Evaluation & Systems

Ling573
Systems & Applications
April 7, 2016

Roadmap

- Evaluation:
  - Scoring without models
- Content selection:
  - Unsupervised word-weighting approaches
- Non-trivial baseline system example:
  - MEAD
- Deliverable #2

Model-free Evaluation

- Techniques so far rely on human model summaries
- How well can we do without?
  - What can we compare summary to instead?
    - Input documents
  - Measures?
    - Distributional: Jensen-Shannon, Kullback-Liebler divergence
      - Vector similarity (cosine)
    - Summary likelihood: unigram, multinomial
    - Topic signature overlap

Assessment

- Correlation with manual score-based rankings
  - Distributional measure well-correlated, sim to ROUGE2

Features	pyramid	respons.
JS div	-0.880	-0.736
JS div smoothed	-0.874	-0.737
% of input topic words	0.795	0.627
KL div summ-inp	-0.763	-0.694
cosine overlap	0.712	0.647
% of summ = topic wd	0.712	0.602
topic overlap	0.699	0.629
KL div inp-summ	-0.688	-0.585
mult. summary prob.	0.222	0.235
unigram summary prob.	-0.188	-0.101
regression	0.867	0.705
ROUGE-1 recall	0.859	0.806
ROUGE-2 recall	0.905	0.873

Shared Task Evaluation

- Multiple measures:
  - Content:
    - Pyramid (recent)
    - ROUGE-n often reported for comparison
  - Focus: Responsiveness
    - Human evaluation of topic fit (1-5 (or 10))
  - Fluency: Readability (1-5)
    - Human evaluation of text quality
    - 5 linguistic factors: grammaticality, non-redundancy, referential clarity, focus, structure and coherence.

Content Selection

- Many dimensions:
  - Information-source based:
    - Words, discourse (position, structure), POS, NER, etc
  - Learner-based:
    - Supervised – classification/regression, unsup, semi-sup
  - Models:
    - Graphs, LSA, ILP, submodularity, Info-theoretic, LDA

Word-Based Unsupervised Models

- Aka “Topic Models” in (Nenkova, 2001)
  - What is the topic of the input?
  - Model what the content is “about”
- Typically unsupervised – Why?
  - Hard to label, no pre-defined topic inventory
- How do we model, identify aboutness?
  - Weighting on surface:
    - Frequency, tf*idf, LLR
  - Identifying underlying concepts (LSA, EM, LDA, etc)

Frequency-based Approach

- Intuitions:
  - Frequent words in doc indicate what it's about
  - Repetition across documents reinforces importance
  - Differences w/background further focus
- Evidence: Human summaries have higher likelihood
- Word weight = p(w) = relative frequency = c(w)/N
- Sentence score: (averaged) weights of its words

$$Score(S) = \frac{1}{|S|} \sum_{w \in S_i} p(w)$$

Selection Methodology

- Implemented in SumBasic (Nenkova et al)
  - Estimate word probabilities from doc(s)

  - Pick sentence containing highest scoring word
    - With highest sentence score
      - Having removed stopwords

  - Update word probabilities
    - Downweight those in selected sentence: avoid redundancy
      - E.g. square their original probabilities

  - Repeat until max length

Word Weight Example

1. Bombing Pan Am…
2. Libya Gadafhi supports…
3. Trail suspects…
4. UK and USA…

Word	Weight
Pan	0.0798
Am	0.0825
Libya	0.0096
Supports	0.0341
Gadafhi	0.0911 ….

Libya refuses to surrender two Pan Am bombing suspects.

Nenkova, 2011

Limitations of Frequency

- Basic approach actually works fairly well
- However, misses some key information
  - No notion of foreground/background contrast
    - Is a word that's frequent everywhere a good choice?
  - Surface form match only
    - Want concept frequency, not just word frequency
      - WordNet, LSA, LDA, etc

Modeling Background

- Capture contrasts between:
  - Documents being summarized
  - Other document content
- Combine with frequency "aboutness" measure
- One solution:
  - TF*IDF
    - Term Frequency: # of occurrences in document (set)
    - Inverse Document Frequency: df = # docs w/word
      - Typically: $IDF = \log(N/df_w)$
  - Raw weight or threshold

Topic Signature Approach

- Topic signature: (Lin & Hovy, 2001; Conroy et al, 2006)
  - Set of terms with saliency above some threshold
- Many ways to select:
  - E.g. tf*idf (MEAD)
- Alternative: Log Likelihood Ratio (LLR) $\lambda(w)$
  - Ratio of:
    - Probability of observing w in cluster and background corpus
      - Assuming same probability in both corpora
        - Vs
      - Assuming different probabilities in both corpora

Log Likelihood Ratio

- k_1= count of w in topic cluster
- k_2= count of w in background corpus
- n_1= # features in topic cluster; n_2=# in background
- $p_1=k_1/n_1$; $p_2=k_2/n_{2;}$ $p= (k_1+k_2)/(n_1+n_2)$

- $L(p,k,n) = p^k (1 - p)^{n-k}$

$$-2log\lambda = 2[logL(p_1, k_1, n_1) + logL(p_2, k_2, n_2) \\ - logL(p, k_1, n_1) - logL(p, k_2, n_2)]$$

Using LLR for Weighting

- Compute weight for all cluster terms
  - weight(w_i) = 1 if $-2\log \lambda > 10$, 0 o.w.
- Use that to compute sentence weights

$$weight(s_i) = \sum_{w \in s_i} \frac{weight(w)}{|\{w | w \in s_i\}|}$$

- How do we use the weights?
  - One option: directly rank sentences for extraction
- LLR-based systems historically perform well
  - Better than tf*idf generally

Deliverable #2

- Goals:
  - Become familiar with shared task summarization data
  - Implement initial base system with all components
  - Focus on content selection
  - Evaluate resulting summaries

TAC 2010 Shared Task

- Basic data:
  - Test Topic Statements:
    - Brief topic description
    - List of associated document identifiers from corpus
  - Document sets:
    - Drawn from AQUAINT/AQUAINT-2 LDC corpora
      - Available on patas
  - Summary results:
    - Model summaries

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>

Notes

- Topic files:
  - Include both docsetA and docsetB
    - Use ONLY *docsetA*
      - “B” used for update task
- IDs reference documents in AQUAINT corpora

Notes

- AQUAINT/AQUAINT-2 corpora
  - Subset of Gigaword
    - Used in many NLP shared tasks
  - Format is SGML
    - Not fully XML compliant
      - Includes non-compliant characters: e.g. with &s
      - May not be "rooted"
    - Some differences between subcorpora
  - Span different date ranges

Tips & Tricks

- Handling SGML with XML tools
  - Elementtree has recover mode:
    - E.g. parser = etree.XMLParser(recover=True)
      data_tree = etree.parse(f, parser)
  - Consider escaping &-prefixed content
  - Varied paragraph structure:
    - .xpath(".//TEXT//P|.//TEXT")
- Non-uniform corpora:
  - You may hard-code corpus handling
    - Or create configuration files

Model Summaries

- Five young Amish girls were killed, shot by a lone gunman.
- At about 1045, on October 02, 2006, the gunman, Charles Carl Roberts IV, age 32, entered the Georgetown Amish School in Nickel Mines, Pennsylvania, a tiny village about 55 miles west of Philadelphia.
- He let the boys and the adults go, before he tied up the girls, ages 6 to 13.
- Police and emergency personnel rushed to the school but the gunman killed himself as they arrived.
- His motive was unclear but in a cell call to his wife he talked about abusing two family members 20 years ago.

Initial System

- Implement end-to-end system
  - From reading in topic files to summarization to eval
- Need at least basic components for:
  - Content selection
  - Information ordering
  - Content realization
- Focus on content selection for D2:
  - Must be non-trivial (i.e. non-random/lead)
  - Others can be minimal (i.e. “copy” for content real.)

Summaries

- Basic formatting:
  - 100 word summaries
  - Just ASCII, English sentences
  - No funny formatting (bullets, etc)
  - May output on multiple lines
  - One file per topic summary
  - All topics in single directory

Summarization Evaluation

- Primarily using ROUGE
  - Standard implementation
  - ROUGE-1, -2, -4:
    - Scores found to have best correlation with responsiveness
  - Primary metric: ROUGE Recall (“R”)
  - Store in results directory

Model & Output Names

- Topic id=D0901A
- Summary file name: D0901-A.M.100.A.A
- 1. Split document id on:
  - id_part1=D0901 and
  - id_part2=A
- 2. Construct filename as:
  - [id_part1]-[docset].M.[max_token_count].[id_part2].[some_unique_alphanum]

Submission

- Code/outputs due 4/22
  - Tag as D2
- Reports due 4/26 am
  - Should tag as D2.1
- Presentations week of 4/26
  - Will do doodle to set times