Special Topics: Natural Language Processing
Adjective Task Results

Adam Meyers
New York University
2018
Method of Evaluation

- My annotation (as the expert annotator) was used as the 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.
- Grade from 1 to 10
  - Grade = round(F-measure/10)
- The diagrams were provided by John Ortega, the class assistant.
- Note that for the Part of Speech task on the Penn Treebank, annotators that have gone through training get about 97% agreement.
  - Other tasks have lower agreement figures.
Precision
Recall
F-measure
F-Score distribution
Error Analysis for 31 Student Annotations

• Missing adjectives > incorrectly marked adjectives

• Fewest “correctly” marked as adjective (with correct assignment of comparative or normal):
  – easy 7, criminal 14, other 15, true 16, last 17, minimum, 18, medical 18, working 19

• Most “incorrectly” marked as adjective:
  – fellow 24 – my fault: answer key differs from sample
  – more 9 – in determiner list – do not mark as per specs
Difficult cases

• *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

• *What was true then can be true now*
  – True is in predicative position
  – The subject is the interrogative pronoun *what*
    • A very special closed class word
    • Stands in for either a noun phrase or a sentence
      – The fact was true
      – That $1 + 1 = 2$ is true
Other Cases of Annotation Disagreement with Possible Reasons

- *last year, medical treatments* – Not sure
- *other leaders* – Unusual meaning for adjective, no predicative usage, possible noun
- *paid leave* – Past Participle classification? Or Collocation?
- *working families* – Pres Participle classification?
- *minimum wage* – collocation or noun?
- *civil rights* – collocation?
- *criminal justice* – collocation or noun?
- *fellow Americans* – *bad example file + noun vs adjective*
- *more people* – *determiner in specifications*
Compatibility Issues

• Version of Mae
  – Old Version of Mae did not work in some systems

• New version of Mae
  – https://github.com/keighrim/mae-annotation/releases
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

• Current specifications may cause annotators to undergenerate more often than overgenerate