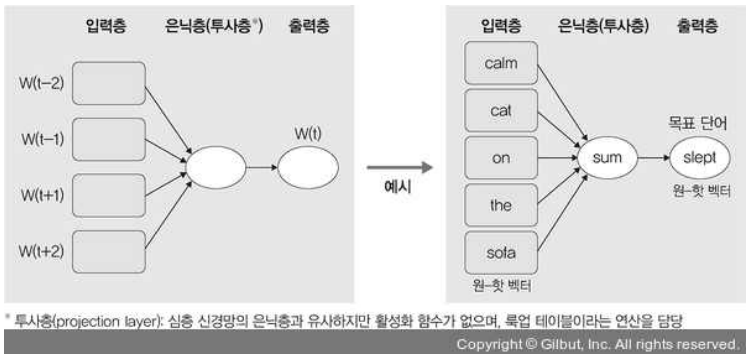
*smoothing

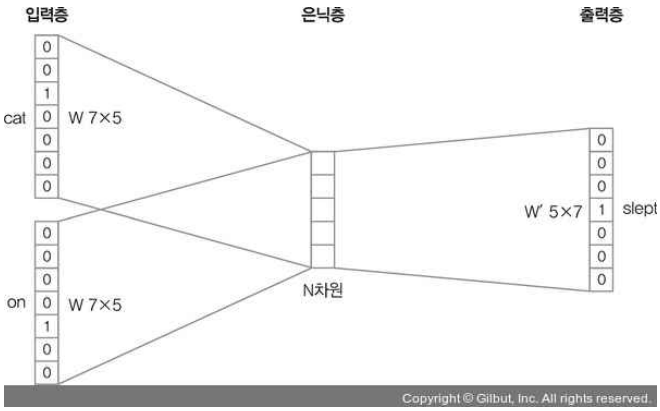
쓰이는 곳:
▫ 키워드 검색을 기반으로 하는 검색 엔진
▫ 중요 키워드 분석
▫ 검색 엔진에서 검색 결과의 순위를 결정

예측 기반 임베딩
•워드투벡터
신경망 알고리즘으로, 주어진 텍스트에서 텍스트의 각 단어마다 하나씩 일련의 벡터를 출력함. 특정 단어의 동의어를 찾을 수 있다(코사인 유사도).



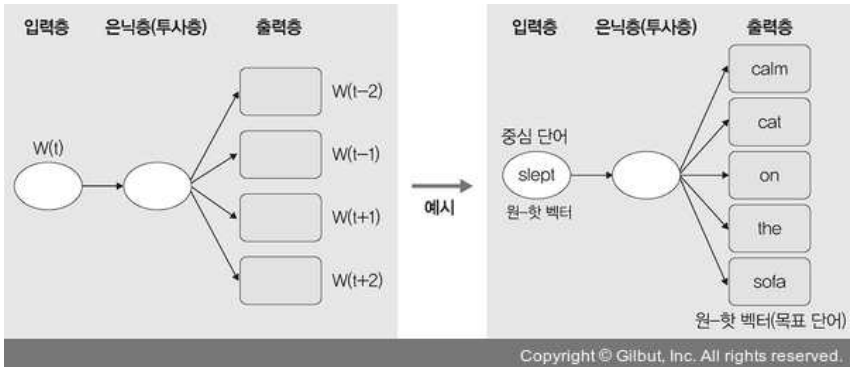
•CBOW(Continuous Bag Of Words)
다음에 등장할 단어를 예측한다.





•skip-gram

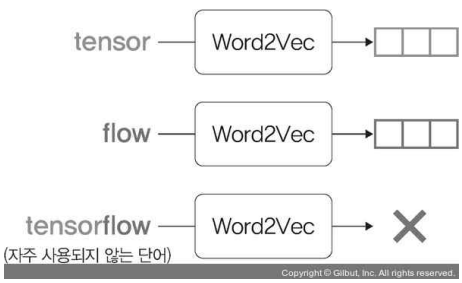
CBOW 방식과 반대로 특정한 단어에서 문맥이 될 수 있는 단어를 예측한다.



•패스트텍스트

워드투벡터의 단점을 보완하고자 페이스북에서 개발함.

“존 워드투벡터의 워드 임베딩 방식은 분산 표현(distributed representation)을 이용하여 단어의 분산 분포가 유사한 단어들에 비슷한 벡터 값을 할당하여 표현합니다. 따라서 워드투벡터는 사전에 없는 단어에 대해서는 벡터 값을 얻을 수 없습니다. 또한, 워드투벡터는 자주 사용되지 않는 단어에 대해서는 학습이 불안정합니다.”

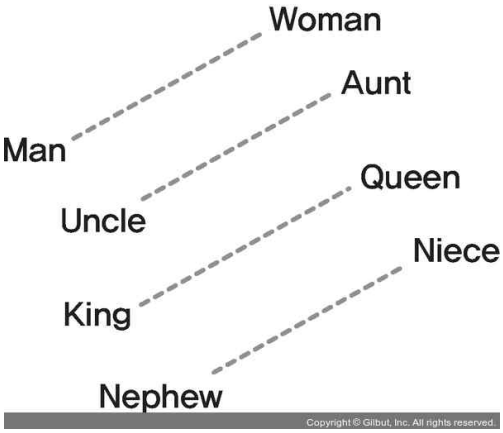


패스트텍스트는 이러한 단점들을 보완하려고 개발된 단어 표현(word representation) 방법을 사용한다. 패스트텍스트는 노이즈에 강하며, 새로운 단어에 대해서는 형태적 유사성을 고려한 벡터 값을 얻기 때문에 자연어 처리 분야에서 많이 사용되는 알고리즘

횟수/예측 기반 임베딩

•글로브

단어에 대한 글로벌 동시 발생 확률(global co-occurrence statistics) 정보를 포함하는 단어 임베딩 방법 (통계 정보와 skip-gram을 합친 방식)

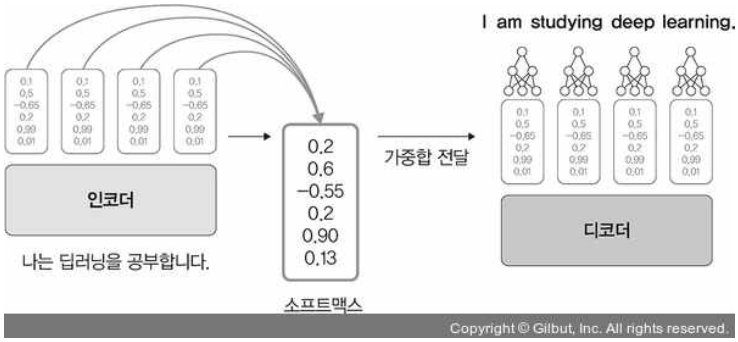


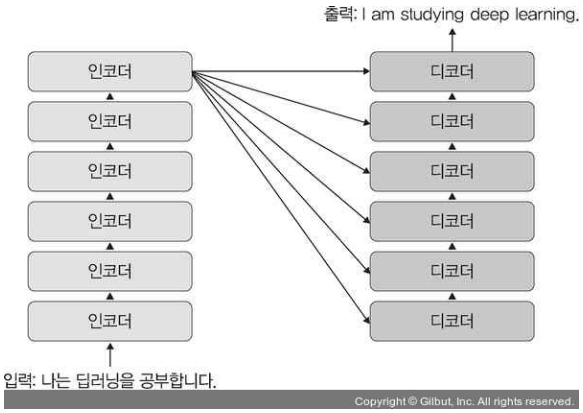
2. 트랜스포머 어텐션

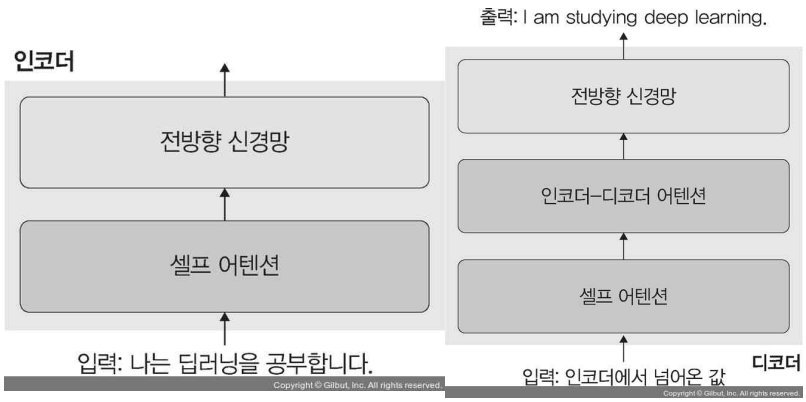
언어 번역에서 사용되기 때문에 인코더와 디코더 네트워크를 사용함.

모든 벡터를 전달하는 이유는 시간이 흐를수록 초기 정보를 잃어버리는 기울기 소멸 문제를 해결하기 위해서이다. 하지만 모든 벡터가 전달되기 때문에 행렬 크기가 굉장히 커지는 단점이 있는데, 이것을 해결하기 위해 소프트맥스 함수를 사용하여 가중합을 구하고 그 값을 디코더에 전달한다.

집중(attention)해서 보아야 할 벡터를 소프트맥스 함수로 점수를 매긴 후 각각을 은닉 상태의 벡터들과 곱합니다. 그리고 이 은닉 상태를 모두 더해서 하나의 값으로 만든다.







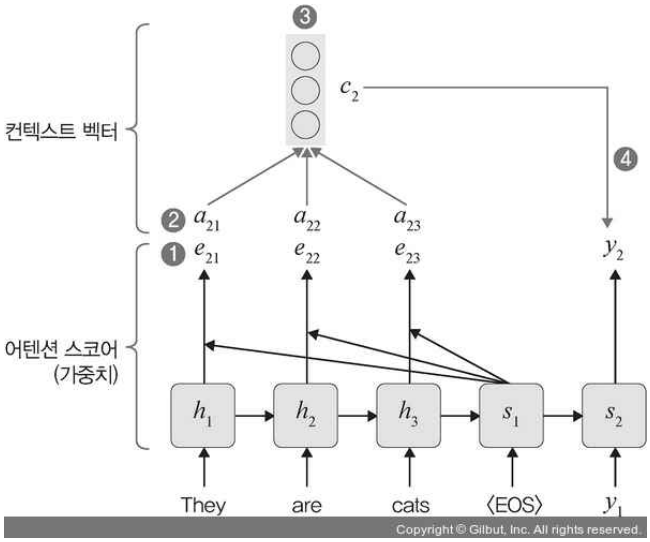


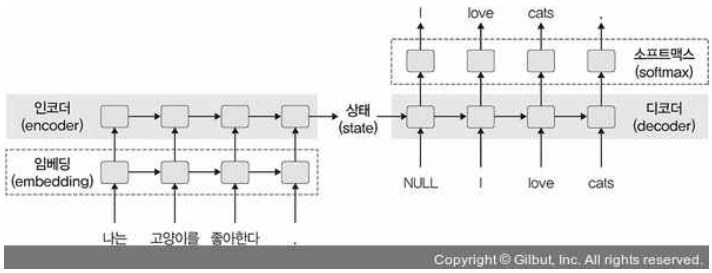
그림 14 어텐션 매커니즘

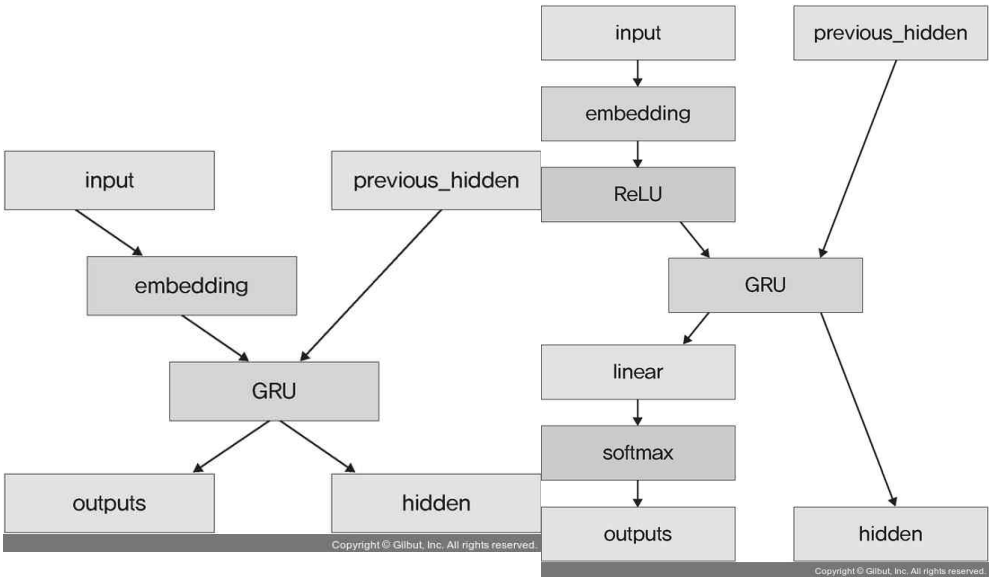
•seq2seq

입력 시퀀스(input sequence)에 대한 출력 시퀀스(output sequence)를 만들기 위한 모델

번역에 초점을 둔 모델

번역은 입력 시퀀스의 x1:n과 의미가 동일한 출력 시퀀스 y1:m을 만드는 것이며, xi, yi 간의 관계는 중요하지 않다. (각 시퀀스 길이도 다를 수 있다)





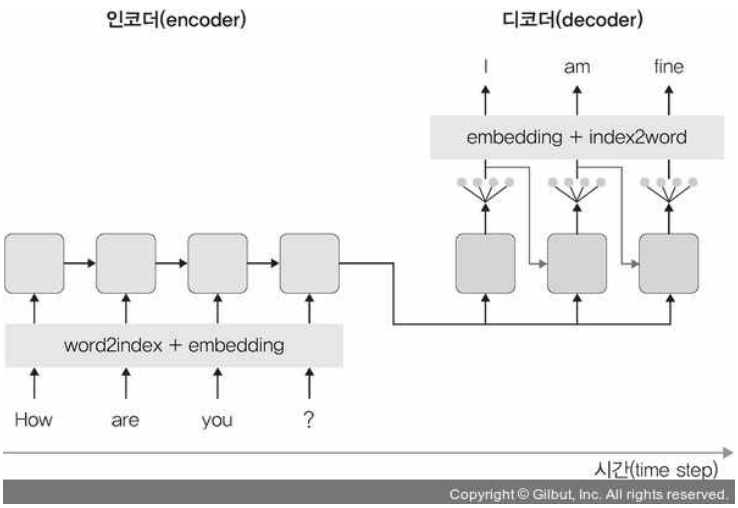
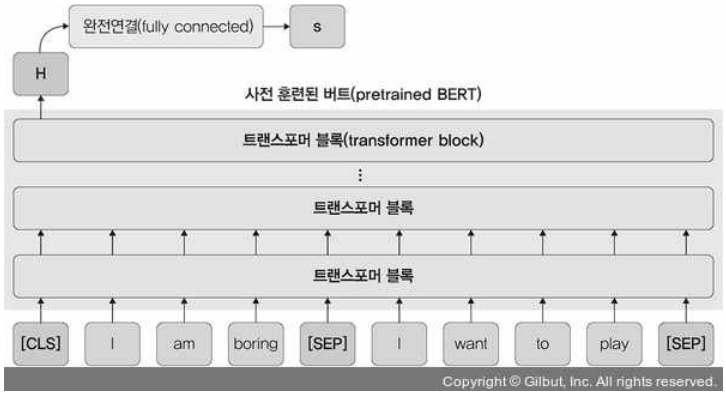


그림 18 seq2seq 네트워크

• BERT

방향 자연어 처리 모델

검색 문장의 단어를 입력된 순서대로 하나씩 처리하는 것이 아니라, 트랜스포머를 이용하여 구현되었으며 방대한 양의 텍스트 데이터로 사전 훈련된 언어 모델이다.


완전연결(fully connected)
s
H
사전 훈련된 버트(pretrained BERT)
트랜스포머 블록(transformer block)
트랜스포머 블록
트랜스포머 블록
[CLS]
I
am
boring
[SEP]
I
want
to
play
[SEP]
Copyright © Gilbut, Inc. All rights reserved.

3. 한국어 임베딩