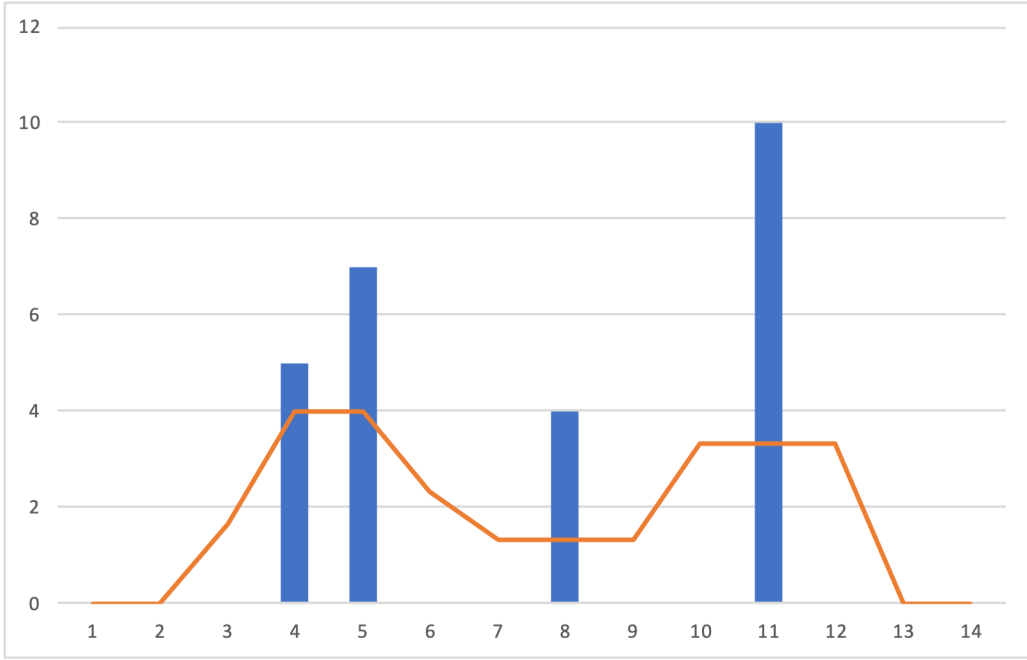
Smoothing and Discounting

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Smoothing

- Markov assumption을 도입하였지만 여전히 문제는 남아있음
- Training corpus에 없는 unseen word sequence의 확률은 0?
- Unseen word sequence에 대한 대처
  - Smoothing or Discounting
- Popular algorithm
  - Modified Kneser-Ney Discounting



Add One Smoothing

- To prevent count becomes zero:

$$
\begin{aligned}
P(w_t|w_{<t}) &\approx \frac{\mathrm{C}(w_{1:t})}{\mathrm{C}(w_{1:t-1})} \\
&\approx \frac{\mathrm{C}(w_{1:t}) + 1}{\mathrm{C}(w_{1:t-1}) + |V|},
\end{aligned}
$$

where $|V|$ is a size of vocabulary.

Generalization of Add One Smoothing

- If we generalize this:

$$
\begin{aligned}
P(w_t|w_{<t}) &\approx \frac{\mathrm{C}(w_{1:t})}{\mathrm{C}(w_{1:t-1})} \\
&\approx \frac{\mathrm{C}(w_{1:t}) + 1}{\mathrm{C}(w_{1:t-1}) + |V|} \\
&\approx \frac{\mathrm{C}(w_{1:t}) + k}{\mathrm{C}(w_{1:t-1}) + k \times |V|} \\
&\approx \frac{\mathrm{C}(w_{1:t}) + \frac{m}{|V|}}{\mathrm{C}(w_{1:t-1}) + m},
\end{aligned}
$$

where $|V|$ is a size of vocabulary.

- Take more generalization:

$$
P(w_t|w_{<t}) \approx \frac{\mathrm{C}(w_{1:t}) + m \times P(w_t)}{\mathrm{C}(w_{1:t-1}) + m},
$$

where $P(w_t)$ is unigram probability.

Kneser-Ney Discounting

- In this lecture,
  - $C(\text{learning}) > C(\text{laptop})$
  - Because of “deep learning”, “machine learning”

- 다양한 단어 뒤에서 나타나는 단어일수록
  unseen word sequence에 등장 할 확률이 높지 않을까?
  - 앞에 등장한 단어의 종류가 다양할 수록 해당 확률이 높을 것 같음

$$P_{continuation}(w) \propto |\{v : C(v, w) > 0\}|$$

Summary

Markov Assumption

- Count 기반의 approximation
- 긴 word sequence는 학습 코퍼스에 존재하지 않을 수 있음
  - 확률 값이 0으로 맵핑
- Markov assumption을 통해 근거리의 단어만 고려

Smoothing and Discounting

- Markov assumption을 통해서도 여전히 확률 값이 0이 될 수 있음
- Smoothing 또는 discounting을 통해 현상을 완화
- 여전히 unseen word sequence에 대한 대처는 미흡