

통계적 언어모델(SLM)

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

“An adorable little boy is spreading smiles” 문장이 등장할 확률

$P(\text{An adorable little boy is spreading smiles}) =$

$P(\text{An}) \times P(\text{adorable}|\text{An}) \times P(\text{little}|\text{An adorable}) \times P(\text{boy}|\text{An adorable little}) \times P(\text{is}|\text{An adorable little boy})$
 $\times P(\text{spreading}|\text{An adorable little boy is}) \times P(\text{smiles}|\text{An adorable little boy is spreading})$

문장의 확률을 구하기 위해서 각 단어에 대한 예측 확률들을 곱합니다.

$$P(\text{is}|\text{An adorable little boy}) = \frac{\text{count}(\text{An adorable little boy is})}{\text{count}(\text{An adorable little boy})}$$

N-gram 언어 모델

- 통계기반 언어모델의 일종. SLM과 같이 카운트 기반 통계적 접근을 사용한다.
- 전통적 SLM과 달리 이전에 등장한 모든 단어가 아닌 **일부 단어만 고려**하는 방법을 사용한다.
 - n-gram에서의 n은 코퍼스 내 단어들을 n개씩 묶어서 빈도를 학습했음을 의미한다.
- **이전 n-1개의 단어를 보고 n번째 단어를 예측**하는 방식
- 임의의 개수만큼의 이전 단어만 참고하여 확률을 근사
 - 코퍼스에서 해당 단어시퀀스를 카운트할 확률이 높아진다.