

Summarization Evaluation & Systems

Ling573
Systems and Applications
April 4, 2017

Roadmap

- Summarization evaluation:
 - Intrinsic:
 - Model-based: ROUGE, Pyramid
 - Model-free
- Content selection
 - Model classes
 - Unsupervised word-based models
 - Sumbasic
 - LLR
 - MEAD

ROUGE

- Pros:
 - Automatic evaluation allows tuning
 - Given set of reference summaries
 - Simple measure
- Cons:
 - Even human summaries highly variable, disagreement
 - Poor handling of coherence
 - Okay for extractive, highly problematic for abstractive

Pyramid Evaluation

- Content selection evaluation:
 - Not focused on ordering, readability
- Aims to address issues in evaluation of summaries:
 - Human variation
 - Significant disagreement, use multiple models
 - Analysis granularity:
 - Not just “which sentence”; overlaps in sentence content
 - Semantic equivalence:
 - Extracts vs Abstracts:
 - Surface form equivalence (e.g. ROUGE) penalizes abstr.

Pyramid Units

- Step 1: Extract Summary Content Units (SCUs)
 - Basic content meaning units
 - Semantic content
 - Roughly clausal
 - Identified manually by annotators from model summaries
 - Described in own words (possibly changing)

Example

- A1. The industrial espionage case ...began with the hiring of Jose Ignacio Lopez, an employee of GM subsidiary Adam Opel, by VW as a production director.
- B3. However, he left GM for VW under circumstances, which ...were described by a German judge as “potentially the biggest-ever case of industrial espionage”.
- C6. He left GM for VW *in March 1993*.
- D6. The issue stems from the alleged recruitment of GM's ...procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's business colleagues.
- E1. *On March 16, 1993, ...* Agnacio Lopez De Arriortua, left his job as head of purchasing at General Motor's Opel, Germany, to become Volkswagen's Purchasing ... director.
- F3. *In March 1993,* Lopez and seven other GM executives moved to VW overnight.

Example SCUs

- SCU1 (w=6): Lopez left GM for VW
 - A1. the hiring of Jose Ignacio Lopez, an employee of GM . . . by VW
 - B3. he left GM for VW
 - C6. He left GM for VW
 - D6. recruitment of GM's . . . Jose Ignacio Lopez
 - E1. Agnacio Lopez De Arriortua, left his job . . . at General Motor's Opel . . .to become Volkswagen's . . . Director
 - F3. Lopez . . . GM . . . moved to VW
- SCU2 (w=3) Lopez changes employers in March 1993
 - C6 in March, 1993
 - E1. On March 16, 1993
 - F3. In March 1993

SCU: A cable car caught fire (Weight = 4)

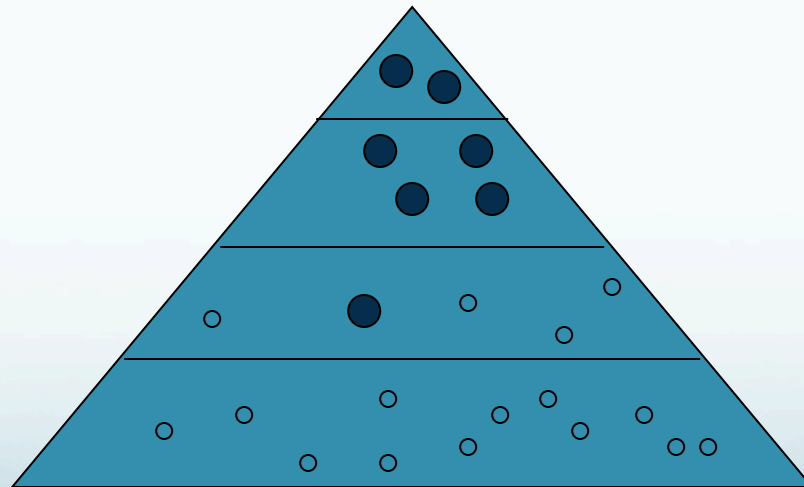
- A. The cause of the fire was unknown.
- B. A cable car caught fire just after entering a mountainside tunnel in an alpine resort in Kaprun, Austria on the morning of November 11, 2000.
- C. A cable car pulling skiers and snowboarders to the Kitzsteinhorn resort, located 60 miles south of Salzburg in the Austrian Alps, caught fire inside a mountain tunnel, killing approximately 170 people.
- D. On November 10, 2000, a cable car filled to capacity caught on fire, trapping 180 passengers inside the Kitzsteinhorn mountain, located in the town of Kaprun, 50 miles south of Salzburg in the central Austrian Alps.

Pyramid Building

- Step 2: Scoring summaries
 - Compute weights of SCUs
 - Weight = # of model summaries in which SCU appears
 - Create “pyramid”:
 - n = maximum # of tiers in pyramid = # of model summ.s
 - Actual # of tiers depends on degree of overlap
 - Highest tier: highest weight SCUs
 - Roughly Zipfian SCU distribution, so pyramidal shape
 - Optimal summary?
 - All from top tier, then all from top -1, until reach max size

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-Leibler 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, 2010)
 - 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.

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: $df = \# \text{ docs w/word}$
 - 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
- $p_1 = k_1/n_1$; $p_2 = k_2/n_2$; $p = (k_1 + k_2)/(n_1 + n_2)$
- $L(p, k, n) = p^k (1 - p)^{n-k}$

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

Using LLR for Weighting

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

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

- How do we use the weights?
 - One option: directly rank sentences for extraction
- LLR-based systems historically perform well
 - Better than $\text{tf} \cdot \text{idf}$ generally