Alternative Summarization: Reviews & Speech

Ling 573
Systems and Applications
May 28, 2015

Roadmap

- Review summarization:
  - Basic approach
  - Learning what users want

  - Speech summarization:
    - Application of speech summarization
    - Speech vs Text

Sentiment Summarization

- Classic approach: (Hu and Liu, 2004)
- Summarization of product reviews (e.g. Amazon)
  - Identify product features mentioned in reviews
  - Identify polarity of sentences about those features
  - For each product,
    - For each feature,
      - For each polarity: provide illustrative examples

Example Summary

- Feature: picture
  - Positive: 12
    - Overall this is a good camera with a really good picture clarity.
    - The pictures are absolutely amazing - the camera captures the minutest of details.
    - After nearly 800 pictures I have found that this camera takes incredible pictures.
    - ...
- Negative: 2
  - The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
  - Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

Learning Sentiment Summarization

- Classic approach is heuristic:
  - May not scale, etc.

Learning Sentiment Summarization

- Classic approach is heuristic:
  - May not scale, etc.
- What do users want?

Learning Sentiment Summarization

- Classic approach is heuristic:
  - May not scale, etc.
- What do users want?
  - Which example sentences should be selected?
    - Strongest sentiment?
    - Most diverse sentiments?
    - Broadest feature coverage?

Review Summarization Factors

- Posed as optimizing score for given length summary
  - Using a sentence extractive strategy

Review Summarization Factors

- Posed as optimizing score for given length summary
  - Using a sentence extractive strategy
- Key factors:
  - Sentence sentiment score
  - Sentiment mismatch: b/t summary and product rating
  - Diversity:
    - Measure of how well diff't "aspects" of product covered
    - Related to both quality of coverage, importance of aspect

Review Summarization Models

- Sentiment Match (SM): Neg(Mismatch)
  - Prefer summaries w/sentiment matching product
  - Issue?

Review Summarization Models

- Sentiment Match (SM): Neg(Mismatch)
  - Prefer summaries w/sentiment matching product
  - Issue?
    - Neutral rating ➔ neutral summary sentences
  - Approach: Force system to select stronger sents first

Review Summarization Models

- Sentiment Match (SM): Neg(Mismatch)
  - Prefer summaries w/sentiment matching product
  - Issue?
    - Neutral rating ➔ neutral summary sentences
  - Approach: Force system to select stronger sents first
- Sentiment Match + Aspect Coverage (SMAC):
  - Linear combination of:
    - Sentiment intensity, mismatch, & diversity

Review Summarization Models

- Sentiment Match (SM): Neg(Mismatch)
  - Prefer summaries w/sentiment matching product
  - Issue?
    - Neutral rating ➔ neutral summary sentences
  - Approach: Force system to select stronger sents first
- Sentiment Match + Aspect Coverage (SMAC):
  - Linear combination of:
    - Sentiment intensity, mismatch, & diversity
  - Issue?
    - Optimizes overall sentiment match, but not per-aspect

Review Summarization Models

- Sentiment-Aspect Match (SAM):
  - Maximize coverage of aspects
    - *consistent* with per-aspect sentiment
  - Computed using probabilistic model
  - Minimize KL-divergence b/t summary, orig documents

Human Evaluation

- Pairwise preference tests for different summaries
  - Side-by-side, along with overall product rating
  - 1-4 symmetric preference
- Also collected comments that justify rating

Human Evaluation

- Pairwise preference tests for different summaries
  - Side-by-side, along with overall product rating
  - 1-4 symmetric preference
- Also collected comments that justify rating
- Usually some preference, but not significant
  - Except between SAM (better) and SMAC
  - And SMAC significantly better than LEAD baseline
    - (70% vs 25%)

Qualitative Comments

- Preferred:

Qualitative Comments

- Preferred:
  - Summaries with list (pro vs con)
- Disliked:

Qualitative Comments

- Preferred:
  - Summaries with list (pro vs con)
- Disliked:
  - Summary sentences w/o sentiment
  - Non-specific sentences
  - Inconsistency b/t overall rating and summary

Qualitative Comments

- Preferred:
  - Summaries with list (pro vs con)
- Disliked:
  - Summary sentences w/o sentiment
  - Non-specific sentences
  - Inconsistency b/t overall rating and summary
- Preferences differed depending on overall rating
  - Prefer SMAC for neutral vs SAM for extremes
    - (SAM excludes low polarity sentences)

Conclusions

- Ultimately, trained meta-classifier to pick model
  - Improved prediction of user preferences

Conclusions

- Ultimately, trained meta-classifier to pick model
  - Improved prediction of user preferences
- Similarities and contrasts w/TAC:
  - Similarities:

Conclusions

- Ultimately, trained meta-classifier to pick model
  - Improved prediction of user preferences
- Similarities and contrasts w/TAC:
  - Similarities:
    - Diversity ~ Non-redundancy
    - Product aspects ~ Topic aspects: coverage, importance
  - Differences:

Conclusions

- Ultimately, trained meta-classifier to pick model
  - Improved prediction of user preferences
- Similarities and contrasts w/TAC:
  - Similarities:
    - Diversity ~ Non-redundancy
    - Product aspects ~ Topic aspects: coverage, importance
  - Differences:
    - Strongly task/user oriented
    - Sentiment focused (overall, per-sentence)
    - Presentation preference: lists vs narratives

Speech Summarization

Speech Summary Applications

- Why summarize speech?
  - Meeting summarization
  - Lecture summarization
  - Voicemail summarization
  - Broadcast news
  - Debates, etc....

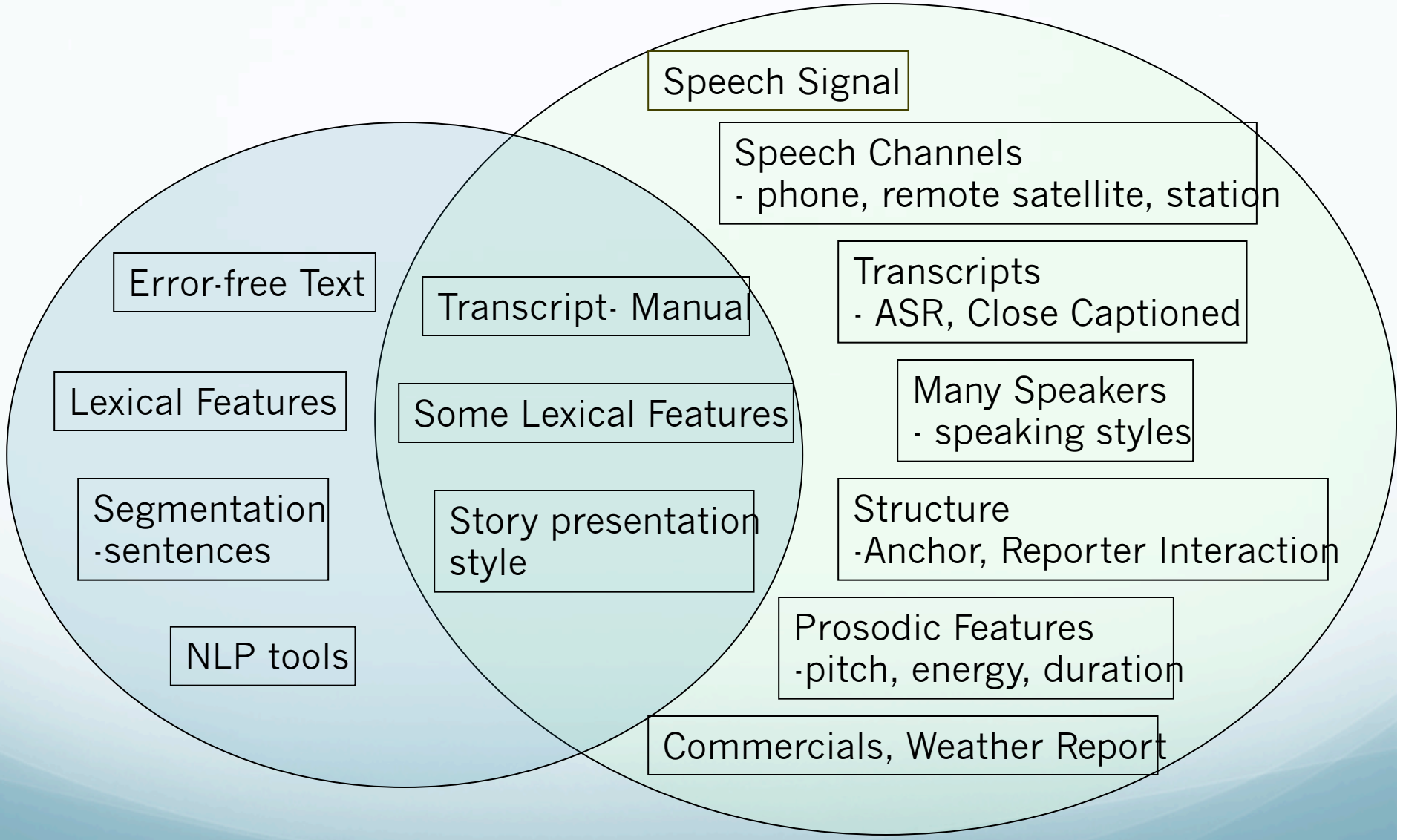
Speech and Text Summarization

- Commonalities:
  - Require key content selection
  - Linguistic cues: lexical, syntactic, discourse structure
  - Alternative strategies: extractive, abstractive

Speech vs Text

- Challenges of speech (summarization):
  - Recognition (and ASR errors)
    - Downstream NLP processing issues, errors
  - Segmentation: speaker, story, sentence
  - Channel issues (anchor vs remote)
  - Disfluencies
  - Overlaps
  - “Lower information density”: off-talk, chitchat, etc
  - Generation: text? Speech? Resynthesis?
  - Other text cues: capitalization, paragraphs, etc
- New information: audio signal, prosody, dialog structure

Text vs. Speech Summarization (NEWS)



Hirschberg, 2006

Current Approaches

- Predominantly extractive
- Significant focus on compression
  - Why?
    - Fluency: raw speech is often messy
    - Speed: speech is (relatively) slow, if using playback
- Integration of speech features

Current Data

- Speech summary data:
  - Broadcast news
  - Lectures
  - Meetings
  - Talk shows
  - Conversations (Switchboard, Callhome)
  - Voicemail

Common Strategies

- Basically, do ASR and treat like text
  - Unsupervised approaches:
    - Tf-idf cosine; LSA; MMR
  - Classification-based approaches:
    - Features include:
      - Sentence position, sentence length, sentence score/weight
      - Discourse & local context features
    - Modeling approaches:
      - SVMs, logistic regression, CRFs, etc

What about “Speech”?

- Automatic sentence segmentation
- Disfluency tagging, filtering
- Speaker-related features:
  - Speaker role (e.g. anchor), proportion of speech
- ASR confidence scores:
  - Intuition: use more reliable content
- Prosody:
  - Pitch, intensity, speaking rate
  - Can indicate: emphasis, new topic, new information

Speech-focused Summarization

- Intuition:
  - **How** something is said is as important as **what** is said
- Hypothesis:
  - Speakers use pitch, intensity, speaking rate to mark important information
- Test:
  - Can we do speech summarization without speech transcription?
    - At least competitively with ASR

Approach

- Maskey & Hirschberg, 2005
- Data: Broadcast News (e.g. CNN)
  - Single-document summarization
- HMM model:
  - Summary vs non-summary states
- Observations:
  - Acoustic-prosodic measures: pitch, intensity,…
  - Speaker features: which speaker, role, etc
  - Lexical: word information
  - Discourse features

Results

- Acoustic, speaker results competitive w/lexical
  - Combined best

Features	ROUGE score
All features	0.8
Lexical	0.7
Acoustic+Speaker	0.68
Acoustic	0.63
Baseline	0.5

Summary

- Speech summarization:
  - Builds on text based models
- Extends to
  - Overcome speech-specific challenges
  - Exploit speech-specific cues
- Can be highly domain/task dependent
- Highly challenging