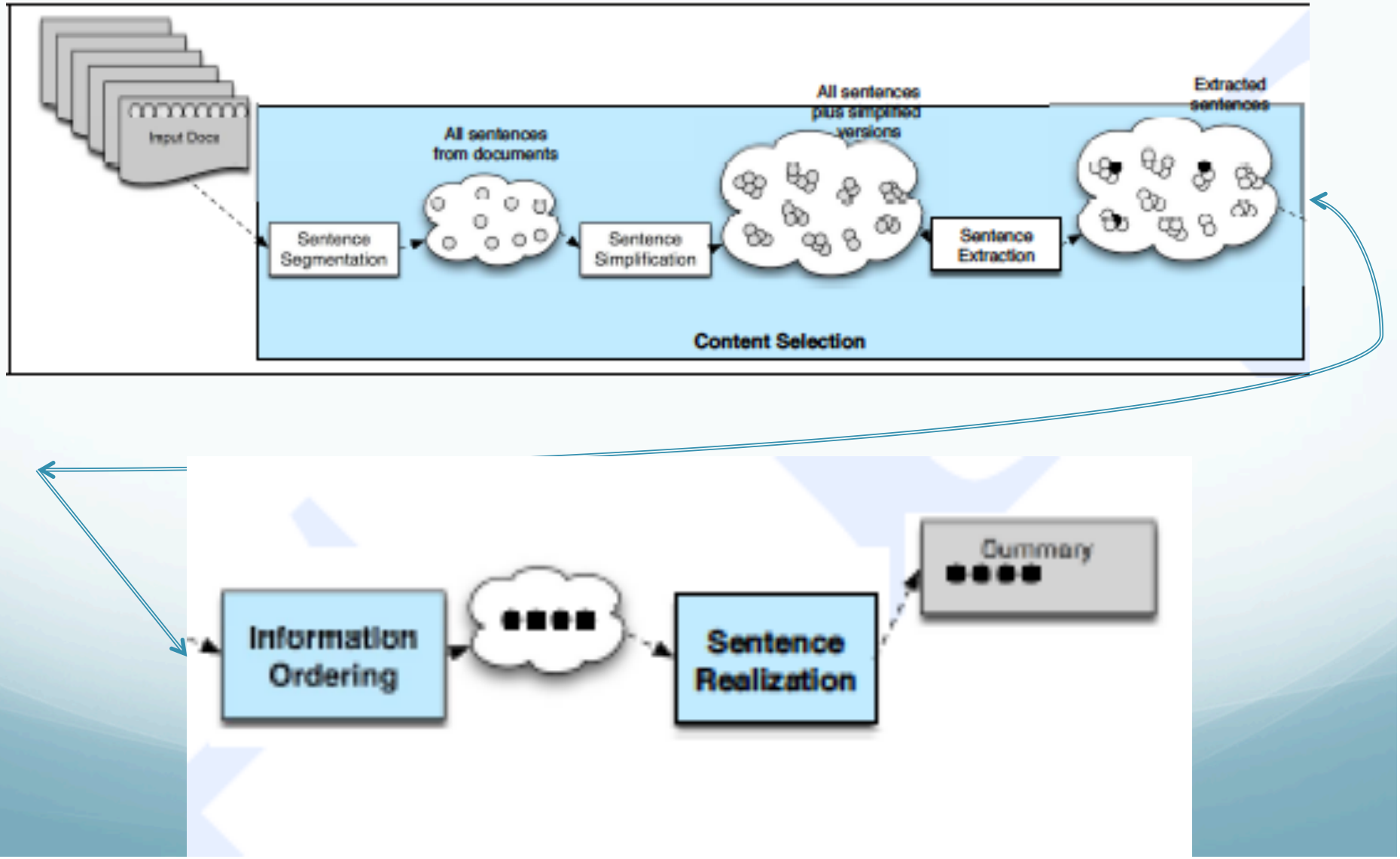
Summarization: Overview

Ling573
Systems & Applications
March 30, 2017

Roadmap

- Architecture of a Summarization system
- Summarization and resources
- Evaluation
- Logistics Check-in, Deliverable #1

General Architecture



General Strategy

- Given a document (or set of documents):
  - Select the key content from the text
  - Determine the order to present that information
  - Perform clean-up or rephrasing to create coherent output
  - Evaluate the resulting summary
- Systems vary in structure, complexity, information

More specific strategy

- For single document, extractive summarization:
  - Segment the text into sentences
  - Identify the most prominent sentences
  - Pick an order to present them
    - Maybe trivial, i.e. document order
  - Do any necessary processing to improve coherence
    - Shorten sentences, fix coref, etc

Content Selection

- Goal: Identify most important/relevant information
- Common perspective:
  - View as binary classification: important vs not
    - For each unit (e.g. sentence in the extractive case)
  - Can be unsupervised or supervised
- What makes a sentence (for simplicity) extract-worthy?

Cues to Saliency

- Approaches significantly differ in terms of cues
- Word-based (unsupervised):
  - Compute a **topic signature** of words above threshold
    - Many different weighting schemes: tf, tf*idf, LLR, etc
  - Select content/sentences with highest weight
- Discourse-based:
  - Discourse saliency ➔ extract-worthiness
- Multi-feature supervised:
  - Cues include position, cue phrases, word salience, ..
  - Training data?

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 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.

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 **O'Neill**, but they recognize a shakeup would help indicate **Bush** was working to improve matters**. U.S. President George W. Bush** pushed out **Treasury Secretary Paul O'Neill** and …

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 …

Our Task

- TAC 2009/10/11 Shared Task
  - Multi-document summarization
    - Newswire text
    - “Guided”
      - Aka topic-oriented
    - ROUGE as primary evaluation metric

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

Dimensions of Summary Evaluation

- Summary evaluation:
  - Inherently hard:
    - Multiple manual abstracts:
      - Surprisingly little overlap; substantial assessor disagreement
  - Developed in parallel with systems/tasks
- Key concepts:
  - Text quality: readability includes sentence, discourse structure
  - Concept capture: Are key concepts covered?
  - Gold standards: model, human summaries
    - Enable comparison, automation, incorporation of specific goals
  - Purpose: Why is the summary created?
    - Intrinsic/Extrinsic evaluation

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
    - 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

$$ROUGE2 = \frac{\sum_{S \in \{\text{Re}\, ferenceSummaries\}} \sum_{bigram \in S} count_{match}(bigram)}{\sum_{S \in \{\text{Re}\, ferenceSummaries\}} \sum_{bigram \in S} count(bigram)}$$

  - Also, ROUGE-L (longest seq), ROUGE-S (skipgrams)

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

Deliverable #1

- Goals:
  - Set up for remainder of course
  - Form teams
  - Set up repository for version control
    - GIT or SVN
  - Create report outline
    - ACL style files
- Mail Glenn (gslayden@uw) with team, repository plan/info
  - By weekend!!
  - Can get repository/extra space on cluster