28-11–2024

Language model evaluation

- Extrinsic evaluation - application specific
- Intrinsic evaluation - independent of any application
- Perplexity ($W = w_1 w_2 \ldots w_N$):
  - $PP(W) = p(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$
  - For 2-gram Model

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(w_i \mid w_{i-1})}}$$

Jurafsky & Martin “Speech and Language Processing, 3rd ed., 2023

Language translation evaluation

- Human evaluation
- Automatic evaluation
- Things to be consider:
  - Adequacy/faithfulness/fidelity
    - How well the translation capture the exact meaning of the source sentence
  - Fluency
    - How fluent the translation is in the target language
      - Grammar, readable, natural

Language translation evaluation (cont.)

- Human evaluation
- Automatic evaluation
  - Bilingual Evaluation Understudy (BLEU), Papineni et al., ACL, 2002
  - $BLEU = BP \times \exp\{\sum_{n=1}^{N} W_n \log p_n\}$
  - BP - Brevity penalty:
    - 1 if $c > r$
    - $\exp(1 - r/c)$ if $c \leq r$
  - p_n - n-gram precision:
    - Number of candidate n-gram matched with the reference n-gram (m_1) divides by the total number of n-grams in the candidate translation (m)
  - W_n - weight factor
  - Original paper- Uniform weight and $N = 4$

BLEU- example

- Example-1:
  - Candidate: the the the the the the the.
  - Reference: the can is on the mat.
- Can you see issues?
- Metric for Evaluation of Translation with Explicit ORdering (METEOR), Banerjee and Davie, ACL, 2005

Language translation evaluation (cont.)

- Human evaluation
- Automatic evaluation
  - Character overlap: character F-score (charFβ)[1]
    - charP: percentage of character 1-gram, …, k-gram in the hypothesis that occur in the reference, averaged
    - charR: percentage of character 1-gram, …, k-gram in the reference that occur in the hypothesis, averaged
    - $\text{charF}\beta = (1+\beta^2)\frac{\text{charP} \times \text{charR}}{\beta^2\text{charP} + \text{charR}}$

[1]Maja Popovic, chrF, in WSMT, 2015

Language translation evaluation (cont.)

- Character overlap: character F-score (charFβ)
  - $\text{charF}\beta = (1+\beta^2)\frac{\text{charP} \times \text{charR}}{\beta^2\text{charP} + \text{charR}}$
- Example:
  - REF: witness for the past,
  - HYP1: witness of the past,
  - HYP2: past witness
  - witnessforthepast, (18 1-grams, 17 2-grams)
  - witnessofthepast, (17 1-grams, 16 2-grams)
  - 1-gram match: 17
  - 2-gram match: 13
  - 1-gramP: 17/17, 1-gramR: 17/18
  - 2-gramP: 13/16, 2-gramR: 13/17
  - charP = (17/17 + 13/16)/2
  - charR = (17/18 + 13/17)/2
  - charF2, 2(REF, HYP1) = 0.86
  - charF2, 2(REF, HYP2) = 0.62

Example: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

Language translation evaluation (cont.)

- Character overlap: character F-score (charFβ)
  - $\text{charF}\beta = (1+\beta^2)\dfrac{\text{charP} \times \text{charR}}{\beta^2\text{charP} + \text{charR}}$
- Limitation:
  - a good translation may use alternate words or paraphrases
- Solution?
  - Word embedding?
    - reference translation: $x = (x_1, x_2, \cdots, x_n)$
    - candidate machine translation: $\bar{x} = (\bar{x_1}, \bar{x_2}, \cdots, \bar{x_m})$
    - human rating: r
    - Train a model[1,2] to predict r based on x and $\bar{x}$
    - Models try to correlates with human labels

Example: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

[1]Rei et al., COMET, in EMNLP. 2020

[2]Sellam et al. BLEURT, in ACL 2020

Language translation evaluation (cont.)

- If human rating is not available!
  - Happen many cases
- Solution?
  - Word embedding?
    - reference translation: $x = (x_1, x_2, \cdots, x_n)$; x_i is a word embedding
    - candidate machine translation: $\bar{x} = (\bar{x_1}, \bar{x_2}, \cdots, \bar{x_m})$
    - Define a similarity[1] between x and $\bar{x}$ as
      - $$Precision_{BERT} = \frac{1}{|\bar{x}|_0} \sum_{\bar{x_j} \in \bar{x}} max_{x_i \in x} x_i \cdot \bar{x}_j$$
      - $$Recall_{BERT} = \frac{1}{|x|_0} \sum_{x_j \in x} max_{\bar{x_j} \in \bar{x}} x_i \cdot \bar{x}_j$$

Example: Jurafsky & Martin "Speech and Language Processing, 3rd ed., 2023

[1] Zhang et al., BERTScore, in ICLR. 2020

Automatic Evaluation of Summaries

- Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Lin, WAS, 2004
  - $$ROUGE - N = \frac{\sum_{S \in Ref.} \sum_{n-gram \in S} Count_{match}(n - gram)}{\sum_{S \in Ref.} \sum_{n-gram \in S} Count(n - gram)}$$
  - $Count_{match}(n - gram)$ - number of n-gram matched with the candidate and reference summaries

Course review