10.3장 한국어 임베딩

작성일시	@2025년 1월 6일 오후 11:50
주제	

버트 토크나이저로 쪼갠 단어들이 정확하지 않다. → 버트 토크나이저가 단어의 가장 작은 조각을 기준으로 쪼개도록 설계되었기 때문이다. 즉, 과하다 싶을 정도로 쪼개져 있다. → 한국어 토크나이저(예: KoBERT) 사용

버트는 문장을 구별하기 위해 1과 0을 사용한다. 문장이 바뀔 때마다 0에서 1로 바뀐다.

예) [0, 0, 1, 1, 1, 0, 0, 0]: 3개의 문장으로 구성되어 있다.

```
model=BertModel.from_pretrained('bert-base-multilingual-cased
model.eval()
```

- `from_pretrained`: 인터넷에서 사전 훈련된 모델을 내려받는다.
  - `'bert-base-multilingual-cased'`: 12개의 계층으로 구성된 심층 신경망. 영어 외에 다국어에 대한 임베딩 처리를 할 때 사용하는 모델
  - `output_hidden_states`: 버트 모델에서 은닉 상태의 값을 가져오기 위해 사용

```
계층 개수: 13   (initial embeddings + 12 BERT layers)
배치 개수: 1
토큰 개수: 33
은닉층의 유닛 개수: 768
```

```
token_embeddings=token_embeddings.permute(1, 0, 2)
token_embeddings.size()
```

- `permute`: transpose와 유사하게 차원을 맞교환할 때 사용한다.

다국어 버트 모델을 사용하더라도 한국어에 대해서는 정확한 판별이 어려운 것을 확인할 수 있다. 또, 사과라는 단어가 한 번 더 쪼개져 있기 때문에 정확한 결과라고 하기 어렵다.