Score Metric: BLEU

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

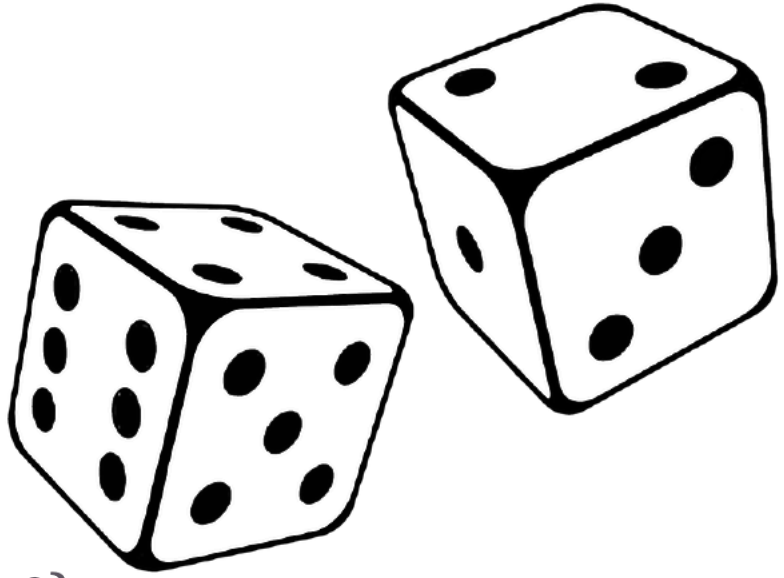
Perplexity

- 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



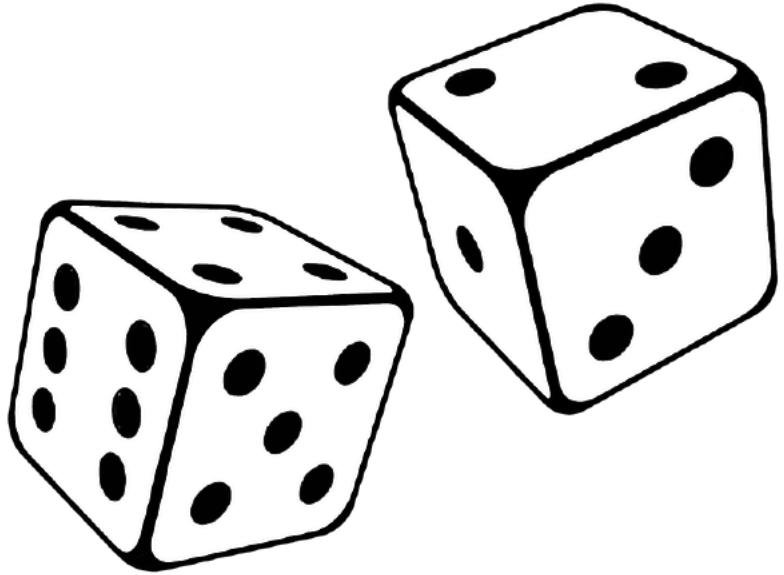
  - uniform distribution

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

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

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
  - PPL이 낮을 수록 확률 분포가 Sharp 하다.
  - PPL이 높을 수록 확률 분포가 Flat 하다.



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

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

However,

- PPL(Cross Entropy)는 정확한 번역의 품질을 반영하지 못함
  - 특히 어순의 변화에 취약함

원문	I	love	to	go	to	school	.
ref	나는	학교에	가는	것을	좋아한다	.	
hyp1	학교에	가는	것을	좋아한다	나는	.	
hyp2	나는	오락실에	가는	것을	싫어한다	.	

BLEU

- 각 n-gram 별 precision의 가중 평균

$$\text{BLEU}(\hat{y}, y) = \text{brevity\_penalty}(\hat{y}, y) \times \prod_{n=1}^{N} p_n^{w_n},$$

$$\text{where brevity\_penalty}(\hat{y}, y) = \min\left(1, \frac{|\hat{y}|}{|y|}\right)$$

$$\text{and } p_n^{w_n} \text{ is precision of } n\text{-gram with weight } w_n = \frac{1}{2^n}.$$

BLEU Example: 2-gram Precision

Hyp1

2-gram	count	hit
<BOS> 학교에	1	0
학교에 가는	1	1
가는 것을	1	1
것을 좋아한다	1	1
좋아한다 나는	1	0
나는 .	1	0
. <EOS>	1	1
합계	**7**	**4**

Hyp2

2-gram	count	hit
<BOS> 나는	1	1
나는 오락실에	1	0
오락실에 가는	1	0
가는 것을	1	1
것을 싫어한다	1	0
싫어한다 .	1	0
. <EOS>	1	1
합계	**7**	**3**

Summary

- Perplexity (Cross Entropy)
  - Lower is better
- BLEU
  - Higher is better

- 하지만 BLEU도 유의어/동의어 등에 대한 대처는 떨어짐

- Open source for BLEU
  - for research:
    - https://github.com/google/seq2seq/blob/master/bin/tools/multi-bleu.perl
  - for integration with python:
    - https://www.nltk.org/_modules/nltk/translate/bleu_score.html