과제명 : 지식 그래프 기반 음식점 광고 콘텐츠 추천 알고리즘 개발
(Title)

과 제 번 호 :
(Project No.)

3. 언어 모델 (Language Model)

① 통계를 이용한 방법 ② 인공 신경망을 이용한 방법

-1) 언어 모델

-1. 언어모델 : 단어 시퀀스에 확률을 할당.
언어 모델링: 이전 단어들이 주어졌을 때 다음 단어 예측

-2. 단어시퀀스의 확률 할당
① 기계 번역 (Machine Translation)
② 오타 교정 (Spell Correction)
③ 음성 인식 (Speech Recognition)

-3. 주어진 이전 단어들로부터 다음 단어 예측하기
① 단어 시퀀스의 확률 (단어: w, 단어시퀀스: W)
$P(W) = P(w_1, w_2, \cdots, w_n)$
② 다음 단어 등장 확률 : $P(w_n | w_1, w_2, \cdots, w_{n-1})$
$P(W) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \cdots, w_{i-1})$

2) 통계적 언어 모델 (Statistical Language Model, SLM)

-1. 조건부확률
$P(x_1, x_2, x_3, \cdots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\cdots P(x_n|x_1, x_2, x_3)$

-2. 문장에 대한 확률
$P(\text{I am the best developer})$
$= P(\text{I}) \times P(\text{am}|\text{I}) \times P(\text{the}|\text{I am}) \cdot P(\text{developer}|\text{I am the best})$

-3. 카운트 기반의 접근과 한계
I가 100번 등장했을 때 am이 30번 등장했다면
$P(\text{am}|\text{I})$는 30%.
코퍼스에 I am the best 가 없었다면 확률은 0 또는 정의되지 않은 확률 ⇒ 희소문제

3) N-gram 언어 모델 (N-gram Language Model)

SLM의 일종. 일부 단어만 고려하는 접근 방법을 사용.

-1. 코퍼스에서 카운트하지 못하는 경우의 감소
$P(\text{developer}|\text{I am the best}) \approx P(\text{developer}|\text{best})$
이렇게 하면 갖고 있는 코퍼스에서 해당 단어의 시퀀스를 카운트 할 확률 증가.

-2. N-gram (unigram, bigram, trigram, n-gram)
n=3) (I am) the best developer

-3. 한계
① 희소문제 ② n을 선택하는 것은 trade-off 문제 ($n \leq 5$ 권장)

-4. 적용될 분야에 맞는 코퍼스 수집
제대로 된 언어 생성을 할 가능성 증가.

4) 한국어에서의 언어 모델 (Language Model for Korean Sentences)

-1. 한국어는 어순이 중요하지 않다.
-2. 한국어는 교착어다.
-3. 한국어는 띄어쓰기가 제대로 지켜지지 않는다.
↳ 한국어 언어모델은 까다롭다.

5) 펄플렉시티 (Perplexity, PPL)

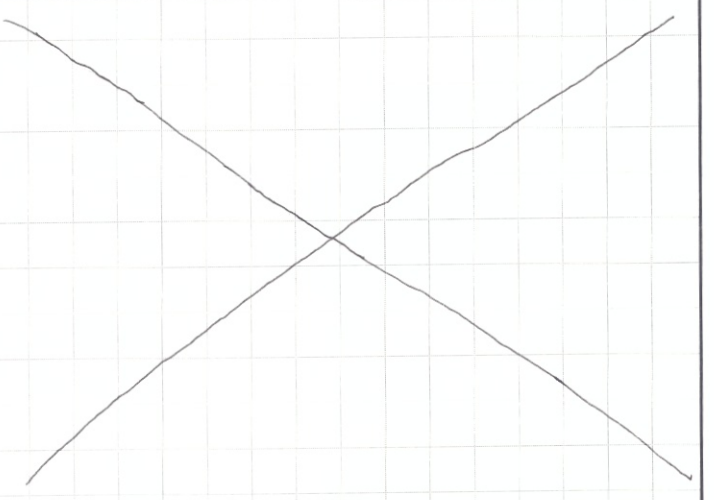
모델 내에서 자신의 성능을 수치화하여 결과를 내놓는다.

-1. 언어모델의 평가 방법 (Evaluation metric : PPL)
PPL은 문장의 길이로 정규화된 문장 확률의 역수
PPL의 수치가 높으면 낮은 성능, 수치가 낮으면 높은 성능.

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

-2. 분기계수 (Branching factor)
PPL은 선택할 수 있는 가능한 경우의 수를 의미하는 분기계수이다.

* 평가 방법이 있어서 주의할 점은 PPL의 값이 낮다는 것은 테스트 데이터 상에서 높은 정확도를 보인다는 것이지, 사람이 직접 느끼기에 좋은 언어 모델이라는 것을 반드시 의미하지는 않는다.
또한 언어 모델의 PPL은 테스트 데이터에 의존하므로 두개 이상의 언어 모델을 비교할 때는 정량적으로 양이 많고 또한 도메인에 알맞은 동일한 테스트 데이터를 사용해야 신뢰도가 높다는 것이다.



기 록 자 (Recorded by)	서 명 (Signature)	일 자 (Date)	
이 상 준	이 상 준	2022. 3. 19	이 노트의 지적재산권은 숭실대학교 산학협력단에 있음 (Intellectual property right for this notebook belongs to Foundation of Soongsil)
점 검 자 (Reviewed by) 정 선 태	서 명 (Signature)	일 자 (Date) 2022. 3. 22	