Evaluation & Systems

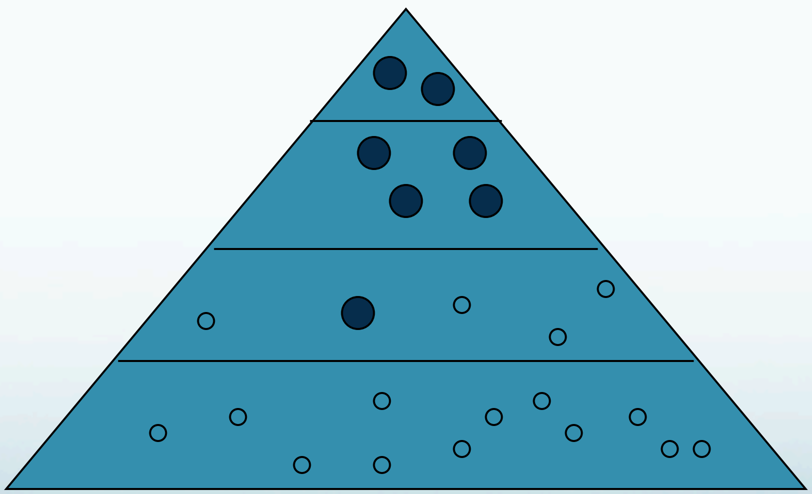
Ling573
Systems & Applications
April 9, 2015

Roadmap

- Evaluation:
  - Pyramid scoring
  - Scoring without models
- Systems:
  - MEAD
  - CLASSY
- Deliverable #2

Ideally informative summary

- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



From Passoneau et al 2005

Pyramid Scores

- T_i = tier with weight i SCUs
  - T_n = top tier; T_1 = bottom tier
- D_i = # of SCUs in summary on T_i
- Total weight of summary $D = \sum_{i=1}^{n} i * D_i$
- Optimal score for X SCU summary: *Max*
  - (j lowest tier in ideal summary)

$$\sum_{i=j+1}^{n} i * |T_i| + j * (X - \sum_{i=j+1}^{n} |T_i|)$$

Pyramid Scores

- Original Pyramid Score:
  - Ratio of D to Max
    - Precision-oriented

- Modified Pyramid Score:
  - X_a = Average # of SCUs in model summaries
  - Ratio of D to Max (using X_a)
    - More recall oriented (most commonly used)

Correlation with Other Scores

Table VI. Pearson's Correlation Between the Different Evaluation Metrics Used in DUC 2005. Computed for 25 Automatic Peers Over 20 Test Sets

	Pyr (mod)	Respons-1	Respons-2	ROUGE-2	ROUGE-SU4
Pyr (orig)	0.96	0.77	0.86	0.84	0.80
Pyr (mod)		0.81	0.90	0.90	0.86
Respons-1			0.83	0.92	0.92
Respons-2				0.88	0.87
ROUGE-2					0.98

- 0.95: effectively indistinguishable
  - Two pyramid models, two ROUGE models
- Two humans only 0.83

Pyramid Model

- Pros:
  - Achieves goals of handling variation, abstraction, semantic equivalence
  - Can be done sufficiently reliably
  - Achieves good correlation with human assessors
- Cons:
  - Heavy manual annotation:
    - Model summaries, also all system summaries
    - Content only

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 (recent): Pyramid
    - 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.

MEAD

- Radev et al, 2000, 2001, 2004

- Exemplar centroid-based summarization system
  - Tf-idf similarity measures
  - Multi-document summarizer
  - Publically available summarization implementation
    - (No warranty)
  - Solid performance in DUC evaluations

Main Ideas

- Select sentences central to cluster:
  - Cluster-based relative utility
    - Measure of sentence relevance to cluster

- Select distinct representative from equivalence classes
  - Cross-sentence information subsumption
    - Sentences including same info content said to subsume
      - A) John fed Spot; B) John gave food to Spot and water to the plants.
        - I(B) subsumes I(A)
      - If mutually subsume, form equivalence class

Centroid-based Models

- Assume clusters of topically related documents
  - Provided by automatic or manual clustering

- Centroid: “pseudo-document of terms with Count * IDF above some threshold”
  - Intuition: centroid terms indicative of topic
  - Count: # of term occurrences in cluster
    - (TF is average # of occurrences)
  - IDF: inverse document frequency
    - Computed over larger side corpus (e.g. full AQUAINT)

MEAD Content Selection

- Input:
  - Sentence segmented, cluster documents (n sents)
  - Compression rate: e.g. 20%
- Output: n * r sentence summary
- Select highest scoring sentences based on:
  - Centroid score
  - Position score
  - First-sentence overlap
  - (Redundancy)

Score Computation

- $Score(s_i) = w_c C_i + w_p P_i + w_f F_i$
  - $C_i = \Sigma_i C_{w,I}$
    - Sum over centroid values of words in sentence
  - $P_i = ((n-i+1)/n)*C_{max}$
    - Positional score: C_{max}:score of highest sent in doc
      - Scaled by distance from beginning of doc
  - $F_i = S_1*S_i$
    - Overlap with first sentence
    - TF-based inner product of sentence with first in doc
- Alternate weighting schemes assessed
  - Diff't optima in different papers

Managing Redundancy

- Alternative redundancy approaches:
  - Redundancymax:
    - Excludes sentences with cosine overlap > threshold
  - Redundancy penalty:
    - Subtracts penalty from computed score
      - R_s = 2 * # overlapping wds/(# wds in sentence pair)
        - Weighted by highest scoring sentence in set

System and Evaluation

- Information ordering:
  - Chronological by document date
- Information realization:
  - Pure extraction, no sentence revision
- Participated in DUC 2001, 2003
  - Among top-5 scoring systems
  - Varies depending on task, evaluation measure
- Solid straightforward system
  - Publicly available; will compute/output weights

CLASSY

- "Clustering, Linguistics and Statistics for Summarization Yield"
  - Conroy et al. 2000-2011
- Highlights:
  - High performing system
    - Often rank 1 in DUC/TAC, commonly used comparison
  - Topic signature-type system (LLR)
  - HMM-based content selection
  - Redundancy handling

Topic Signature Approach

- Topic signature:
  - 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>

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:
  - 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
  - Store in results directory

Submission

- Code/outputs due 4/24
- Reports due 4/28 am
  - Should tag as D2.1
- Presentations week of 4/28
  - Will do doodle to set times