Wrap-up

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

PPL cannot reflect correct quality

- PPL은 left-to-right 관점에 다음 time-step의 단어에 대한 확률 분포를 수치화
  - 따라서 문장 전체의 의미를 아우르는 문장 생성 task에서는 적절하지 못한 metric

- 물론 거시적인 관점에서는 낮은 PPL이 더 좋은 모델 성능을 가리킨다.
  - 하지만 미시적인 관점에서, 작은 차이의 PPL로는 모델의 우열을 가리기 힘듦

- 그럼 BLEU를 maximize하도록 하면 안될까?
  - 미분 불가능한 함수이므로 MLE에 적용하기 어려움

Wrap-up

- 테스트셋 선정은 매우 중요한 문제
  - 난이도
  - 적합성 (e.g. 도메인)

- 정량 평가(extrinsic evaluation)를 통해 최종 후보 모델을 선정하고,
  정성 평가(intrinsic evaluation)를 통해 최종 배포를 결정
  - 정성 평가 시에는 일관성과 객관성(e.g. blind test) 확보가 가장 중요