5. Tasks and Applications in NLP

Overview

By purpose:

- Capabilities: test key abilities (linguistic, social, cultural, etc.) of language understanding

  e.g., parts-of-speech tagging, parsing, commonsense
- Application: a use case with potential products in mind

  e.g., machine translation, question answering
- NLP + X: new dimensions of NLP

  e.g., multilingual, multimodal, social NLP etc.

By modeling:

- Classifcation: output is a categorical variable
- Structured prediction: output is a chain, a tree, a graph
- Generation: output is free-form text

Evaluation

BLEU

Weighted mean of n-gram precision for each n (typically up to 4):

$\exp(\sum_{i=1}^{N} w_n \log p_n)$

Brevity penalty:

$$BP = \begin{cases} 1 & \text{if candidate length} \geq \text{ref length} \quad (\text{no penalty}) \\ e^{1-R/C} & \text{otherwise} \quad (\text{downweight score}) \end{cases}$$

BLEU: putting everything together

$\text{BLEU} = BP \cdot \exp(\sum_{i=1}^{N} w_n \log p_n)$

$\log \text{BLEU} = \min(1 - \frac{R}{C}, 0) + \sum_{i=1}^{N} w_n \log p_n$

ROUGE

Task: given a candidate summary and a set of reference summaries, evaluate the quality of the candidate.

ROUGE-n: n-gram recall

ROUGE-L: measures longest common subsequence between a candidate and areference (doesn't require consecutive match)

- Precision = $\text{LCS}(c, r)/len(c)$
- Recall = $\text{LCS}(c, r)/len(r)$
- F-measure = $\frac{(1+\beta^2)PR}{\beta^2 P+R}$