Discourse & Topic-orientation

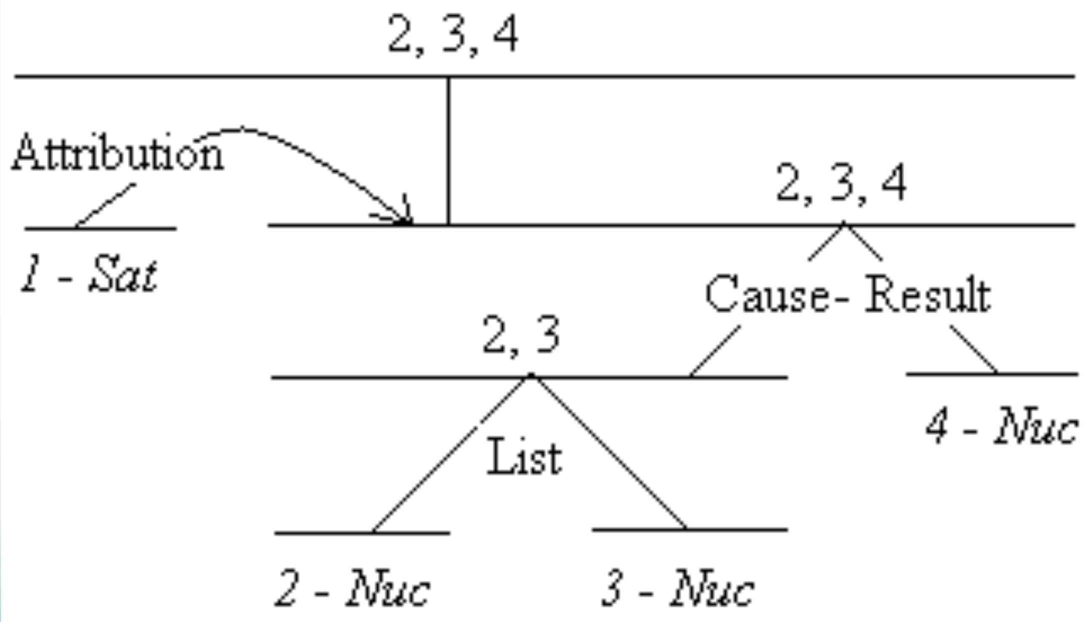
Ling 573
Systems & Applications
April 21, 2015

Roadmap

- Discourse for content selection:
  - Discourse Structure
  - Discourse Relations
  - Results
- Topic-orientation
  - Key idea
  - Common strategies

Discourse Structure Example

- 1. [Mr. Watkins said] 2. [volume on Interprovincial's system is down about 2% since January] 3. [and is expected to fall further,] 4. [making expansion unnecessary until perhaps the mid-1990s.]



Discourse Structure Features

- Satellite penalty:
  - For each EDU: # of satellite nodes b/t it and root
    - 1 satellite in tree: (1), one step to root: penalty = 1

Discourse Structure Features

- Satellite penalty:
  - For each EDU: # of satellite nodes b/t it and root
    - 1 satellite in tree: (1), one step to root: penalty = 1
- Promotion set:
  - Nuclear units at some level of tree
    - At leaves, EDUs are themselves nuclear

Discourse Structure Features

- Satellite penalty:
  - For each EDU: # of satellite nodes b/t it and root
    - 1 satellite in tree: (1), one step to root: penalty = 1
- Promotion set:
  - Nuclear units at some level of tree
    - At leaves, EDUs are themselves nuclear
- Depth score:
  - Distance from lowest tree level to EDUs highest rank
    - 2,3,4: score= 4; 1: score= 3

Discourse Structure Features

- Satellite penalty:
  - For each EDU: # of satellite nodes b/t it and root
    - 1 satellite in tree: (1), one step to root: penalty = 1
- Promotion set:
  - Nuclear units at some level of tree
    - At leaves, EDUs are themselves nuclear
- Depth score:
  - Distance from lowest tree level to EDUs highest rank
    - 2,3,4: score= 4; 1: score= 3
- Promotion score:
  - # of levels span is promoted:
    - 1: score = 0; 4: score = 2; 2,3: score = 3

Converting to Sentence Level

- Each feature has:
  - Raw score
  - Normalized score: Raw/sentence_length

Converting to Sentence Level

- Each feature has:
  - Raw score
  - Normalized score: Raw/sentence_length
- Sentence score for a feature:
  - Max over EDUs in sentence

“Semantic” Features

- Capture specific relations on spans
- Binary features over tuple of:
  - Implicit vs Explicit
  - Name of relation that holds
    - Top-level or second level
  - If relation is between sentences,
    - Indicate whether Arg1 or Arg2
- E.g. “contains Arg1 of Implicit Restatement relation”
- Also, # of relations, distance b/t args w/in sentence

Non-discourse Features

- Typical features:

Non-discourse Features

- Typical features:
  - Sentence length
  - Sentence position
  - Probabilities of words in sent: mean, sum, product
  - # of signature words (LLR)

Significant Features

- Associated with summary sentences
  - Structure: depth score, promotion score

Significant Features

- Associated with summary sentences
  - Structure: depth score, promotion score
  - Semantic: Arg1 of Explicit Expansion, Implicit Contingency, Implicit Expansion, distance to arg

Significant Features

- Associated with summary sentences
  - Structure: depth score, promotion score
  - Semantic: Arg1 of Explicit Expansion, Implicit Contingency, Implicit Expansion, distance to arg
  - Non-discourse: length, 1st in para, offset from end of para, # signature terms; mean, sum word probabilities

Significant Features

- Associated with non-summary sentences
  - Structural: satellite penalty

Significant Features

- Associated with non-summary sentences
  - Structural: satellite penalty

  - Semantic: Explicit expansion, explicit contingency, Arg2 of implicit temporal, implicit contingency,...
    - # shared relations

Significant Features

- Associated with non-summary sentences
  - Structural: satellite penalty
  - Semantic: Explicit expansion, explicit contingency, Arg2 of implicit temporal, implicit contingency,…
    - # shared relations
  - Non-discourse: offset from para, article beginning; sent. probability

Observations

- Non-discourse features good cues to summary
- Structural features match intuition

- Semantic features:
  - Relatively few useful for selecting summary sentences
    - Most associated with non-summary, but most sentences are non-summary

Evaluation

- Structural best:
  - Alone and in combination
- Best overall combine all types
  - Both F-1 and ROUGE

Features used	**Acc**	**P**	**R**	**F**
structural	78.11	63.38	22.77	33.50
semantic	75.53	44.31	5.04	9.05
non-discourse (ND)	77.25	67.48	11.02	18.95
ND + semantic	77.38	59.38	20.62	30.61
ND + structural	78.51	63.49	26.05	36.94
semantic + structural	77.94	58.39	30.47	40.04
structural + semantic + ND	78.93	61.85	34.42	44.23

Graph-Based Comparison

- Page-Rank-based centrality computed over:
  - RST link structure
  - Graphbank link structure
  - LexRank (sentence cosine similarity)

Graph-Based Comparison

- Page-Rank-based centrality computed over:
  - RST link structure
  - Graphbank link structure
  - LexRank (sentence cosine similarity)
- Quite similar:
  - F1: LR > GB > RST
  - ROUGE: RST > LR > GB

Notes

Notes

- Single document, short (100 wd) summaries
  - What about multi-document? Longer?

- Structure relatively better, all contribute

Notes

- Single document, short (100 wd) summaries
  - What about multi-document? Longer?
- Structure relatively better, all contribute
- Manually labeled discourse structure, relations
  - Some automatic systems, but not perfect
    - However, better at structure than relation ID
      - Esp. implicit

Topic-Orientation

Key Idea

- (aka ”query-focused”, “guided”)

- Motivations:

Key Idea

- (aka ”query-focused”, “guided”)

- Motivations:
  - Extrinsic task vs generic
    - Why are we creating this summary?
      - Viewed as complex question answering (vs factoid)

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

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..

Basic Strategies

- Adapt existing generic summarization strategies
  - Augment techniques to focus on query/topic
    - E.g. query-focused LexRank, query-focused CLASSY

Basic Strategies

- Adapt existing generic summarization strategies
  - Augment techniques to focus on query/topic
    - E.g. query-focused LexRank, query-focused CLASSY
- Information extraction strategies
  - View topic category + aspects as template
    - Similar to earlier MUC tasks
  - Identify entities, sentences to complete
  - Generate summary

Basic Strategies

- **Most common approach →**
- **Adapt existing generic summarization strategies**
  - Augment techniques to focus on query/topic
    - E.g. query-focused LexRank, query-focused CLASSY
- Information extraction strategies
  - View topic category + aspects as template
    - Similar to earlier MUC tasks
  - Identify entities, sentences to complete
  - Generate summary

Focusing LexRank

- Original Continuous LexRank:
  - Compute sentence centrality by similarity graph
  - Weighting:

Focusing LexRank

- Original Continuous LexRank:
  - Compute sentence centrality by similarity graph
  - Weighting: cosine similarity between sentences
  - Damping factor ‘d’ to jump to other clusters (uniform)

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj(u)} \frac{\cos sim(u,v)}{\sum_{z \in adj(v)} \cos sim(z,v)} p(v)$$

Focusing LexRank

- Original Continuous LexRank:
  - Compute sentence centrality by similarity graph
  - Weighting: cosine similarity between sentences
  - Damping factor ‘d’ to jump to other clusters (uniform)

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj(u)} \frac{\cos sim(u,v)}{\sum_{z \in adj(v)} \cos sim(z,v)} p(v)$$

- Given a topic (American Tobacco Companies Overseas)
  - How can we focus the summary?

Query-focused LexRank

- Focus on sentences relevant to query
  - Rather than uniform jump

Query-focused LexRank

- Focus on sentences relevant to query
  - Rather than uniform jump
- How do we measure relevance?

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

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

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
    - 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

TAC 2010 Results

- For context:
  - LEAD baseline: first 100 words of chron. last article

System	ROUGE-2
LEAD baseline	0.06410
MEAD	0.08682
Best (NUS)	0.13440