Perplexity

正田 備也

masada@rikkyo.ac.jp

良い言語モデルとは？

- 未知の文を最も良く予測できる言語モデル
- その未知の文に最も高い確率を与える言語モデル
- 文Wのperplexityは以下のように定義される：

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \dots w_{i-1})}}$$

perplexityの直感的な意味

- perplexityは、直後に現れる単語の候補をどれだけ絞り込めているかを表す
- perplexityが小さいほど、候補を絞り込めている
- つまり、 perplexityが小さいほど、より良い予測ができている

良くない言語モデルの例

- 簡単のため、０～９の数字だけからなる文を考える
- どんなコンテクスト$w_1 \ldots w_{i-1}$に対しても、全ての数字に等しく1/10の確率を割り振るような言語モデルを考える
  - つまり、どんなコンテクスト$w_1 \ldots w_{i-1}$に対しても、また、どんな数字w_iに対しても $P(w_i|w_1 \ldots w_{i-1}) = \frac{1}{10}$と予測するような言語モデル
- このときperplexityは10
  - 次にどの数字が来るかについて、候補を全く絞り込めていない！

cross entropyとの関係

- RNNやTransformersを使って得られる各トークンの予測確率、つまり$P(w_i|w_1 \dots w_{i-1})$を先ほどの式に当てはめればよい
- これは、cross entropyの計算と実質的には同じ計算

$$CE(W) = \frac{1}{N}\sum_{i=1}^{N}(-\log P(w_i|w_1 \dots w_{i-1}))$$

$$= \log \sqrt[N]{\prod_{i=1}^{N}\frac{1}{P(w_i|w_1 \dots w_{i-1})}} = \log PP(W)$$

自然言語は非ゼロのエントロピーを持つ

- これまでの議論からすると、perplexityが1の言語モデルが最高の言語モデルのように思えるが・・・
- 自然言語そのものが非ゼロのエントロピーを持つ
  - https://arxiv.org/abs/2001.08361
- つまり、1より大きいperplexityを持つ
  - これを、自然言語の冗長性と呼んだりもする。
- perplexityは、言語モデルの唯一正しい評価尺度だろうか？
- 言語モデルは、自然言語の唯一正しいモデルだろうか？