Language & Statistics
Final Project

Callie Vaughn
Mario Piergallini
Tom van Drunen
Wes Feely

Approach

- We tagged and parsed the data using TurboParser
- Features
  - Document length (number of words)
  - # of DEP relations in parse trees in doc
    - DEP = unknown dependency relation
  - # of "bad" consecutive tags
    - DT-DT, CC-CC, POS-POS
    - Normalized by doc length
  - # of repeated words in doc
    - Normalized by doc length

Approach

- Features
  - Topic model features with Latent Dirichlet Allocation (LDA)
    - Model trained 20 topics on 30-line segments from the 100mil word corpus
    - Ran inference on the train and dev sets, outputs P(Topic|Doc) for each topic
    - Took the median topic probability
    - Possible to achieve >98% accuracy on the train set in cross-validation, but performance degrades significantly on the shorter documents in the dev set

Results

- We use linguistic features to train a classifier in Weka ML toolkit

Feature set	Accuracy
POS + Repeated words	56.5%
LDA Topic features + Doc Length	76.0%
Combination	85.0%

Questions?