

2018학년도 2학기 언어와 컴퓨터

제14강 N-그램 언어 모형 (1)

박수지

서울대학교 인문대학 언어학과

2018년 11월 12일 월요일

오늘의 목표

- 1 N -그램의 개념을 설명할 수 있다.
- 2 N -그램을 사용하는 이유를 설명할 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

⇒ 단어 연쇄에 (조건부) 확률을 할당하기

예시

산에 _____

- 놀라
- 올라

구글 검색으로 “산에 놀라”와 “산에 올라”를 각각 세어 보자!

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

- 알 길이 없다
- 알 도로가 없다

어떻게 알 수 있는가?

언어 모형

단어 연쇄에 확률을 할당하는 모형

- N -그램 언어 모형
- 신경망 언어 모형
- ...

N -그램

N 개 단어의 연쇄

- 단어, 형태소, 품사, 문자, ...

단어의 연쇄의 확률

$P(w_1w_2\cdots w_n)$ 를 어떻게 알아내는가?

조건부확률과 연쇄법칙

$$P(W|H) = \frac{P(H, W)}{P(H)}$$

$$\Rightarrow P(H, W) = P(H)P(W|H)$$

$$\Rightarrow \cdots$$

$$\Rightarrow P(w_1w_2w_3) = \cdots = P(w_3|w_1w_2)P(w_2|w_1)P(w_1)$$

단어의 조건부확률 계산: 상대 빈도로 추정하기

$P(w_3|w_1, w_2)$ 등을 어떻게 알아내는가?

$$P(w_n|w_1w_2\cdots w_{n-1}) = \frac{C(w_1w_2\cdots w_{n-1}w_n)}{C(w_1w_2\cdots w_{n-1})}$$

예시

“알 길이 _____”에 “없다”가 출현할 확률

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

다른 예시

$P(\text{“하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고”})$ 에서...

$$P(\text{“노랗고”} | \text{“하늘은 ...은행잎은”}) = \frac{C(\text{“하늘은 ...노랗고”})}{C(\text{“하늘은 ...은행잎은”})}$$

⇒ 구글 검색에서 “하늘은 파랗고 단풍잎은 빨갛고 은행잎은”의 결과를 세어 보자.

문제

코퍼스에 나타나지 않은 단어 연쇄의 확률은 0이 된다.

해결

N-그램으로 근사 ($N = 2$)

$$\begin{aligned}
 &P(\text{하, 파, 단, 빨, 은, 노}) \\
 &= P(\text{하})P(\text{파}|\text{하})P(\text{단}|\text{하, 파}) \cdots P(\text{노}|\text{하, 파, 단, 빨, 은}) \\
 &\approx P(\text{하})P(\text{파}|\text{하})P(\text{단}|\text{파}) \cdots P(\text{노}|\text{은})
 \end{aligned}$$

N-그램 단어 모형

N-개 단어의 연쇄만 세어 문장의 확률을 계산한다.

확률 계산의 다른 문제

Underflow

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

해결

확률의 곱 \Rightarrow 확률의 로그값의 합

$$p_1 \times p_2 \times \dots p_n = \exp(\log p_1 + \log p_2 + \dots + \log p_n)$$