Summarization: Overview

Ling573
Systems & Applications
March 31, 2016

Roadmap

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

Structuring the Summarization Task

- Summarization Task: (Mani and Mayberry 1999)
  - Process of distilling the most important information from a text to produce an abridged version for a particular task and user

- Main components:
  - Content selection
  - Information ordering
  - Sentence realization

Dimensions of Summarization

- Rich problem domain:
  - Tasks and Systems vary on:
    - Use purpose
    - Audience
    - Derivation
    - Coverage
    - Reduction
    - Input/Output form factors

Dimensions of Summarization

- Purpose:
  - What is the goal of the summary? How will it be used?
    - Often surprisingly vague
    - Generic “reflective” summaries:
      - Highlight prominent content
    - Relevance filtering:
      - “Indicative”: Quickly tell if document covers desired content
    - Browsing, skimming
    - Compression for assistive tech
    - Briefings: medical summaries, to-do lists; definition Q/A

Dimensions of Summarization

- Audience:
  - Who is the summary for?
    - Also related to the content
    - Often contrasts experts vs novice/generalists
  - News summaries:
    - ‘Ordinary’ vs analysts
      - Many funded evaluation programs target analysts
  - Medical:
    - Patient directed vs doctor/scientist-directed

Dimensions of Summarization

- "Derivation":
  - Continuum
    - Extractive: Built from units extracted from original text
    - Abstractive: Concepts from source, generated in final form
  - Predominantly extractive
- Coverage:
  - Comprehensive (generic) vs query-/topic-oriented
    - Most evaluations focused
- Units: single vs multi-document
- Reduction (aka compression):
  - Typically percentage or absolute length

Extract vs Abstract

Extract from the Gettysburg Address:
Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field. But the brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that government of the people, by the people for the people shall not perish from the earth.
Abstract of the Gettysburg Address:
This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

Figure 23.13 An extract versus an abstract from the Gettysburg Address (abstract from Mani (2001)).

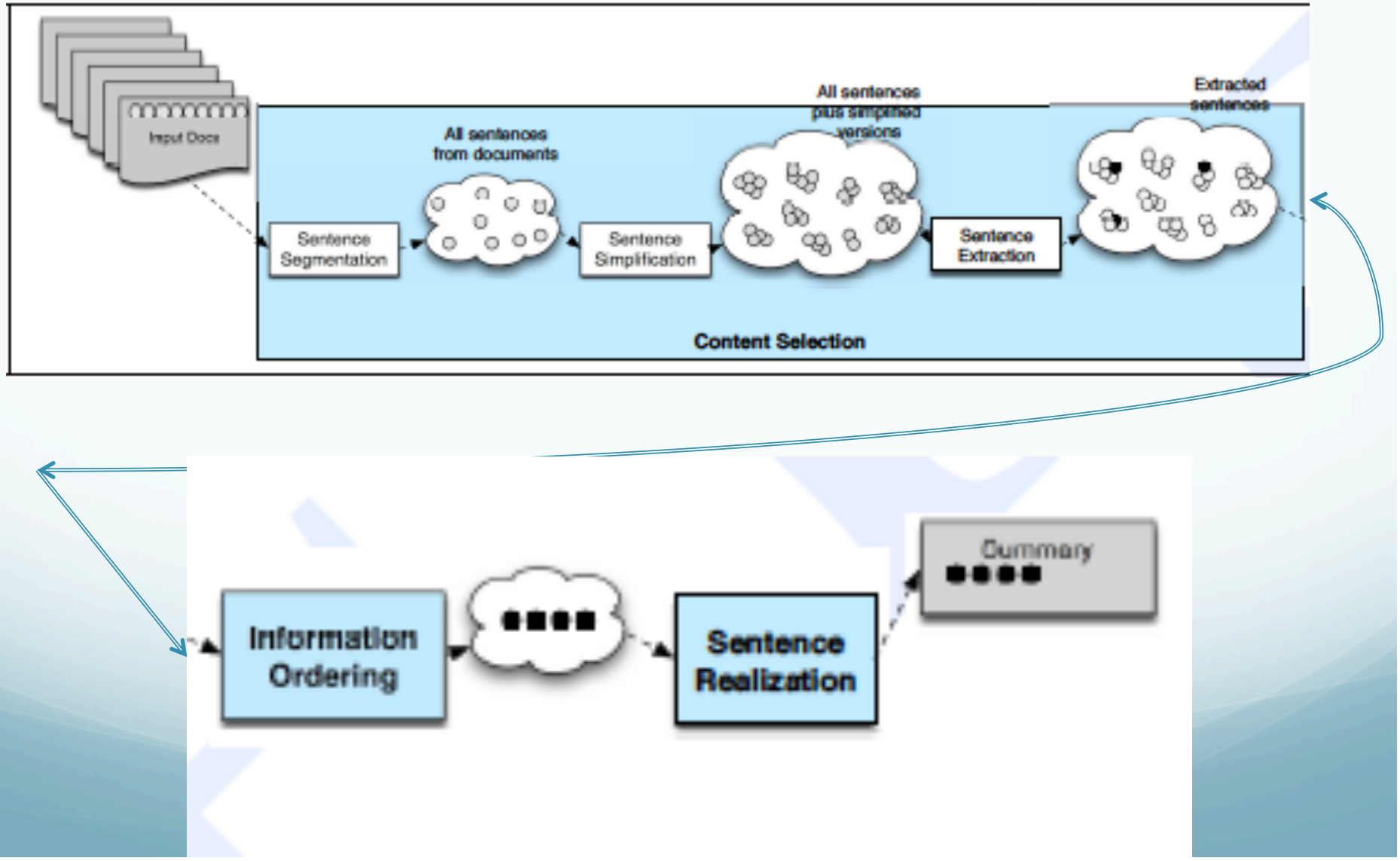
Dimensions of Summarization

- Input/Output form factors:
  - Language: Evaluations include:
    - English, Arabic, Chinese, Japanese, multilingual
  - Register: Formality, style
  - Genre: e.g. News, sports, medical, technical,....
  - Structure: forms, tables, lists, web pages
  - Medium: text, speech, video, tables
  - Subject

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

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

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