Alternate Perspectives on Summarization

LING 573 — NLP Systems and Applications

May 24, 2018

Announcements

- Since 5/28 is memorial day, D4 code deadline is extended until 5/30.
  - …everyone is welcome to the extension *so long as you have results for your presentation*
  - ***you must have evaluation completed*** by your presentation date.
- …you are allowed to use an off-the-shelf sentence compressor!
  - ***As long as*** *you can explain what it's doing in your presentation + report!*
  - One group found: [this github project]

Dimensions of TAC Summarization

- **Use purpose** — Reflective summaries
- **Audience** — Analysts
- **Derivation** (extractive vs. abstractive) — Largely extractive
- **Coverage** (generic vs focused) — "Guided"
- **Units** (single vs. multi) — Multi-document
- **Reduction** — 100 words

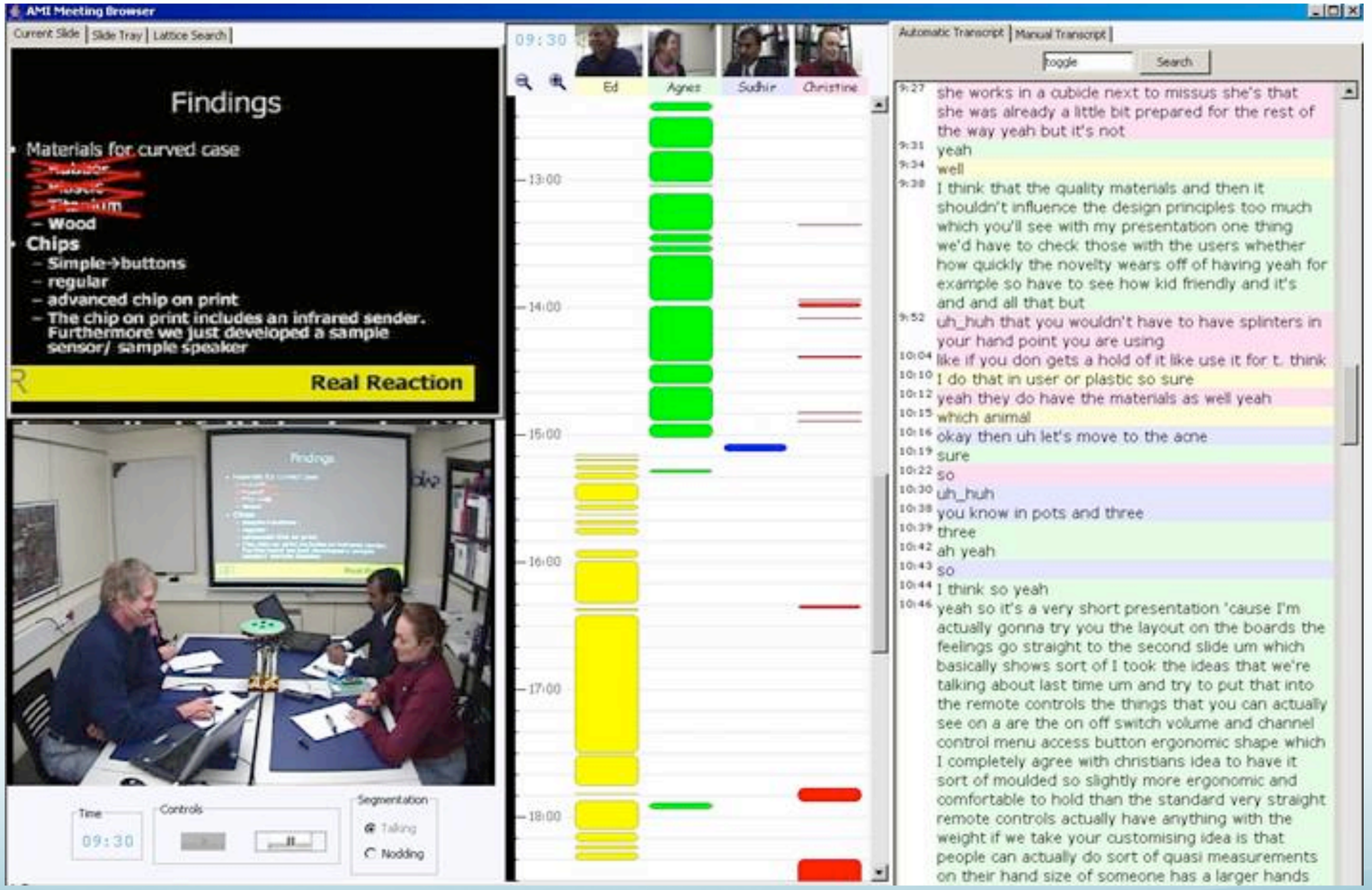
Dimensions of TAC Summarization

- **Input Form Factor**
  - English
  - Newswire text
  - Multiple documents, multiple paragraphs
- **Output Form Factor**
  - English
  - Single paragraph

Meeting Summaries

Other Types of Summaries: Meeting Summaries

- Renals, 2010 & AMI Consortium



Meeting Summaries

- What do you want out of the summary?
  - Minutes?
  - Agenda?
  - To-do list?
    - (Think: automatic sprint management based on team stand-ups!)
  - Points of (dis)agreement?

Dimensions of Meeting Summaries

- **Use purpose**
  - Catch up on missed meetings
- **Audience**
  - Ordinary attendees
- **Derivation**
  - Either abstractive or abstractive

Dimensions of Meeting Summaries

- **Coverage (generic vs. focused)**
  - Depends on user
    - Team member — "what do ***I*** have to focus on?"
    - PM — How is the team doing as a whole?
- **Units (single vs. multi)**
  - Single meeting?
  - Recurring problems over project?

Dimensions of Meeting Summaries

- **Input Form Factor**:
  - Speech
  - …maybe list items, whiteboard diagrams?
- **Output Form Factor**:
  - Lists, bullets, todos
  - 100-200-word summary?

Examples

- Decision summary:
  1. The remote will resemble the potato prototype
  2. There will be no feature to help find the remote when misplaced
  3. Instead, remote will be in bright color
  4. Corporate logo WILL be on the remote
  5. One of the colors will contain the corporate colors
  6. Remote will have six buttons
  7. Buttons will all be one color
  8. The case will be single curve
  9. The case will be rubber
  10. The case will have special color

Examples

- Action items:
  - Each team member receives specific instructions for next meeting by email

Examples

- Abstractive summary:
  - When this functional design meeting opens the project manager tells the group about the project restrictions he received from management by email. The marketing expert is first to present, summarizing user requirements data from a questionnaire given to 100 respondents. The marketing expert explains various user preferences and complaints about remotes as well as different interests among age groups. He prefers that they aim users from ages 16-45, improve the most-used functions, and make a placeholder for the remote…

Abstractive Summarization Using AMR

Abstractive Summarization: Recap

- 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

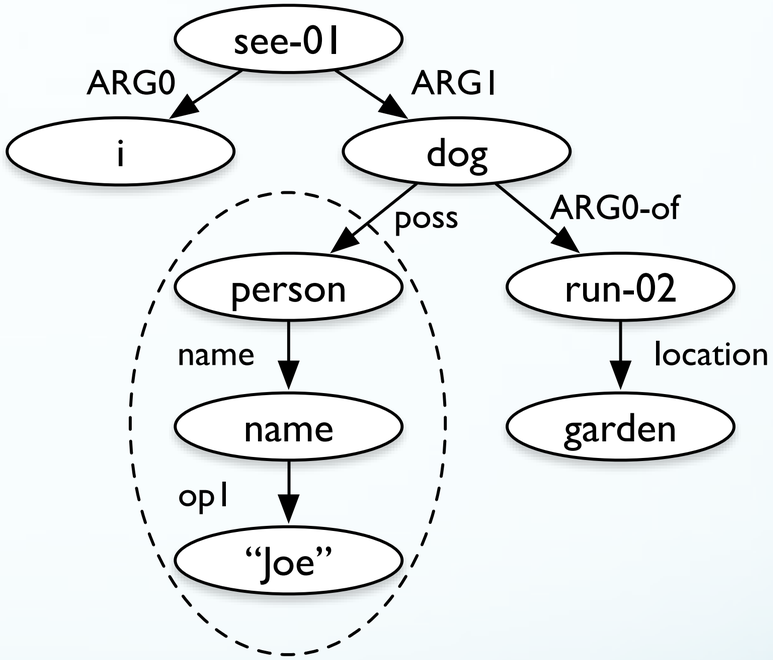
- Different levels of Representation
  - Syntax, Semantics, Discourse
- All embed:
  - Some nodes/substructure capturing concepts
  - Some arcs, etc. capturing relations
  - In some sort of graph representation (maybe a tree)
- What's the right level of representation?

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

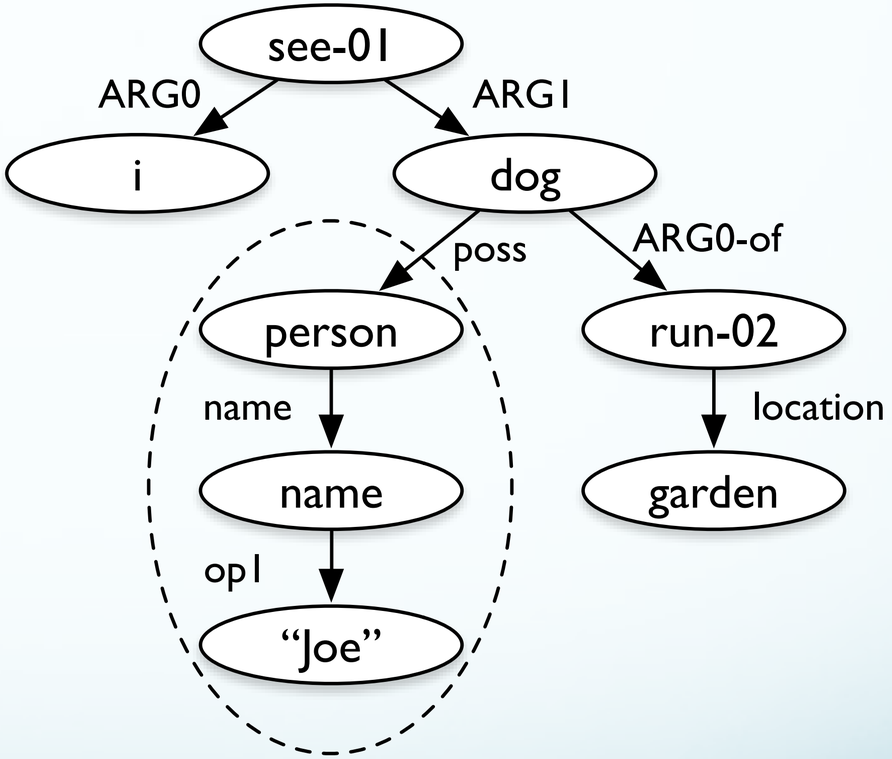
- Sentence-level semantic representation
- Nodes = Concepts
  - English words, PropBank predicates, or keywords ('person')
- Edges = Relations:
  - PropBank thematic roles (ARG0-ARG5)
  - Others including 'location', 'name', 'time', etc…
  - ~100 in total



Liu et al, 2015, p.1078

AMR

- AMR Bank: (now) ~40K annotated sentences
- JAMR parser: 63% F-measure (2015)
  - Alignments between word spans & graph fragments
- Example:
  - *"I saw Joe's dog, which was running in the garden."*



Liu et al, 2015, Fig. 1 p.1078

Summarization Using AMR

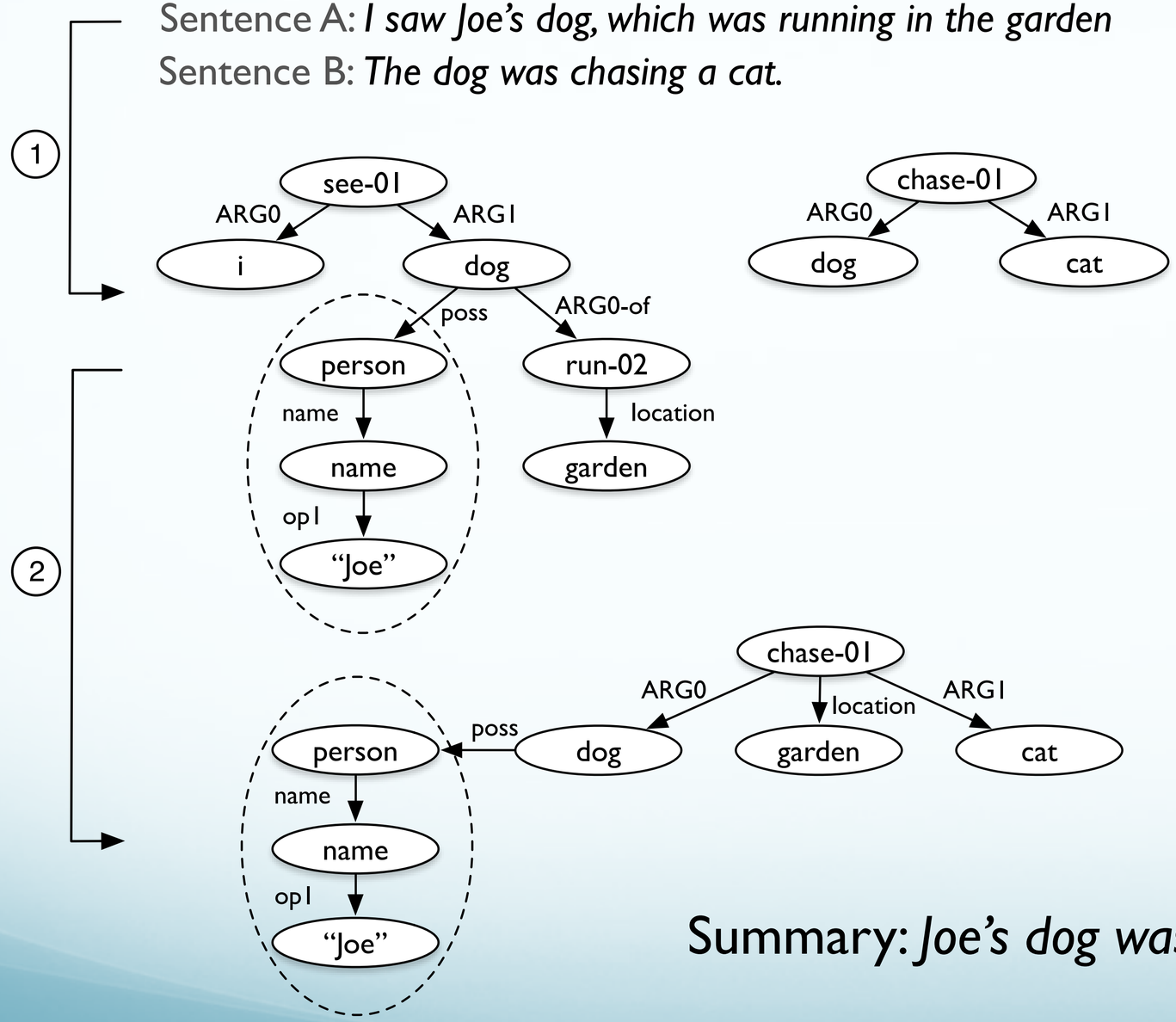
Liu et al, (2015)

- 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)

Toy Example

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

Sentence B: *The dog was chasing a cat.*



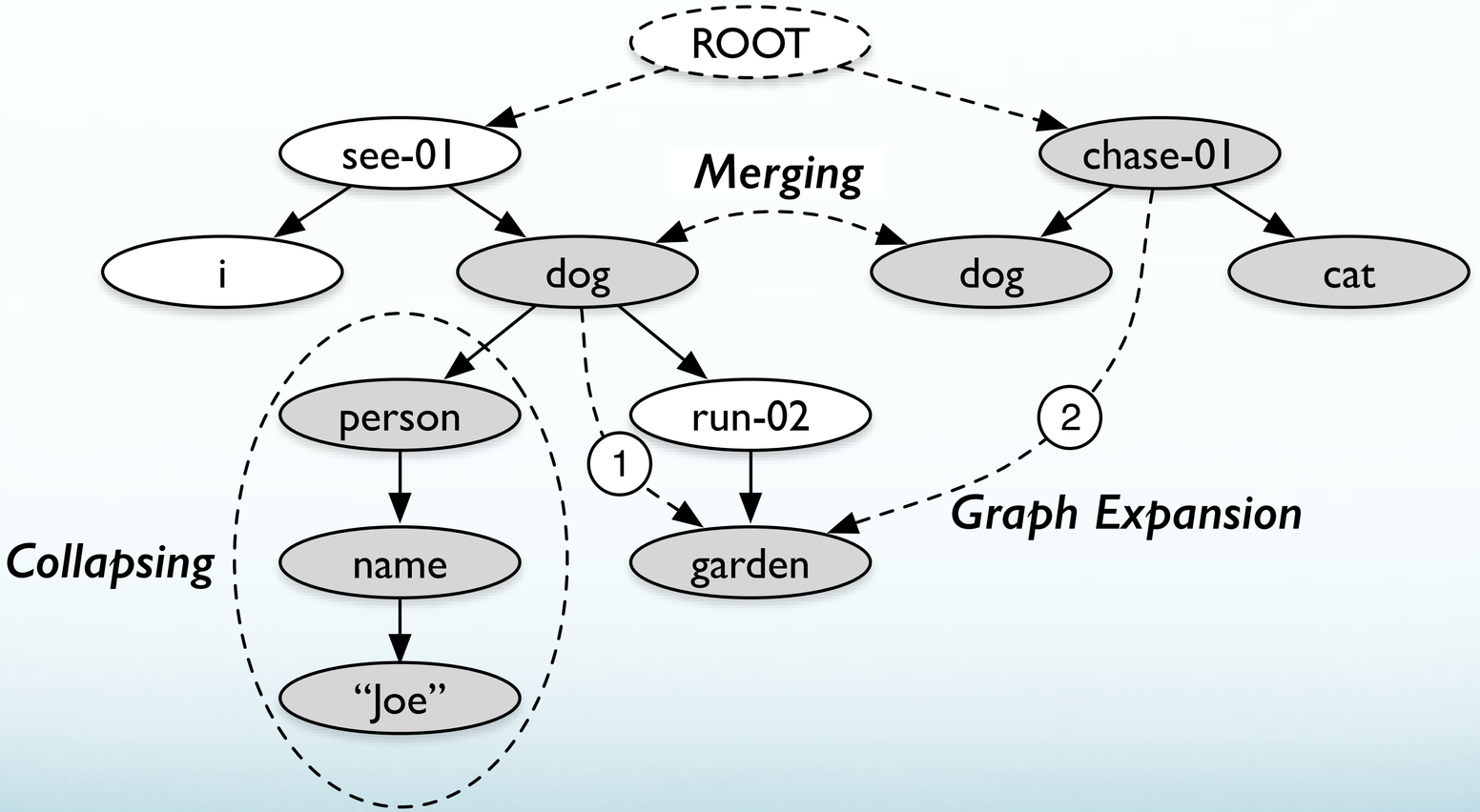
Summary: *Joe's dog was chasing a cat in the garden.*

Liu et al, 2015, Fig. 1 p.1078

Creating a Unified Document Graph

- Concept Merging
  - **Idea**: Combine nodes for same entity in different sentences
  - Highly constrained
  - Applies ONLY to named entities & dates
  - Collapse multi-node entities to single node
  - Merge **ONLY** identical nodes (coreference saved for future work)
    - Barack Obama = Barack Obama
    - Barack Obama ≠ Obama
  - Replace multiple edges between two nodes with an unlabeled edge

Merged Graph Example



Liu et al, 2015; Fig. 3

Content Selection

- Formulated as subgraph selection
  - Modeled as Integer Linear Programming
- 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 Constant (Size of graph in edges)

Content Selection

- Node Features:
  - Concept/label
  - Frequency — Concept frequency in the input sentence
  - Depth — Average and smallest depth of node to the root
  - Position — Average and foremost position of sentences containing concept
  - Span — Average and longest word span of concept
  - Entity/Date — Binary features indicating entity or date

Content Selection

- Edge Features:
  - Label — Most frequent edge labels between concepts
  - Freq — edge frequency in the document sentences
  - Position — average position of sentences containing edge
  - Nodes — Node features extracted from the source and target
  - IsExpanded — is the edge due to graph expansion or not

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
    - Generate Bag-of-Phrases via most frequent subspans
      - Associate with graph fragments
    - Compute ROUGE-1 (word overlap)

Evaluation: Results

- ROUGE-1:
  - P–0.5; R–0.4; F–0.44
    - Both for manual and automatic parses
- Upper Bound:
  - Oracle: P–0.85; R–0.44; F–0.58

Summary

- Interesting strategy based on semantic representation
  - Builds on graph structure over deep model
- Limitations
  - Single-document — does extension to multi-document make sense?
  - Literal matching — based on reference, lexical content
  - Generation aspect missing

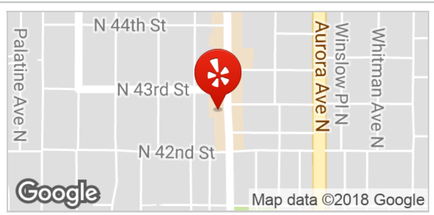
Review Summaries

Review Summaries

Paseo Caribbean Food - Fremont Claimed

4644 reviews Details

$ · Caribbean, Cuban, Sandwiches Edit



4225 Fremont Ave N
Seattle, WA 98103
b/t 42nd St & Motor Pl
Fremont

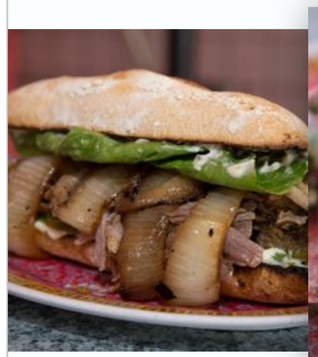
Edit

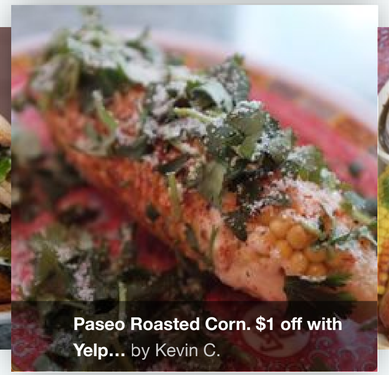
Get Directions

(206) 545-7440

paseorestaurants.com

Send to your Phone





Paseo Roasted Corn. $1 off with Yelp… by Kevin C.

See all 1749

Ask about catering, event space & parties today

"Lined up early to place my order: **Caribbean Roast** sandwich, Smokin' Thighs sandwich, and a side of scallops in red sauce over rice." in 678 reviews

$9.75 Caribbean Roast

"Since I can never decide what to order, I like to go with a friend and split the Caribbean Roast and **Paseo Press**." in 229 reviews

$11.50 Paseo Press

"It's definitely not easy being sexy while stuffing your face with juicy chunks of pork and monster sized **carmelized onions**." in 58 reviews

$1 Carmelized Onions

Show more review highlights

Order Delivery or Takeout

Takeout

Enter your delivery address

1 Yelp St., San Francisco, CA 94105

Today **11:00 am - 9:00 pm**
Closed now

Full menu

$$$$ Price range **Under $10**

$1 off Paseo Roasted Corn or 1st Beer Send to your phone

Review Summary Dimensions

- **Use purpose** — Product selection, comparison
- **Audience** — Non-experts, customers
- **Derivation** — Extractive/abstractive: Extractive+
- **Coverage (Generic vs. Focused)** — Aspect-oriented
- **Units (single vs. multi)** — Multi-document
- **Reduction** — Varies

Review Summary Dimensions

- **Input form factors**
  - Language: ???
  - Genre — less formal
- **Output form factors**:
  - Pros/cons
  - Tables (formatted data)
  - etc…

Sentiment Summarization

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
    - 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 between summary and product rating
  - Diversity
    - Measure of how well different "aspects" of product covered
    - Related to both quality of coverage, importance of aspect

Review Summarization Models (Lerman et al., 2009)

- **Sentiment Match (SM)**
  - Prefer summaries where average sentiment most closely matches overall review
  - Issue?
    - Neutral rating → model prefers sentences with no opinion whatsoever
  - Approach:
    - Force system to select "stronger" sentences first

Review Summarization Models (Lerman et al., 2009)

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

$$L(S) = \alpha \cdot Intensity(S) - \beta \cdot Mismatch(S) + \gamma \cdot Diversity(S)$$

Review Summarization Models (Lerman et al., 2009)

- **Sentiment-Aspect Match (SAM)**
  - Maximize coverage of aspects
    - *consistent* with per-aspect sentiment
  - Computed using probabilistic model
  - Minimize KL-divergence between summary, original docs

Human Evaluation

- Pairwise prefernece tests for different summaries
  - Side by side, along with overall product rating
  - Judged: no preference, strongly, weakly prefer A/B
- Also collected comments justifying rating
- Usually some preference, but not significant
- Except between SAM (highest) and SMAC (lowest)
- Do users care?
  - YES! — SMAC significantly better than LEAD baseline (70% vs 25%)

Qualitative Comments

- Preferred:
  - Summaries with list (pro vs. con)
- Disliked:
  - Summary sentence without sentiment
  - Non-specific sentences
  - Inconsistency between overall rating and Summary
- Preferences differed depending on overall ratings
  - 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 with 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 vs. Text Summarization

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

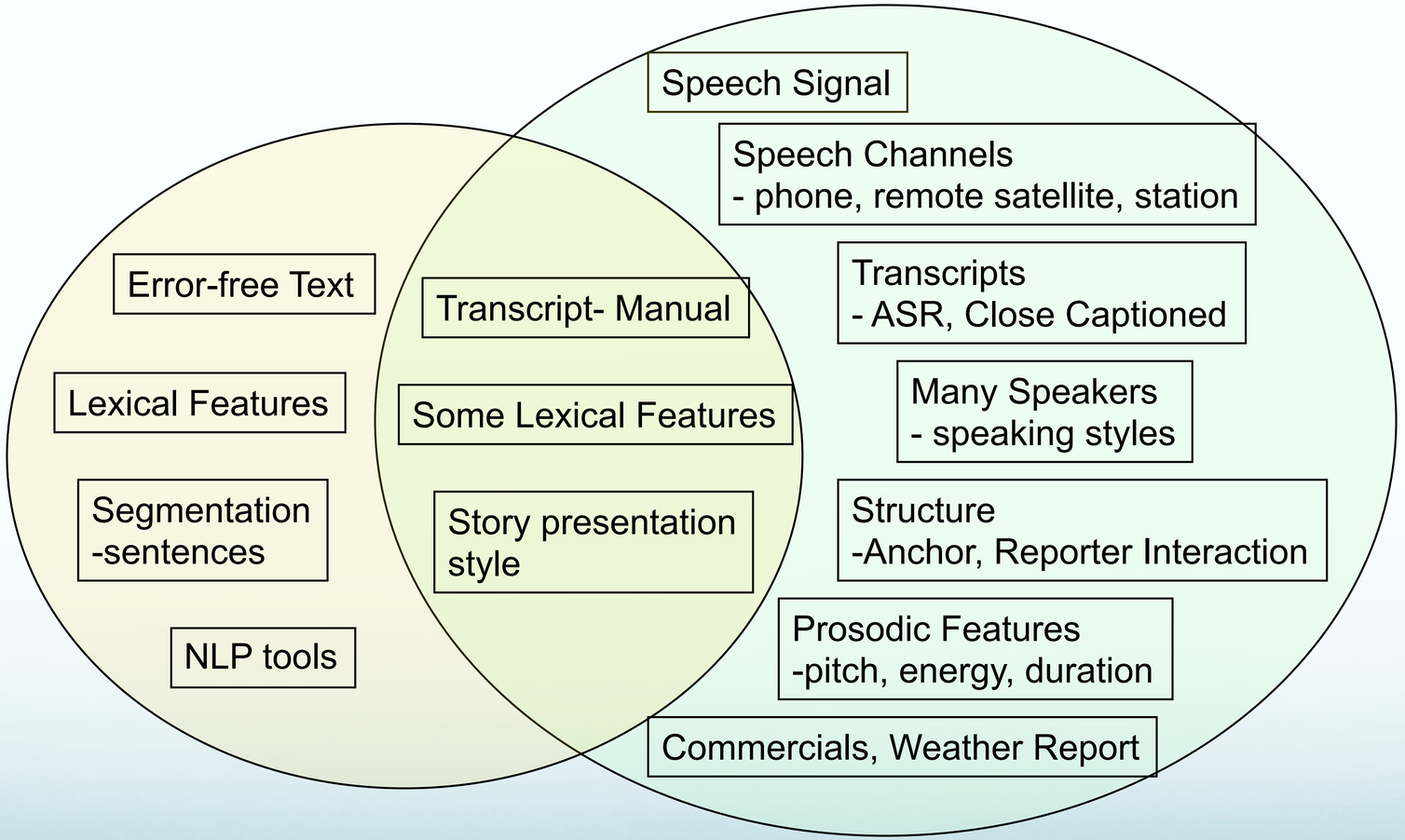
Speech vs. Text Summarization

- Challenges of speech (summarization)
  - Recognition (and ASR errors)
    - Downstream NLP processing issues
  - Segmentation: speaker – story – sentence
  - Channel issues
  - Disfluencies
  - Overlaps
  - "Lower information density": off-talk, chitchat, etc
  - Generation: Text? Speech?
  - Other text cues — capitalization, paragraphs, etc

Speech vs. Text Summarization

- New information:
  - Intonation
  - Prosody
  - Dialogue structure

Text vs. Speech Summarization (News Domain)



via Hirschberg, 2006

Current Approaches

- Predominantly Extractive
- Significant focus on compression
  - Raw speech often "messy"
  - Speech is (relatively) slow

Current Data

- Speech Summary Data:
  - Broadcast news
  - Lectures
  - Meetings
  - Talk Shows
  - Conversations (Switchboard, Callhome)
  - Voicemail

Common Strategies

- Typically — do ASR and treat like text
- Unsupervised approaches
  - tf•idf cosine — LSA — MMR
- Classification-based approaches:
  - Sentence position, sentence length, sentence score/weight
  - Discourse & local context features
- Modeling approaches:
  - SVMs, logistic regression, CRFs, RNNs, 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, rate, to mark important information
- Test:
  - Can we do speech summarization ***without transcription***?
    - Jauhar, Chen, and Metze 2013
    - Maskey & Hirschberg, '05, '06

Approach

Maskey & Hirschberg (2006)

- Data: Broadcast News (e.g. CNN)
  - "Single-document" summarization
    - Sentence, "turn," topic annotation
- Bayesian Network model here:
  - Later, used HMM
    - Summary vs. non-summary states

Observations

Maskey & Hirschberg (2006)

- Acoustic-prosodic measures
  - pitch, intensity
- Structural Features:
  - Which Speaker?
  - Speaker role?
- Lexical Features:
  - Word information
- Discourse features:
  - Ratio of given/new information

Results

- Acoustic + Speaker results competitive w/lexical
- Combined is best
- Baseline: Lead sentences

Features	ROUGE
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
  - Exploits speech-specific cues
- Can be highly domain/task dependent
- Highly challenging

Conclusions

- Summarization:
  - Broad range of applications
  - Differ across many dimensions
  - Here, we've looked at TAC summarization in depth
- Draws on wide range of:
  - Shallow, deep NLP methods
  - Machine learning models
- Many remaining challenges, opportunities