Perplexity: How to evaluate LM

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

How to Evaluate

- Test set
  ① 나는 학교에 갑니다.
  ② 나는 학교를 갑니다.

- Intrinsic evaluation(정성평가)
  - 정확함
  - 시간과 비용이 많이 들어감

- Extrinsic evaluation(정량평가)
  - 시간과 비용을 아낄 수 있음
  - Intrinsic evaluation과 비슷할 수록 좋은 방법!

What is Good Language Model?

- 실제 사용하는 언어의 분포를 가장 잘 근사한 모델
  - 실제 사용하는 언어 → 테스트 시의 입력 문장들
  - 분포를 잘 근사 → 문장의 likelihood가 높을 것

- 잘 정의된 테스트셋의 문장에 대해서 높은 확률을 반환하는 언어모델이 좋은 모델!

Evaluation

- Perplexity (PPL)
  - 테스트 문장에 대해서 언어모델을 이용하여 확률(likelihood)을 구하고
  - PPL 수식에 넣어 언어모델의 성능 측정
    - 문장의 확률을 길이에 대해서 normalization (기하평균)

$$\begin{aligned}\mathrm{PPL}(x_1, \cdots, x_n; \theta) &= P(x_1, \cdots, x_n; \theta)^{-\frac{1}{n}} \\ &= \sqrt[n]{\frac{1}{P(x_1, \cdots, x_n; \theta)}}\end{aligned}$$

Evaluation

- Chain rule에 의해서

$$
\begin{aligned}
\mathrm{PPL}(x_1, \cdots, x_n; \theta) &= P(x_1, \cdots, x_n; \theta)^{-\frac{1}{n}} \\
&= \sqrt[n]{\frac{1}{P(x_1, \cdots, x_n; \theta)}} \\
&= \sqrt[n]{\frac{1}{\prod_{i=1}^{n} P(x_i | x_{<i}; \theta)}}
\end{aligned}
$$

Evaluation

- Markov assumption이 적용 될 경우,

$$
\begin{aligned}
\mathrm{PPL}(x_1, \cdots, x_n; \theta) &= P(x_1, \cdots, x_n; \theta)^{-\frac{1}{n}} \\
&= \sqrt[n]{\frac{1}{P(x_1, \cdots, x_n; \theta)}} \\
&= \sqrt[n]{\frac{1}{\prod_{i=1}^{n} P(x_i | x_{<i}; \theta)}} \\
&\approx \sqrt[n]{\frac{1}{\prod_{i=1}^{n} P(x_i | x_{i-1}, \cdots, x_{i-k}; \theta)}}
\end{aligned}
$$

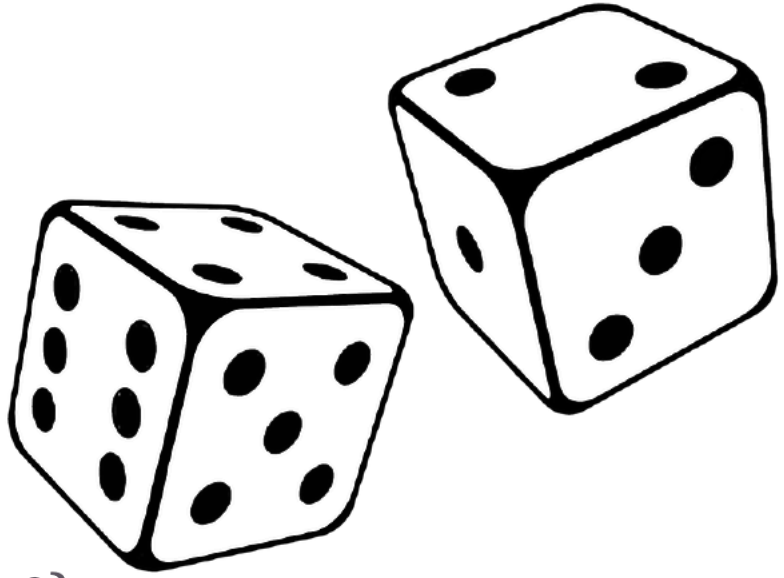
Perplexity

- 테스트 문장에 대해서 **확률을 높게 반환할수록** 좋은 언어모델

- 테스트 문장에 대한 **PPL이 작을수록** 좋은 언어모델

Perplexity

- 주사위를 던져 봅시다.
  - 1부터 6까지의 6개의 숫자로 이루어진 수열
  - 1부터 6까지 6개의 숫자의 출현 확률은 모두 같다



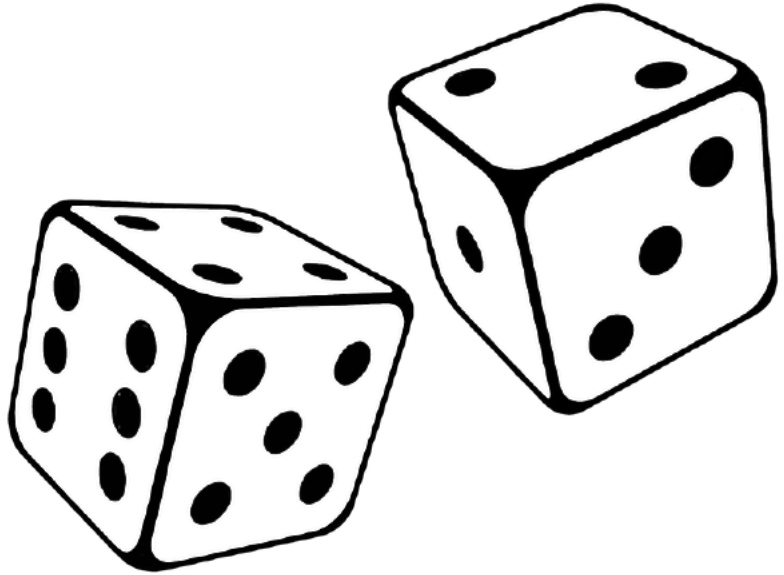
  - uniform distribution

$$\mathcal{D} = \{x_i\}_{i=1}^{n}, \text{ where } x_i \sim P(\mathrm{x}) \text{ and } \forall x \in \{1, 2, 3, 4, 5, 6\}.$$

$$\begin{aligned}
\mathrm{PPL}(x_1, \cdots, x_n) &= \sqrt[n]{\frac{1}{P(x_1, \cdots, x_n)}} \\
&= \sqrt[n]{\frac{1}{\prod_{i=1}^{n} P(x_i)}} \\
&= \sqrt[n]{\frac{1}{(\frac{1}{6})^n}} = 6
\end{aligned}$$

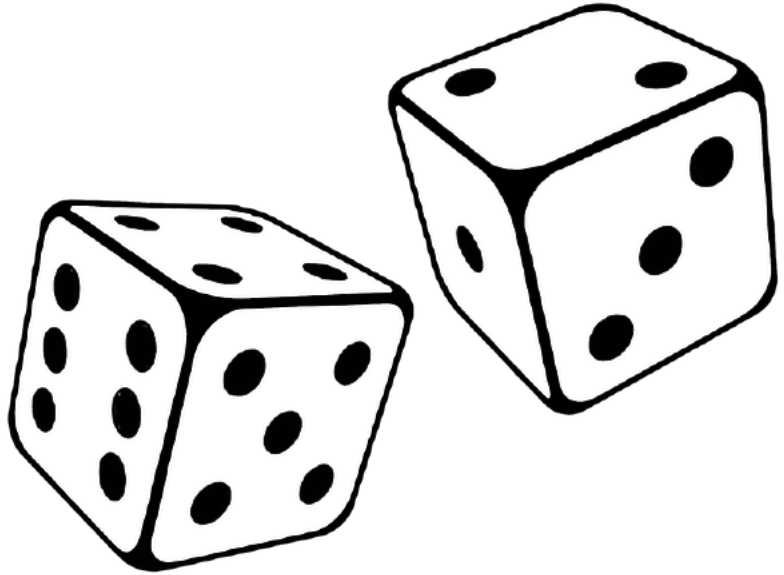
Perplexity

- Perplexity를 해석하는 방법
  - 주사위 PPL: 매 time-step 가능한 가짓수인 6
  - 뻗어나갈 수 있는 branch(가지)의 숫자를 의미
  - Time-step 별 평균 branch의 수



Perplexity

- Perplexity를 해석하는 방법
  - 주사위 PPL: 매 time-step 가능한 가짓수인 6
  - 뻗어나갈 수 있는 branch(가지)의 숫자를 의미
  - Time-step 별 평균 branch의 수
  - PPL이 낮을 수록 확률 분포가 Sharp 하다.
  - PPL이 높을 수록 확률 분포가 Flat 하다.



Summary

- 좋은 언어모델:
  - 잘 정의된 테스트셋 문장에 대해서 높은 확률(=낮은 PPL)을 갖는 모델

- Perplexity (PPL)
  - Lower is better
  - 확률의 역수에 문장 길이로 기하 평균
  - 매 time-step 마다 평균적으로 헷갈리고(no clue) 있는 단어의 수