Language Model 2 : 통계적 언어 모델

조건부 확률

- **정의** : 어떤 사건 A가 일어났을 때 사건 B가 일어날 확률을. 사건 B가 발생하는 도수(혹은 수량)는 사건 A의 영향을 받아 변하는데 이를 조건부 확률이라 함.

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

즉, $P(A,B) = P(A)P(B|A)$

문장에 대한 확률

- 각 단어는 문맥이라는 관계로 인해 이전 단어의 영향을 받아 나온 단어
- 즉, 문장에 대한 확률은 아래와 같이 조건부 확률의 곱으로 구성됨 → n개 단어가 동시에 나타날 확률

$$P(w_1, w_2, w_3, w_4, w_5, \ldots, w_n) = \prod_{n=1}^{n} P(w_n|w_1, \ldots, w_{n-1})$$

- ex) 문장 '내 마음 속에 영원히 기억될 최고의 명작이다'의 확률

$$P(\text{내}, \text{마음}, \text{속에}, \text{영원히}, \text{기억될}, \text{최고의}, \text{명작이다}) =$$
$$P(\text{내}) \times P(\text{마음}|\text{내}) \times P(\text{속에}|\text{내}, \text{마음}) \times P(\text{영원히}|\text{내}, \text{마음}, \text{속에}) \times P(\text{기억될}|\text{내}, \text{마음}, \text{속에}, \text{영원히})$$
$$\times P(\text{최고의}|\text{내}, \text{마음}, \text{속에}, \text{영원히}, \text{기억될}) \times P(\text{명작이다}|\text{내}, \text{마음}, \text{속에}, \text{영원히}, \text{기억될}, \text{최고의})$$

카운트 기반의 접근

- 이전 단어로부터 다음 단어에 대한 확률(조건부 확률) 구하는 방법 → 해당 문자열 시퀀스가 말뭉치에서 나타난 빈도(frequency)를 사용
- ex) 문장 '내 마음 속에 영원히 기억될 최고의' 다음에 '명작이다'가 나타날 확률

$$P(\text{명작이다}|\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의}) = \frac{Freq(\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의},\text{명작이다})}{Freq(\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의})}$$

- 만약 기계가 학습한 코퍼스 데이터에서 '내 마음 속에 영원히 기억될 최고의'가 100번 등장하였고, 그 다음에 '명작이다'가 등장한 경우가 30번이라면
  $P(\text{명작이다}|\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의})$는 30%

카운드 기반 접근의 한계 : 희소 문제 (Sparsity Problem)

- $P(\text{명작이다}|\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의})$를 구하는 경우에서 기계가 훈련한 코퍼스 데이터에 '내 마음 속에 영원히 기억될 최고의 명작이다'라는 단어 시퀀스가 없다면 분자가 0이 되어서 전체 값이 0이 된다. 즉, 매끄러운 문장임에도 불구하고 확률이 0이 된다.
- **희소 문제** : 이와 같이 충분한 데이터를 관측하지 못하여 언어를 정확히 모델링하지 못하는 문제
  → 완화 방법 : **N-gram**, **스무딩과 백오프** (일반화 기법)

  → But, 희소 문제에 대한 근본적인 해결책이 되지 못함 → 이러한 통계 모델의 한계로 인해 언어 모델의 트렌드는 **통계적 언어 모델**에서 **인공 신경망 언어 모델**로 넘어가게 됨