A Survey on Machine Reading Comprehension: Tasks, Evaluation Metrics and Benchmark Datasets -metrics-

허 진규

Accuracy(=Exact Match)

$$\text{Accuracy} = \frac{M}{N} \qquad \text{Exact Match} = \frac{M}{N}$$

질문의 개수 N, 맞춘 개수 M

- 거의 span prediction task에만 사용

Precision & Recall

- Precision : 전체 예측한 값에서 얼마나 정답을 맞췄는 지를 판별
- Recall : 예측한 값이 정답 값이랑 얼마나 일치하는 지를 판별
- 질문 query에서 punctuation과 article words(a, an, the ...) 빼고 계산

Precision

- Token-level precision
  - TP = pred answer과 correct answer 둘 다 있는 경우
  - FP = correct answer에는 없고 pred answer에는 있는 경우
  - e.g. “a cat in the garden”, “a dog in the garden”
  - ‘a’, ‘the’ 를 빼고 3개의 토큰 중에 일치하는 건 ‘in’, ‘garden’ 2개 이므로 2/3가 된다.

$$Precision_{TS} = \frac{Num(TP_T)}{Num(TP_T) + Num(FP_T)}$$

Precision

- Question-level precision
  - 1개의 answer가 1개의 entity, 몇 단어인지는 상관 없음
  - TP = correct answer과 pred answer에 둘 다 있는 entity
  - FP = pred answer에는 있지만, correct answer에는 없는 entity

$$Precision_Q = \frac{Num(TP_Q)}{Num(TP_Q) + Num(FP_Q)}$$

Recall

- Token-level Recall
  - TF = 예측한 값이 정답인 token
  - FN = 정답이지만 예측에 실패한 token

$$Recall_{TS} = \frac{Num(TP_T)}{Num(TP_T) + Num(FN_T)}$$

Recall

- Question-level Recall
  - TP = 정답이고 예측에 성공한 entity
  - FN = 정답이지만 예측에 실패한 entity

$$Recall_Q = \frac{Num(TP_Q)}{Num(TP_Q) + Num(FN_Q)}$$

F1 score

- Token-level F1 score
  - MRC task에서 일반적으로 사용 되는 evaluation metrics
  - single question에서 사용된다.

$$F1_{TS} = \frac{2 \times Precision_{TS} \times Recall_{TS}}{Precision_{TS} + Recall_{TS}}$$

- 더 신뢰도 높은 연산을 위해, 평균 값을 구해준다.
  - 분자는 모든 질문들에 대해서 가장 큰 token-level F1값들의 합을 의미한다.
  - 분모는 질문의 개수를 의미한다.

$$F1_T = \frac{\sum Max\left(Precision_{TS}\right)}{Num(Questions)}$$

F1 score

- Question-level F1 score
  - Token level과 연산은 동일하나 question-level의 값들을 가지고 연산한다.

$$F1_Q = \frac{2 \times Precision_Q \times Recall_Q}{Precision_Q + Recall_Q}$$

ROUGE(Recall-Oriented Understudy for Gisting Evaluation)

- Text summary system의 성능을 평가하기 위해 만들어졌다.
- n-gram에서 reference(Gold standard)와 candidate(모델 생성) 사이의 recall을 구한다.
  - n은 n-gram의 길이를 의미
  - Count(gram)은 n-gram이candidate text, predicted text에서 가장 많이 나온 횟수

$$\text{ROUGE-N} = \frac{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count(gram_n)}$$

BLEU (Bilingual Evaluation Understudy)

- Machine translation system의 성능을 평가하기 위해 만들어짐
- BP (Brevity Penalty) : 문장 길이에 대한 과적합 보정
- Clip : 같은 단어가 연속적으로 나올 때 과적합 되는 것을 보정

- Pn = modified n-gram precision
  - (중복 제거 n-gram 개수) / (n-gram 개수)
- Wn = weight
- N = n-gram 길이

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c = candidate sentence 길이
r = effective reference corpus 길이

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Meteor

- machine translation system 평가하기 위해 고안
- Fmean = precision과 recall의 조화 평균에 weight를 주었다.
- m = 전체 matched words 수
- ch = chunk의 수

$$F_{mean} = \frac{Precision \times Recall}{\alpha \times Precision + (1-\alpha) \times Recall}$$

$$Meteor = F_{mean} \times (1 - Penalty) \qquad Penalty = \gamma \times \left(\frac{ch}{m}\right)^{\beta}$$

HEQ (Human Equivalence Score)

- 대화식 기계 독해 dataset에 사용 e.g. QuAC
- 한 개의 질문에 대해서 human level F1 만큼 헷갈리는 질문이 많은 경우에 유용하다.
- questions의 수가 N, 헷갈리는 human level F1을 초과하는 질문의 수가 M

$$HEQ = \frac{M}{N}$$

question 타입에 따른 metrics 선호도