Alternative Perspectives on Summarization

Systems & Applications
Ling 573
May 25, 2017

Roadmap

- Abstractive summarization example
  - Using Abstract Meaning Representation

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

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

Abstractive Summarization

- Basic components:
  - Content selection
  - Information ordering
  - Content realization
    - Comparable to extractive summarization
- Fundamental differences:
  - What do the processes operate on?
    - Extractive? Sentences (or subspans)
    - Abstractive? Major question
      - Need some notion of concepts, relations in text

Levels of Representation

- How can we represent concepts, relations from text?
  - Ideally, abstract away from surface sentences
- Build on some deep NLP representation:
  - Dependency trees: (Cheung & Penn, 2014)
  - Discourse parse trees: (Gerani et al, 2014)
  - Logical Forms
  - Abstract Meaning Representation (AMR): (Liu et al, 2015)

Representations

- Different levels of representation:
  - Syntax, Semantics, Discourse
- All embed:
  - Some nodes/substructure capturing concepts
  - Some arcs, etc capturing relations
  - In some sort of graph representation (maybe a tree)
- What's the right level of representation??

Typical Approach

- Parse original documents to deep representation

- Manipulate resulting graph for content selection
  - Splice dependency trees, remove satellite nodes, etc

- Generate based on resulting revised graph

- All rely on parsing/generation to/from representation

AMR

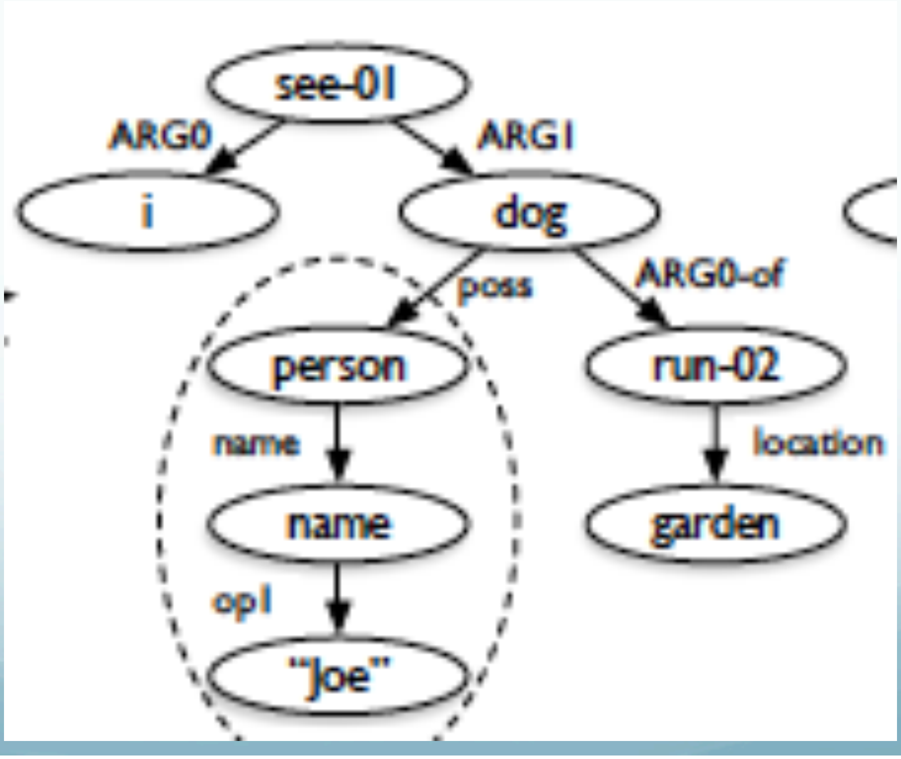
- “Abstract Meaning Representation”
  - Sentence-level semantic representation

  - Nodes: Concepts:
    - English words, PropBank predicates, or keywords (‘person’)

  - Edges: Relations:
    - PropBank thematic roles (ARG0-ARG5)
    - Others including ‘location’, ‘name’, ‘time’, etc…
    - ~100 in total

AMR 2

- AMR Bank: (now) ~40K annotated sentences
- JAMR parser: 63% F-measure (2015)
  - Alignments b/t word spans & graph fragments
- Example: “I saw Joe’s dog, which was running in the garden.”



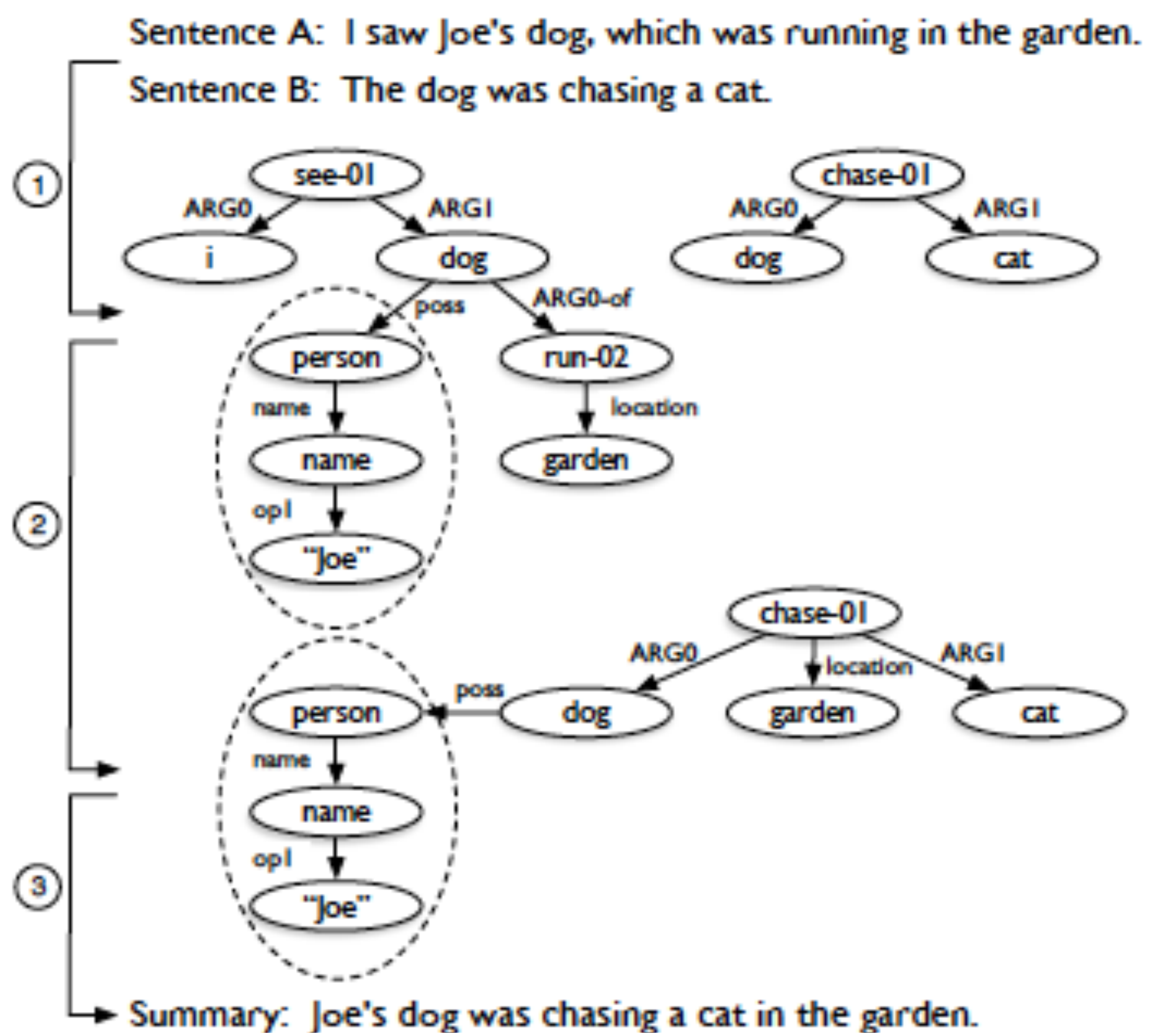
Liu et al, 2015.

Summarization Using Abstract Meaning Representation

- Use JAMR to parse input sentences to AMR
- Create unified document graph
  - Link coreferent nodes by “concept merging”
  - Join sentence AMRs to common (dummy) ROOT
  - Create other connections as needed
- Select subset of nodes for inclusion in summary
- *Generate surface realization of AMR (future work)

Liu et al, 2015.

Toy Example



Sentence A: I saw Joe's dog, which was running in the garden.

Sentence B: The dog was chasing a cat.

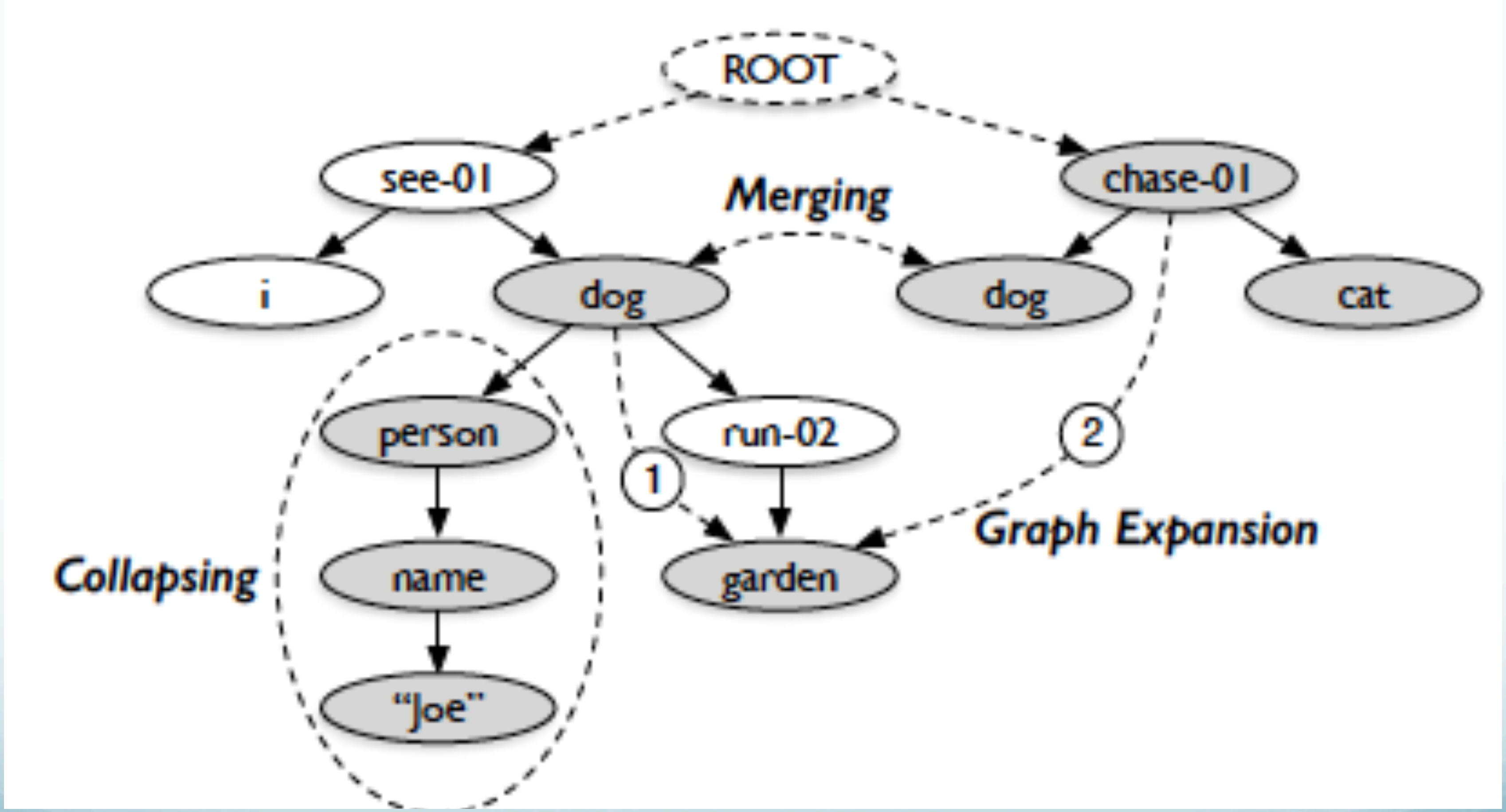
Summary: Joe's dog was chasing a cat in the garden.

Liu et al, 2015.

Creating a Unified Document Graph

- Concept merging:
  - Idea: Combine nodes for same entity in diff’t sentences
    - Highly Constrained
    - Applies ONLY to Named entities & dates
    - Collapse multi-node entities to single node
    - Merge ONLY identical nodes
      - Barak Obama = Barak Obama; Barak Obama ≠ Obama
    - Replace multiple edges b/t two nodes with unlabeled edge

Merged Graph Example



Liu et al, 2015; Fig 3.

Content Selection

- Formulated as subgraph selection
  - Modeled as Integer Linear Programming (ILP)

- Maximize the graph score (over edges, nodes)
  - Inclusion score for nodes, edges
  - Subject to:
    - Graph validity: edges must include endpoint nodes
    - Graph connectivity
    - Tree structure (one incoming edge/node)
    - Compression constraint (size of graph in edges)

- Features: Concept/label, frequency, depth, position,
  - Span, NE?, Date?

Evaluation

- Compare to gold-standard “proxy report”
  - ~ Single document summary In style of analyst’s report
    - All sentences paired w/AMR
  - Fully intrinsic measure:
    - Subgraph overlap with AMR
  - Slightly less intrinsic measure:
    - Generate Bag-of-Phrases via most frequent subspans
      - Associated with graph fragments
    - Compute ROUGE-1, aka word overlap

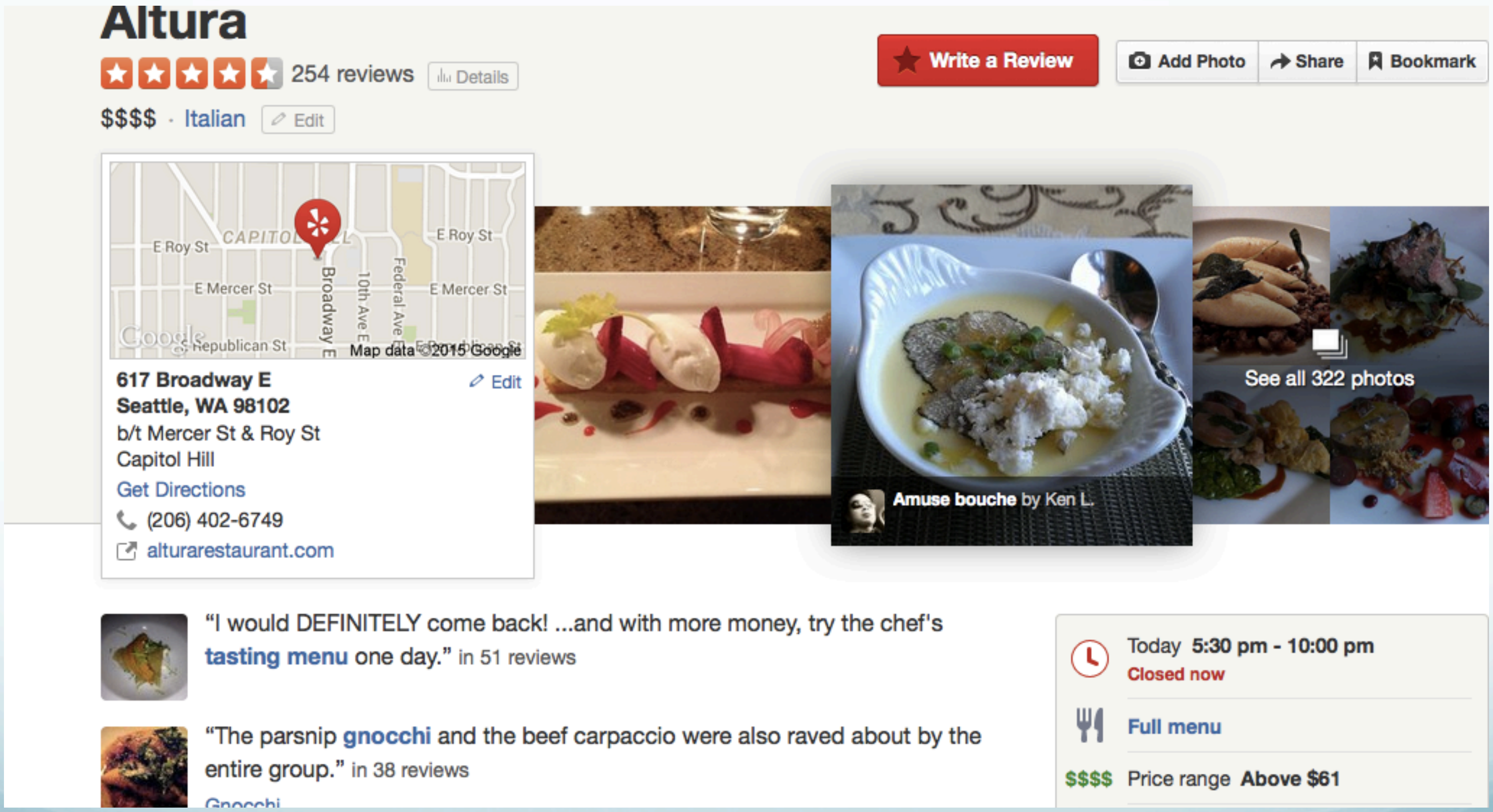
Evaluation

- Results:
  - ROUGE-1: P: 0.5; R: 0.4; F: 0.44
    - Similar for manual AMR and automatic parse
  - Topline:
    - Oracle: P: 0.85; R: 0.44; F: 0.58
    - Based on similar bag-of-phrase generation from gold AMR

Summary

- Interesting strategy based on semantic represent'n
  - Builds on graph structure over deep model
  - Promising strategy
- Limitations:
  - Single-document
    - Does extension to multi-doc make sense?
  - Literal matching:
    - Reference, lexical content
  - Generation

Review Summaries



Altura

★★★★½ 254 reviews Details

$$$$ · Italian Edit

617 Broadway E
Seattle, WA 98102
b/t Mercer St & Roy St
Capitol Hill
Get Directions
☎ (206) 402-6749
alturarestaurant.com

Edit

"I would DEFINITELY come back! ...and with more money, try the chef's **tasting menu** one day." in 51 reviews

"The parsnip **gnocchi** and the beef carpaccio were also raved about by the entire group." in 38 reviews

Gnocchi

Today **5:30 pm - 10:00 pm**
Closed now

Full menu

$$$$ Price range **Above $61**

Review Summary Dimensions

- Use purpose: Product selection, comparison
- Audience: Ordinary people/customers
- Derivation (extactive vs abstractive): Extractive+
- Coverage (generic vs focused): Aspect-oriented
- Units (single vs multi): Multi-document
- Reduction: Varies
- Input/Output form factors (language, genre, register, form)
  - ??, user reviews, less formal, pros & cons, tables, etc

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.

- 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
- 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 I

- 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 II

- Sentiment Match + Aspect Coverage (SMAC):
  - Linear combination of:
    - Sentiment intensity, mismatch, & diversity
  - Issue?
    - Optimizes overall sentiment match, but not per-aspect

Review Summarization Models III

- 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
  - Judged: No pref, Strongly – Weakly prefer A/B
- Also collected comments that justify rating
- Usually some preference, but not significant
  - Except between SAM (highest) and SMAC (lowest)
- Do users care at all?
  - **Yes!!** SMAC significantly better than LEAD baseline
    - (70% vs 25%)

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
- 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....

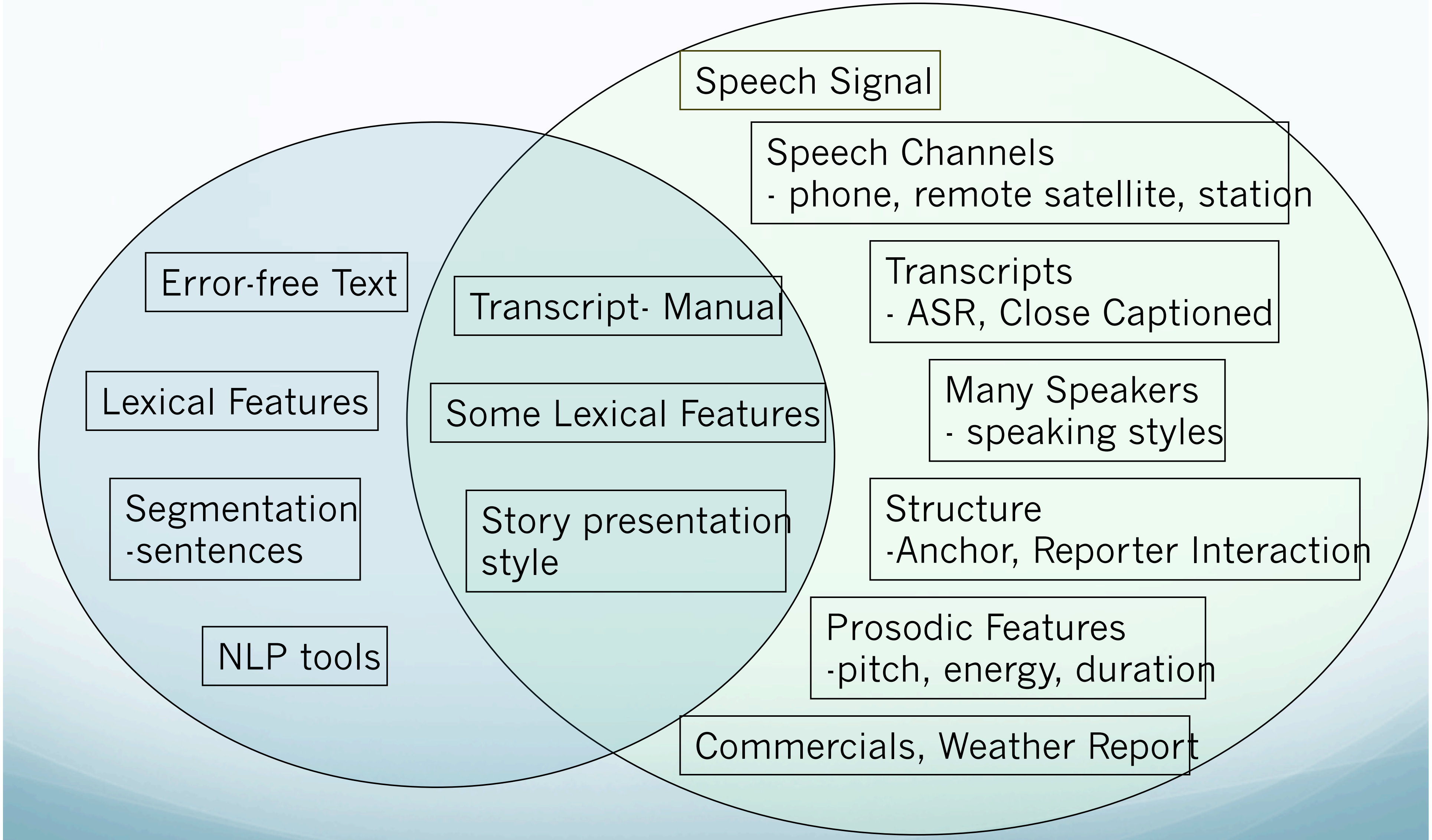
Speech and Text Summarization

- 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)



Error-free Text

Lexical Features

Segmentation
-sentences

NLP tools

Transcript- Manual

Some Lexical Features

Story presentation
style

Speech Signal

Speech Channels
- phone, remote satellite, station

Transcripts
- ASR, Close Captioned

Many Speakers
- speaking styles

Structure
-Anchor, Reporter Interaction

Prosodic Features
-pitch, energy, duration

Commercials, Weather Report

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

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
      - Jauhar, Chen, and Metze 2013; Maskey & Hirschberg, ‘05,’06

Approach

  - Maskey & Hirschberg, 2006

- Data: Broadcast News (e.g. CNN)
  - Single-document summarization
    - Has sentence, turn, topic annotation

- Bayesian Network model here:
  - Later HMM model:
    - Summary vs non-summary states

Approach

  - Maskey & Hirschberg, 2006

- Data: Broadcast News (e.g. CNN)
  - Single-document summarization
    - Has sentence, turn, topic annotation

- Bayesian Network model here:
  - Later used HMM model:
    - Summary vs non-summary states

- Observations:
  - Acoustic-prosodic measures: pitch, intensity,...
  - Structural features: which speaker, role, position, etc
  - Lexical: word information
  - Discourse features: Ratio of given/new information

Results

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

Features	ROUGE score
All features	0.8
Lexical	0.7
Acoustic+Structural	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

Conclusions

- Summarization:
  - Broad range of applications
    - Differ across dimensions
  - Delved into TAC summarization in depth
  - Draws on wide range of:
    - Shallow, deep NLP methods
    - Machine learning models
  - Many remaining challenges, opportunities

Reminders

- Final code deliverable due Sunday
- Doodle for presentation times
- Manual evaluation instructions/data out Monday