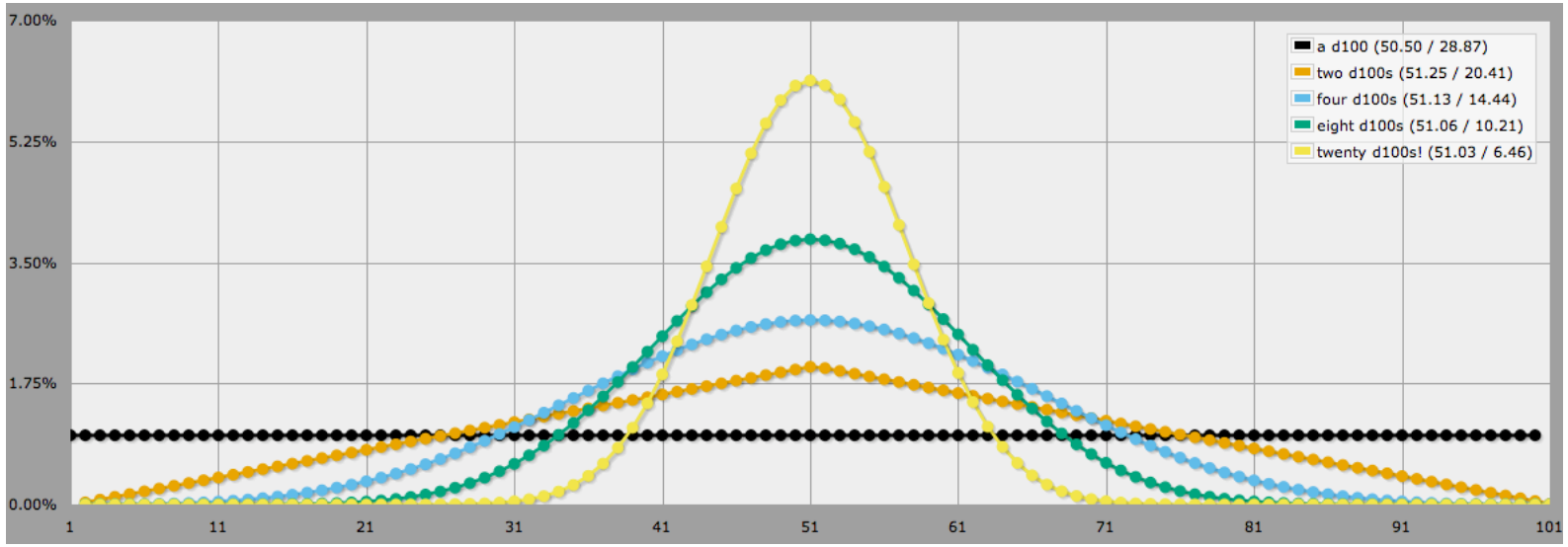
Perplexity and Entropy

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Perplexity

- Sharp vs Flat distribution



Information and Entropy

- 정보이론에서 엔트로피는 어떤 정보의 불확실성을 나타냄

- 불확실성은 일어날 것 같은 사건(likely event)의 확률
  - 자주 발생하는(일어날 확률이 높은) 사건은 낮은 정보량을 가진다.
  - 드물게 발생하는(일어날 확률이 낮은) 사건은 높은 정보량을 가진다.

- 불확실성 $\propto$ 1/확률 $\propto$ 정보량

Information and Entropy

① 내일 아침에는 해가 동쪽에서 뜬다.

② 내일 아침에는 해가 서쪽에서 뜬다.

a. 대한민국 올 여름의 평균 기온은 섭씨 28도로 예상 된다.

b. 대한민국 올 여름의 평균 기온은 섭씨 5도로 예상 된다.

Information and Entropy

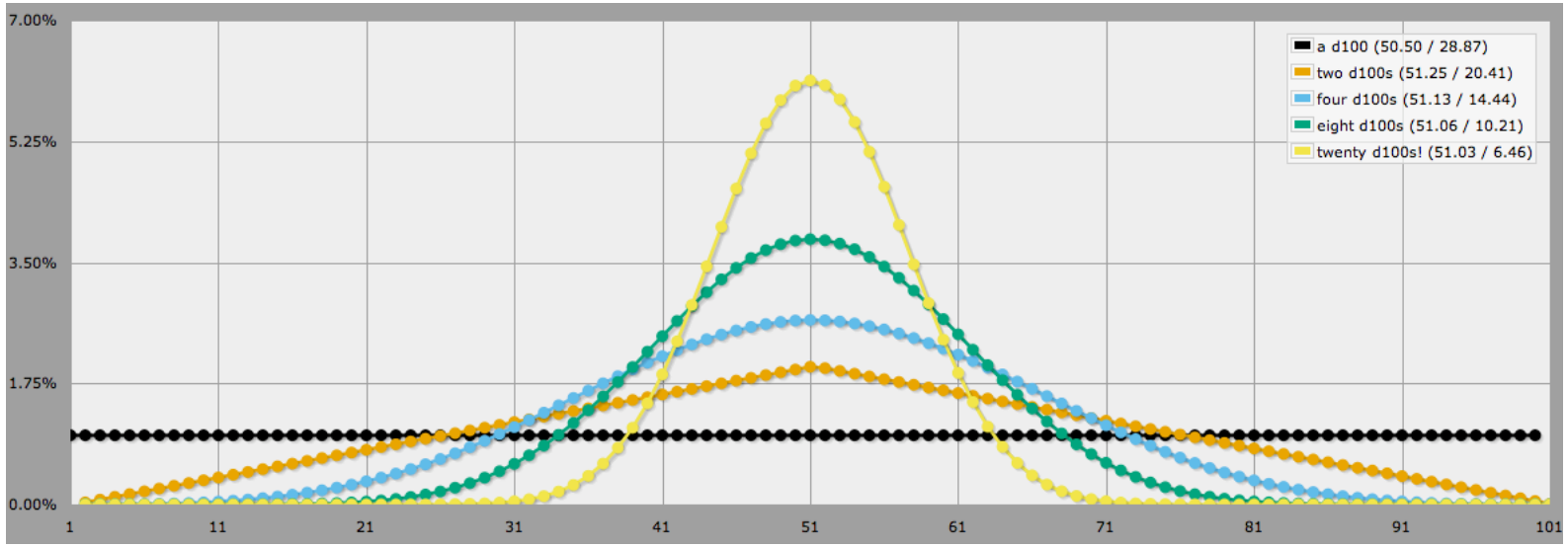
- 정보량
  - $-\log$ 때문에, 확률이 0에 가까워질수록 높은 정보량

$$I(x) = -\log P(x)$$

- 언어모델 관점
  - 흔히 나올 수 없는 문장(확률이 낮은 문장)일수록 더 높은 정보량

Entropy

- Sharp vs Flat distribution



Perplexity

- 확률값 역수의 기하평균

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

Entropy and Perplexity

- Cross Entropy

$$
\begin{aligned}
H(P, P_\theta) &= -\mathbb{E}_{x_{1:n} \sim P}[\log P(x_{1:n}; \theta)] \\
&\approx -\frac{1}{n} \sum_{x_{1:n} \in \mathcal{X}} P(x_{1:n}) \log P(x_{1:n}; \theta), \text{ defined as per-word entropy} \\
&\approx -\frac{1}{n \times N} \sum_{i=1}^{N} \log P(x^i_{1:n}; \theta), \text{ by Monte-carlo} \\
&\approx -\frac{1}{n} \log P(x_{1:n}; \theta), \text{ where } N = 1 \\
&\approx -\frac{1}{n} \sum_{i=1}^{n} \log P(x_i | x_{<i}; \theta) \\
&= \mathcal{L}(x_{1:n}; \theta)
\end{aligned}
$$

Entropy and Perplexity

$$
\begin{aligned}
\mathcal{L}(x_{1:n};\theta) &\approx -\frac{1}{n}\sum_{i=1}^{n}\log P(x_i|x_{<i};\theta) \\
&= -\frac{1}{n}\log\prod_{i=1}^{n}P(x_i|x_{<i};\theta) \\
&= \log\sqrt[n]{\frac{1}{\prod_{i=1}^{n}P(x_i|x_{<i};\theta)}} \\
&= \log \mathrm{PPL}(x_{1:n};\theta)
\end{aligned}
$$

Summary

- Objective: minimize perplexity
  - equivalent to minimize cross entropy
  - is also same as minimizing negative log-likelihood

- 문장의 likelihood를 maximize하는 파라미터를 찾고 싶음
  - Ground-truth 확률 분포(실제 사람이 가진 언어 모델)에
    언어모델을 근사(approximate)하고 싶음

- GT 분포와 LM 분포 사이의 cross entropy를 구하고 minimize.
  - 문장의 perplexity를 minimize.