# 언어 모델

## 1) 언어 모델 (Language Model)

: 단어 시퀀스에 확률을 할당하는 모델

↳ 가장 자연스러운 단어 시퀀스를 찾아내는 모델.

보편적 방법 : 이전 단어들이 주어졌을 때 다음 단어 예측

· 스탠포드에선 '문법'이라 비유.

· 확률을 할당 이유

1. 기계 번역 : $P$(나는 버스를 탔다) > $P$( " 태웠다)
2. 오타 교정 : 부리나케 $P$(달려갔다) > $P$(잘려갔다)
3. 음성 인식 : $P$( 메론을 먹는다) < $P$ ( 메론을 먹는다)

→ 더 적절한 문장을 선택.

· 확률.

$P(W) = P(w_1, w_2, \cdots w_n)$ $W$ : 단어시퀀스, $w$ : 단어

다음단어 등장 확률 : $P(w_n | w_1, w_2 \cdots w_{n-1})$ → 조건부 확률.

$P(W) = P(w_1, w_2, \cdots w_n)$

$= \prod_{i=1}^{n} P(w_n | w_1, \cdots w_{n-1})$

→ 지금까지 나온 단어를 고려해서 다음 단어 예측.

## 2) 통계적 언어 모델 (SLM)

: 문장에 대한 확률을 카운트 기반으로 접근.

ex) An adorable little boy 뒤에 is 나올 확률.

$P(\text{is} | \text{An} \cdots \text{boy}) =$ An … boy 100개 중 is가 30개

→ 30%.

엄청난 양의 데이터가 요구된다

→ 희소문제 (Sparsity problem)

· 충분한 데이터를 관측하지 못해 정확히 모델링 못함.

: N-gram 같이 완화가능 but 근본 해결X

⇒ 트렌드 : 통계적 언어모델 → 인공신경망

## 3) N-gram

: 카운트 할 단어 개수를 n-1개로 줄인다.

ex) 4-gram. ~~An adorable little~~ boy is spreading ?

↳ 고려.

· 한계.

1. 희소문제
2. n 선택은 trade-off → 최대 5로 권장

⇒ 적용 분야에 맞는 코퍼스 수집이 중요하다.

## 4) 한글 언어 모델

· 특징

(1) 어순이 중요 X

(2) 교착어 → 토큰화 시 접사나 조사 분리 중요

(3) 띄어쓰기 잘 지켜 X

## 5) Perplexity

: 모델 내에서 성능 수치화. 내부 평가.

· PPL (PerPLexity)

$$PPL(W) = P(w_1, \cdots w_n)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{\prod_{i=1}^{n} P(w_i | w_1 \cdots w_{i-1})}}$$

→ 분기계수 (branching factor) : 선택 가능한 경우의 수.

→ 낮으면 성능↑.