Revisiting Summarization Evaluation for Scientific Articles

Arman Cohan and Nazli Goharian
Information Retrieval Lab, Department of Computer Science
Georgetown University
{arman, nazli}@ir.cs.georgetown.edu

Summarization Evaluation

- Evaluation of text summarization
  - Human assessors quantify the quality
    - Expensive and not reproducible
  - Using evaluation metrics
    - Human generated summaries as gold standard
    - Compare the system summary with the gold standard
    - ROUGE (Lin 2004)

Motivation & Background

- How to evaluate scientific summarization?
- How reliable is ROUGE (the most widely used metric) in this context?

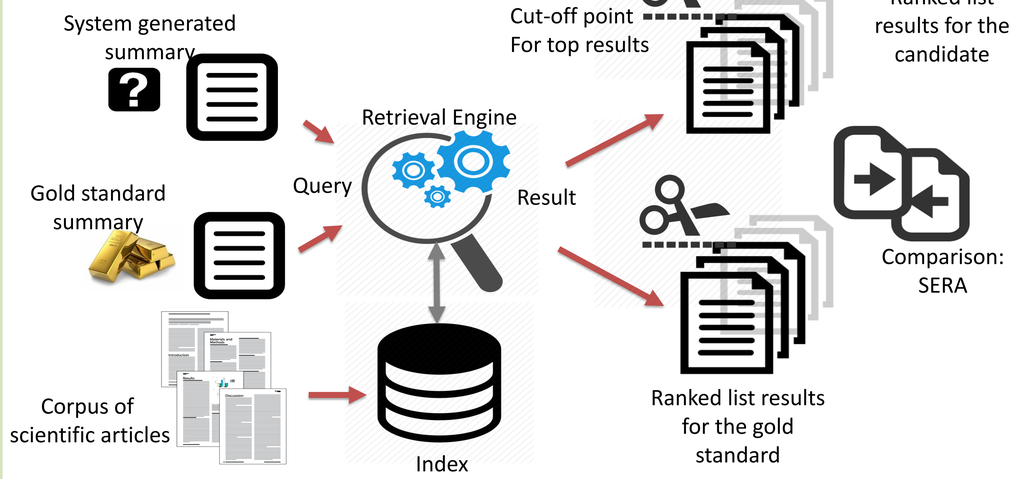
- ROUGE: based on textual overlaps
  - Many variants (often arbitrarily chosen)

- Scientific summarization is different
  - Longer articles and higher compression rate
  - Paraphrasing and terminology variations

- ROUGE has shown good correlations with human judgments on DUC 2001-2003 collections
  - DUC is composed of News articles (different with scientific papers)
- Can we still rely on ROUGE?

Proposed metric: SERA

- SERA = Summarization Evaluation by Relevance Analysis
- Based on the linguistic premise meaning comes from the context
- Based on finding common context for the summaries
- Rewards terms that are semantically related but not lexically equivalent



- Comparison by intersection of results (SERA)
- Comparison by discounted ranking difference (SERA-DIS)
- Variants:
  - Plain: Using the entire summary as query
  - Using only the key words of the summary as query (SERA-KW)
  - Using only the noun phrases of the summary as query (SERA-NP)

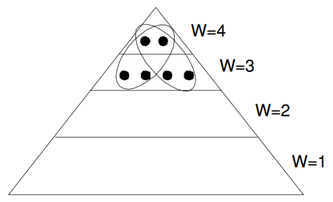
Evaluation & Results

- Data: TAC 2014 scientific summarization benchmark
- Evaluation: Semi-Manual evaluation method: Pyramid (Nenkova, et al 2007)
  - Uses gold-standard summaries to find important content in an ideal summary

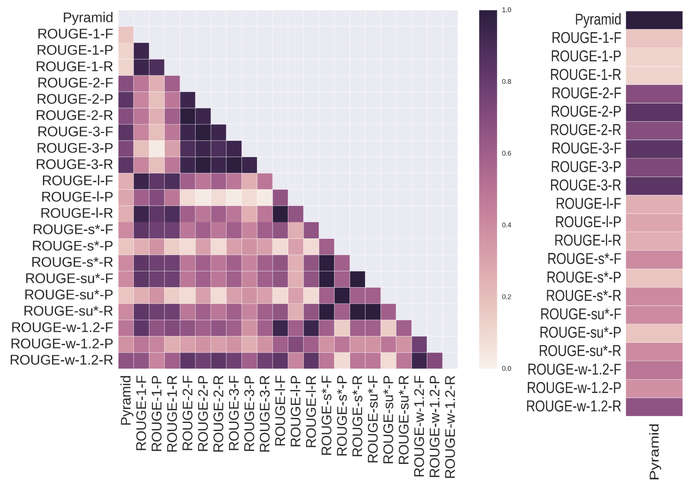
Example:

Id	Nugget	Tier
n_1	miRNA	3
n_2	IDH1/2	1
n_3	cell mutation	4

$$P = \frac{1}{P_{\max}} \sum_{i=1}^{n} i \times N_i$$



- Evaluation framework: Correlation analysis
- Compare ROUGE and SERA with Pyramid manual scores



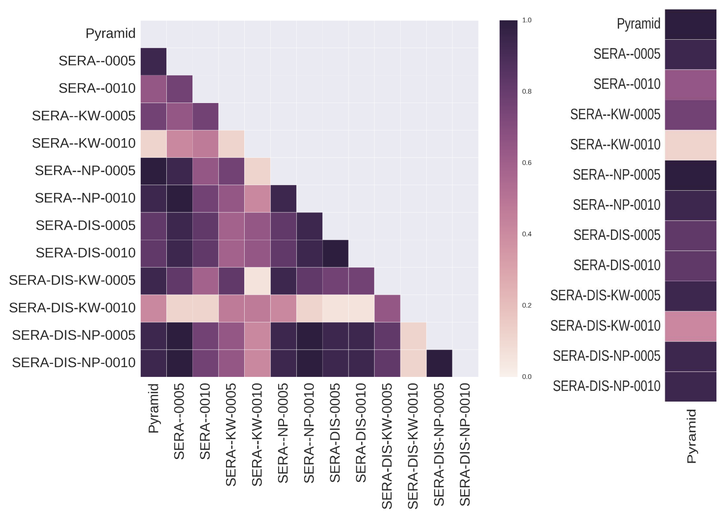
ROUGE: Weak Correlations

- Most variants are weakly correlated with pyramid manual judgments
- Rouge variants are not consistant
  They are weakly correlated
- ROUGE-2 and ROUGE-3 are the best performing

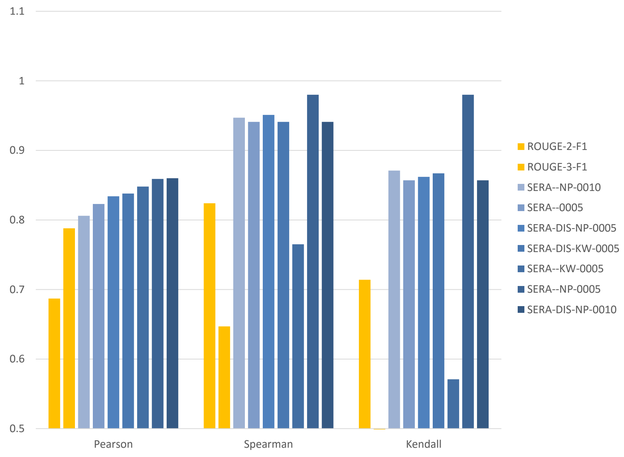
SERA: Strong Correlations

- Most vairants have strong correlation with manual judgements

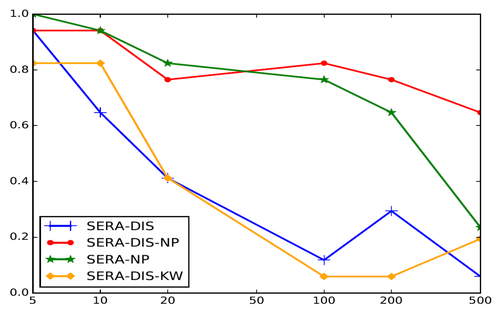
- SERA is robust as most variants correlate well



Direct Comparison with ROUGE



Effect of the Cut-off point



Conclusions

- We studied scientific summarization evaluation through correlation analysis
- We showed that most of ROUGE variants are not reliable for evaluating scientific summarization
- Among all ROUGE variants, ROUGE-2 and ROUGE-3 show the best results
- We proposed an alternative metric, SERA, which outperforms all ROUGE variants

Acknowledgements: This research was partially supported by NSF through grant CNS-1204347

https://github.com/acohan/TAC-Pyramid-Annotations