Evaluation

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Evaluation?

- Test set의 적격성
  - 적절한 난이도인가?
    - e.g. 불수능 vs 물수능
  - Noise가 포함되지 않았는가?
  - 실제 deploy 할 때와 같은 domain인가?

- Scoring Metric의 정확도
  - 해당 task를 채점하기에 적절한 metric인가?
    - e.g. Discrimination vs Generation

- 다양한 정답이 가능한 경우
  - 기계번역: 번역 문장의 정답은 굉장히 다양함
  - 챗봇(QnA): 질문에 대한 대답은 굉장히 다양함

Evaluation Methods

Intrinsic Evaluations

- Manual (Human evaluation)
- Pros
  - 정확함
- Cons
  - 큰 비용
  - 느린 속도
  - 주관 개입 가능

- 절대 평가
- 상대 평가

Extrinsic Evaluations

- Automatic
- Pros
  - 저렴한 비용
  - 빠른 속도
  - 객관적 평가 가능
- Cons
  - 정확도가 낮을 수 있음

- Loss (cross entropy & PPL)
- BLEU, METEOR, ROUGE

Summary

- 많은 고민/노력을 통해 test set을 만들어야 함
  - 적격성(e.g. domain, 난이도)
  - 정확도(e.g. noise 여부, 채점 방식)

- 실제 deploy를 위해서는 intrinsic evaluation을 꼭 거칠 것
  - 잦은 모델 업데이트: extrinsic evaluation
  - 현재 업데이트 버전의 최종 모델 선정 후, intrinsic evaluation을 통해 모델 배포 여부 결정