

2장. 벡터가 어떻게 의미를 가지는가

2.1 자연어 계산과 이해

- 임베딩을 만드는 세 가지 철학

구분	백 오브 워즈 가정	언어 모델	분포 가정
내용	어떤 단어가 (많이) 쓰였는가	단어가 어떤 순서로 쓰였는가	어떤 단어가 같이 쓰였는가
대표 통계량	TF-IDF	-	PMI
대표 모델	Deep Averaging Network	ELMo, GPT	Word2Vec

2.2 어떤 단어가 많이 쓰였는가

2.2.1 백오브워즈 임베딩

“저자가 생각한 주제 → 단어 빈도”

별 하나에 추억과
별 하나에 사랑과
별 하나에 쓸쓸함과
별 하나에 동경과
별 하나에 시와
별 하나에 어머니, 어머니



시 하나 동경 별에
별과 하나에 사랑
추억과 하나별
에, 하나어머니하나함
별와에별쓸쓸
어머니에쓸쓸하나별
하나과하나과



별	하나	에	추억	과	사랑	쓸쓸	함	동경	시	와	어머니	,
6	6	6	1	4	1	1	1	1	1	1	2	1

2.2 어떤 단어가 많이 쓰였는가

2.2.2 TF-IDF (Term Frequency - Inverse Document Frequency)

“여러 문서에 많이 등장하는 단어는 쓸모 없는 단어다”

$$TF - IDF(w) = TF(w) \times \log\left(\frac{N}{DF(w)}\right)$$

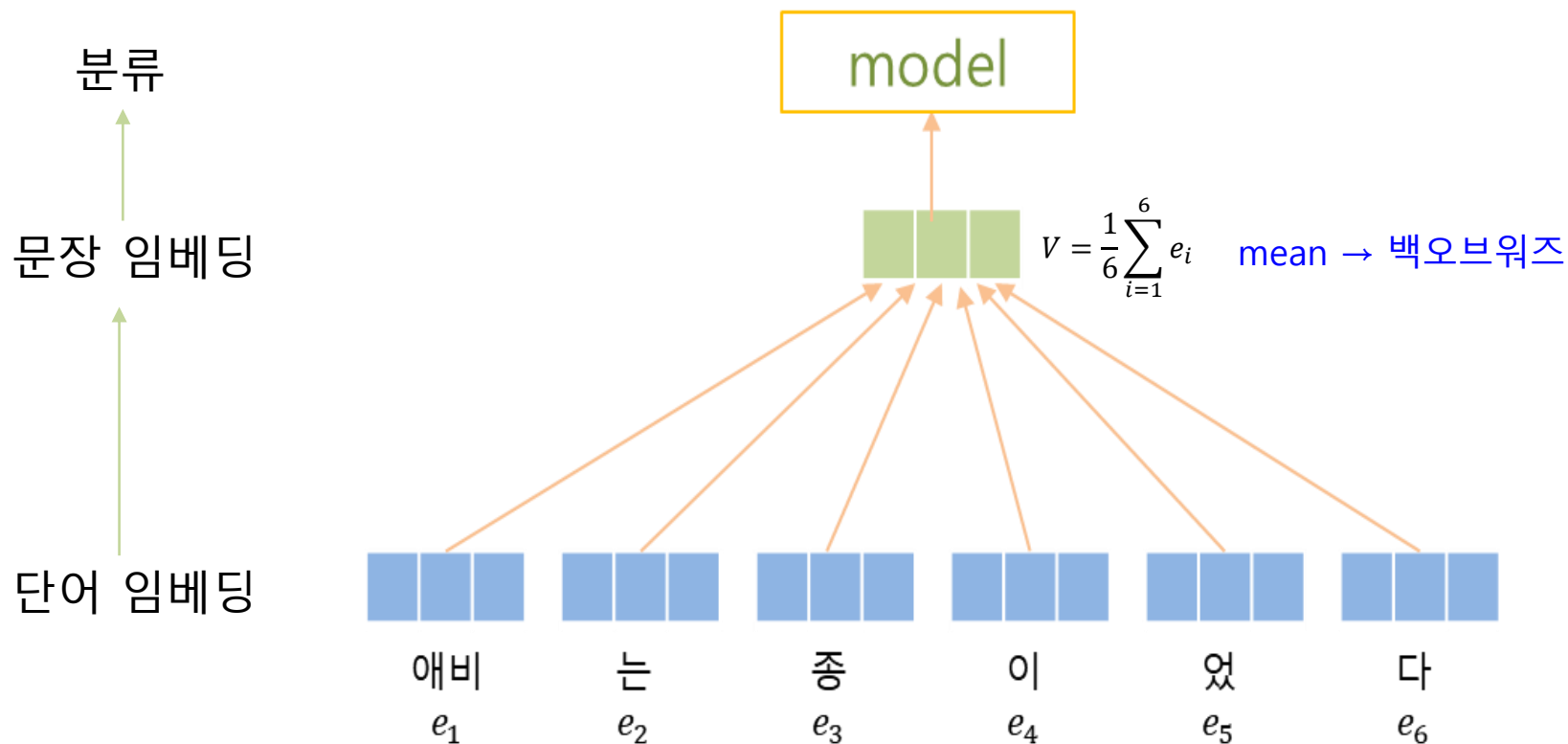
N : 전체문서수

ex) TF-IDF 행렬

구분	메밀꽃 필 무렵	운수 좋은 날	사랑 손님과 어머니	삼포 가는 길
담배	0.2603	0.2875	0.0364	0.2932
를	0.0	0.0034	0.0	0.0

2.2 어떤 단어가 많이 쓰였는가

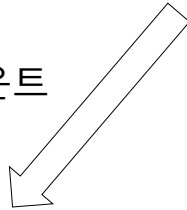
2.2.3 Deep Averaging Network



2.3 단어가 어떤 순서로 쓰였는가

언어 모델
(단어 시퀀스에 확률 부여)

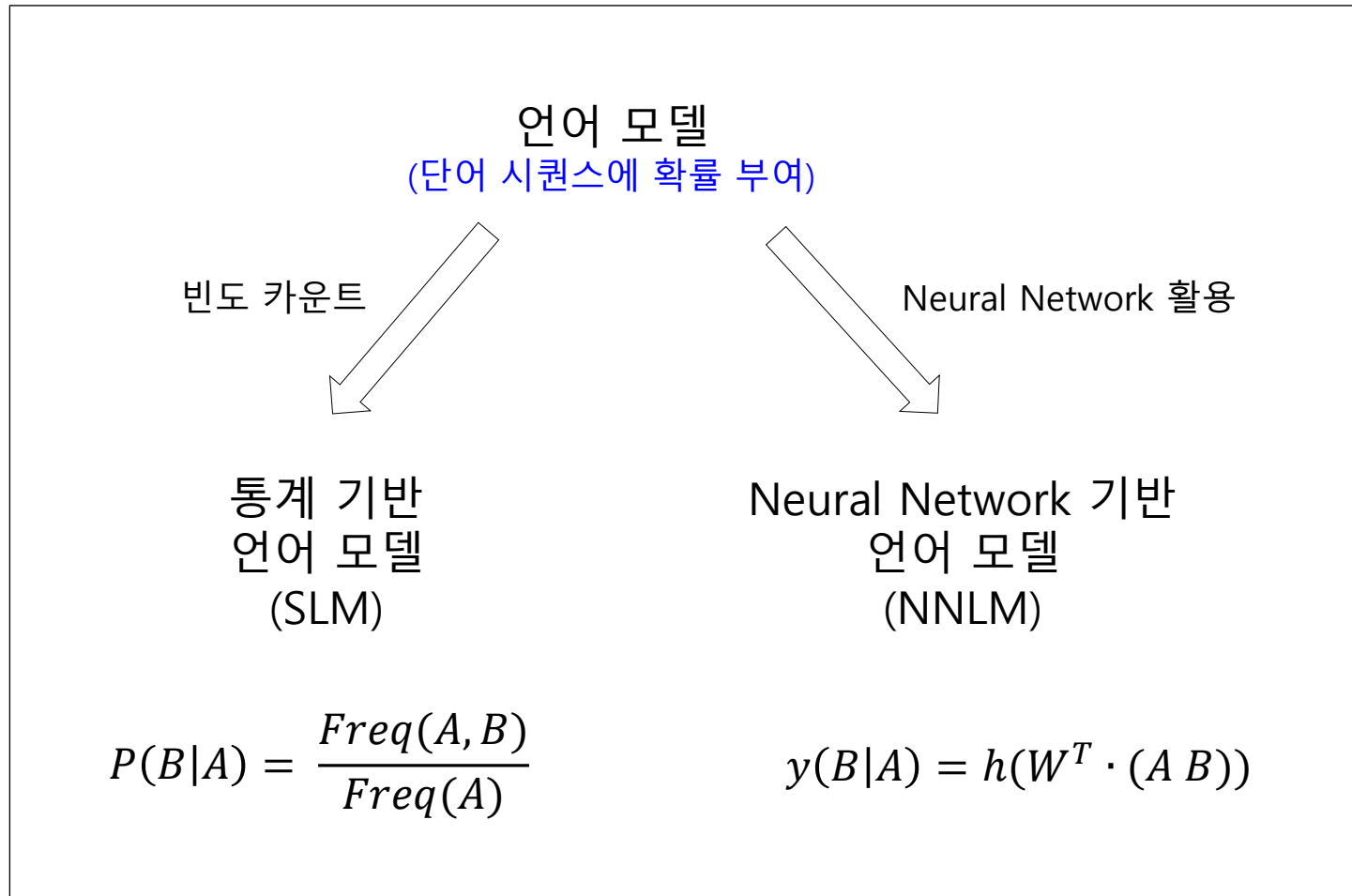
빈도 카운트



통계 기반
언어 모델
(SLM)

$$P(B|A) = \frac{Freq(A, B)}{Freq(A)}$$

2.3 단어가 어떤 순서로 쓰였는가



2.3 단어가 어떤 순서로 쓰였는가

2.3.1 통계 기반 언어 모델

- 단어가 n 개 주어졌을 때 언어 모델은 주어진 단어가 자연스러울 확률을 $P(w_1, \dots, w_n)$ 로 반환함
- 이 때, 확률은 해당 단어 시퀀스의 빈도로 계산한다.

$$P(w_1, \dots, w_n) = \frac{\text{Freq}(w_1, \dots, w_n)}{\text{Freq}(\text{All word sequences})}$$

ex) 네이버 영화 말뭉치의 각 표현별 등장 횟수

표현	빈도
내	1309
마음	172
⋮	⋮
영원히 기억될 최고의 명작이다	1
내 마음 속에 영원히 기억될 최고의 명작이다	0

2.3 단어가 어떤 순서로 쓰였는가

2.3.1 통계 기반 언어 모델

- 말뭉치에 "내 마음 속에 영원히 기억될 최고의 명작이다" 라는 문장이 없다면..?


$$\begin{aligned} &P(\text{명작이다} | \text{내 마음 속에 영원히 기억될 최고의}) \\ &= \frac{\text{Freq}(\text{내 마음 속에 영원히 기억될 최고의 명작이다})}{\text{Freq}(\text{내 마음 속에 영원히 기억될 최고의})} \\ &= \mathbf{0} \end{aligned}$$

2.3 단어가 어떤 순서로 쓰였는가

2.3.1 통계 기반 언어 모델

- 말뭉치에 "내 마음 속에 영원히 기억될 최고의 명작이다" 라는 문장이 없다면..?

$$\begin{aligned} &P(\text{명작이다} | \text{내 마음 속에 영원히 기억될 최고의}) \\ &= \frac{\text{Freq}(\text{내 마음 속에 영원히 기억될 최고의 명작이다})}{\text{Freq}(\text{내 마음 속에 영원히 기억될 최고의})} \\ &= \mathbf{0} \end{aligned}$$

n-gram model  n-gram : n개의 단어 ex) (누명,을,쓰다) → 3-gram

$$\begin{aligned} &P(\text{명작이다} | \text{내 마음 속에 영원히 기억될 최고의}) \\ &\approx P(\text{명작이다} | \text{최고의}) \\ &= \frac{\text{Freq}(\text{최고의 명작이다})}{\text{Freq}(\text{최고의})} = \frac{23}{3503} \end{aligned}$$

마코프 가정
ex) bigram model

2.3 단어가 어떤 순서로 쓰였는가

2.3.1 통계 기반 언어 모델

- 말뭉치에 여전히 존재하지 않는 단어가 있다면..?

$$\begin{aligned} &P(\text{그 아이는 또바기 인사를 잘한다}) \\ &= P(\text{그}) \times P(\text{아이는}|\text{그}) \times \frac{P(\text{또바기}|\text{아이는})}{P(\text{인사를}|\text{또바기}) \times P(\text{잘한다}|\text{인사를})} = 0 \\ &= 0 \end{aligned}$$

2.3 단어가 어떤 순서로 쓰였는가

2.3.1 통계 기반 언어 모델

- 말뭉치에 여전히 존재하지 않는 단어가 있다면..?

$$\begin{aligned} &P(\text{그 아이는 또바기 인사를 잘한다}) \\ &= P(\text{그}) \times P(\text{아이는}|\text{그}) \times \frac{P(\text{또바기}|\text{아이는})}{P(\text{인사를}|\text{또바기}) \times P(\text{잘한다}|\text{인사를})} = 0 \\ &= 0 \end{aligned}$$

⇒ ① 백오프 : n-gram 빈도를 n보다 작은 단어시퀀스의 빈도로 근사

$$\begin{aligned} &Freq(\text{내 마음 속에 영원히 기억될 최고의 명작이다}) \\ &\approx \alpha Freq(\text{영원히 기억될 최고의 명작이다}) + \beta \end{aligned}$$

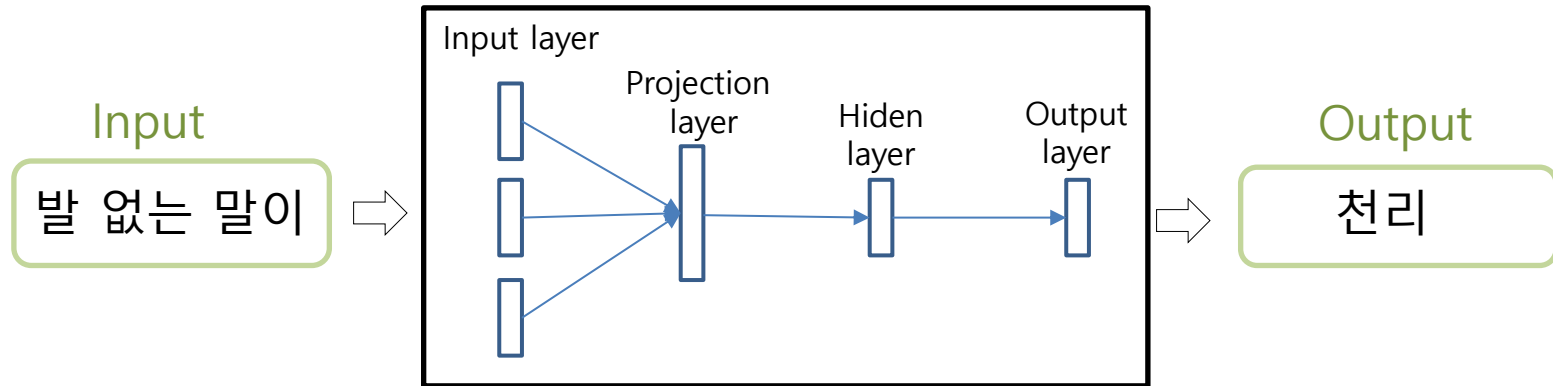
② 스무딩 : 모든 단어의 빈도를 k 만큼 더함

$$\begin{aligned} &Freq(\text{내 마음 속에 영원히 기억될 최고의 명작이다}) = 0 \\ &\quad \downarrow \\ &Freq(\text{내 마음 속에 영원히 기억될 최고의 명작이다}) = k \end{aligned}$$

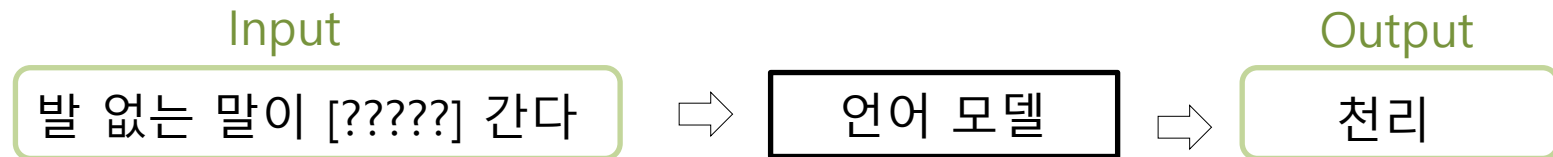
2.3 단어가 어떤 순서로 쓰였는가

2.3.2 Neural Network 기반 언어 모델

① Network model (ELMo, GPT)



② Masked language model (BERT)



2.4 어떤 단어가 같이 쓰였는가

2.4.1 분포 가정

- 분포 가정의 전제 :

어떤 단어 쌍이 비슷한 문맥 환경에서 자주 등장한다면, 그 의미 또한 유사할 것이다.

... 특기 는 자칭 청소와 빨래 지만 요리 는 절망 적 ...
... 재 를 우려낸 물 로 빨래 할 때 나 ...
... 개울가 에서 속옷 빨래 를 하 는 남녀 ...



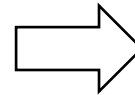
빨래 - 청소, 요리, 물, 속옷

Target - Context

세탁 - 청소, 요리, 물, 옷



... 찬 물 로 옷 을 세탁 한다 ...
... 세탁, 청소, 요리 와 가사 는 ...

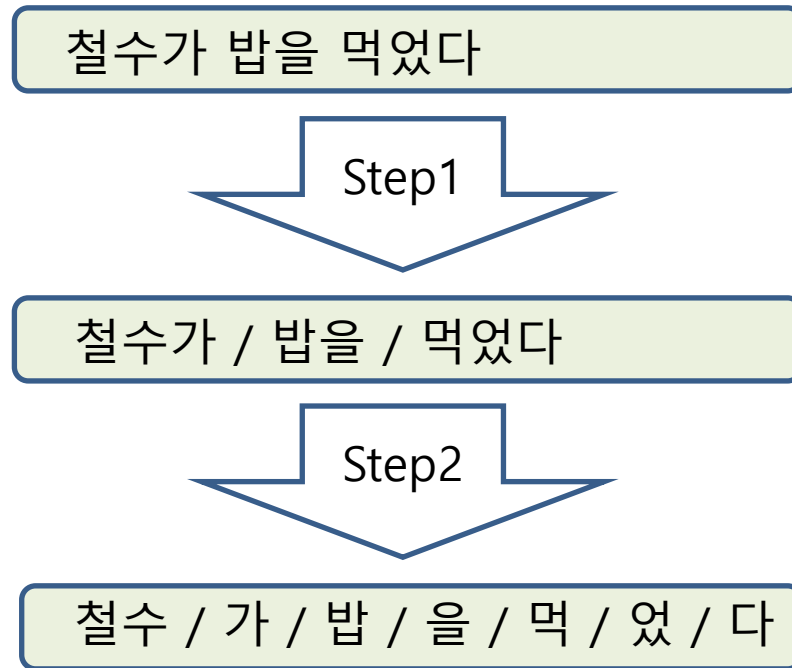


빨래 \approx **세탁**

2.4 어떤 단어가 같이 쓰였는가

2.4.2 분포와 의미(1) : 형태소

- 형태소 : 의미를 가지는 최소 단위 (더 이상 쪼갤 수 없는 최소 의미 단위)

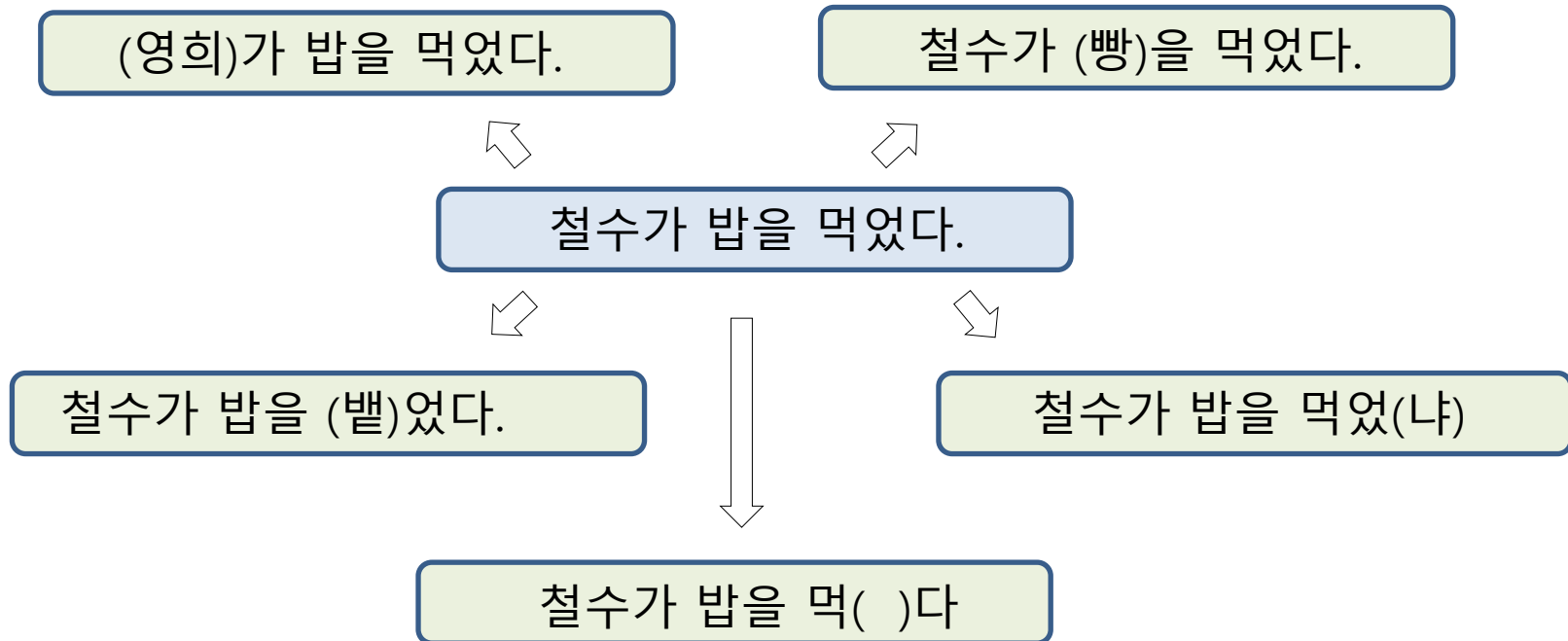


일반인st.

2.4 어떤 단어가 같이 쓰였는가

2.4.2 분포와 의미(1) : 형태소

- 형태소 : 의미를 가지는 최소 단위 (더 이상 쪼갤 수 없는 최소 의미 단위)



2.4 어떤 단어가 같이 쓰였는가

2.4.3 분포와 의미(2) : 품사

- 품사 : 단어를 문법적 성질의 공통성에 따라 묶어 놓은 것
- 세가지 품사 분류 기준

① (형식적) 의미 : 사물의 이름을 나타내느냐 (ex. 깊이),
움직임이나 성질 또는 상태를 나타내느냐 (ex. 깊다)

공부하다 vs 공부 “공부”는 움직임이 아닐까?



2.4 어떤 단어가 같이 쓰였는가

2.4.3 분포와 의미(2) : 품사

- 품사 : 단어를 문법적 성질의 공통성에 따라 묶어 놓은 것
- 세가지 품사 분류 기준

① (형식적) 의미 : 사물의 이름을 나타내느냐 (ex. 깊이),
움직임이나 성질 또는 상태를 나타내느냐 (ex. 깊다)

공부하다 vs 공부 “공부”는 움직임이 아닐까?



② 형태 : (a) 영수가 학교에 간다. “영수” → 명사
(b) 영수! 조용히 해. “영수” → 감탄사



2.4 어떤 단어가 같이 쓰였는가

2.4.3 분포와 의미(2) : 품사

- 품사 : 단어를 문법적 성질의 공통성에 따라 묶어 놓은 것
- 세가지 품사 분류 기준

① (형식적) 의미 : 사물의 이름을 나타내느냐 (ex. 깊이),
움직임이나 성질 또는 상태를 나타내느냐 (ex. 깊다)

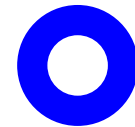


공부하다 vs 공부 “공부”는 움직임이 아닐까?

② 형태 : (a) 영수가 학교에 간다. “영수” → 명사
(b) 영수! 조용히 해. “영수” → 감탄사



③ 기능 : 단어가 문장 내에서 점하는 역할 ⇔ 단어의 분포



체언(명사) : 관형사가 그 앞에 올 수 있고 조사가 그 뒤에 올 수 있음
용언(동사/형용사) : 부사가 그 앞에 올 수 있고 선어말어미가 그 뒤에 올 수 있고 어말어미가 그 뒤에 와야 함
관형사 : 명사가 그 뒤에 와야 함
부사 : 용언, 부사, 절이 그 뒤에 와야 함
조사 : 체언 뒤에 와야 함
어미 : 용언 뒤에 와야 함
감탄사(간투사) : 특별한 결합 제약 없이 즉, 문장 내의 다른 단어와 문법적 관계를 맺지 않고 따로 존재함

2.4 어떤 단어가 같이 쓰였는가

2.4.4 점별 상호 정보량(Pointwise Mutual Information)

- PMI : 두 단어의 등장 빈도가 독립일 때 대비해 얼마나 자주 같이 등장하는지를 수치화

$$PMI(A, B) = \log \frac{P(A, B)}{P(A) \times P(B)}$$

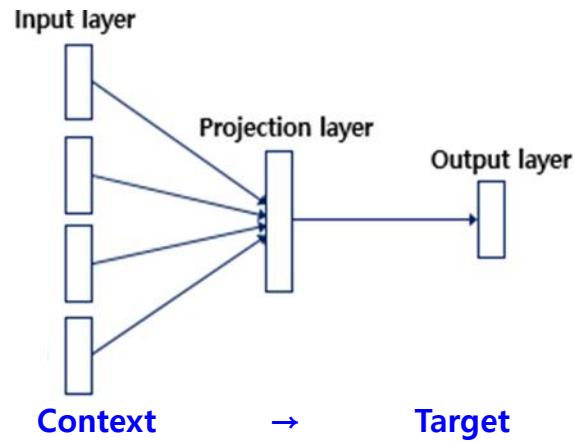
window = 2 인 단어-문맥 행렬

개울가, 에서, 속옷, 빨래, 를, 하는, 남녀								
문맥 \ 단어	개울가	에서	속옷	빨래	를	하는	남녀	total
개울가								
⋮								
빨래		+1	+1		+1	+1		20
⋮								
total			15					1000

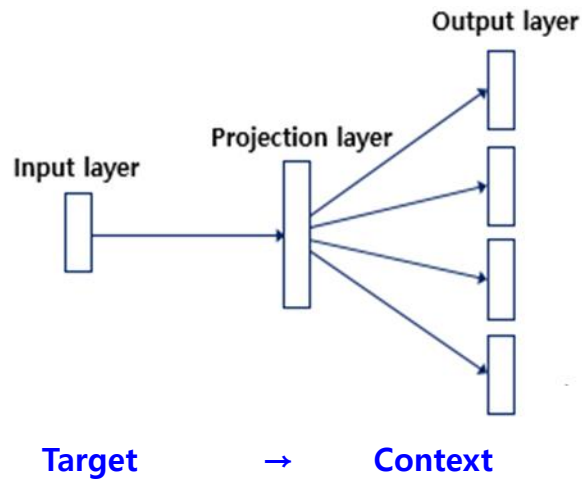
2.4 어떤 단어가 같이 쓰였는가

2.4.5 Word2Vec

CBOW



Skip-gram



2.5 Summary

- ① 백오브워즈 가정은 단어의 **빈도 정보**를 중시하고 순서는 무시한다.
 - ② 언어모델은 **단어의 순서를 학습**해 주어진 단어 시퀀스가 얼마나 자연스러운지 **확률을 부여**한다.
 - ③ 분포가정에서는 문장에서 **어떤 단어가 같이 쓰였는지를** 중시한다.
 - ④ 백오브워즈 가정, 언어모델, 분포 가정은 말뭉치의 통계적 패턴을 다른 각도에서 분석하는 방법으로 **상호 보완적**이다.
-

끝
