2021-03-26

언어 모델

김민주

목차
INDEX

언어 모델 (Language Model)

단어 시퀀스(문장)에 확률을 할당하는 모델

I eat breakfast.

I am breakfast

구분

언어 모델을 만드는 방법

통계를 이용한 방법

Statistical Languagel Model, SLM

인공 신경망을 이용한 방법

- GPT
- BERT

01 언어 모델

기본 개념

'단어 시퀀스에 확률을 할당'

기계 번역(Machine Translation)

P(나는 버스에 탔다) > P(나는 버스에 태운다)

오타 교정(Spell Correction)

선생님이 교실로 부리나케

P(달려갔다) > P(잘려갔다)

음성 인식(Speech Recognition)

P(나는 메롱을 먹는다) < P(나는 메론을 먹는다)

01 언어 모델

다음 단어 예측

주어진 이전 단어들로부터 다음 단어 예측

단어 시퀀스의 확률

$$P(W) = P(w_1, w_2, w_3, w_4, w_5, \ldots, w_n)$$

다음 단어 등장 확률

$$P(w_n | w_1, \ldots, w_{n-1})$$

전체 단어 시퀀스 W의 확률

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

01 언어 모델

예시

검색 엔진에서 언어모델의 예

조건부 확률

조건부 확률의 연쇄법칙(chain rule)

$$p(B|A) = P(A,B)/P(A)$$

$$P(A,B) = P(A)P(B|A)$$

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$

$$P(x_1, x_2, x_3 \ldots x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1 \ldots x_{n-1})$$

02 통계적 언어모델

개념

문장에 대한 확률

P(An adorable little boy is spreading smiles)

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

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

02 통계적 언어모델

Statistical Language Model, SLM

카운트 기반의 접근

이전 단어로부터 다음 단어에 대한 확률? 카운트에 기반

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

희소 문제(sparsity problem)?

충분한 데이터를 관측하지 못하여 언어를 정확히 모델링하지 못하는 문제

코퍼스에서 카운트하지 못하는 경우의 감소

앞 단어 중 임의의 개수만 포함해서 카운트하여 근사

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

$$P(\text{is}|\text{boy}) \quad P(\text{is}|\text{little boy})$$

N-gram

An adorable little boy is spreading smiles

Unigrams

An, adorable, little, boy, is, spreading, smiles

bigrams

an adorable, adorable little, little boy, boy is, is spreading, spreading smiles

trigrams

an adorable little, adorable little boy, little boy is, boy is spreading, is spreading smiles

4-grams

an adorable little boy, adorable little boy is, little boy is spreading, boy is spreading smiles

한국어에서의 언어모델

1. 한국어는 어순이 중요하지 않음

2. 한국어는 교착어

3. 한국어는 띄어쓰기가 제대로 지켜지지 않음

5 펄플렉서티

Perplexity

언어모델을 평가하기 위한 내부 평가 지표(PPL)

PPL은 단어의 수로 정규화(normalization) 된 테스트 데이터에 대한 확률의 역수

PPL을 최소화함 = 문장의 확률을 최대화 함

$$PPL(W) = P(w_1, w_2, w_3, \ldots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \ldots, w_N)}}$$

$$PPL(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \ldots, w_N)}} = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} P(w_i|w_1, w_2, \ldots, w_{i-1})}}$$

5 펄플렉서티

Perplexity

분기 계수(Branching factor)

PPL?

이 언어 모델이 특정 시점에서 평균적으로 몇 개의 선택지를 가지고 고민하고 있는지를 의미

ex) PPL=10

$$PPL(W) = P(w_1, w_2, w_3, \ldots, w_N)^{-\frac{1}{N}} = (\frac{1}{10}^{N})^{-\frac{1}{N}} = \frac{1}{10}^{-1} = 10$$

언어 모델 성능 판단의 지표

5 펄플렉서티

Perplexity

인공신경망과의 비교

Model	Perplexity
Interpolated Kneser-Ney 5-gram (Chelba et al., 2013)	67.6
RNN-1024 + MaxEnt 9-gram (Chelba et al., 2013)	51.3
RNN-2048 + BlackOut sampling (Ji et al., 2015)	68.3
Sparse Non-negative Matrix factorization (Shazeer et al., 2015)	52.9
LSTM-2048 (Jozefowicz et al., 2016)	43.7
2-layer LSTM-8192 (Jozefowicz et al., 2016)	30
Ours small (LSTM-2048)	43.9
Ours large (2-layer LSTM-2048)	39.8