Language Modeling

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Language Modeling

- Objective:

$$\mathcal{D} = \{x^i\}_{i=1}^{N}$$

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(x_{1:n}^i; \theta)$$

$$\text{where } x_{1:n} = \{x_1, \cdots, x_n\}.$$

Chain Rule

- We can convert joint probability to conditional probability.

$$
\begin{aligned}
P(A,B,C,D) &= P(D|A,B,C)P(A,B,C) \\
&= P(D|A,B,C)P(C|A,B)P(A,B) \\
&= P(D|A,B,C)P(C|A,B)P(B|A)P(A)
\end{aligned}
$$

By Chain Rule,

- We can re-write the equation,

$$
\begin{aligned}
P(x_{1:n}) &= P(x_1, \cdots, x_n) \\
&= P(x_n | x_1, \cdots, x_{n-1}) \cdots P(x_2 | x_1) P(x_1) \\
&= \prod_{i=1}^{n} P(x_i | x_{<i})
\end{aligned}
$$

$$
\log P(x_{1:n}) = \sum_{i=1}^{N} \log P(x_i | x_{<i})
$$

Chain Rule Example

$$
\begin{aligned}
P(A,B,C,D) &= P(D|A,B,C)P(A,B,C) \\
&= P(D|A,B,C)P(C|A,B)P(A,B) \\
&= P(D|A,B,C)P(C|A,B)P(B|A)P(A)
\end{aligned}
$$

$$
\begin{aligned}
P(\text{<BOS>, I, love, to, play, <EOS>}) &= P(\text{<EOS>}|\text{<BOS>, I, love, to, play})P(\text{<BOS>, I, love, to, play}) \\
&= P(\text{<EOS>}|\text{<BOS>, I, love, to, play})P(\text{play}|\text{<BOS>, I, love, to})P(\text{<BOS>, I, love, to}) \\
&= P(\text{<EOS>}|\text{<BOS>, I, love, to, play})P(\text{play}|\text{<BOS>, I, love, to})P(\text{to}|\text{<BOS>, I, love})P(\text{<BOS>, I, love}) \\
&= P(\text{<EOS>}|\text{<BOS>, I, love, to, play})P(\text{play}|\text{<BOS>, I, love, to})P(\text{to}|\text{<BOS>, I, love})P(\text{love}|\text{<BOS>, I})P(\text{<BOS>, I}) \\
&= P(\text{<EOS>}|\text{<BOS>, I, love, to, play})P(\text{play}|\text{<BOS>, I, love, to})P(\text{to}|\text{<BOS>, I, love})P(\text{love}|\text{<BOS>, I})P(\text{I}|\text{<BOS>})P(\text{<BOS>})
\end{aligned}
$$

By Chain Rule,

- We can re-write objective,

$$
\begin{aligned}
\mathcal{D} &= \{x^i\}_{i=1}^N \\
\hat{\theta} &= \underset{\theta\in\Theta}{\text{argmax}} \sum_{i=1}^N \log P(x_{1:n}^i;\theta) \\
&= \underset{\theta\in\Theta}{\text{argmax}} \sum_{i=1}^N \sum_{j=1}^n \log P(x_j^i|x_{<j}^i;\theta) \\
&\text{where } x_{1:n} = \{x_1,\cdots,x_n\}.
\end{aligned}
$$

Using Language Model

- Pick better(fluent) sentence.

- Predict next word given previous words.

$$\hat{x}_t = \underset{\mathbf{x}_t \in \mathcal{X}}{\operatorname{argmax}} \log P(\mathbf{x}_t | x_{<t}; \theta)$$

Summary

- 언어모델은 주어진 코퍼스 문장들의 likelihood를 최대화 하는 파라미터를 찾아내, 주어진 코퍼스를 기반으로 언어의 분포를 학습한다.
  - 즉, 코퍼스 기반으로 문장들에 대한 확률 분포 함수를 근사(approximate)한다.

- 문장의 확률은 단어가 주어졌을 때, 다음 단어를 예측하는 확률을 차례대로 곱한 것과 같다.

- 따라서 언어모델링은 주어진 단어가 있을 때, 다음 단어의 likelihood를 최대화하는 파라미터를 찾는 과정이라고도 볼 수 있다.
  - 주어진 단어들이 있을 때, 다음 단어에 대한 확률 분포 함수를 근사하는 과정