NLP
Introduction to NLP

Parsing Evaluation
Parsing Model

- GEN/EVAL framework
- GEN maps the input to a set of candidate parses
- EVAL ranks the candidate parses
  \[ y^* = \arg\max_{y \in \text{GEN}(X)} \text{EVAL} (X,Y) \]
Evaluation Methodology (1/2)

• Classification tasks
  – Document retrieval
  – Part of speech tagging
  – Parsing

• Data split
  – Training
  – Dev–test
  – Test
Evaluation Methodology (2/2)

• Baselines
  – Dumb baseline
  – Intelligent baseline
  – Human performance (ceiling)

• New method

• Evaluation methods
  – Accuracy
  – Precision and Recall

• Multiple references
  – Interjudge agreement
Kappa

\[ \kappa = \frac{P(A) - P(E)}{1 - P(E)} \]

- **Agreement vs. expected agreement**
  - \( P(A) \) is the level of agreement of the judges
  - \( P(E) \) is the expected probability of agreement by chance

- **When \( \kappa > .7 \) – agreement is considered high**

- **Question**
  - Judge agreement on a binary classification task is 60%, is this high?
Answer

\[ \kappa = \frac{P(A) - P(E)}{1 - P(E)} \]

- **Data**
  - \( P(A) = .6 \)
  - \( P(E) = .5 \)

- **Kappa**
  - \( k = \frac{1}{.5} = .2 \)
  - not high
Parsing Evaluation

• Parseval: precision and recall
  – get the proper constituents
• Labeled precision and recall
  – also get the correct non-terminal labels
• F1
  – harmonic mean of precision and recall
• Crossing brackets
  – (A (B C)) vs ((A B) C)
• PTB corpus
  – training 02–21, development 22, test 23
Evaluation Example

GOLD = (S (NP (DT The) (JJ Japanese) (JJ industrial) (NNS companies))
     (VP (MD should) (VP (VB know) (ADVP (JJR better))))
     (.) (. )

CHAR = (S (NP (DT The) (JJ Japanese) (JJ industrial) (NNS companies))
     (VP (MD should) (VP (VB know)) ((ADVP (RBR better))))
     (.) (. )

Bracketing Recall = 80.00
Bracketing Precision = 66.67
Bracketing FMeasure = 72.73
Complete match = 0.00
No crossing = 100.00
Tagging accuracy = 87.50
NLP