Ecologically Valid Evaluation of Speech Summarization

Anthony McCallum
University of Toronto
mccallum@cs.toronto.edu
Spoken audio archives

• Recording equipment
• Compression and Storage
• Distribution
  – Widespread broadband Internet
How can we use archived speech?

• Given recorded meetings: How was a given decision made?
• Given webcasted lectures: I missed a lecture and want to get up to speed
• Given broadcast news: Tell me what happened
• Etc.
How do we accomplish this?

• Natural access to speech is linear
• We want an aid that can
  – Tell us the ‘gist’ of what is in the archive / webcast
  – Direct us to the portion of the archive where that content is
  – Be used in place of the original audio if time is limited
• Our solution: Summarize!
Extractive summarization

- Split audio into utterances
  - Separated by 200 ms pauses
- Choose a percentage (5-30%) of utterances
  - Most salient / important
  - Determined by acoustic and lexical features
- How do we know when we have chosen the correct utterances?
  - Evaluation
    - Intrinsic versus extrinsic
Intrinsic evaluation

• Summary quality judged based on the content of the summary compared to a gold standard

• Subjective

• “Interesting” results:
  – Simple baselines perform very well
    • Length baseline
    • Even when compared to prosodic features
Extrinsic evaluation

• Summary quality judged based on how well an external task can be completed given the summary

• Ecologically valid!
Our study: ecologically valid extrinsic evaluation

- Domain: university lectures
  - Spontaneous speech

- Task: quizzes
  - TA designed
  - Time constrained (to ensure that summaries are useful)
  - Similar to university quizzes or exams (therefore ecologically valid)
What will we accomplish?

- Provide an ecologically valid extrinsic evaluation framework for extractive speech summarization
  - Will prosodic features outperform simple baselines?

- Determine what a validated gold standard summary should contain
Thank you!