Topic-Orientation & Information Ordering

Ling573
Systems & Applications
April 21, 2016

Notes

- Deliverable 2:
  - Code/results
  - Updated project report
  - Presentations next week:
    - Doodle poll will be sent after class
    - Please email me slide deck (or pointer) by noon
    - If planning to present remotely, contact me to check audio

Deliverable #3

- Goals:
  - Focus on information ordering
    - Using one or more of:
      - Chronology, Cohesion, Coherence
  - Continue to improve content selection
    - Incorporate some guided/topic-orientation
- Same deliverable structure as D#2
  - Due in 3 weeks:
    - Code/results; Updated report

Roadmap

- Topic-focused summarization
  - Focusing existing approaches
    - LexRank
    - CLASSY, FastSumm
- Information Ordering:
  - Basic approaches
    - Variants on chronological ordering
  - Enhancing cohesion

Key Idea

- (aka ”query-focused”, “guided”)
- Motivations:
  - Extrinsic task vs generic
    - Why are we creating this summary?
      - Viewed as complex question answering (vs factoid)
  - High variation in human summaries
    - Depending on perspective different content focused
- Idea:
  - Target response to specific question, topic in docs
    - Later TACs identify topic categories and aspects
      - E.g Natural disasters: who, what, where, when..

Query-focused LexRank

- Focus on sentences relevant to query
  - Rather than uniform jump
- How do we measure relevance?
  - Tf*idf-like measure over sentences & query
    - Compute sentence-level "idf"
      - N = # of sentences in cluster; sf_w = # of sentences with w

$$idf_w = \log\left(\frac{N+1}{0.5+sf_w}\right)$$

$$rel(s \mid q) = \sum_{w \in q} \log(tf_{w,s}+1) * \log(tf_{w,q}+1) * idf_w$$

Updated LexRank Model

- Combines original similarity weighting w/query

Updated LexRank Model

- Combines original similarity weighting w/query
  - Mixture model of query relevance, sentence similarity

Updated LexRank Model

- Combines original similarity weighting w/query
  - Mixture model of query relevance, sentence similarity

$$p(s \mid q) = d \frac{rel(s \mid q)}{\sum_{z \in C} rel(z \mid q)} + (1-d) \sum_{v \in C} \frac{sim(s,v)}{\sum_{z \in C} sim(z,v)} p(v \mid q)$$

  - d controls ‘bias’: i.e. relative weighting

Tuning & Assessment

- Parameters:
  - Similarity threshold: filters adjacency matrix
  - Question bias: Weights emphasis on question focus

Tuning & Assessment

- Parameters:
  - Similarity threshold: filters adjacency matrix
  - Question bias: Weights emphasis on question focus
- Parameter sweep:
  - Best similarity threshold: 0.14-0.2
    - As before
  - Best question bias: high: 0.8-0.95

Tuning & Assessment

- Parameters:
  - Similarity threshold: filters adjacency matrix
  - Question bias: Weights emphasis on question focus
- Parameter sweep:
  - Best similarity threshold: 0.14-0.2
    - As before
  - Best question bias: high: 0.8-0.95
- Question bias in LexRank can improve

Other Strategies

- Methods depend on base system design
  - All aim to incorporate similarity with query/topic

Other Strategies

- Methods depend on base system design
  - All aim to incorporate similarity with query/topic
- CLASSY HMM:

Other Strategies

- Methods depend on base system design
  - All aim to incorporate similarity with query/topic
- CLASSY HMM:
  - Add question overlap feature to HMM vector

Other Strategies

- Methods depend on base system design
  - All aim to incorporate similarity with query/topic
- CLASSY HMM:
  - Add question overlap feature to HMM vector
    - Log (# query tokens in sentence + 1)
      - Query tokens: tagged as noun, verb, adj, adv, or proper nouns

Other Strategies

- Methods depend on base system design
  - All aim to incorporate similarity with query/topic
- CLASSY HMM:
  - Add question overlap feature to HMM vector
    - Log (# query tokens in sentence + 1)
      - Query tokens: tagged as noun, verb, adj, adv, or proper nouns
  - Other, more aggressive approach detrimental
- FastSumm: SVM regression on sentences

Other Strategies

- Methods depend on base system design
  - All aim to incorporate similarity with query/topic
- CLASSY HMM:
  - Add question overlap feature to HMM vector
    - Log (# query tokens in sentence + 1)
      - Query tokens: tagged as noun, verb, adj, adv, or proper nouns
  - Other, more aggressive approach detrimental
- FastSumm: SVM regression on sentences
  - Adds topic title frequency feature:
    - Proportion of words in sent which appear in title

Overview

- Many similar strategies:
  - Features, weighting, ranking: overlap based

Overview

- Many similar strategies:
  - Features, weighting, ranking: overlap based
- Actual evaluation impact:
  - Not necessarily very large (e.g. 0.003 ROUGE)
    - But can be useful

Overview

- Many similar strategies:
  - Features, weighting, ranking: overlap based
- Actual evaluation impact:
  - Not necessarily very large (e.g. 0.003 ROUGE)
    - But can be useful
  - Aggressive approaches can have large negative impact
    - I.e. explicitly adding NER spans

Optimization Approaches to Reducing Redundancy

- DPP: Determinantal Point Processes (Kulesza &Taskar, '12)
  - Set models balancing information importance w/diversity
- ICSISumm: Uses Integer Linear Programming frame
  - Optimizes coverage of key bigrams weighted by doc freq
- OCCAMS_V
  - Uses LSA (Latent Semantic Analysis) to weight terms
  - Sentence selection via optimization problems:
    - Budgeted maximal coverage; knapsack

Information Ordering

Basics

- Content selection:
  - Identified sentences or information units for summary

Basics

- Content selection:
  - Identified sentences or information units for summary
- Information ordering:
  - Linearize selected content into a smooth-flowing text

Basics

- Content selection:
  - Identified sentences or information units for summary
- Information ordering:
  - Linearize selected content into a smooth-flowing text
- Factors:
  - Semantics

Basics

- Content selection:
  - Identified sentences or information units for summary
- Information ordering:
  - Linearize selected content into a smooth-flowing text
- Factors:
  - Semantics
    - Chronology: respect sequential flow of content (esp. events)
  - Discourse

Basics

- Content selection:
  - Identified sentences or information units for summary
- Information ordering:
  - Linearize selected content into a smooth-flowing text
- Factors:
  - Semantics
    - Chronology: respect sequential flow of content (esp. events)
  - Discourse
    - Cohesion: Adjacent sentences talk about same thing
    - Coherence: Adjacent sentences naturally related (PDTB)

Single vs Multi-Document

- Strategy for single-document summarization?

Single vs Multi-Document

- Strategy for single-document summarization?
  - Just keep original order
  - Chronology? Cohesion? Coherence?
- Multi-document

Single vs Multi-Document

- Strategy for single-document summarization?
  - Just keep original order
  - Chronology? Cohesion? Coherence?
- Multi-document
  - “Original order” can be problematic
  - Chronology?

Single vs Multi-Document

- Strategy for single-document summarization?
  - Just keep original order
  - Chronology? Cohesion? Coherence?
- Multi-document
  - “Original order” can be problematic
  - Chronology?
    - Publication order vs document-internal order
    - Differences in document ordering of information

Single vs Multi-Document

- Strategy for single-document summarization?
  - Just keep original order
  - Chronology? Cohesion? Coherence?
- Multi-document
  - “Original order” can be problematic
  - Chronology?
    - Publication order vs document-internal order
    - Differences in document ordering of information
  - Cohesion?
  - Coherence?

Single vs Multi-Document

- Strategy for single-document summarization?
  - Just keep original order
  - Chronology? Ok Cohesion? Ok Coherence? Iffy
- Multi-document
  - “Original order” can be problematic
  - Chronology?
    - Publication order vs document-internal order
    - Differences in document ordering of information
  - Cohesion? Probably poor
  - Coherence? Probably poor

Example

- Hemingway, 69, died of natural causes in a Miami jail after being arrested for indecent exposure.
- A book he wrote about his father, “Papa: A Personal Memoir”, was published in 1976.
- He was picked up last Wednesday after walking naked in Miami.
- “He had a difficult life.”
- A transvestite who later had a sex-change operation, he suffered bouts of drinking, depression and drifting according to acquaintances.
- “It’s not easy to be the son of a great man,” Scott Donaldson, told Reuters.

A Bad Example

- Hemingway, 69, died of natural causes in a Miami jail after being arrested for indecent exposure.
- A book he wrote about his father, “Papa: A Personal Memoir”, was published in 1976.
- He was picked up last Wednesday after walking naked in Miami.
- “He had a difficult life.”
- A transvestite who later had a sex-change operation, he suffered bouts of drinking, depression and drifting according to acquaintances.
- “It’s not easy to be the son of a great man,” Scott Donaldson, told Reuters.

A Basic Approach

- Publication chronology:
- Given a set of ranked extracted sentences
- Order by:

A Basic Approach

- Publication chronology:
- Given a set of ranked extracted sentences
- Order by:
  - Across articles
    -

A Basic Approach

- Publication chronology:
- Given a set of ranked extracted sentences
- Order by:
  - Across articles
    - By publication date
  - Within articles

A Basic Approach

- Publication chronology:
- Given a set of ranked extracted sentences
- Order by:
  - Across articles
    - By publication date
  - Within articles
    - By original sentence ordering
- Clearly not ideal, but used in some eval. submissions