Special Topics: Natural Language Processing
Adjective Task Results

Adam Meyers
New York University
2017
Method of Evaluation

- Cheng Ni (the grader) created a pooled result consisting of adjectives that were, in effect, voted for by most of the annotations. This pooled annotation was used as one answer key.
- My annotation (as the expert annotator) was used as another answer key.
- Annotation matched the answer key if the strings matched and the normal/comparative/superlative classification matched.
- Precision, Recall and F-measure were used for evaluation. The results were similar, but the score against the pooled annotation was slightly higher.
- Grade = 8 points plus 0-2 points based on how much F-measure exceeds 80%.
- All the statistics and histograms in this talk were provided by Cheng.
Prof's Annotation vs Pooled Result

• 43 adjectives marked by the professor were in the pool

• 1 adjective (easy) marked by the professor was not in the pool, i.e., only 2/28 students marked this one
Precision Against Pooled Results
Recall Against Pooled Results

![Graph showing recall against pooled results.](image)
F-measure Against Pooled Results
The Most Difficult Case

- I want to go **easy** on the traditional list of proposals
  - Not covered by the specifications
  - Part of a sort-of idiomatic expression (it is productive)
    - **go** + certain, but not all adjectives
    - **go** +
      
      easy/hard/soft/crazy/bonkers/bananas/ballistic/Dutch/futuristic/native
    
    - Key factors: it is productive, but only adjectives can fill in that slot
    - In this context, **go** means something like **be** or **become**, **verbs** that normally preceded predicative adjective position
Other Cases of Annotation Disagreement with Possible Reasons

• **What was true then can be true now** – Not Sure (16/28)
• **Raising the minimum wage** – Collocation or NN (18/28)
• **The right thing to do** – Unusual Construction (21/28)
• **Amazing medical breakthroughs** – Not Sure (21/28)
• **Paid leave** – Past Participle classification (21/28)
• **some bipartisan priorities** – Not Sure (22/28)
• **movements to expand civil rights** – collocation (22/28)
• **the next frontier** – Not Sure (23/28)
Procedures for Annotation Projects

• Do several test runs, to train annotators and to fine-tune specifications

• Last semester more students missed some examples, but I corrected parts of the specifications, e.g.,
  – most remote villages – not superlative

• Marking adjectives is not hard to get consistency

• Harder Tasks (for writing specifications):
  – Marking arguments of nouns (NomBank)
    • America has been through economic changes
  – Sentiment Analysis: positive/negative/neutral?
    • Rayovac lasted longer than Duracell
Conclusions

• Consistency is difficult because most tasks are interpreted differently by different annotators

• Specifications should:
  – Include details
  – Include examples
  – Address annotator's assumptions
  – Be tested against real world data and refined
  – Be tested with multiple annotators

• Annotators should be trained