Summarization: Overview

Ling573
Systems & Applications
April 2, 2015

Roadmap

- Deliverable #1
- Dimensions of the problem
- A brief history: Shared tasks & Summarization
- Architecture of a Summarization system
- Summarization and resources
- Evaluation
- Logistics Check-in

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

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

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

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

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:

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:

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

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

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

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

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?

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

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

Shared Tasks: Perspective

- Late ‘80s-90s:

Shared Tasks: Perspective

- Late ‘80s-90s:
  - ATIS: spoken dialog systems
  - MUC: Message Understanding: information extraction

Shared Tasks: Perspective

- Late ‘80s-90s:
  - ATIS: spoken dialog systems
  - MUC: Message Understanding: information extraction
- TREC (Text Retrieval Conference)
  - Arguably largest (often >100 participating teams)
  - Longest running (1992-current)
  - Information retrieval (and related technologies)
    - Actually hasn’t had ‘ad-hoc’ since ~2000, though
  - Organized by NIST

TREC Tracks

- Track: Basic task organization

TREC Tracks

- Track: Basic task organization
- Previous tracks:
  - Ad-hoc – Basic retrieval from fixed document set

TREC Tracks

- Track: Basic task organization
- Previous tracks:
  - Ad-hoc – Basic retrieval from fixed document set
  - Cross-language – Query in one language, docs in other
    - English, French, Spanish, Italian, German, Chinese, Arabic

TREC Tracks

- Track: Basic task organization
- Previous tracks:
  - Ad-hoc – Basic retrieval from fixed document set
  - Cross-language – Query in one language, docs in other
    - English, French, Spanish, Italian, German, Chinese, Arabic
  - Genomics

TREC Tracks

- Track: Basic task organization
- Previous tracks:
  - Ad-hoc – Basic retrieval from fixed document set
  - Cross-language – Query in one language, docs in other
    - English, French, Spanish, Italian, German, Chinese, Arabic
  - Genomics
  - Spoken Document Retrieval

TREC Tracks

- Track: Basic task organization
- Previous tracks:
  - Ad-hoc – Basic retrieval from fixed document set
  - Cross-language – Query in one language, docs in other
    - English, French, Spanish, Italian, German, Chinese, Arabic
  - Genomics
  - Spoken Document Retrieval
  - Video search

TREC Tracks

- Track: Basic task organization
- Previous tracks:
  - Ad-hoc – Basic retrieval from fixed document set
  - Cross-language – Query in one language, docs in other
    - English, French, Spanish, Italian, German, Chinese, Arabic
  - Genomics
  - Spoken Document Retrieval
  - Video search
  - Question Answering

Other Shared Tasks

- International:
  - CLEF (Europe); FIRE (India)

Other Shared Tasks

- International:
  - CLEF (Europe); FIRE (India)
- Other NIST:
  - Machine Translation
  - Topic Detection & Tracking

Other Shared Tasks

- International:
  - CLEF (Europe); FIRE (India)
- Other NIST:
  - Machine Translation
  - Topic Detection & Tracking
- Various:
  - CoNLL (NE, parsing,..); SENSEVAL: WSD; PASCAL (morphology); BioNLP (biological entities, relations)

Other Shared Tasks

- International:
  - CLEF (Europe); FIRE (India)
- Other NIST:
  - Machine Translation
  - Topic Detection & Tracking
- Various:
  - CoNLL (NE, parsing,..); SENSEVAL: WSD; PASCAL (morphology); BioNLP (biological entities, relations)
  - Mediaeval (multi-media information access)

Summarization History

- "The Automatic Creation of Literature Abstracts"
  - Luhn, 1956
    - Early IBM system based on word, sentence statistics
- 1993 Dagstuhl seminar:
  - Meeting launched renewed interest in summarization
- 1997 ACL summarization workshop

Summarization Campaigns

- SUMMAC: (1998)
  - Initial cross-system evaluation campaign
- DUC (Document Understanding Conference)
  - 2001-2007
    - Increasing complexity, including multi-document, topic-oriented, multi-lingual
    - Developed systems and evaluation in tandem
- NTCIR (3 years)
  - Single, multi-document; Japanese

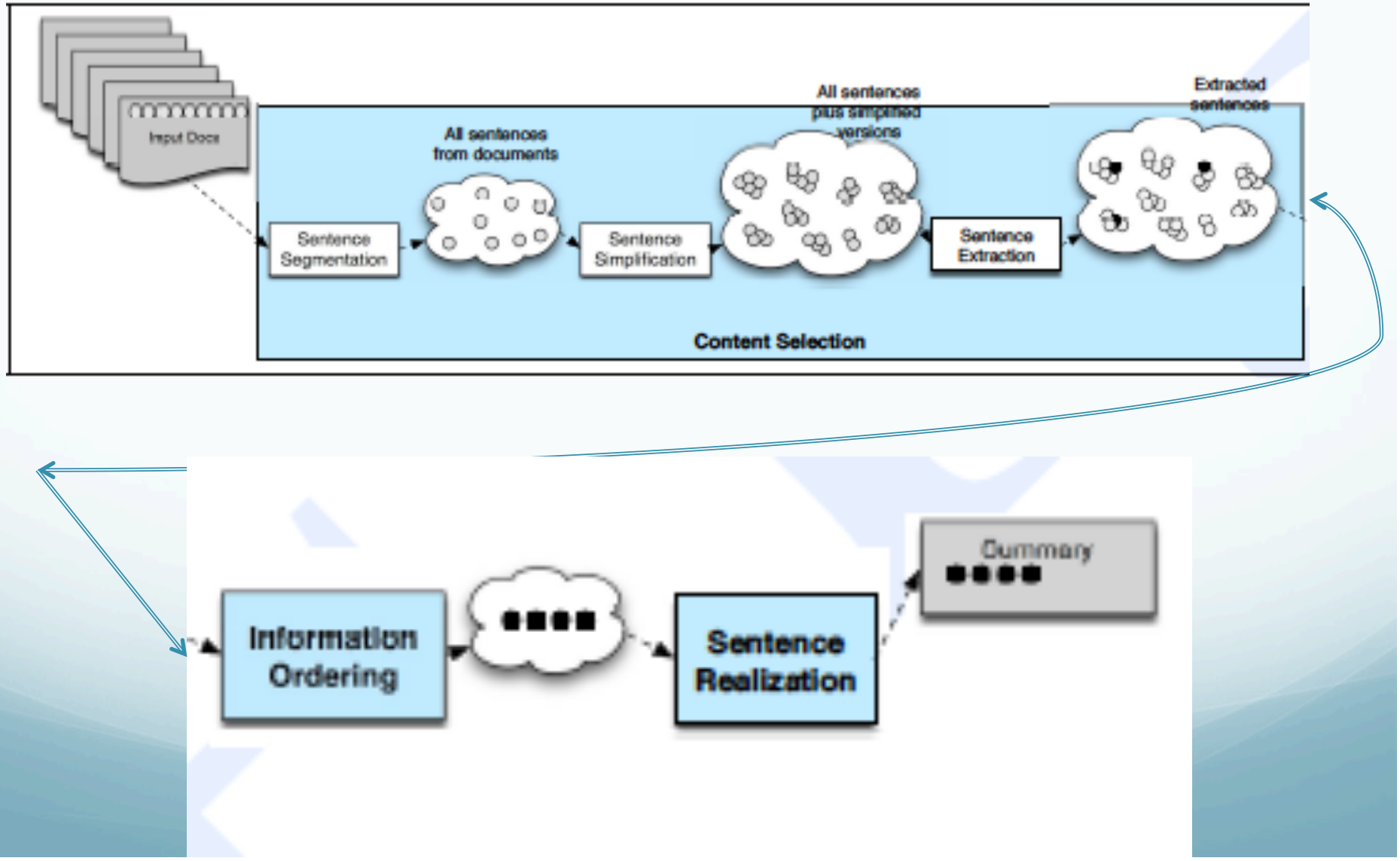
Most Recent Summarization Campaigns

- TAC (Text Analytics Conference): 2008---current
  - Variety of tasks
    - Summarization systems:
      - Opinion
      - Update
      - Guided
      - Multi-lingual
    - Automatic evaluation methodology

Summarization Tasks

- Provide:
  - Lists of topics (e.g.”guided” summarization)
  - Document collections (licensed via LDC, NIST)
  - Lists of relevant documents
  - Validation tools
  - Evaluation tools: Model summaries, systems
  - Derived resources:
  - Reams of related publications

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

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
  - Do any necessary processing to improve coherence

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

Cues to Saliency

- Approaches significantly differ in terms of cues
- Word-based (unsupervised):
  - Compute a **topic signature** of words above threshold

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

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

More Complex Settings

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

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

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

Information Ordering

- Goal: Determine presentation order for salient content
- Relatively trivial for single document extractive case:

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?

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

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

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

Content Realization

- Goal: Create a fluent, readable, compact output
- Abstractive approaches range from templates to full NLG

Content Realization

- Goal: Create a fluent, readable, compact output
- Abstractive approaches range from templates to full NLG
- Extractive approaches focus on:

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

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**.

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 …

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:

Evaluation

- Extrinsic evaluations:
  - Does the summary allow users to perform some task?
    - As well as full docs? Faster?

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

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?

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)

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:

Intrinsic Evaluation

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

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)

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

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:

ROUGE

- Pros:
  - Automatic evaluation allows tuning
    - Given set of reference summaries
  - Simple measure
- Cons:

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

Topics

- <topic id = "D0906B" category = "1">
  - <title> Rains and mudslides in Southern California </title>
    - <docsetA id = "D0906B-A">
      - <doc id = "AFP_ENG_20050110.0079" />
      - <doc id = "LTW_ENG_20050110.0006" />
      - <doc id = "LTW_ENG_20050112.0156" />
      - <doc id = "NYT_ENG_20050110.0340" />
      - <doc id = "NYT_ENG_20050111.0349" />
      - <doc id = "LTW_ENG_20050109.0001" />
      - <doc id = "LTW_ENG_20050110.0118" />
      - <doc id = "NYT_ENG_20050110.0009" />
      - <doc id = "NYT_ENG_20050111.0015" />
      - <doc id = "NYT_ENG_20050112.0012" />
    - </docset> <docsetB id = "D0906B-B">
      - <doc id = "AFP_ENG_20050221.0700" />
      - .....

Documents

- <DOC><DOCNO> APW20000817.0002 </DOCNO>
- <DOCTYPE> NEWS STORY </DOCTYPE><DATE_TIME> 2000-08-17 00:05 </ DATE_TIME>
- <BODY> <HEADLINE> 19 charged with drug trafficking </HEADLINE>
- <TEXT><P>
- UTICA, N.Y. (AP) - Nineteen people involved in a drug trafficking ring in the Utica area were arrested early Wednesday, police said.
- </P><P>
- Those arrested are linked to 22 others picked up in May and comprise "a major cocaine, crack cocaine and marijuana distribution organization," according to the U.S. Department of Justice.
- </P>

Model Summaries

- <SUM>
- <aid="1.2">In January 2005</aid="1.2">, <aid="1.7">rescue workers <aid="1.3">in southern California</aid="1.3"> used snowplows, snowcats and snowmobiles to free <aid="1.5">people</aid="1.5"> from a highway where</aid="1.7"> <aid="1.1">snow, sleet, rain and fog caused a 200-vehicle logjam</aid="1.1">. <aid="1.1">A fourth day of storms took a heavy toll as saturated hillsides gave way</aid="1.1">, <aid="1.6">mudslides inundating houses and closing highways</aid="1.6">. <aid="1.5">People fled neighborhoods up and down the coast.</aid="1.5"> Eight of nine horse races at Santa Anita were canceled for the first time in 10 years. <aid="1.6">More than 6,000 houses were without power</aid="1.6"> <aid="1.3">in Los Angeles</aid="1.3">. A scientist said Los Angeles had not seen such intensity of winter downpours since 1889-90.
- </SUM>

Reminder

- Team up!