03 언어 모델

12기 이가영

KU-BIG

언어 모델

- **언어 모델**(Language Model) : 가장 자연스러운 단어 시퀀스를 찾아내는 모델
- **언어 모델링**(Language Modeling): 주어진 단어들로부터 아직 모르는 단어를 예측하는 작업,

  단어 시퀀스에 확률을 할당하는 일

  - 기계 번역, 오타 교정, 음성 인식

- 단어 시퀀스의 확률 $P(W) = P(w_1, w_2, w_3, w_4, w_5, \ldots, w_n)$
- 다음 단어가 등장할 확률 $P(w_n|w_1, \ldots, w_{n-1})$
- 전체 단어 시퀀스의 확률 $P(W) = P(w_1, w_2, w_3, w_4, w_5, \ldots w_n) = \prod_{i=1}^{n} P(w_n|w_1, \ldots, w_{n-1})$

통계적 언어 모델

- **문장에 대한 확률** : 각 단어에 대한 예측 확률들의 곱

  P(An adorable little boy is spreading smiles)

  = P(An) × P(adorable|An) × P(little|An adorable) × P(boy|An adorable little) × P(is|An adorable little boy)

  × P(spreading|An adorable little boy is) × P(smiles|An adorable little boy is spreading)

- **카운트 기반 접근**: $P(\text{is}|\text{An adorable little boy}) = \frac{\text{count}(\text{An adorable little boy is})}{\text{count}(\text{An adorable little boy })}$

- **희소 문제(Sparsity problem)** : 기계가 훈련한 코퍼스에 해당 단어 시퀀스가 없는 경우

N-gram 언어 모델

- **N-gram**: n개의 연속적인 단어 나열

- 코퍼스에서 n개의 단어 단위로 끊은 것을 하나의 토큰으로 간주

→ n-1개 단어에 의존해 다음에 나올 단어 예측

ex. N=4 $P(w|\text{boy is spreading}) = \frac{\text{count}(\text{boy is spreading } w)}{\text{count}(\text{boy is spreading})}$

- **한계점**

- n 선택의 trade-off : n ↑ ⇨ 언어모델 성능 ↑,

훈련 코퍼스에서 해당 n-gram 카운트할 수 있는 확률 ↓ (희소문제)

⇨ n≤ 5, 적용 분야에 맞는 코퍼스 수집

Perplexity

- PPL: 언어 모델을 평가하기 위한 내부 평가 지표

$$PPL(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \ldots, w_N)}} = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} P(w_i|w_1, w_2, \ldots, w_{i-1})}}$$

- 단어의 수로 정규화된 테스트 데이터에 대한 확률의 역수
- 해당 언어 모델이 특정 시점에 평균적으로 몇 개의 선택지를 가지고 고민하고 있는지를 의미

ex. PPL = 10

$$PPL(W) = P(w_1, w_2, w_3, \ldots, w_N)^{-\frac{1}{N}} = (\frac{1}{10}^{N})^{-\frac{1}{N}} = \frac{1}{10}^{-1} = 10$$

KU-BIG