

2021학년도 2학기 언어와 컴퓨터

제16강 N-그램 언어 모형 (2)

박수지

서울대학교 인문대학 언어학과

2021년 11월 8일 월요일

오늘의 목표

- 1 주어진 코퍼스로부터 N-그램 언어 모델을 훈련시킬 수 있다.
- 2 N-그램 언어 모델로부터 문장을 생성할 수 있다.

표기법 주의

SLP3 3장에서 N은 세 가지 다른 의미로 사용된다.

- N-그램 (3.1절)
- 실험 집합 $W = w_1 w_2 \cdots w_N$ (3.2절)
 - 강의 자료에서는 K로 표기: $PP(W) = P(w_1 w_2 \dots w_K)^{-\frac{1}{K}}$
- 코퍼스의 크기 N (3.4절)
 - 강의 자료에서는 M으로 표기

예시: 훈련 코퍼스

```
i 'd like to eat dinner .  
show me the list again .  
i like to get a hamburger .  
i 'd like to go to a japanese restaurant .
```

예시: 바이그램 모형 훈련

$$P(i|<s>) = \frac{C(<s> i)}{C(<s>)} = \frac{3}{4}, \quad P('d|<s>) = \frac{C(<s> 'd)}{C(<s>)} = \frac{0}{4}, \quad \dots$$

<s> 문장 시작 기호

예시: 훈련 코퍼스

i 'd like to eat dinner .
show me the list again .
i like to get a hamburger .
i 'd like to go to a japanese restaurant .

예시: 바이그램 모형 훈련

$$P('d|i) = \frac{C(i \text{ 'd})}{C(i)} = \frac{2}{3}, \quad P(\text{like}|i) = \frac{C(i \text{ like})}{C(i)} = \frac{1}{3}, \quad \dots$$

예시: 훈련 코퍼스

```
i 'd like to eat dinner .  
show me the list again .  
i like to get a hamburger .  
i 'd like to go to a japanese restaurant .
```

예시: 바이그램 모형 훈련

$$P(</s>|.)=\frac{C(. </s>)}{C(.)}=\frac{4}{4}, P(</s>|i)=\frac{C(i </s>)}{C(i)}=\frac{0}{3}, \dots$$

</s> 문장 끝 기호

예시: 바이그램(N = 2) 모형의 조건부확률

	i	'd	like	to	eat	dinner	.	</s>	show	...
<s>	0.75	0	0	0	0	0	0	0	0.25	...
i	0	0.67	0.33	0	0	0	0	0	0	...
'd	0	0	1	0	0	0	0	0	0	...
like	0	0	0	1	0	0	0	0	0	...
to	0	0	0	0	<u>0.25</u>	0	0	0	0	...
...										

$$P(\text{eat}|\text{to}) = \frac{C(\text{to eat})}{C(\text{to})} = \frac{1}{4}$$

예시: 트라이그램(N = 3) 모형의 조건부확률

	i	'd	like	to	eat	dinner	.	</s>	show	...
<s> <s>	0.75	0	0	0	0	0	0	0	0.25	...
<s> i	0	0.67	0.33	0	0	0	0	0	0	...
i 'd	0	0	1	0	0	0	0	0	0	...
'd like	0	0	0	1	0	0	0	0	0	...
like to	0	0	0	0	<u>0.33</u>	0	0	0	0	...
...										

$$P(\text{eat}|\text{like to}) = \frac{C(\text{like to eat})}{C(\text{like to})} = \frac{1}{3}$$

모형 훈련

- 0 N-그램의 N을 정한다.
- 1 훈련 코퍼스의 각 문장 앞에 문장 시작 기호 $\langle s \rangle$ 를 $(N - 1)$ 개만큼 채워 넣는다.
- 2 훈련 코퍼스에 등장한 N-그램을 모두 센다.
- 3 $h = w_1 w_2 \cdots w_{N-1}$ 가 주어졌을 때 $w = w_N$ 의 조건부확률을 계산한다.

문장 생성

- 1 $P(w_1 | \langle s \rangle \cdots \langle s \rangle)$ 의 확률분포에 따라 첫 번째 단어 w_1 을 추출한다.
 - 예시 코퍼스에서는 0.75의 확률로 i, 0.25의 확률로 show를 추출한다.
- 2 $P(w_n | w_{n-(N-1)} \cdots w_{n-1})$ 의 확률분포에 따라 w_n 을 추출한다.
- 3 추출된 w_n 이 문장 끝 기호 $\langle /s \rangle$ 이면 종료한다.

언어 모형 훈련용 코퍼스

- BeRP (Berkely Restaurant Project) corpus — SLP3 3장의 예시 코퍼스
 - <https://web.stanford.edu/~jurafsky/icslp-red.pdf>
 - <https://github.com/wooters/berp-trans>
- Brown Corpus — 최초의 대규모 영어 코퍼스
 - 파이썬 nltk 모듈에서 사용 가능

오늘 배운 것

- 훈련 코퍼스에서 N-그램 확률 구하기
- N-그램 확률에 따라 문장 생성하기

남은 문제

훈련 집합에 없는 N-그램이 실험 집합에 나타난 경우 0보다 큰 확률값을 부여하기

다음 시간에 배울 것

- 1 라플라스 평탄화(Laplace smoothing)
- 2 보간법(Interpolation)
- 3 N-그램 언어 모형의 한계