Penalty

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Motivations

- 확률이 높은 문장을 생성하는 것이 목표

- 긴 문장일수록 확률이 작아짐

Length Penalty

- 길이(n)가 긴 문장일수록 큰 음의 값(로그 확률 값)을 지닐 것
  - 따라서 긴 문장에 대해 더 작은 값을 곱해주어, 큰 확률 값이 되도록 함

$$\log \tilde{P}(\hat{y}_{1:n}|x_{1:m};\theta) = \log P(\hat{y}_{1:n}|x_{1:m};\theta) \times \text{penalty}(n)$$
$$\text{where } \hat{y}_{1:n} \sim \log P(\cdot|x_{1:m};\theta) \text{ and}$$
$$\text{penalty}(n) = \left(\frac{1+\beta}{1+n}\right)^{\alpha}.$$

Coverage Penalty

- 디코더의 각 time-step마다 attention이 다른 곳에 집중해야 좋은 번역일 것
  - 골고루 attention이 분배되어야 함

$$\log \tilde{P}(\hat{y}_{1:n}|x_{1:m};\theta) = \log P(\hat{y}_{1:n}|x_{1:m};\theta) \times \text{penalty}_{\text{length}}(n) + \text{penalty}_{\text{coverage}}(x_{1:m}, \hat{y}_{1:n})$$

$$\text{penalty}_{\text{coverage}}(x_{1:m}, \hat{y}_{1:n}) = \beta \times \sum_{i=1}^{m} \log\left(\min(\sum_{j=1}^{n} w_{i,j}, 1.0)\right),$$

$$\text{where } w_{i,j} = \text{softmax}(h_j^{\text{dec}} \cdot W_{\text{a}} \cdot h_i^{\text{enc}T}).$$

Wrap-up

- 각 penalty 마다 hyper-parameter가 존재
  - Google의 논문 또는 open-source를 참고
  - 튜닝에 따른 성능의 변화가 미미함