2021학년도 2학기 언어와 컴퓨터

제17강 N-그램 언어 모형 (3)

박수지

서울대학교 인문대학 언어학과

2021년 11월 10일 수요일

오늘의 목표

1. N-그램 모형에서 발생하는 0 문제를 평탄화로 해결할 수 있다.
2. N-그램 언어 모형의 한계를 설명할 수 있다.

Toy data: 훈련 코퍼스

문장	빈도
귀여운 고양이	25
귀여운 강아지	22
귀여운 다람쥐	2
귀여운 망아지	1
병아리	1

코퍼스 크기

- M = 101 (Number of word tokens)
  - “귀여운” 50회
  - “고양이” 25회
  - “강아지” 22회
  - “다람쥐” 2회
  - “망아지” 1회
  - “병아리” 1회
- |V| = 6 (Vocabulary size)
  - V= {귀여운, 고양이, 강아지, 다람쥐, 망아지, 병아리}

Toy data: 훈련 코퍼스

history = “귀여운”

바이그램	빈도
귀여운 귀여운	0
귀여운 고양이	25
귀여운 강아지	22
귀여운 다람쥐	2
귀여운 망아지	1
귀여운 병아리	0
귀여운	50

$$P(\text{“고양이”}|\text{“귀여운”}) = \frac{C(\text{“귀여운 고양이”})}{C(\text{“귀여운”})} = \frac{25}{50} = 0.5$$

$$P(\text{“병아리”}|\text{“귀여운”}) = \frac{C(\text{“귀여운 병아리”})}{C(\text{“귀여운”})} = \frac{0}{50} = 0.0$$

문제

실험 집합이 “귀여운 병아리”라면?

평탄화(Smoothing)

문제 훈련 코퍼스에 **없는** N-그램이 실험 코퍼스에 나타날 수 있다.

목표 훈련 코퍼스에 **없는** N-그램에도 양의 확률을 부여한다.

사실 확률의 합은 1이다.

종합 훈련 코퍼스에 **있는** N-그램의 확률을 깎아야 한다.

라플라스 평탄화(Laplace smoothing, Add-1 smoothing)

(0을 포함한) 모든 빈도에 1을 더해서 확률을 추정한다.

Toy data: 훈련 코퍼스

history = “귀여운”

바이그램	빈도	평탄화
귀여운 귀여운	0	0+1
귀여운 고양이	25	25+1
귀여운 강아지	22	22+1
귀여운 다람쥐	2	2+1
귀여운 망아지	1	1+1
귀여운 병아리	0	0+1
귀여운	50	50 + 6

P(“고양이”|“귀여운”)

$$= \frac{26}{56} = 0.4643 < 0.5$$

P(“병아리”|“귀여운”)

$$= \frac{1}{56} = 0.0179 > 0$$

Add-k smoothing

1 대신 k(< 1)을 더한다.

훈련 코퍼스의 어휘

$V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ ={귀여운, 고양이, 강아지, 다람쥐, 망아지, 병아리}

주변확률분포

$$\begin{aligned} C(\text{“귀여운”}) &= C(\text{“귀여운 귀여운”}) + C(\text{“귀여운 고양이”}) + \cdots C(\text{“귀여운 병아리”}) \\ &= C(\text{“귀여운 } v_1\text{”}) + C(\text{“귀여운 } v_2\text{”}) + \cdots C(\text{“귀여운 } v_6\text{”}) \\ &= \sum_{i=1}^{|V|} C(\text{“귀여운 } v_i\text{”}) \end{aligned}$$

Add-k 평탄화

$$
\begin{aligned}
P_{\text{Add-k}^*}(\text{“다람쥐”}|\text{“귀여운”}) &= \frac{C(\text{“귀여운 다람쥐”}) + k}{\sum_{i=1}^{|V|} [C(\text{“귀여운 } v_i\text{”}) + k]} \\
&= \frac{C(\text{“귀여운 다람쥐”}) + k}{\sum_{i=1}^{|V|} C(\text{“귀여운 } v_i\text{”}) + \sum_{i=1}^{|V|} k} \\
&= \frac{C(\text{“귀여운 다람쥐”}) + k}{C(\text{“귀여운”}) + k|V|}
\end{aligned}
$$

0에 대처하는 또다른 방법

N-그램이 없으면 $(N-1)$-그램을 동원하자!

back-off N-그램이 없을 때만 $(N-1)$-그램을 사용한다.

보간법 항상 N-그램과 $(N-1)$-그램을 함께 사용한다.

보간법(interpolation)

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n|w_{n-2}w_{n-1})$$

$\lambda_1, \lambda_2, \lambda_3$ 모형 매개변수(model parameters)

λ_1 유니그램의 가중치

λ_2 바이그램의 가중치

λ_3 트라이그램의 가중치

0의 문제

실험 코퍼스의 N-그램이 훈련 코퍼스에 없는 경우에 발생한다.

해결 방법

- $N \geq 2 \Rightarrow$ 평탄화, 보간법
- $N = 1 \Rightarrow$???

남은 문제

실험 집합이 “귀여운 햄스터”라면?

문제

훈련 코퍼스에 없는 단어가 실험 코퍼스에 나올 때는 어떻게 처리하는가?

해결

훈련 코퍼스에 미리 <UNK>(unknown word)의 자리를 잡아 놓는다.

<UNK> 설정 방법

어휘 V의 크기나 최소 출현 횟수를 미리 정해 놓는다.

- 예1: $|V| = 10,000$으로 한정하고 나머지 단어는 <UNK>으로 처리한다.
- 예2: 훈련 코퍼스에서 3회 이하 출현한 단어는 모두 <UNK>으로 처리한다.

Toy data: 훈련 코퍼스

문장	빈도
귀여운 고양이	25
귀여운 강아지	22
귀여운 다람쥐	2
귀여운 ~~망아지~~	1
~~병아리~~	1

<UNK> 도입 후 코퍼스 크기

- M = 101 ... 변하지 않는다.
  - “귀여운” 50회
  - “고양이” 25회
  - “강아지” 22회
  - “다람쥐” 2회
  - ~~“망아지” 1회, “병아리” 1회~~ “<UNK>” 2회
- |V| = 5 ... 줄어든다.
  - V= {귀여운, 고양이, 강아지, 다람쥐, <UNK>}

확률 추정 방법

실험 집합 “귀여운 햄스터”의 확률 추정값은 “귀여운 <UNK>”의 확률과 같다.

N-그램의 한계

- 장거리 의존을 반영하지 못한다.
  예: The **computers** which I had just put into the machine room on the fifth floor **are** crashing
  - 확률 추정치: P(“is”|“floor”) $\gg$ P(“are”|“floor”)
- 자유어순언어에 대해서는 잘 작동하지 않는다.
  예: 돈을 그에게 주었다. 그에게 돈을 주었다.
  - 확률 추정치: P(“돈을”|“그에게”) $\simeq$ P(“그에게”|“돈을”)
    $\Rightarrow$ “돈을 그에게 돈을 그에게 돈을 …” 생성 가능

오늘 배운 것

- N-그램 확률이 0이 되는 문제 및 해결 방법
  - 평탄화
  - 보간법
  - 미지의 단어 <UNK> 처리

다음 주에 배울 것

- 기계학습 — 단순 베이즈 분류
- 감정분석
  - SLP3 Ch. 4