Summarization Systems & Evaluation

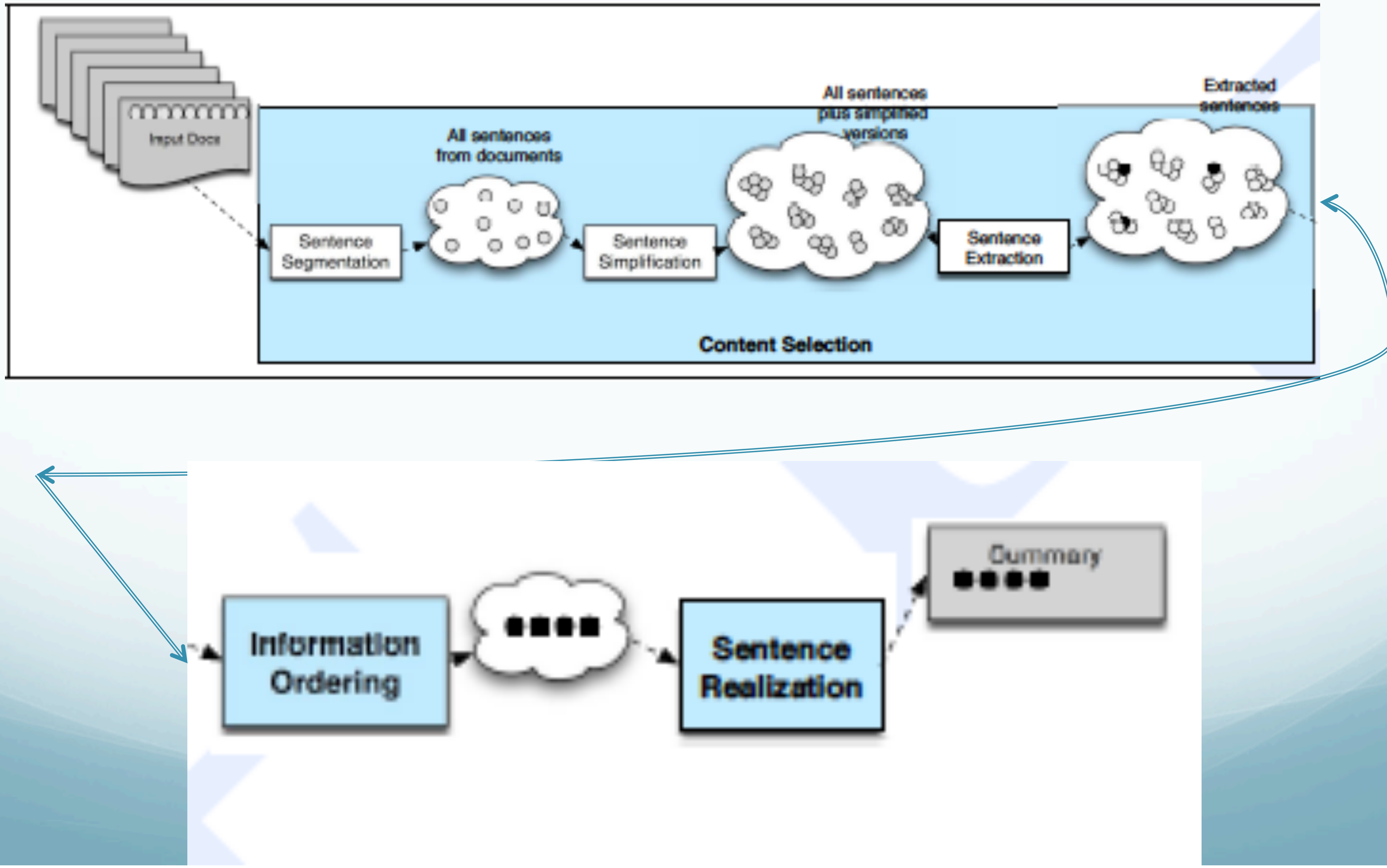
Ling573
Systems and Applications
April 5, 2016

Roadmap

- Summarization components:
  - Complex content selection
  - Information ordering
  - Content realization

- Summarization evaluation:
  - Extrinsic
  - Intrinsic:
    - Model-based: ROUGE, Pyramid
    - Model-free

General Architecture



More Complex Settings

- Multi-document case:
  - Key issue: redundancy
    - General idea:
      - Add salient content that is least similar to that already there

- Topic-/query-focused:
  - Ensure salient content related to topic/query
  - Prefer content more similar to topic
  - Alternatively, when given specific question types,
    - Apply more Q/A information extraction oriented approach

Information Ordering

- Goal: Determine presentation order for salient content
- Relatively trivial for single document extractive case:
  - Just retain original document order of extracted sentences
- Multi-document case more challenging: Why?
  - Factors:
    - Story chronological order – insufficient alone
    - Discourse coherence and cohesion
      - Create discourse relations
      - Maintain cohesion among sentences, entities
- Template approaches also used with strong query

Content Realization

- Goal: Create a fluent, readable, compact output
- Abstractive approaches range from templates to full NLG
- Extractive approaches focus on:
  - Sentence simplification/compression:
    - Manipulation of parse tree to remove unneeded info
      - Rule-based, machine-learned
  - Reference presentation and ordering:
    - Based on saliency hierarchy of mentions

Examples

- Compression:
  - ~~When it arrives sometime next year in new TV sets,~~ **the V-chip will give parents a** ~~new and potentially revolutionary~~ **device to block out programs they don't want their children to see**.
- Coreference:
  - Advisers do not blame **Treasury Secretary Paul O'Neill**, but they recognize a shakeup would help indicate **U.S. President George W. Bush** was working to improve matters**. Bush** pushed out **O'Neill** and ...

Systems & Resources

- System development requires resources
  - Especially true of data-driven machine learning
- Summarization resources:
  - Sets of document(s) and summaries, info
    - Existing data sets from shared tasks
    - Manual summaries from other corpora
  - Summary websites with pointers to source
  - For technical domain, almost any paper
    - Articles require abstracts…

Component Resources

- Content selection:
  - Documents, corpora for term weighting
  - Sentence breakers
  - Semantic similarity tools (WordNet sim)
  - Coreference resolver
  - Discourse parser
  - NER, IE
  - Topic segmentation
  - Alignment tools

Component Resources

- Information ordering:
  - Temporal processing
  - Coreference resolution
  - Lexical chains
  - Topic modeling
  - (Un)Compressed sentence sets
- Content realization:
  - Parsing
  - NP chunking
  - Coreference

Evaluation

- Extrinsic evaluations:
  - Does the summary allow users to perform some task?
    - As well as full docs? Faster?
  - Example:
    - Time-limited fact-gathering:
      - Answer questions about news event
        - Compare with full doc, human summary, auto summary
    - Relevance assessment: relevant or not?
    - MOOC navigation: raw video vs auto-summary/index
      - Task completed faster w/summary (except expert MOOCers)
- Hard to frame in general, though

Intrinsic Evaluation

- Need basic comparison to simple, naïve approach
- Baselines:
  - Random baseline:
    - Select N random sentences

  - Leading sentences:
    - Select N leading sentences
      - Or LASTEST (N leading sentences from chrono last doc)
    - For news, surprisingly hard to beat
      - (For reviews, last N sentences better.)

Intrinsic Evaluation

- Most common automatic method: ROUGE
  - "Recall-Oriented Understudy for Gisting Evaluation"
  - Inspired by BLEU (MT)
  - Computes overlap b/t auto and human summaries
  - E.g. ROUGE-2: bigram overlap

$$ROUGE2 = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{bigram \in S} count_{match}(bigram)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{bigram \in S} count(bigram)}$$

  - Also, ROUGE-L (longest seq), ROUGE-S (skipgrams)
  - ROUGE-BE: dependency path overlap

ROUGE

- Pros:
  - Automatic evaluation allows tuning
    - Given set of reference summaries
  - Simple measure
- Cons:
  - Even human summaries highly variable, disagreement
  - Poor handling of coherence
  - Okay for extractive, highly problematic for abstractive

Pyramid Evaluation

- Content selection evaluation:
  - Not focused on ordering, readability
- Aims to address issues in evaluation of summaries:
  - Human variation
    - Significant disagreement, use multiple models
  - Analysis granularity:
    - Not just "which sentence"; overlaps in sentence content
  - Semantic equivalence:
  - Extracts vs Abstracts:
    - Surface form equivalence (e.g. ROUGE) penalizes abstr.

Pyramid Units

- Step 1: Extract Summary Content Units (SCUs)
  - Basic content meaning units
    - Semantic content
    - Roughly clausal
    - Identified manually by annotators from model summaries
    - Described in own words (possibly changing)

Example

- A1. The industrial espionage case …began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as a production director.
- B3. However, he left GM for VW under circumstances, which …were described by a German judge as "potentially the biggest-ever case of industrial espionage".
- C6. He left GM for VW *in March 1993*.
- D6. The issue stems from the alleged recruitment of GM's …procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues.
- E1. *On March 16, 1993*, … Agnacio Lopez De Arriortua, left his job as head of purchasing at *General Motor's* Opel, Germany, to become Volkswagen's Purchasing … director.
- F3. *In March 1993*, Lopez and seven other GM executives moved to VW overnight.

Example SCUs

- SCU1 (w=6): Lopez left GM for VW
  - A1. the hiring of Jose Ignacio Lopez, an employee of GM . . . by VW
  - B3. he left GM for VW
  - C6. He left GM for VW
  - D6. recruitment of GM's . . . Jose Ignacio Lopez
  - E1. Agnacio Lopez De Arriortua, left his job . . . at General Motor's Opel . . .to become Volkswagen's . . . Director
  - F3. Lopez . . . GM . . . moved to VW
- SCU2 (w=3) Lopez changes employers in March 1993
  - C6 in March, 1993
  - E1. On March 16, 1993
  - F3. In March 1993

SCU: A cable car caught fire (Weight = 4)

A. The cause of the fire was unknown.

B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.

C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.

D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

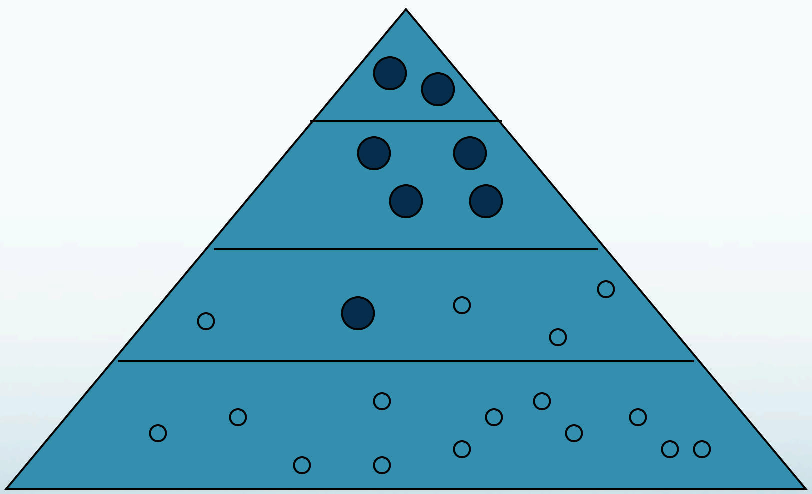
Pyramid Building

- Step 2: Scoring summaries
  - Compute weights of SCUs
    - Weight = # of model summaries in which SCU appears

  - Create “pyramid”:
    - n = maximum # of tiers in pyramid = # of model summ.s
    - Actual # of tiers depends on degree of overlap
    - Highest tier: highest weight SCUs
      - Roughly Zipfian SCU distribution, so pyramidal shape
  - Optimal summary?
    - All from top tier, then all from top -1, until reach max size

Ideally informative summary

- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



From Passoneau et al 2005

Pyramid Scores

- T_i = tier with weight i SCUs
  - T_n = top tier; T_1 = bottom tier
- D_i = # of SCUs in summary on T_i
- Total weight of summary $D = \sum_{i=1}^{n} i * D_i$
- Optimal score for X SCU summary: *Max*
  - (j lowest tier in ideal summary)

$$\sum_{i=j+1}^{n} i * |T_i| + j * (X - \sum_{i=j+1}^{n} |T_i|)$$

Correlation with Other Scores

Table VI. Pearson's Correlation Between the Different Evaluation Metrics Used in DUC 2005. Computed for 25 Automatic Peers Over 20 Test Sets

	Pyr (mod)	Respons-1	Respons-2	ROUGE-2	ROUGE-SU4
Pyr (orig)	0.96	0.77	0.86	0.84	0.80
Pyr (mod)		0.81	0.90	0.90	0.86
Respons-1			0.83	0.92	0.92
Respons-2				0.88	0.87
ROUGE-2					0.98

- 0.95: effectively indistinguishable
  - Two pyramid models, two ROUGE models
- Two humans only 0.83

Pyramid Model

- Pros:
  - Achieves goals of handling variation, abstraction, semantic equivalence
  - Can be done sufficiently reliably
  - Achieves good correlation with human assessors
- Cons:

Pyramid Model

- Pros:
  - Achieves goals of handling variation, abstraction, semantic equivalence
  - Can be done sufficiently reliably
  - Achieves good correlation with human assessors
- Cons:
  - Heavy manual annotation:
    - Model summaries, also all system summaries
    - Content only