Scientific Article Summarization Using Citation-Context & Article's Discourse Structure

Arman Cohan

Nazli Goharian

EMNLP '15

Introduction: Scientific Summarization

- Summarizing scientific articles
  - Motivation: Keeping up with the developments in each scientific field
- Characteristics
  - Long articles
  - Coherence is not the focus
  - Follow a common inherent discourse
    - hypotheses, methods, experiments, findings

Scientific Summarization

- Abstracts
  - Biased towards author's viewpoint
  - Contributions are over[/under]stated
  - Do not contain all contributions
  - Stated contributions may not have real impact
  - Usually stated in a general and less focused fashion

DB-CSC: A Density-Based Approach for Subspace Clustering in Graphs with Feature Vectors

Stephan Günnemann, Brigitte Boden, and Thomas Seidl

RWTH Aachen University, Germany
{guennemann,boden,seidl}@cs.rwth-aachen.de

Abstract. Data sources representing attribute information in combination with network information are widely available in today's applications. To realize the full potential for knowledge extraction, mining techniques like clustering should consider both information types simultaneously. Recent clustering approaches combine *subspace clustering* with *dense subgraph mining* to identify groups of objects that are similar in subsets of their attributes as well as densely connected within the network. While those approaches successfully circumvent the problem of full-space clustering, their limited cluster definitions are restricted to clusters of certain shapes.

In this work, we introduce a density-based cluster definition taking the attribute similarity in subspaces and the graph density into account. This novel cluster model enables us to detect clusters of arbitrary shape and size. We avoid redundancy in the result by selecting only the most interesting non-redundant clusters. Based on this model, we introduce the clustering algorithm DB-CSC. In thorough experiments we demonstrate the strength of DB-CSC in comparison to related approaches.

1 Introduction

In the past few years, data sources representing attribute information in combination with network information have become more numerous. Such data describes single objects via attribute vectors and also relationships between different objects via edges. Examples include social networks, where friendship

Citation based summarization (Qazvinian, et al 2008, 2013)

- Using citations to an article for generating its summary
  - Capture various contributions
  - Include contributions with real impacts
  - More focused

  - Biased towards citing authors
  - Not accurate
  - Inconsistency b/w degree of certainty of findings

The proposed approach

- Improve the citation based approach towards scientific summarization
  - Overcome the problem of inaccuracy of the citations in citing the original work
  - Find the context of the citations in the reference article
  - Use inherent scientific article's discourse structure

The proposed approach: Overview

- Four steps:
  - Extract citation-context in the reference article
  - Group citation-contexts
  - Rank sentences in each group
  - Select sentences from each group

Citation-context

Citing article:

... The general impression that has emerged is that transformation of human cells by Ras requires the inactivation of both the pRb and p53 pathways, typically achieved by introduc[citation]r virus oncoproteins such as SV40 large tumor antigen (T-Ag) or human papillomavirus E6 and E7 proteins (Serrano et al., 1997).
To address this question, we have been investigating the ...

Reference article (Serrano et al., 1997):

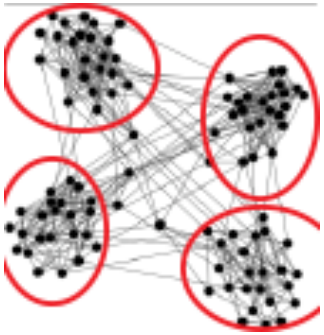
... continued to incorporate BrdU and proliferate following introduction of H-ras V12. In agreement with previous reports (66 and 60), both p53/ and p16/ MEFs expressing H-ras V12 displayed features of oncogenic transformation (e.g., refractile morp[citation-context]bition), which were apparent almost immediately [citation-context]uced (data not shown). These results indicate that [citation-context]or ras-induced arrest in MEFs, and that inactivation of either p53 or p16 alone is sufficient to circumvent arrest. In REF52 and IMR90 fibroblasts, a different approach was ...

1- Extracting citation-context

- Vector Space Model
  - Citation as source vector
  - Text spans in the reference article as target vector
  - Cosine similarity, tf-idf weighting, pivoted normalized vectors
- 4 different ways:
  - Preprocessed citation
  - Noun phrases
  - Keywords
  - Biomedical concepts

2- Grouping citation-contexts

- 1- Community detection (unsupervised)
  - Graph of citation-contexts
    - Edges: Cosine similarity b/w citation-contexts
  - Partition the graph into communities
  - Find sub-graphs that maximize modularity
  - Graph modularity:
    - Quantifies the denseness of the sub-graphs w.r.t. randomly distributed edges
    - Higher modularity: Denser sub-graphs



2- Grouping citation-contexts

- Find sub-graphs that maximize modularity
  - (Computationally intractable)
- Use a heuristic (Blondel et al., 2008)
  - Gain in modularity
  - Iteratively add nodes to subgraphs while there is a positive gain

2- Grouping citation-contexts

- 2- Using scientific articles inherent discourse structure (Supervised)
  - Relate each citation-context to one of the discourse facets of the article
    - (e.g. Hypothesis, Method, Results, Discussion)
  - Diversify the selection based on discourse
  - SVM classifier with ngram & verb features

3 & 4 - Sentence Ranking and Selection

- Rank most central sentences in each group
  - Using eigenvectors – Power method (Erkan 2004)
- Selecting sentences for final summary
  - Iterative
  - Diversity and Centrality (Greedy)
    - Linear interpolation of novelty and informativeness

Experiments: Data

- Data: TAC 2014 Biomedical Summarization dataset

	Mean	Std
# topics	20	0
# Gold summaries per topic	4	0
# citing articles in each topic	15.65	2.7
# citations in each citing article	1.57	1.17
Length of summaries (words)	235.64	31.24
Length of articles (words)	9759.86	2199.48

Experiments: Evaluation Metrics & Baselines

- ROUGE evaluation framework (Lin, 2004).
  - ROUGE-N
  - ROUGE-L
- Baselines
  - LexRank (Erkan and Radev, 2004)
  - LSA (Steinberger and Jezek, 2004)
  - MMR (Carbonell and Goldstein, 1998)
  - Citation summary (Qazvinian and Radev, 2008)

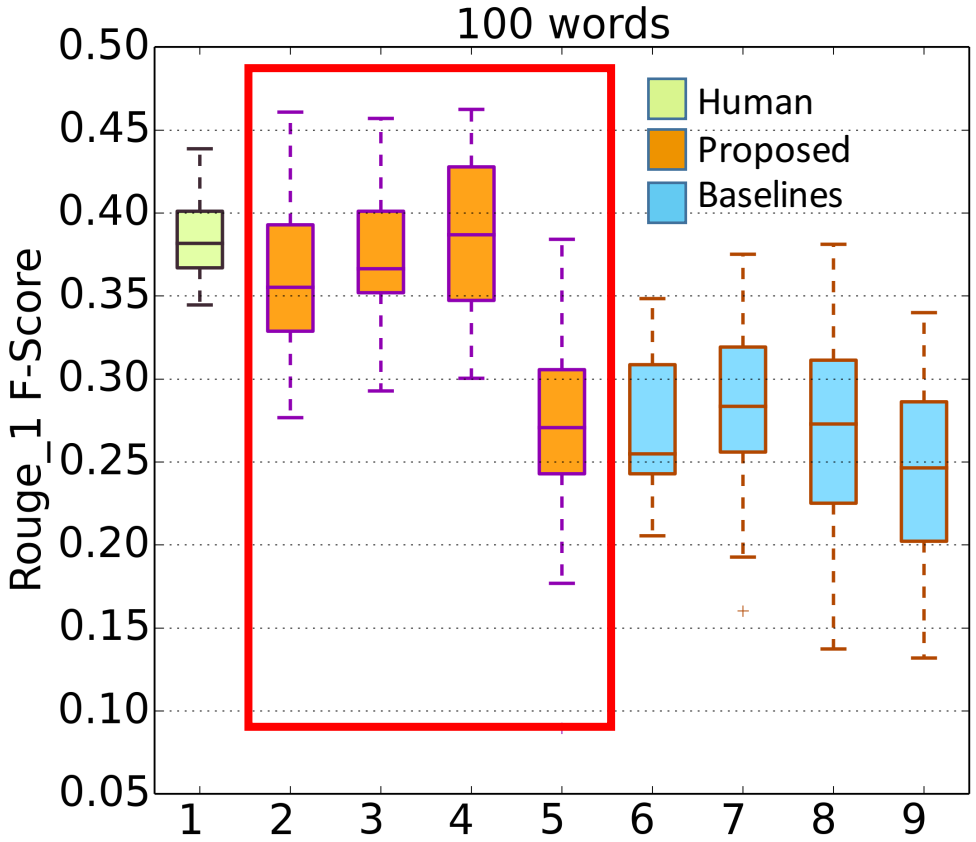
Research Questions

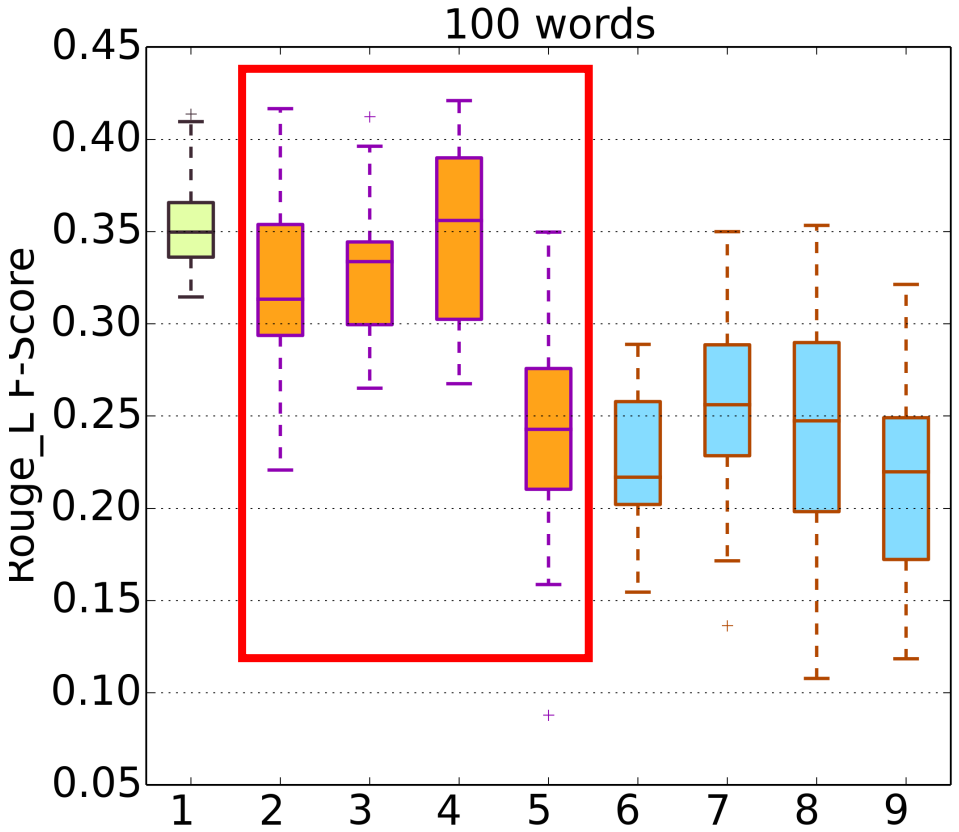
- Can citation-context improve over the existing baselines?
- How well does the discourse method perform?
- What is the effect of different citation-extraction approaches on the final summary?
- How well different sentence selection strategies work?

Results: In a nutshell

- Citation-context approach towards scientific summarization improves over the baselines
- Citations by themselves are not as effective
- Discourse based diversification improves performance
- In short summaries, sentence selection strategy matters more than longer summaries

Results

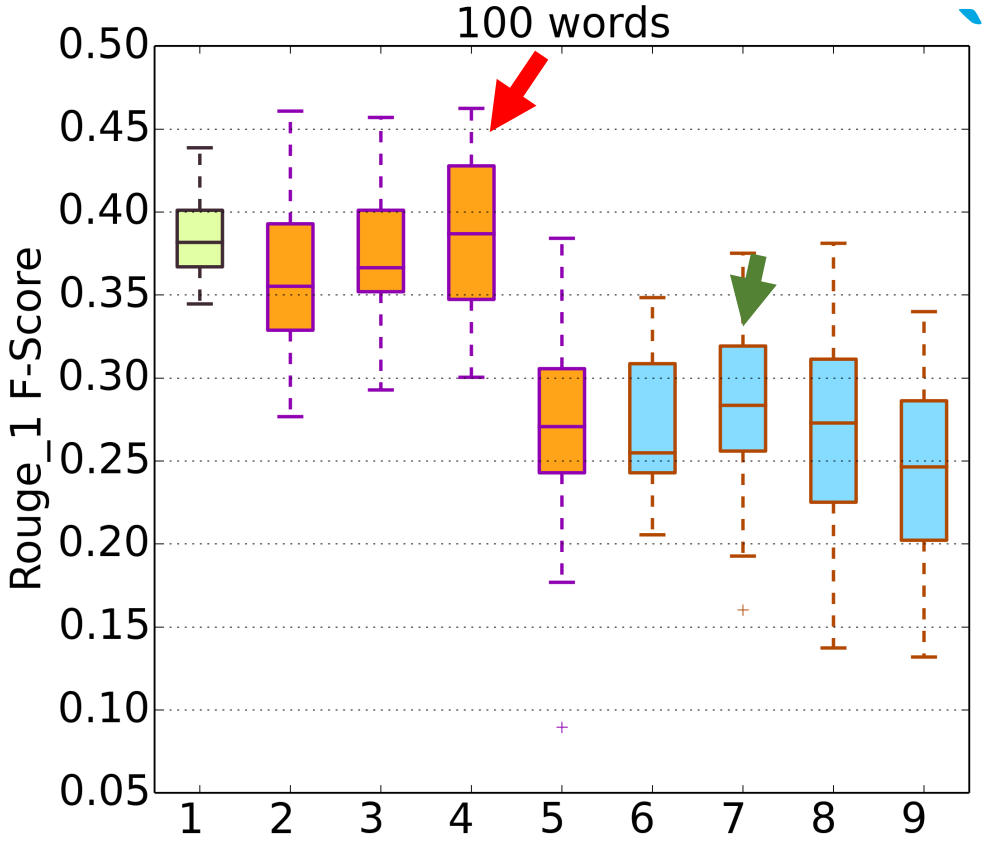


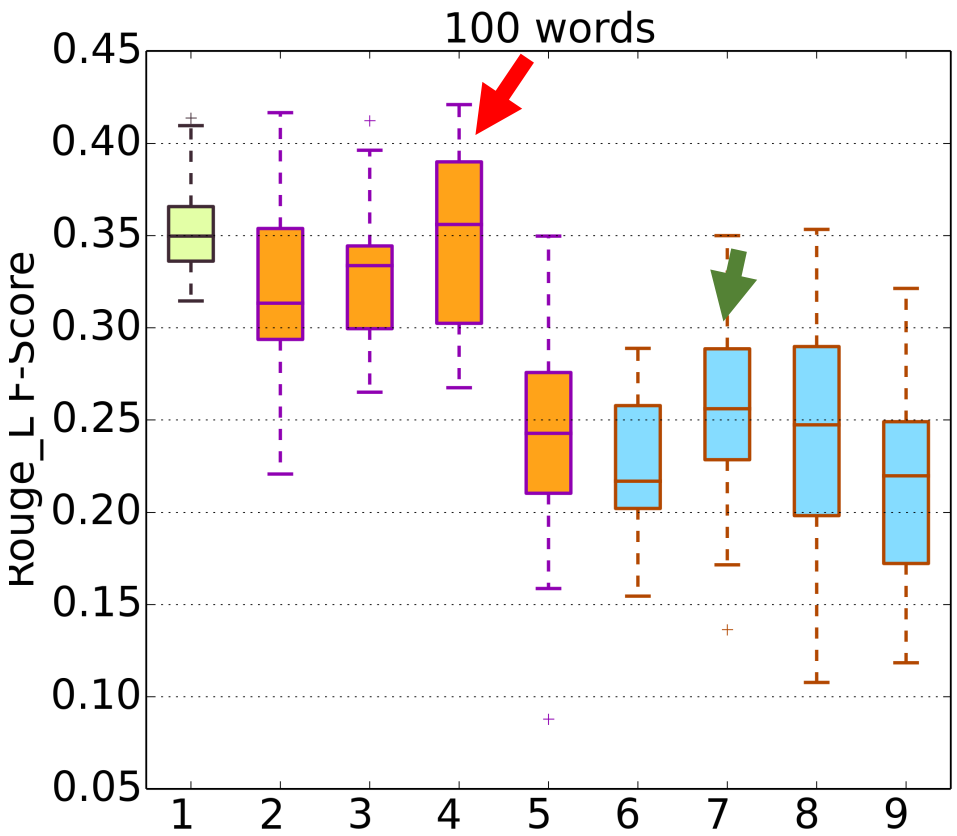


1: 'Human'
2: 'Community_Iterative'
3: 'Community_Diversity'
4: 'Discourse_Diversity'
5: 'Discourse_Iterative'
6: 'Citation Summary'
7: 'LexRank'
8: 'LSA'
9: 'MMR'

- Relatively encouraging performance of the proposed methods

Results: Short summaries

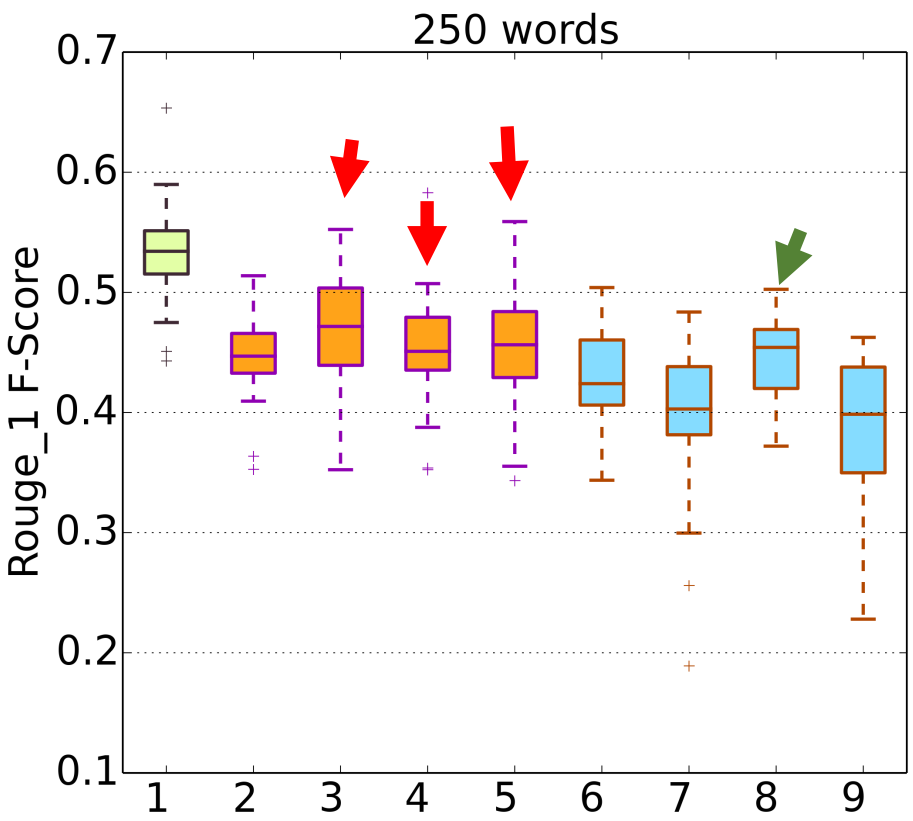


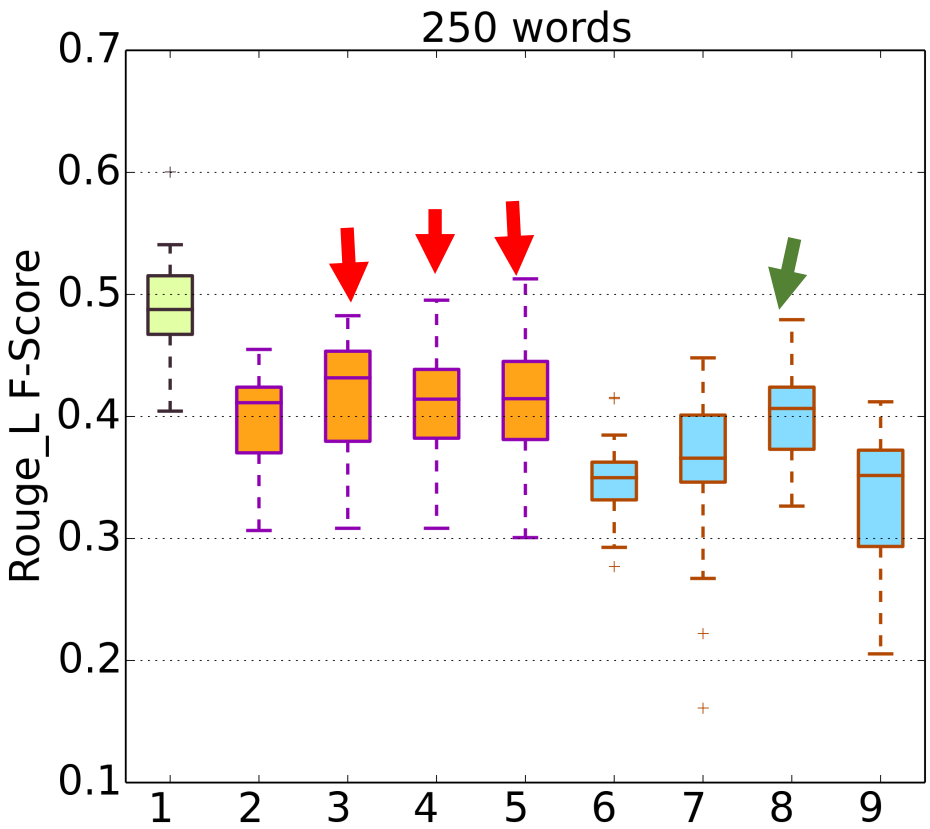


1: 'Human'
2: 'Community_Iterative'
3: 'Community_Diversity'
4: 'Discourse_Diversity'
5: 'Discourse_Iterative'
6: 'Citation Summary'
7: 'LexRank'
8: 'LSA'
9: 'MMR'

- In short summaries: Discourse based approach with selection by diversity performs the best
- LexRank: The best performing baseline

Results: Long summaries

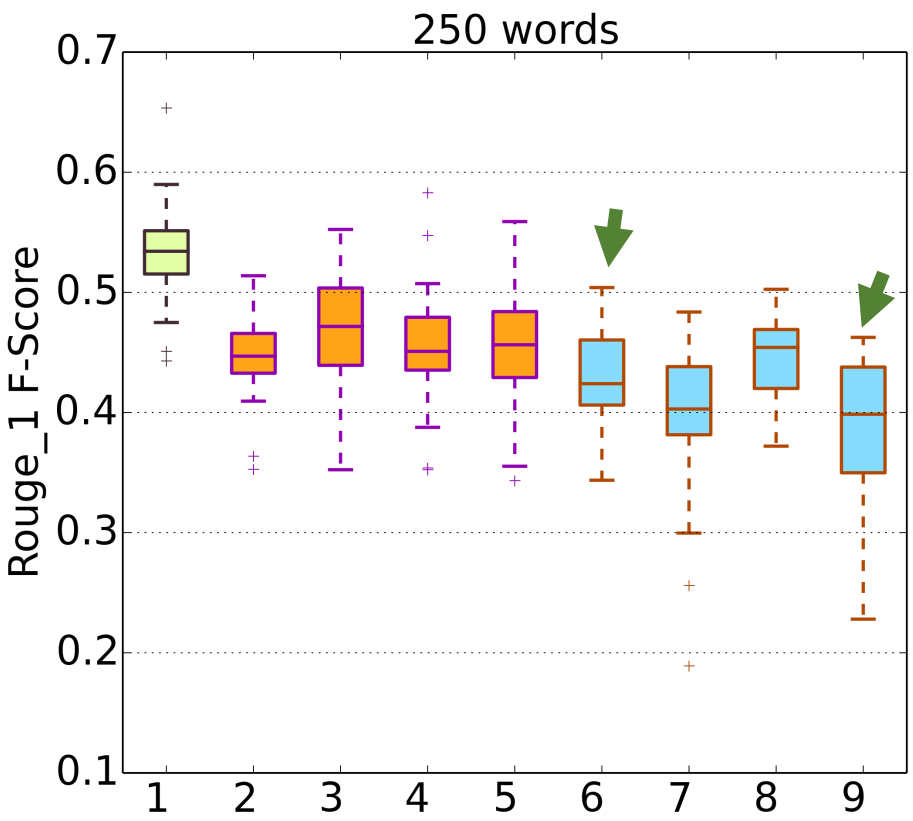


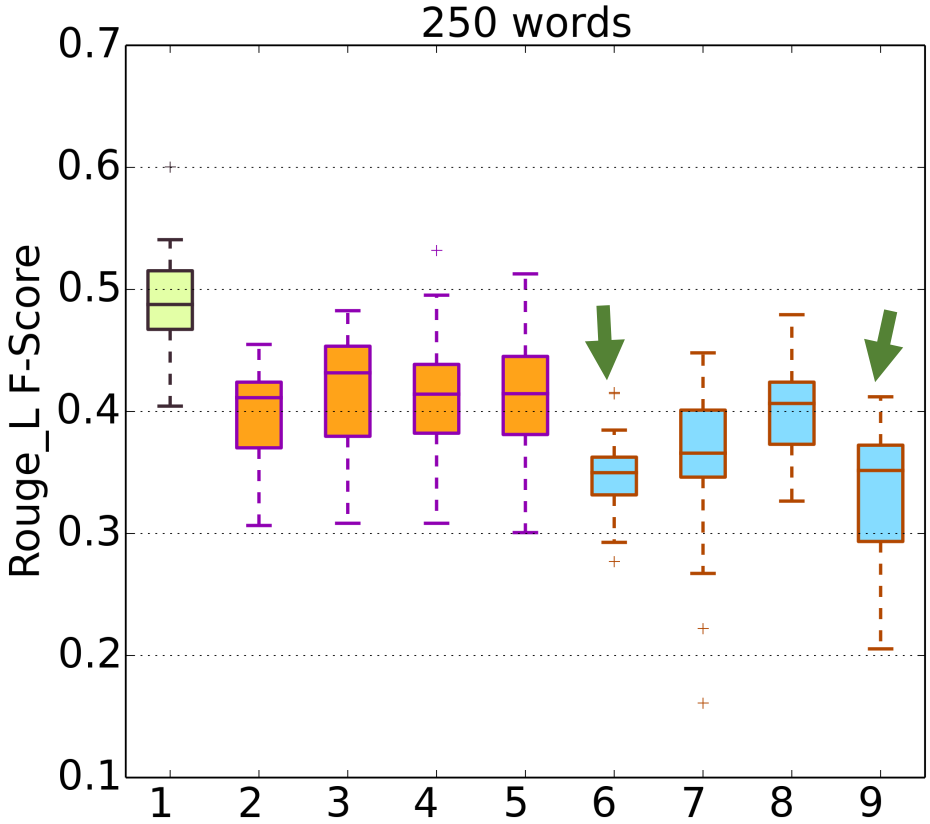


1: 'Human'
2: 'Community_Iterative'
3: 'Community_Diversity'
4: 'Discourse_Diversity'
5: 'Discourse_Iterative'
6: 'Citation Summary'
7: 'LexRank'
8: 'LSA'
9: 'MMR'

- In longer summaries: Both community based and discourse based approaches show relatively superior performance
- LSA: Best baseline

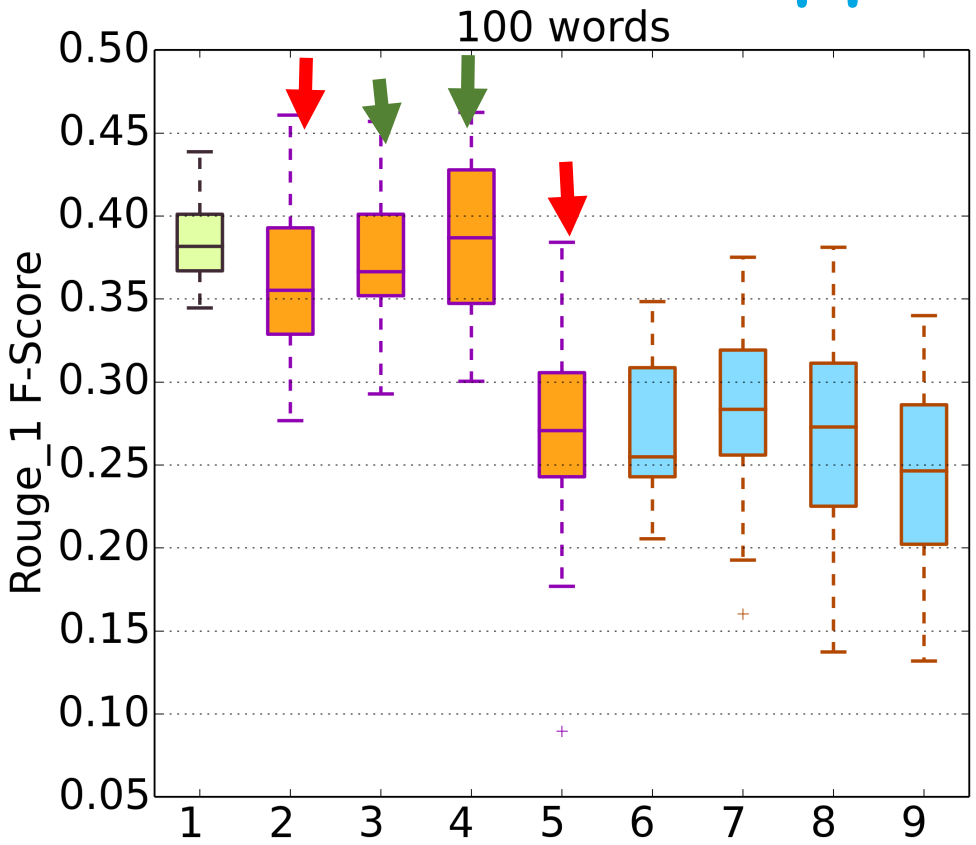
Results: Baselines

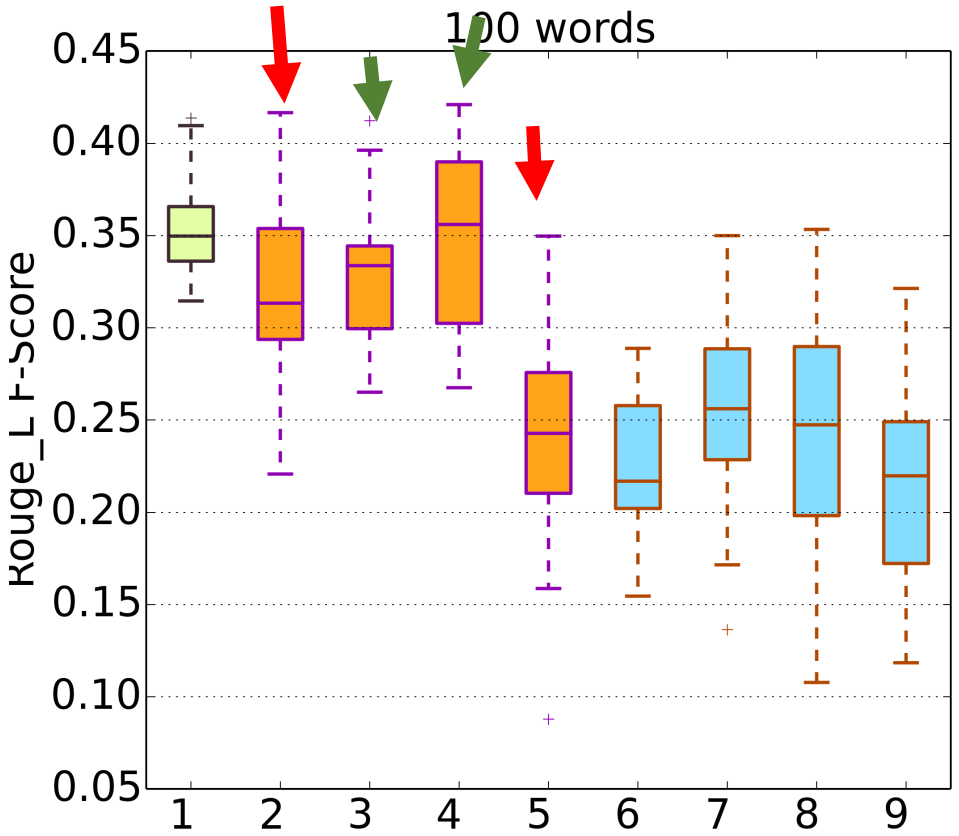




1: 'Human'
2: 'Community_Iterative'
3: 'Community_Diversity'
4: 'Discourse_Diversity'
5: 'Discourse_Iterative'
6: 'Citation Summary'
7: 'LexRank'
8: 'LSA'
9: 'MMR'

Results: Comparison of sentence selection approaches

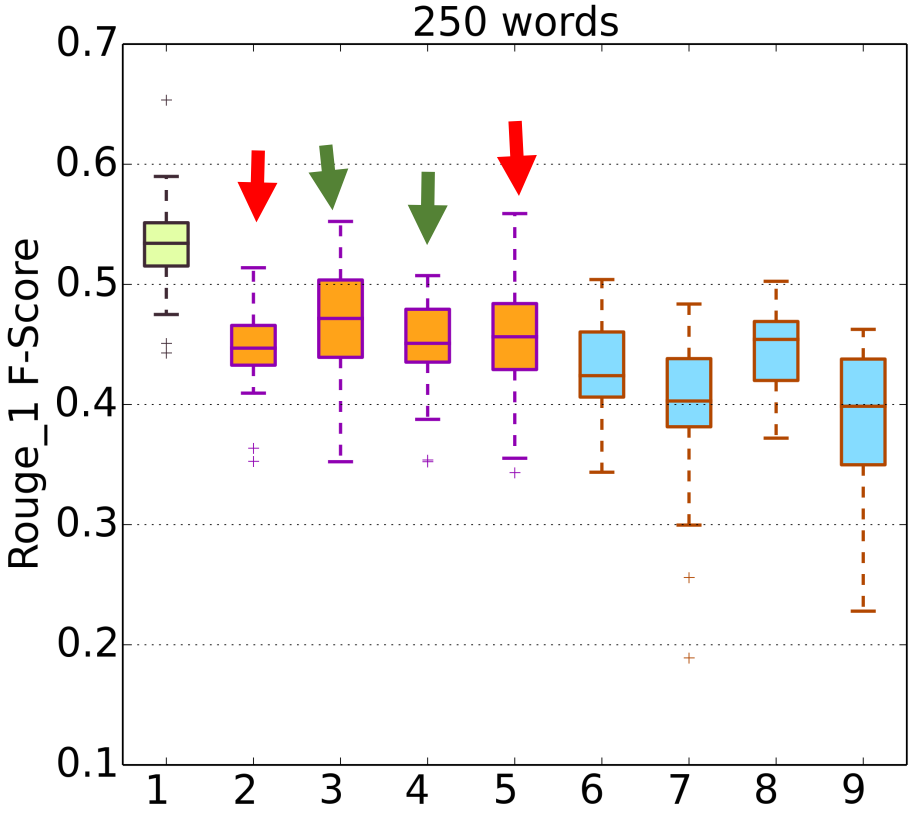


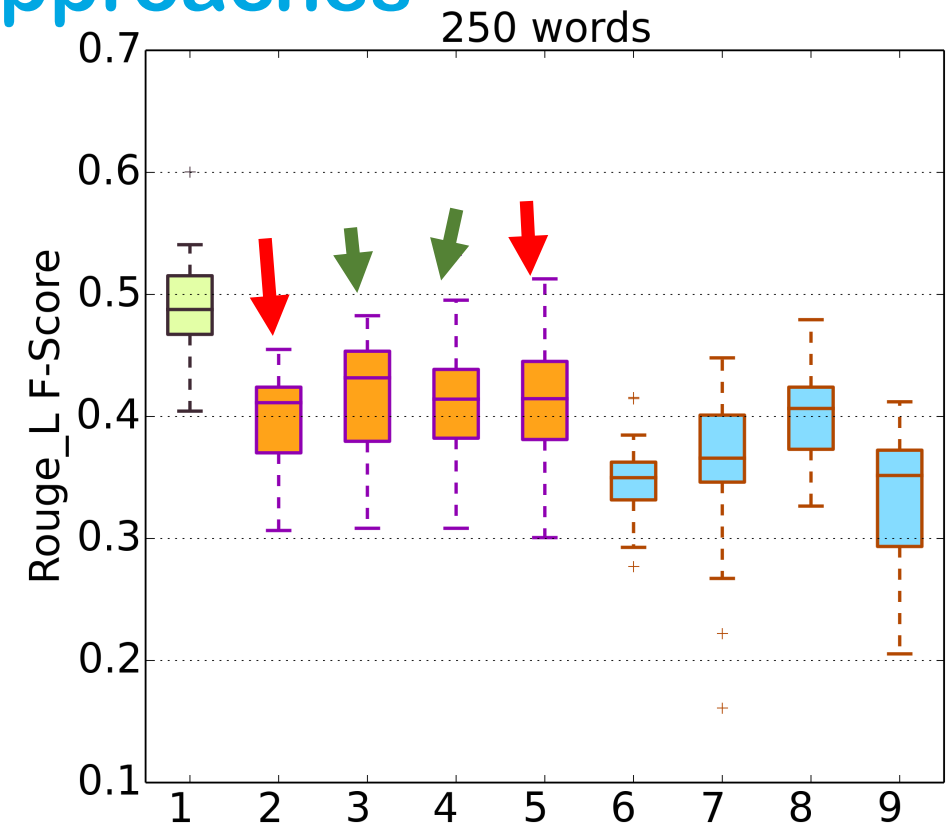


1: 'Human'
2: 'Community_Iterative'
3: 'Community_Diversity'
4: 'Discourse_Diversity'
5: 'Discourse_Iterative'
6: 'Citation Summary'
7: 'LexRank'
8: 'LSA'
9: 'MMR'

- **Short summaries: diversity based heuristic outperforms the iterative approach**

Results: Comparison of sentence selection approaches

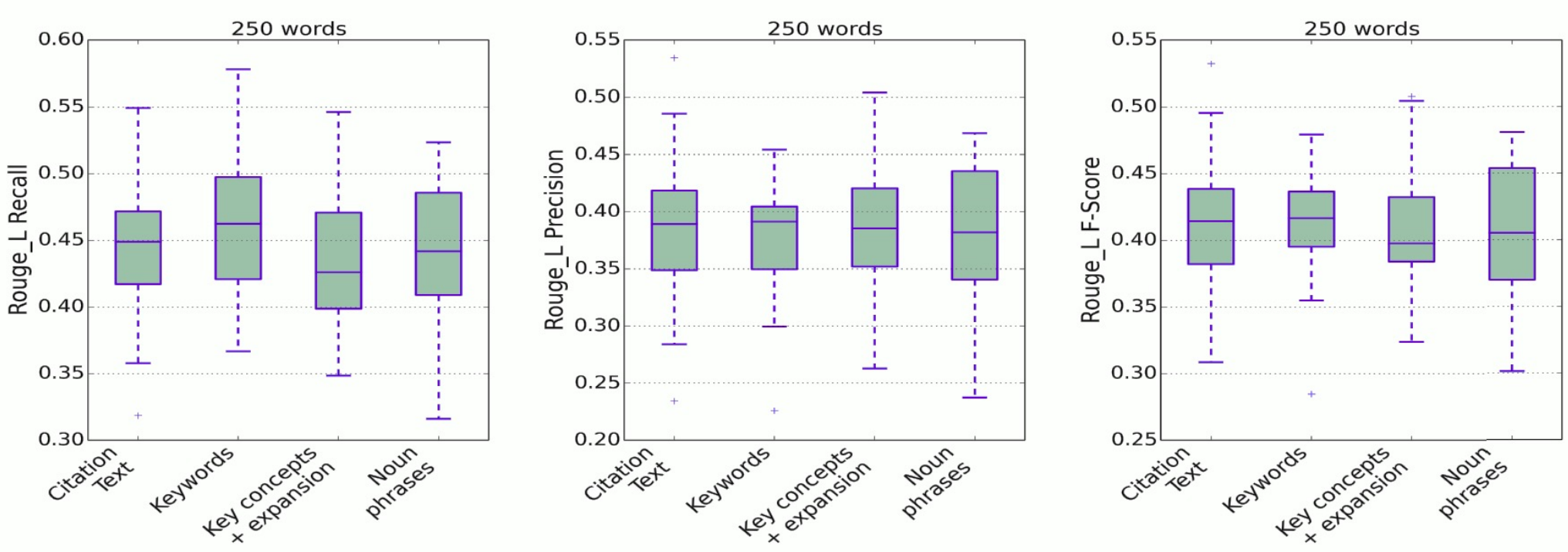




1: 'Homan'
2: 'Community_Iterative'
3: 'Community_Diversity'
4: 'Discourse_Diversity'
5: 'Discourse_Iterative'
6: 'Citation Summary'
7: 'LexRank'
8: 'LSA'
9: 'MMR'

- **In Long summaries iterative approach is as good as the diversity based heuristic**

Results: Comparison of citation-context extraction methods



- Relatively comparable performance
- Citation text and keywords performed the best
- Domain specific concepts did not result in improvement

Conclusions

- Citation-context approach towards scientific summarization improves over the baselines
- Citations by themselves are not as effective
- Discourse based diversification improves performance

Thank you