Ordering by Optimization &Content Realization

Ling573
Systems and Applications
May 10, 2016

Roadmap

- Ordering by Optimization
- Content realization
  - Goals
  - Broad approaches
  - Implementation exemplars

Ordering as Optimization

- Given a set of sentences to order
- Define a local pairwise coherence score b/t sentences
- Compute a total order optimizing local distances
- Can we do this efficiently?
  - Optimal ordering of this type is equivalent to TSP
    - Traveling Salesperson Problem: Given a list of cities and distances between cities, find the shortest route that visits each city exactly once and returns to the origin city.
    - TSP is NP-complete (NP-hard)

Ordering as TSP

- Can we do this practically?
  - Summaries are 100 words, so 6-10 sentences
    - 10 sentences have how many possible orders? O(n!)
    - Not impossible

  - Alternatively,
    - Use an approximation methods
    - Take the best of a sample

CLASSY 2006

- Formulates ordering as TSP
- Requires pairwise sentence distance measure
  - Term-based similarity: # of overlapping terms
  - Document similarity:
    - Multiply by a weight if in the same document (there, 1.6)
  - Normalize to between 0 and 1 (sqrt of product of selfsim)
    - Make distance: subtract from 1

Practicalities of Ordering

- Brute force: O(n!)
  - "there are **only** 3,628,800 ways to order 10 sentences plus a lead sentence, so exhaustive search is feasible." (Conroy)
- Still,..
  - Used sample set to pick best
    - Candidates:
      - Random
      - Single-swap changes from good candidates
  - 50K enough to consistently generate minimum cost order

Conclusions

- Many cues to ordering:
  - Temporal, coherence, cohesion
    - Chronology, topic structure, entity transitions, similarity
- Strategies:
  - Heuristic, machine learned; supervised, unsupervised
  - Incremental build-up versus generate & rank
- Issues:
  - Domain independence, semantic similarity, reference

Content Realization

Goals of Content Realization

- Abstractive summaries:
  - Content selection works over concepts
  - Need to produce important concepts in fluent NL
- Extractive summaries:
  - Already working with NL sentences
  - Extreme compression: e.g 60 byte summaries: headlines
  - Increase information:
    - Remove verbose, unnecessary content
    - More space left for new information
  - Increase readability, fluency
    - Present content from multiple docs, non-adjacent sents
  - Improve content scoring
    - Remove distractors, boost scores: i.e. % signature terms in doc

Broad Approaches

- Abstractive summaries:
  - Complex Q-A: template-based methods
  - More generally: full NLG: concept-to-text
- Extractive summaries:
  - Sentence compression:
    - Remove "unnecessary" phrases:
      - Information? Readability?
  - Sentence reformulation:
    - Reference handling
      - Information? Readability?
  - Sentence fusion: Merge content from multiple sents

Sentence Compression

- Main strategies:
  - Heuristic approaches
    - Deep vs Shallow processing
    - Information- vs readability- oriented
  - Machine-learning approaches
    - Sequence models
      - HMM, CRF
    - Deep vs Shallow information
  - Integration with selection
    - Pre/post-processing; Candidate selection: heuristic/learned

Form	CLASSY	ISCI	UMd	SumBasic+	Cornell
Initial Adverbials	Y	M	Y	Y	Y
Initial Conj	Y		Y	Y	
Gerund Phr.	Y	M	M	Y	M
Rel clause appos	Y		M	Y	Y
Other adv	Y				
Numeric: ages,	Y				
Junk (byline, edit)	Y				Y
Attributives	Y	Y		Y	Y
Manner modifiers	M	Y	M		Y
Temporal modifiers	M	Y	Y		Y
POS: det, that, MD			Y		
XP over XP			Y		
PPs (w/, w/o constraint)			Y		
Preposed Adjuncts			Y		
SBARs			Y		M
Conjuncts			Y		
Content in parentheses		Y			Y

Shallow, Heuristic

- CLASSY 2006
  - Pre-processing! Improved ROUGE
    - Previously used automatic POS tag patterns: error-prone
- Lexical & punctuation surface-form patterns
  - “function” word lists: Prep, conj, det; adv, gerund; punct
- Removes:
  - Junk: bylines, editorial
  - Sentence-initial adv, conj phrase (up to comma)
  - Sentence medial adv (“also”), ages
  - Gerund (-ing) phrases
  - Rel. clause attributives, attributions w/o quotes
- Conservative: < 3% error (vs 25% w/POS)

Deep, Minimal, Heuristic

- ICSI/UTD:
  - Use an Integer Linear Programming approach to solve
- Trimming:
  - Goal: Readability (not info squeezing)
  - Removes temporal expressions, manner modifiers, “said”
    - Why?: “next Thursday”
  - Methodology: Automatic SRL labeling over dependencies
    - SRL not perfect: How can we handle?
    - Restrict to high-confidence labels
- Improved ROUGE on (some) training data
  - Also improved linguistic quality scores

Example

A ban against bistros providing plastic bags free of charge will be lifted at the beginning of March.

A ban against bistros providing plastic bags free of charge will be lifted.

Deep, Extensive, Heuristic

- Both UMD & SumBasic+
  - Based on output of phrase structure parse
  - UMD: Originally designed for headline generation
  - Goal: Information squeezing, compress to add content
- Approach: (UMd)
  - Ordered cascade of increasingly aggressive rules
    - Subsumes many earlier compressions
    - Adds headline oriented rules (e.g. removing MD, DT)
    - Adds rules to drop large portions of structure
      - E.g. halves of AND/OR, wholescale SBAR/PP deletion

Integrating Compression & Selection

- Simplest strategy: (Classy, SumBasic+)
  - Deterministic, compressed sentence replaces original
- Multi-candidate approaches: (most others)
  - Generate sentences at multiple levels of compression
    - Possibly constrained by: compression ratio, minimum len
      - E.g. exclude: < 50% original, < 5 words (ICSI)
  - Add to original candidate sentences list
  - Select based on overall content selection procedure
    - Possibly include source sentence information
    - E.g. only include single candidate per original sentence

Multi-Candidate Selection

- (UMd, Zajic et al. 2007, etc)

- Sentences selected by tuned weighted sum of feats
  - Static:
    - Position of sentence in document
    - Relevance of sentence/document to query
    - Centrality of sentence/document to topic cluster
      - Computed as: IDF overlap or (average) Lucene similarity
    - # of compression rules applied
  - Dynamic:
    - Redundancy: $S=\Pi_{wi \text{ in } S} \lambda P(w|D) + (1-\lambda)P(w|C)$
    - # of sentences already taken from same document

- Significantly better on ROUGE-1 than uncompressed
  - Grammaticality lousy (tuned on headlinese)