벡터가 어떻게 의미를 가지게 되는가

한국어임베딩 57p~69p

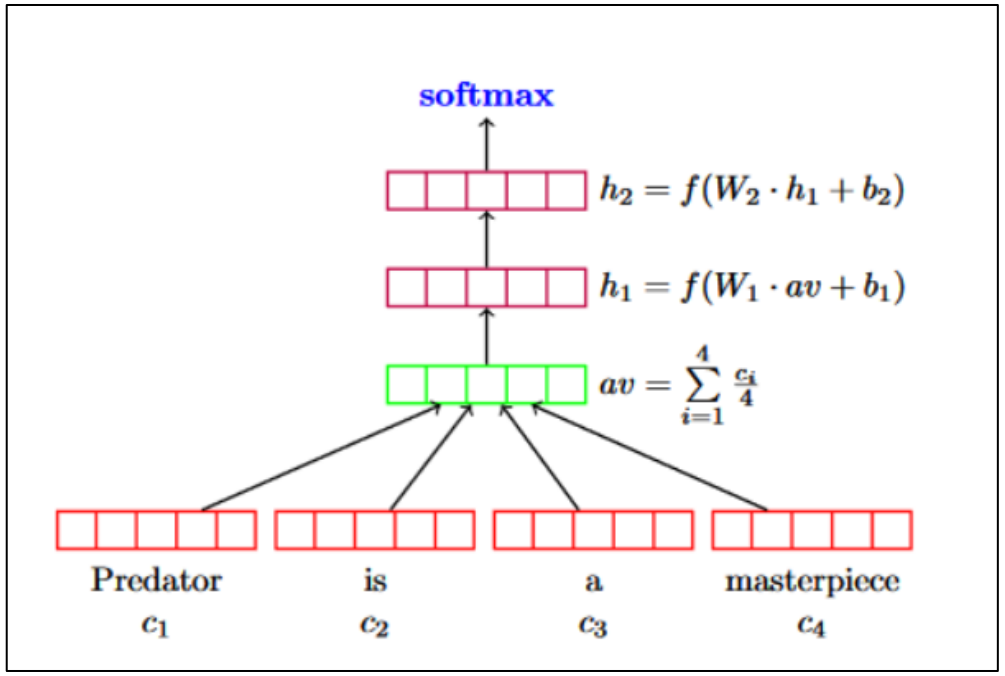
NLP2팀 황인택

차 례

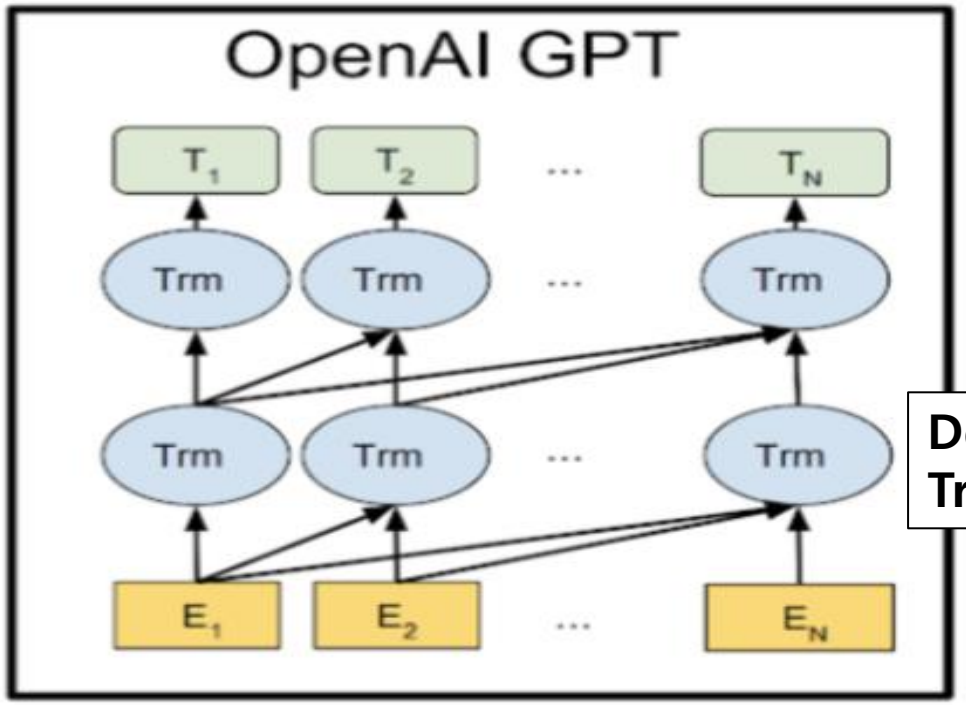
2.1 자연어 계산과 이해

구분	백오브워즈 가정	언어 모델	분포 가정
내용	어떤 단어가 (많이) 쓰였는가	단어가 어떤 순서로 쓰였는가	어떤 단어와 같이 쓰였는가
대표 통계량	TF-IDF	-	PMI
대표 모델	Deep Averaging Network	ELMo, GPT	Word2Vec
특징	순서 무시, 빈도 중시	순서, 문맥 중시	문맥 중시

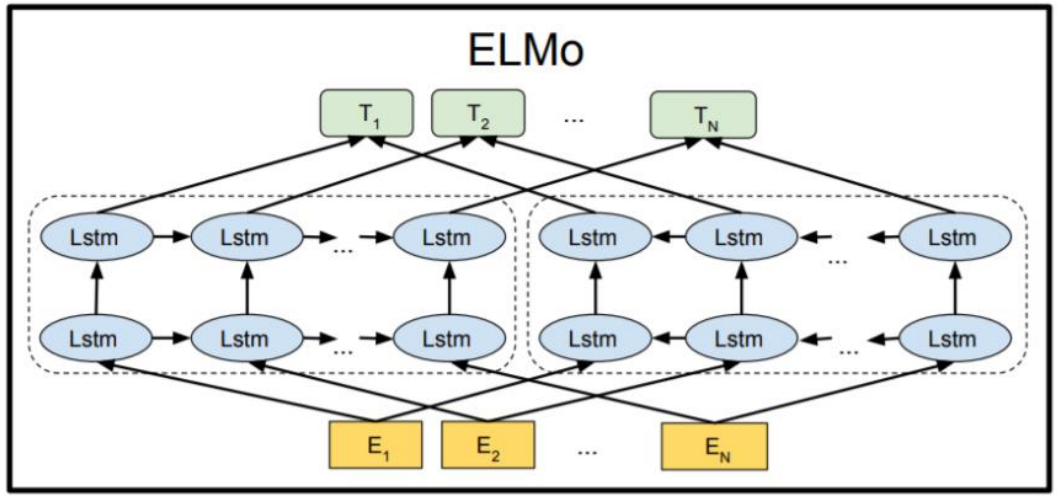
TF-IDF: **단어 빈도-역 문서 빈도, Term Frequency-Inverse Document Frequency**
PMI: **점별 상호 정보량, Pointwise mutual information**
ELMo: **Embeddings from Language Model**
GPT: **Generative Pre-Training**

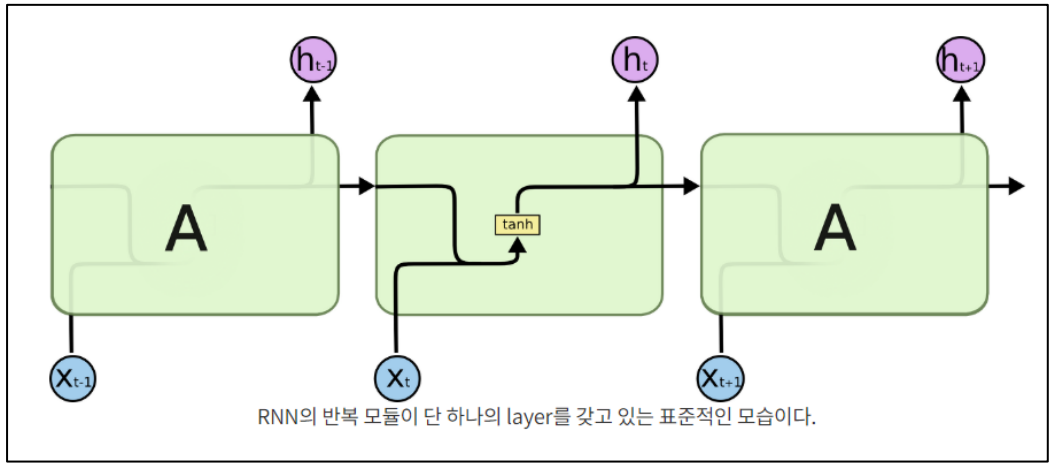


Deep Averaging Network(DAN)



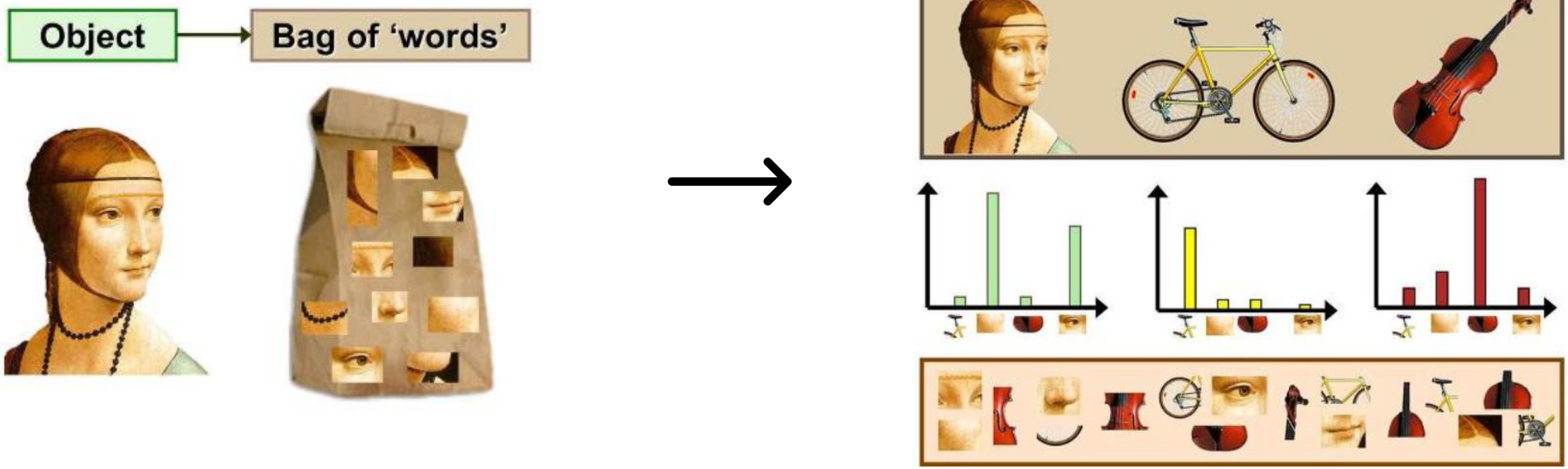
Decoding Transformer





Long Short-Term Memory

2.2 어떤 단어가 많이 쓰였는가
2.2.1 백오브워즈 가정



2.2 어떤 단어가 많이 쓰였는가
2.2.2 TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

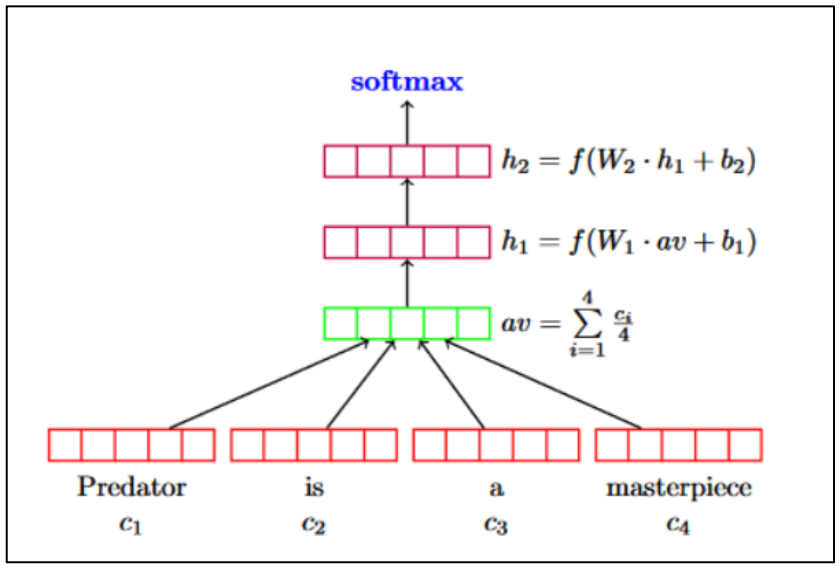
N = total number of documents

term	df	idf
calpurnia	1	6
animal	100	4
sunday	1,000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

최소 빈도를 나타내는 'Calpurnia'가 최대 빈도를 나타내는 'the'보다 더 중요함.

여전히 count 기반의 vectorizer라는 점에서 한계를 지님.

2.2 어떤 단어가 많이 쓰였는가
2.2.3 Deep Averaging Network



Deep unordered model that obtains near state of art accuracy on sentence and document level tasks with very less training time works in three steps:

(a) take the vector average of the embeddings associated with an input sequence of tokens

(b) pass that average through one or more feed-forward layer

(c) perform (linear) classification on the final layers representation

(d) Loss function is cross entropy.

2.3 단어가 어떤 순서로 쓰였는가
2.3.1 통계 기반 언어 모델

최대우도추정법(maximum likelihood estimation)

모수(parameter)가 미지의 θ인 확률분포에서 뽑은 표본(관측치) x들을 바탕으로 θ를 추정하는 기법

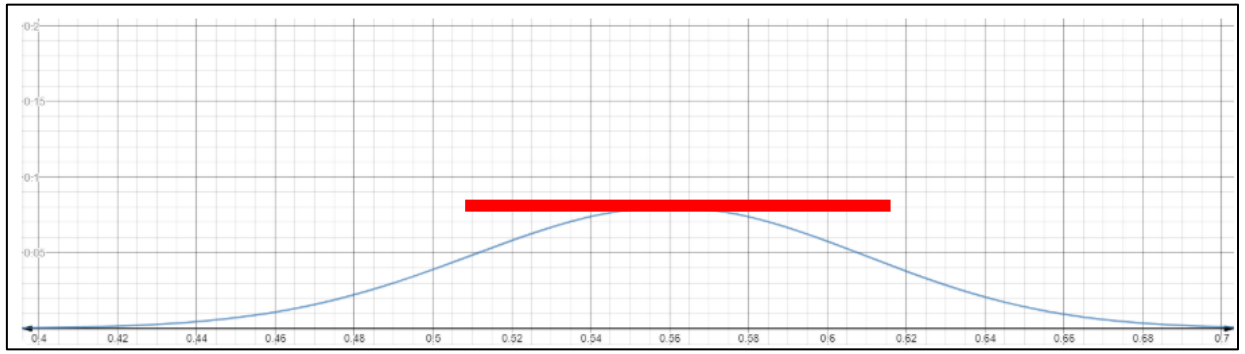
우도(likelihood): 이미 주어진 표본 xx들에 비추어 봤을 때 모집단의 모수 θ에 대한 추정이 그럴듯한 정도를 가리킵니다.

예) 동전 던지기 100번 시행에 56번 앞면일 경우, 최대우도는?

θ가 0.5로 추정할 때,

$$p(X = 56|\theta = 0.5) = \binom{100}{56} 0.5^{56} 0.5^{44} \approx 0.0389$$

θ	likelihood
0.48	0.0222
0.50	0.0389
0.52	0.0587
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378



2.3 단어가 어떤 순서로 쓰였는가
2.3.1 통계 기반 언어 모델

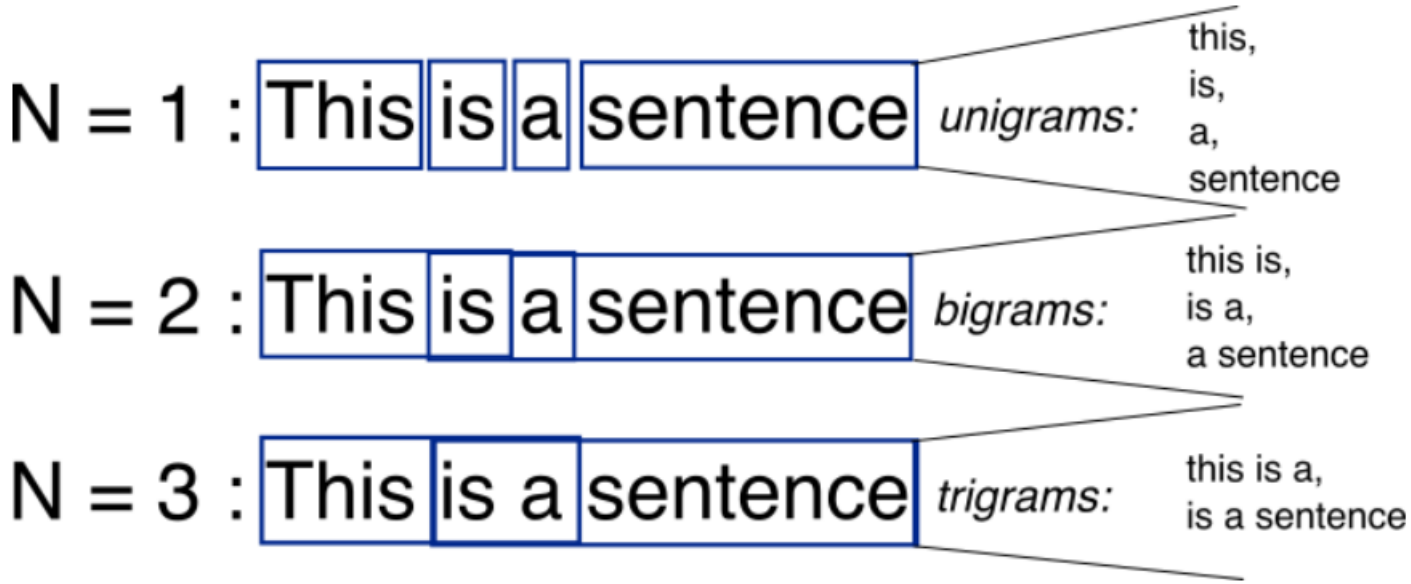
N-gram 언어 모델

~~An adorable little~~ boy is spreading ?

무시됨!

n-1개의 단어

$$P(w|\text{boy is spreading}) = \frac{\text{count}(\text{boy is spreading } w)}{\text{count}(\text{boy is spreading})}$$



2.3 단어가 어떤 순서로 쓰였는가
2.3.1 통계 기반 언어 모델

Zero count problem에 대해...

백오프 – n=3 일 때

$$\hat{P}(w_i \mid w_{i-2}w_{i-1}) = \begin{cases} P(w_i \mid w_{i-2}w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 P(w_i \mid w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) = 0 \\ & \text{and } C(w_{i-1}w_i) > 0 \\ \alpha_2 P(w_i), & \text{otherwise.} \end{cases}$$

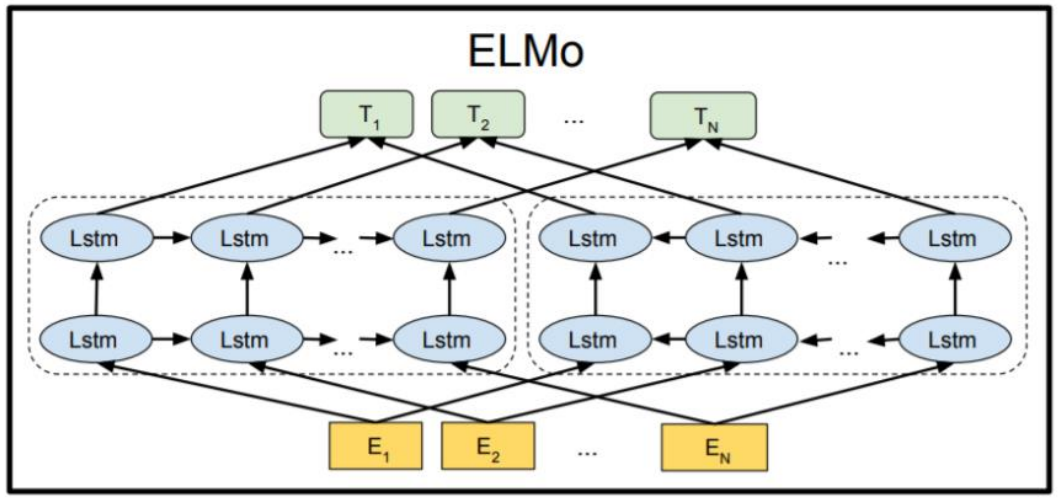
스무딩 – Laplace smoothing

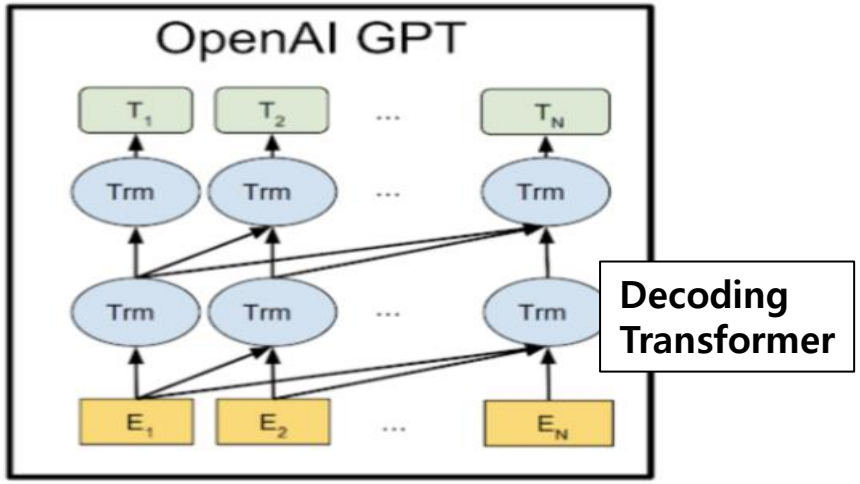
Golf dataset

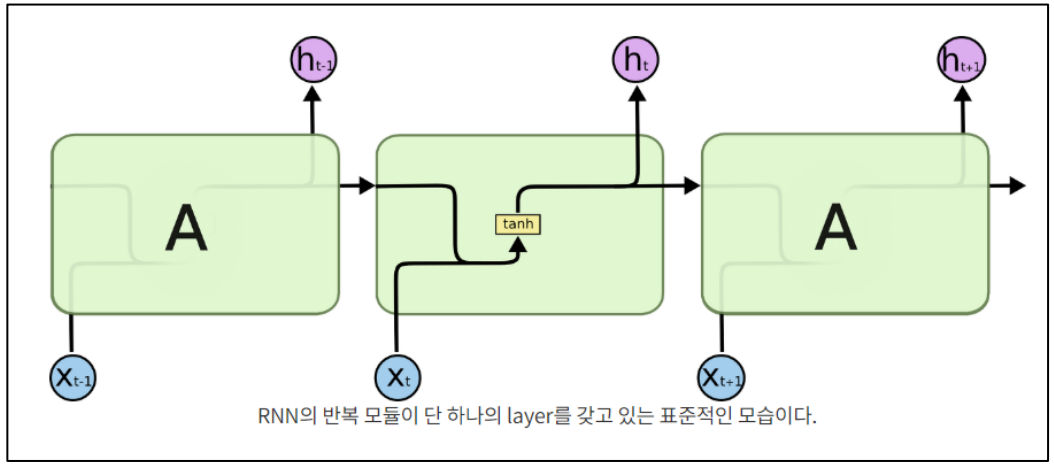
	outlook			temperature			humidity			windy	
	yes	no		yes	no		yes	no		yes	no
overcast	4/9	0/5	hot	2/9	2/5	high	3/9	4/5	TRUE	3/9	3/5
rainy	3/9	2/5	cool	3/9	1/5	normal	6/9	1/5	FALSE	6/9	2/5
sunny	2/9	3/5	mild	4/9	2/5						

	outlook			temperature			humidity			windy	
	yes	no		yes	no		yes	no		yes	no
overcast	5/12	1/8	hot	3/12	3/8	high	4/11	5/7	TRUE	4/11	4/7
rainy	4/12	3/8	cool	4/12	2/8	normal	7/11	2/7	FALSE	7/11	3/7
sunny	3/12	4/8	mild	5/12	3/8						

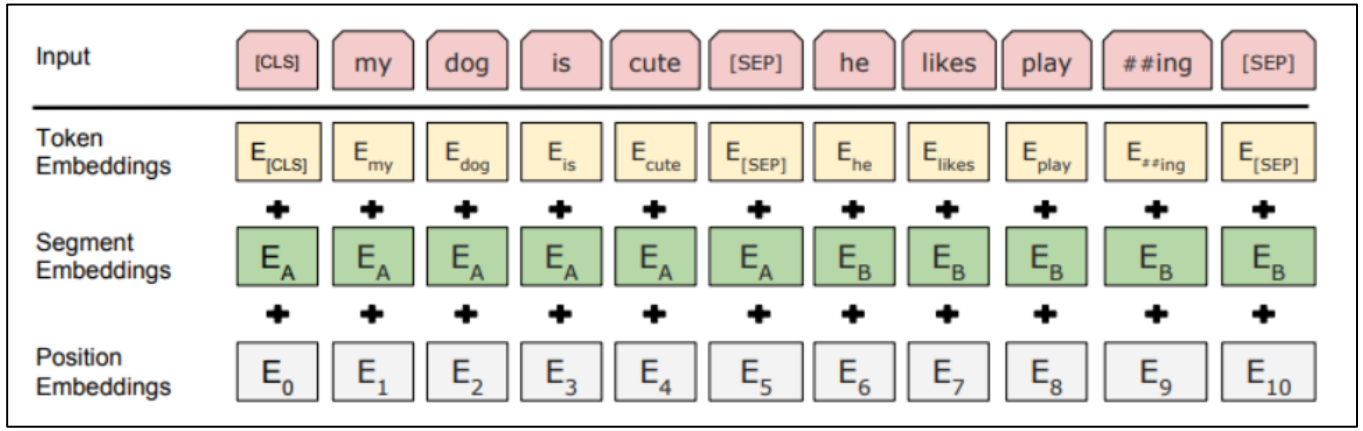
2.3 단어가 어떤 순서로 쓰였는가
2.3.2 뉴럴 네트워크 기반 언어 모델





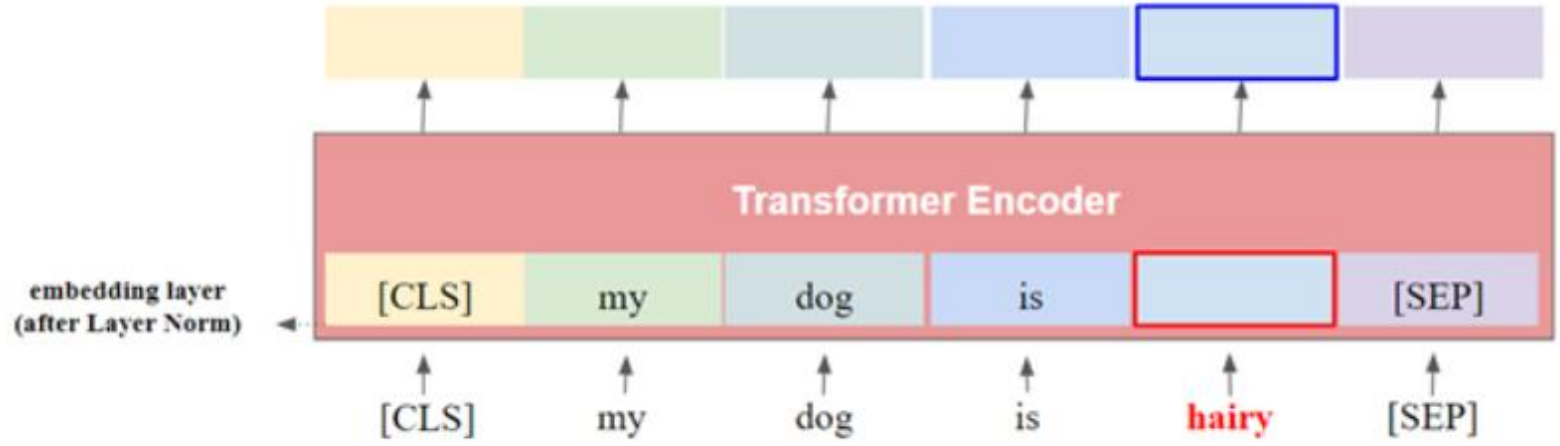


Long Short-Term Memory



BERT(Bidirectional Encoder Representations from Transformers)

2.3 단어가 어떤 순서로 쓰였는가
2.3.2 뉴럴 네트워크 기반 언어 모델



Mask **15%** of all WordPiece tokens in each sequence at **random**. (e.g., **hairy**)

[MASK] 80% of the time : Replace **[MASK]** token.

apple 10% of the time : Replace the word with a **random** word

hairy 10% of the time : Keep the word **unchanged**.

MLM MASK 선정 방식