2021학년도 2학기 언어와 컴퓨터

제15강 N-그램 언어 모형 (1)

박수지

서울대학교 인문대학 언어학과

2021년 11월 3일 수요일

오늘의 목표

1. N-그램의 개념을 설명할 수 있다.
2. N-그램을 사용하는 이유를 설명할 수 있다.
3. N-그램의 확률을 추정할 수 있다.
4. 언어 모형의 성능을 복잡도라는 척도로 평가할 수 있다.

단어 예측

작업

현재까지 나온 단어를 보고 다음에 어떤 단어가 나올지 예측하기

- 예측하기: 확률이 가장 높은 단어를 선택하기
- 현재까지 나온 단어: 단어가 출현할 조건

$\Rightarrow$ 단어 연쇄에 (조건부)확률을 할당하기

예시

산에 _____

- 놀라
- 올라 $\leftarrow$ 아무래도 이쪽의 확률이 더 높을 것이다.

단어 예측

활용

음성 인식, 필기 인식, 철자 교정, 기계 번역, 보완·대체의사소통

- 여러 후보 중에서 실제로 나올 확률이 가장 높은 것을 선택한다.

예시: 기계 번역

I have no **way** of knowing

1. 알 **길**이 없다
2. 알 **도로**가 없다

단어의 확률을 어떻게 알 수 있는가?

언어 모형

단어 연쇄(ex. 문장)에 확률을 할당하는 모형

- N-그램 언어 모형
- 신경망 언어 모형
- …

N-그램

N 개 단어의 연쇄

- 단어, 형태소, 품사, 문자, …

단어의 연쇄의 확률

$P(w_1 w_2 \cdots w_n)$를 어떻게 알아내는가?

조건부확률과 연쇄법칙

$$P(w|h) = \frac{P(h,w)}{P(h)} \Rightarrow P(h,w) = P(h)P(w|h)$$

길이 3인 예시

$$P(w_1 w_2 w_3) = P(w_1 w_2)P(w_3|w_1 w_2) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)$$

단어의 조건부확률 계산

$P(w_3|w_1, w_2)$ 등을 어떻게 알아내는가? — 상대 빈도를 계산한다.

$$P(w_n|w_1 w_2 \cdots w_{n-1}) = \frac{C(w_1 w_2 \cdots w_{n-1} w_n)}{C(w_1 w_2 \cdots w_{n-1})}$$

최대가능도추정(MLE: Maximum likelihood estimation)

예시

"알 길이 _____"에 "없다"가 출현할 확률

$$P(\text{“없다”}|\text{“알 길이”}) = \frac{C(\text{“알 길이 없다”})}{C(\text{“알 길이”})}$$

단어 연쇄의 확률 계산

(1) 조건부확률의 연쇄법칙으로 단어 연쇄의 확률 표현하기

$$
\begin{aligned}
&\mathrm{P}(\text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고”}) \\
&= \mathrm{P}(\text{“하늘은”}) \times \mathrm{P}(\text{“파랗고”} \mid \text{“하늘은”}) \\
&\quad \times \mathrm{P}(\text{“단풍잎은”} \mid \text{“하늘은 파랗고”}) \\
&\quad \times \mathrm{P}(\text{“빨갛고”} \mid \text{“하늘은 파랗고 단풍잎은”}) \\
&\quad \times \mathrm{P}(\text{“은행잎은”} \mid \text{“하늘은 파랗고 단풍잎은 빨갛고”}) \\
&\quad \times \mathrm{P}(\text{“노랗고”} \mid \text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은”})
\end{aligned}
$$

단어 연쇄의 확률 계산

(2) 단어 연쇄의 코퍼스 출현 빈도로 조건부확률 추정하기

$$P(\text{“은행잎은”}|\text{“하늘은 파랗고 단풍잎은 빨갛고”}) = \frac{\text{Count}(\text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은”})}{\text{Count}(\text{“하늘은 파랗고 단풍잎은 빨갛고”})}$$

문제

코퍼스에 한 번도 출현하지 않은 부분이 있으면 전체 확률을 0으로 추정한다.

단어 연쇄의 확률 계산

(2) 단어 연쇄의 코퍼스 출현 빈도로 조건부확률 추정하기

문제

"하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고"가 한국어에서 불가능한 문장인가?

단어 연쇄의 확률 계산 — N-그램 모형

(3) 조건부확률의 근삿값을 사용하기

문제

단어 연쇄가 길수록 코퍼스에 출현할 가능성이 낮으므로 확률 추정치가 0이 된다.
$\Rightarrow$ 언어의 창조성을 반영하지 못한다.

N-그램 단어 모형의 해결

N개 단어의 연쇄만 세어 문장의 확률을 계산한다.

- $N = 1$ 유니그램(Unigram)
- $N = 2$ 바이그램(Bigram)
- $N = 3$ 트라이그램(Trigram)

단어 연쇄의 확률 계산 — N-그램 모형

(3) 조건부확률의 근삿값을 사용하기

예시

트라이그램($\mathrm{N} = 3$) 근사

$$
\begin{aligned}
&\mathrm{P}(\text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고”}) \\
&\approx \mathrm{P}(\text{“하늘은”}) \times \mathrm{P}(\text{“파랗고”} \mid \text{“하늘은”}) \\
&\quad \times \mathrm{P}(\text{“단풍잎은”} \mid \text{“하늘은 파랗고”}) \\
&\quad \times \mathrm{P}(\text{“빨갛고”} \mid \text{“파랗고 단풍잎은”}) \\
&\quad \times \mathrm{P}(\text{“은행잎은”} \mid \text{“단풍잎은 빨갛고”}) \\
&\quad \times \mathrm{P}(\text{“노랗고”} \mid \text{“빨갛고 은행잎은”})
\end{aligned}
$$

단어의 직전 $(\mathrm{N} - 1)$개만 보자!

단어 연쇄의 확률 계산 — N-그램 모형

(3) 조건부확률의 근삿값을 사용하기

예시

바이그램($N = 2$) 근사

$$
\begin{aligned}
&\mathrm{P}(\text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고”}) \\
&\approx \mathrm{P}(\text{“하늘은”}) \times \mathrm{P}(\text{“파랗고”} | \text{“하늘은”}) \\
&\quad \times \mathrm{P}(\text{“단풍잎은”} | \text{“파랗고”}) \\
&\quad \times \mathrm{P}(\text{“빨갛고”} | \text{“단풍잎은”}) \\
&\quad \times \mathrm{P}(\text{“은행잎은”} | \text{“빨갛고”}) \\
&\quad \times \mathrm{P}(\text{“노랗고”} | \text{“은행잎은”})
\end{aligned}
$$

단어 연쇄의 확률 계산 — N-그램 모형

(3) 조건부확률의 근삿값을 사용하기

예시

바이그램($N = 2$) 근사

$$
\begin{aligned}
&P(\text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고”}) \\
&\approx P(\text{“하늘은”}) \times \frac{C(\text{“하늘은 파랗고”})}{C(\text{“하늘은”})} \\
&\quad \times \frac{C(\text{“파랗고 단풍잎은”})}{C(\text{“파랗고”})} \times \frac{C(\text{“단풍잎은 빨갛고”})}{C(\text{“단풍잎은”})} \\
&\quad \times \frac{C(\text{“빨갛고 은행잎은”})}{C(\text{“빨갛고”})} \times \frac{C(\text{“은행잎은 노랗고”})}{C(\text{“은행잎은”})}
\end{aligned}
$$

단어 연쇄의 확률 계산 — N-그램 모형

(3) 조건부확률의 근삿값을 사용하기

유니그램	구글 검색결과
”하늘은”	
”파랗고”	
”단풍잎은”	
”빨갛고”	
”은행잎은”	

바이그램	구글 검색결과
”하늘은 파랗고”	
”파랗고 단풍잎은”	
”단풍잎은 빨갛고”	
”빨갛고 은행잎은”	
”은행잎은 노랗고”	

문제

첫 단어의 확률 P(“하늘은”)은 어떻게 계산하는가?

단어 연쇄의 확률 계산 — N-그램 모형

(3) 조건부확률의 근삿값을 사용하기

문제

"하늘은"의 확률이 실제로 의미하는 것

- 문장이나 구가 "하늘은"으로 시작할 확률

→ (임의의 위치가 아니라) 첫 단어로 "하늘은"이 나올 확률

→ 조건부확률 P("하늘은"|<s>)

문장이 경계지어져 있어야 한다.

확률 계산의 다른 문제

Underflow

```
>>> 10. ** -323, 10. ** -324
(1e-323, 0.0)
```

해결

확률의 곱 $\Rightarrow$ 확률의 로그값의 합

$$p_1 \times p_2 \times \cdots p_n = \exp\left(\log p_1 + \log p_2 + \cdots + \log p_n\right)$$

로그함수와 지수함수

$e^a = b$가 성립할 때(자연상수 $e \approx 2.71828$)

- 로그함수 $a = \log(b)$
  - $\log(xy) = \log(x) + \log(y)$
  - $\log(\exp(x)) = x$
  - $\log(1) = 0$
  - 양의 실수 $x > 0$에 대해서만 $\log(x)$ 정의 가능
- 지수함수 $b = \exp(a)$
  - $\exp(x + y) = \exp(x)\exp(y)$
  - $\exp(\log(x)) = x$
  - $\exp(0) = 1$
  - 모든 실수 x에 대하여 $\exp(x) > 0$ 성립

$$\exp\left(\log p_1 + \log p_2 + \cdots + \log p_n\right) = \exp\left[\log(p_1 p_2 \cdots p_n)\right] = p_1 p_2 \cdots p_n$$

언어 모형 평가 방법

외재적 모형을 다른 과제에 응용했을 때 성능이 얼마나 향상되는가?

- 과제 수행을 위한 시간과 비용이 필요하다.

내재적 외부 과제와 무관하게 모형의 품질이 얼마나 좋은가?

- 모형의 품질을 평가할 척도가 필요하다.

내재적 평가의 원칙 혹은 전제

실제로 존재하는 문장에 높은 확률을 부여해야 한다.

복잡도(perplexity)

정의

실험 집합 $W = w_1 w_2 \dots w_K$의 확률의 역수의 K제곱근

$$PP(W) = P(w_1 w_2 \dots w_K)^{-\frac{1}{K}} = \sqrt[K]{\frac{1}{P(w_1 w_2 \dots w_K)}}$$

역수를 취했으므로, 확률이 커지면 복잡도가 낮아진다.

의문

왜 K제곱근을 취하는가?

- 길이 K인 연쇄의 확률: 조건부확률 K개의 곱
- 효과: 문장이 길어질수록 확률이 낮아지는 현상을 보완해 준다.

오늘 배운 내용

1. 언어 모형이란 무엇인가?
2. 단어 연쇄의 확률을 어떻게 계산하는가?
3. N-그램 언어 모형이란 무엇인가?
4. N-그램 언어 모형을 왜 사용하는가?
5. N-그램 확률을 어떻게 추정하는가?
6. 언어 모형을 어떻게 평가하는가?

다음 시간에 할 일

- 언어 모형을 사용하여 문장을 생성하기 (SLP3 3.3)
- 확률 추정치가 0이 되는 또 다른 문제 해결하기 (SLP3 3.4)