Language Model 3 : N-gram 언어 모델

일반적인 통계 언어 모델의 단점인 희소문제를 완화하는 모델로서, 일반 언어 모델처럼 이전에 등장한 모든 단어를 고려한 것이 아니라 일부 단어만을 고려하는 방법. 이때 일부 단어를 몇개 보느냐에 따라 N이 바뀐다.
example : 3개의 단어만 보고 다음 단어를 예측할 경우 3-gram 언어모델

코퍼스에서 카운트하지 못하는 경우의 감소

- 즉, 희소문제를 완화시킴.
- 모든 단어를 고려하는 경우 코퍼스에서 그 sequence(모든 단어)가 존재하지 않을 가능성 ↑
  → 일부 단어만을 고려하면 코퍼스에서 그 sequence(일부 단어)가 존재할 가능성 ↑
- bigram 모델 예제
  $P(\text{명작이다}|\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의}) \approx P(\text{명작이다}|\text{최고의})$
  "내 마음 속에 영원히 기억될 최고의 명작이다"가 있을 가능성보다는 "최고의 명작이다"라는 더 짧은 단어 시퀀스가 존재할 가능성 ↑
- trigram 모델 예제
  $P(\text{명작이다}|\text{내},\text{마음},\text{속에},\text{영원히},\text{기억될},\text{최고의}) \approx P(\text{명작이다}|\text{기억될},\text{최고의})$
  "내 마음 속에 영원히 기억될 최고의 명작이다"가 있을 가능성보다는 "기억될 최고의 명작이다"라는 더 짧은 단어 시퀀스가 존재할 가능성 ↑

N-gram

- n-gram을 통한 언어 모델에서는 다음에 나올 단어의 예측은 오직 **n-1**개의 단어에만 의존
- example :
  An adorable little boy is spreading 다음에 나올 단어 예측을 위한, 4-gram을 이용한 언어 모델 사용 →
  3(=4-1)개의 단어만을 고려하여 다음 단어 예측

  $P(w|boy\ is\ spreading) = \frac{count(boy\ is\ spreading\ w)}{count(boy\ is\ spreading)}$

  $count(boy\ is\ spreading) = 1000, count(boy\ is\ spreading\ insults) = 500, count(boy\ is\ spreading\ smiles) = 200$
  라면 $P(insults|boyisspreading) = 50\%$, $P(smiles|boyisspreading) = 20\%$

N-gram Language Model의 한계

N-gram의 예제에서 An adorable little boy라는 수식어는 제거되어, 반영되지 않음. 하지만, An adorable little boy라는 수식어가 제거되지 않았더라면 **insults** 대신 **smiles**라는 단어가 좀 더 적합할 수 도 있음 → n-gram은 단어 몇개만을 고려하여 다음 단어를 예측하기에 좀 더 **매끄럽지 못한 문장**을 만들어내기도 함

1. 희소 문제
   - 데이터에 한번도 등장하지 않은 n-gram이 존재할 때 생기는 희소 문제 존재
   - example : **또바기**라는 단어가 학습데이터에 한번도 등장하지 않았다면, 이 언어 모델은 예측 단계에서 **그 아이는 또바기 인사를 잘한다**라는 자연스러운 문장이 등장할 확률을 0으로 부여
2. n을 선택하는 것은 trade-off 문제
   - n을 크게 할 경우 : 희소 문제가 점점 심각해짐
   - n을 작게 할 경우 : 근사 정확도가 낮아짐
   - **적절할 n**을 선택하는 것이 관건 → n은 최대 5를 넘게 잡지 않도록 권장됨

적용 분야(Domain)에 맞는 코퍼스 수집

- 해당 도메인의 코퍼스를 언어 모델의 코퍼스로 사용 → 언어 모델이 제대로 된 언어 생성을 할 가능성 ↑

- **훈련에 사용된 도메인 코퍼스**에 따라, 언어모델의 성능이 비약적으로 달라짐

인공 신경망을 이용한 언어 모델

- N-gram Language Model보다 성능이 우수한 인공 신경망을 이용한 언어 모델 주로 사용됨 → **ELMo, GPT**

출처 : https://wikidocs.net/21692