AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

언어모델

Seokhwan Yang

언어 모델(Language Model)

- 언어 모델이란?
  - 언어 모델이란 문장의 확률을 나타내는 모델이다.
  - 문장은 단어들로 이루어진 시퀀셜 데이터이다.
  - 따라서 언어 모델이란 단어 시퀀스에 확률을 부여하는 모델이다.

    → 단어 시퀀스를 입력받아서 해당 시퀀스가 얼마나 그럴듯한지 확률을 출력
  - 한국어 말뭉치로 학습된 언어모델은 자연스러운 한국어 문장에 높은 확률값 부여 → 한국어 언어 모델이 필요한 이유

언어 모델(Language Model)

- 다음 문장에서 어떤 것이 가장 자연스러운가?
  - 나는 버스정류장에서 방금 버스를 _____.

1. 사랑해
2. 고양이
3. 놓쳤다
4. 사고남

우리는 아주 쉽게 정답을 맞출 수 있다. 3번

그렇다면 컴퓨터는? AI는?

문제가 “버스를”이 아니라 “버스가”였다면 쉽게 4번을 선택할 것이다.

문제가 “버스를”인 경우 4번을 선택하면? 뜻은 전달된다. 어색하지만..

언어 모델(Language Model)

- 다음의 두 문장 중에서 우리가 살아가면서 더 자주 접할 문장은?

  1. 저는 어제 점심을 먹었습니다.
  2. 저는 2015년 3월 18일 점심을 먹었습니다.

1, 2번 중에서 틀린 문장은 없다.

그러나 2번 같은 문장을 살면서 쉽게 접하지는 않는다.

문장을 구성하는 단어의 조합은 매우 다양하지만
우리가 자연스럽게 받아들이는 조합은
동등한 확률보다는 평소에 자주 사용되는 단어나 표현의 조합이
훨씬 높은 확률로 발생한다.

언어 모델(Language Model)

- 인간의 경우는…
  - 우리는 살아오면서 수많은 문장을 접해왔고
  - 머릿속에는 단어와 단어 사이의 확률이 자신도 모르게 학습되어 있다.
  - 그래서 대화 도중에 몇 단어 정도 알아듣지 못해도 대화에는 큰 지장이 없다.
  - 추가적으로 문맥 정보를 이용하는 것도 큰 도움이 된다.

언어 모델(Language Model)

- 인간과 같은 능력을 가진 언어 모델을 만들기 위하여...
  - 우리는 인터넷, 책 등 다양한 경로로 수많은 문장을 수집하고
  - 단어와 단어 사이의 출현 빈도를 세어 확률을 계산한다.
  - 특정 분야의 문장 분포를 파악하기 위해 전문 분야에 대한 코퍼스를 수집하기도 한다.
  - 이러한 과정의 궁극적인 목표는
  - 일상 생활에서 사용하는 실제 언어(문장)의 분포를 정확하게 근사하는 것이다.

언어 모델(Language Model)

- 언어 모델의 분류
  - 확률에 기초한 통계적 언어 모델(Statistical Language Model, SLM)
  - 신경망에 기초한 딥러닝 언어모델(Deep Neural Network Language Model, DNN LM)
- 두 가지의 분류 모두 기본적으로는
  - 주어진 단어를 바탕으로 다음 단어, 혹은 단어들의 조합을 예측하며
  - 이를 바탕으로 문장 생성, 기계 번역, 음성 인식, 문서 요약 등과 같은 다양한 자연어 처리 문제를 해결한다

언어 모델(Language Model)

- 통계적 언어 모델은
  - 단어열이 가지는 확률 분포를 기반으로 각 단어의 조합을 예측하는 전통적인 언어 모델
- 모델의 목표는
  - 실제로 많이 사용하는 단어열(문장)의 분포를 정확하게 근사하는 것

언어 모델(Language Model)

- 확률을 기반으로 단어의 조합을 예측한다는 것은
  - 주어진 단어를 통해 다음 단어로 돌 확률이 가장 높은 단어를 예측하는 일련의 과정을 의미
  - 쉽게 접할 수 있는 예시는?
    - 스마트 폰의 "자동 완성" 기능
  - 조건부 확률(Conditional Probabilities)을 언어 현상에 적용해 보는 것에서 출발함

언어 모델(Language Model)

- **문장의 확률 표현**
  - **문장에서 i 번째로 등장하는 단어를 w_i로 표시한다면**
  - **n개의 단어로 구성된 문장이 해당 언어에서 등장할 확률(=언어모델의 출력)**
    - $P(w_1, w_2, w_3, w_4, \dots, w_n)$
    - n개의 단어가 **동시에** 나타날 결합 확률
  - **예: 잘 학습된 한국어 모델이 있다면**
    - $P(\text{무모}, \text{운전}) < P(\text{난폭}, \text{운전})$
    - "난폭"이 나타난 다음에 "운전"이 나타날 확률 $\rightarrow P(\text{운전}|\text{난폭}) = \frac{P(\text{난폭}, \text{운전})}{P(\text{난폭})}$: 조건부 확률

조건부 확률 표기:
결과가 되는 사건을 앞에,
조건이 되는 사건을 뒤에 표기함

언어 모델(Language Model)

- **결합 확률과 조건부 확률**
  - **3개의 단어가 동시에 등장할 결합 확률**
    - $P(w_1, w_2, w_3) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2)$
    - 다음의 3가지 사건이 동시에 발생하여야 함
      - 첫 번째 단어(w_1) 등장
      - 첫 번째 단어(w_1) 등장 후 두 번째 단어 (w_2) 등장
      - 첫 번째 단어(w_1)와 두 번째 단어 (w_2) 등장 후 세 번째 단어 (w_3) 등장
  - **조건부 확률로 다시 쓰면**
    - $P(w_1, w_2, w_3, \dots, w_n) = \prod_{i=1}^{n} P(w_i|w_1, \dots, w_{i-1})$

언어 모델(Language Model)

- 우리는 임의의 단어 시퀀스가 해당 언어에서 얼마나 자연스러운지를 이해하고 있는 언어 모델을 구축하고자 한다.
- 조건부 확률의 정의에 따라... 수식의 좌변, 우변이 같으므로

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

- 이전 단어(컨텍스트)들이 주어졌을 때 다음 단어를 맞히는 문제로도 목표를 달성할 수 있다.

언어 모델(Language Model)

- **순방향 언어 모델(Forward Language Model)**
  - **문장의 앞에서 뒤로, 사람이 이해하는 순서대로 계산하는 모델**
    - 어제 카페 갔었어 거기 사람 많더라: 어제→카페→갔었어→거기→사람→많더라
  - **GPT(Generative Pretrained Transformer), ELMo 등**
- **역방향 언어 모델(Backward Language Model)**
  - **문장의 뒤부터 앞으로 계산하는 모델**
    - 어제 카페 갔었어 거기 사람 많더라: 많더라→사람→거기→갔었어→카페→어제
  - **ELMo(Embeddings from Language Models) 등**

ELMo는 순방향, 역방향 모두 사용

언어 모델(Language Model)

- 넓은 의미의 언어 모델
  - 전통적인 의미의 언어 모델은 조건부확률의 정의를 따르는 수식으로 표현
  - 최근에는...

$$P(w|context)$$

    - 컨텍스트(주변 맥락 정보)가 전제된 상태에서 특정 단어(w)가 나타날 조건부 확률로 표현하기도 한다.

언어 모델(Language Model)

- **잘 학습된 언어 모델**
  - **어떤 문장이 자연스러운지 가려낼 수 있으므로 그 자체로 가치가 있다.**
  - **학습 대상 언어의 풍부한 맥락을 표현하고 있다.**

  → 기계 번역, 문법 교정, 문장 생성 등 다양한 태스크 수행 가능

    - 기계 번역: $P($? | You can't be free from death $)$
    - 문법 교정: $P($ 두 시 삼십 이분 $) > P($ 이시 서른 두분 $)$
    - 문장 생성: $P($? | 발 없는 말이 $)$

언어 모델(Language Model)

- 언어 모델
  - 언어를 이루는 구성 요소(글자, 형태소, 단어, 단어열(문장), 문단 등)에 확률 값을 부여하여 이를 바탕으로 다음 구성요소를 예측하거나 생성하는 모델

언어 모델(Language Model)

- **언어 모델의 변화 추세**
  - **전통적인 언어처리 연구에서 딥러닝 패러다임으로 전환**
  - **다양한 모델의 조합**
    - 예: 단어의 문자를 인식하는 CNN 모델 + 시퀀스 처리를 위한 LSTM(RNN) + MLP를 이용한 LSTM의 출력 분류 등
  - **시퀀스를 위한 합성곱 연산(학습속도 향상)**
  - **어텐션 모델(주의모델)의 주도**
  - **전이학습 활용**