The Computational Linguistics Summarization Task @ TAC 2014, BIRNDL 2016, SIGIR 2017

- Summarization Challenge
- 3 years, 7 countries, 17 participating teams

Kokil Jaidka[1], Muthu Kumar Chandrasekaran[2], Min-Yen Kan[2],

[1]*University of Pennsylvania*

[2]*National University of Singapore*

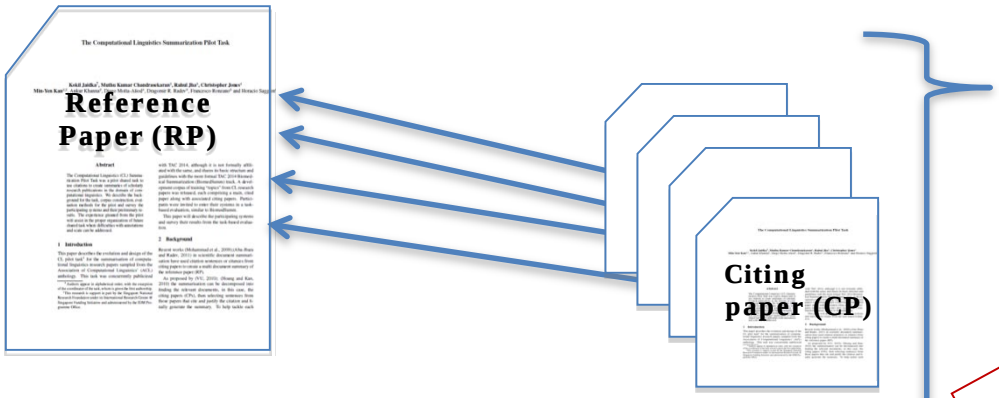
Corpus Highlights

- Continuing effort to advance scientific document summarization by encouraging the incorporation of semantic and citation information.

- Corpus of 30 articles; 500 citing papers
- Annotation by 6 paid and trained annotators (Master in Linguistics students) from U-Hyderabad
- Sponsorship from Microsoft Research Asia
- https://github.com/WING-NUS/scisumm-corpus/

The CL-SciSumm Shared Task

Task 1A: Identify the text span in the RP which corresponds to the *citances* from the CP.



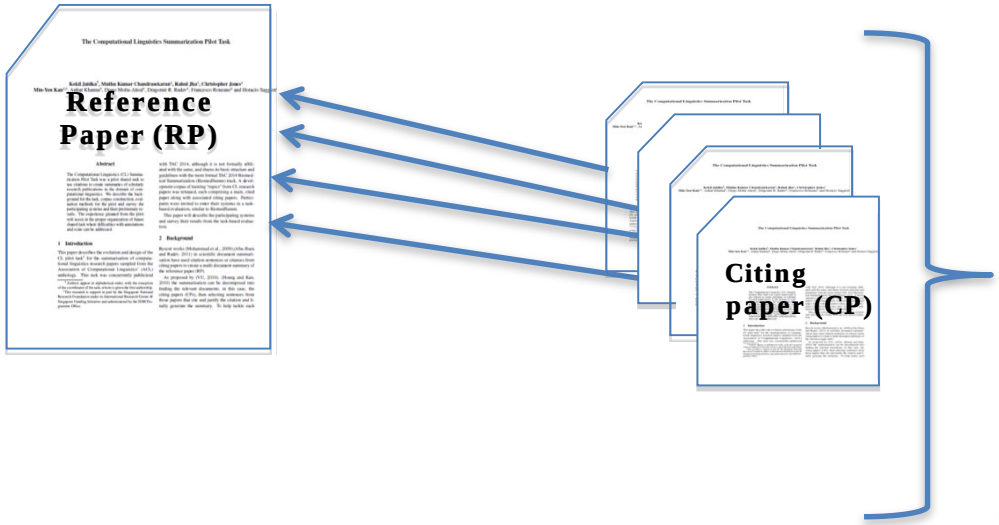
Citing papers
Citing text is called *citance*

Task 1A

Match the citing text in the CP to text in the RP

The CL-SciSumm Shared Task

Task 1B: Identify the discourse facet for every cited text span from a predefined set of facets.



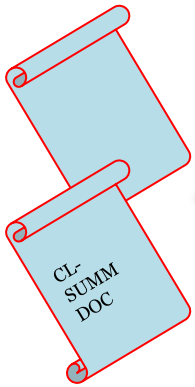
Classify the cited text in RP into one of several facets

Task 1B

Annotating the SciSumm corpus

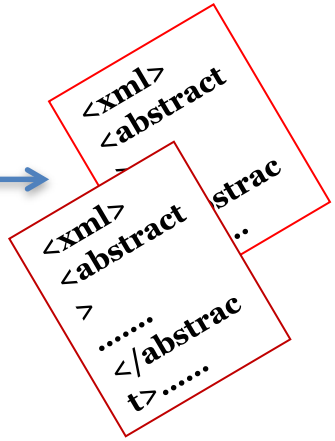
- 6 annotators selected from a pool of 25
- 6 hours of training
- Gold standard annotations for Task 1A and 1B, per topic or reference paper
- Community and hand-written summaries for Task 2, per topic

Annotation Pipeline

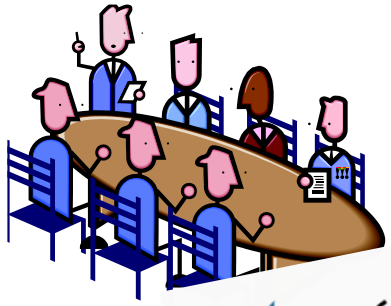


OCR & Section Parse

CLAIR - Umich's Python module



Annotation!



protégé

Post Processing with U-Colorado's python scripts

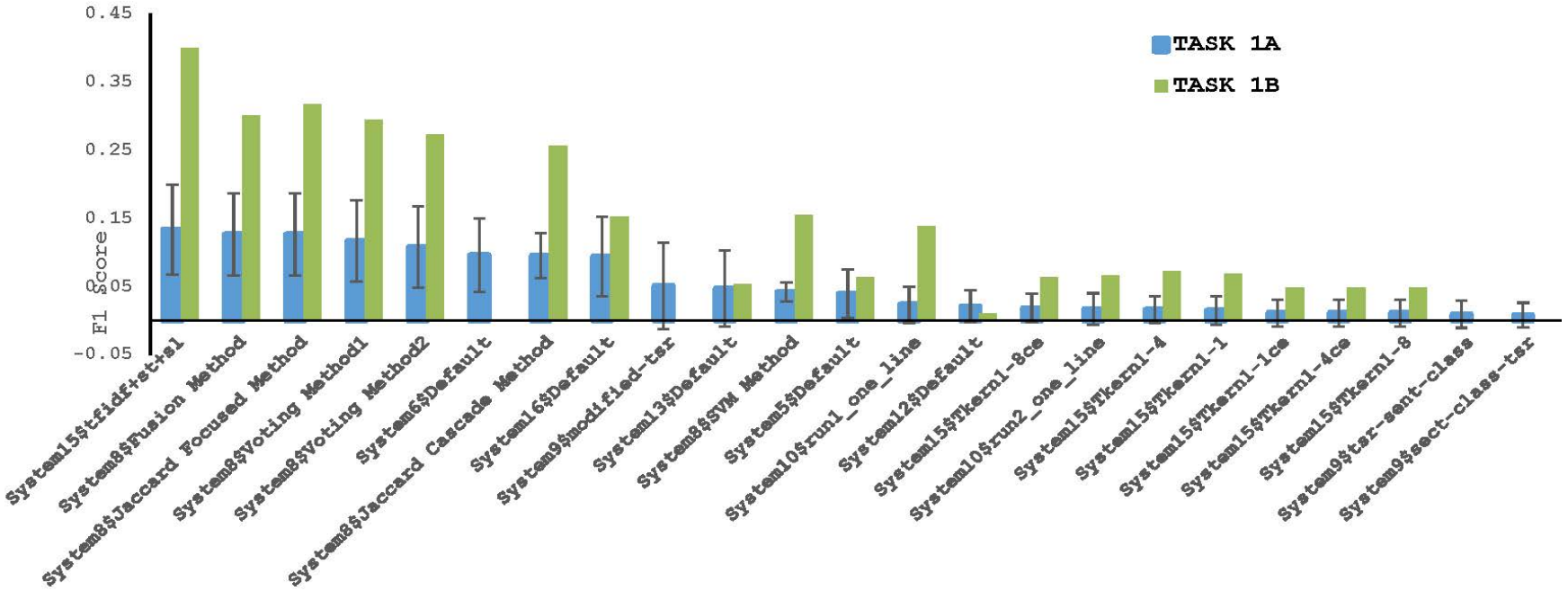
The CL-SciSumm Shared Task

- Task 2: Generate a structured summary of the RP from the cited text spans of the RP. The length of the summary should not exceed 250 words.
  - Compare with abstractive summary, human summary and community summary

Evaluation

- Task 1A – Exact sentence id match
- Task 1B –
  - conditional on Task 1A
  - BoW overlap between discourse facets
- Task 2 - ROUGE-SU2 and ROUGE-SU4

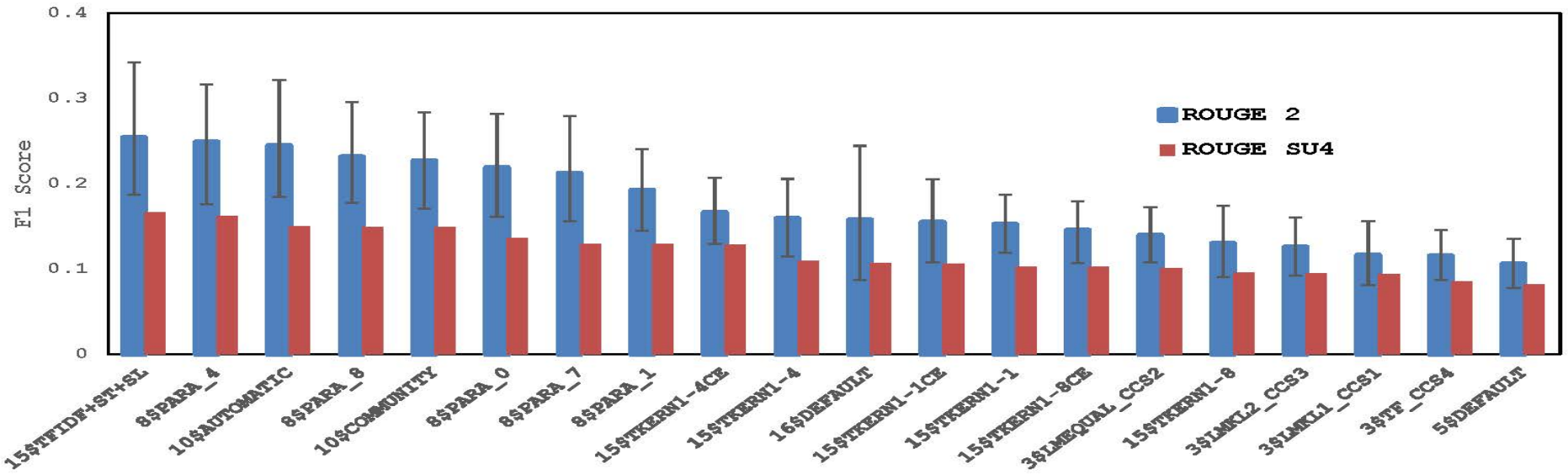
Best Performing Approaches: Task 1



- Tfidf [15]
- Combinations of SVM Classifier + term frequencies + surface features [16]
- Tfidf + embeddings-based neural network [17]
- SVM with similarity-based convolution kernel [18]

Best Performing Approaches (Task 2)

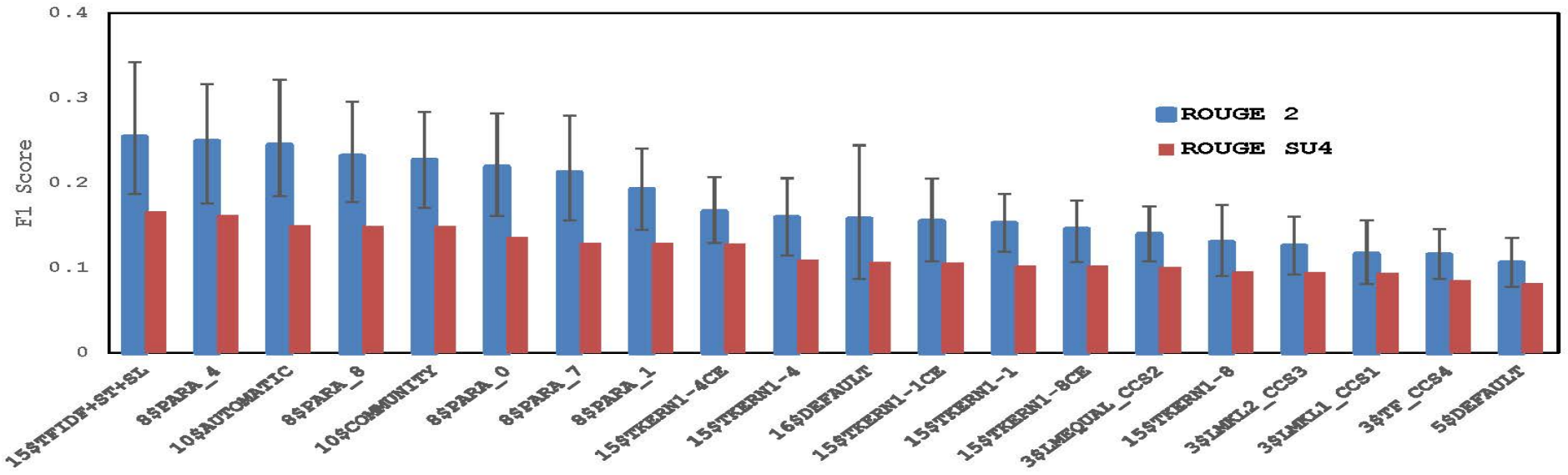
Human summaries



- Hierarchical LDA [19]
- Term frequency + NNMF [20]
- WEKA + feature relevance scores [19]

Best Performing Approaches (Task 2)

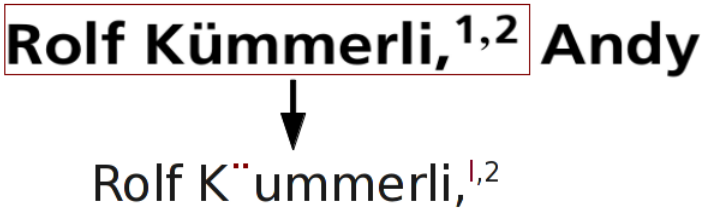
Community summaries



- WEKA + feature relevance scores [19]
- SVM with convolution kernel [21]
- Manifold Ranking Method on inter- and intra-document similarities [22]

Dataset Limitations

- Task 1B: limited number of samples for most (e.g., hypothesis) discourse facets, inconsistent labeling
- Preprocessing: OCR + Parsing



- Software: Protégé w/ manual alignment and post-processing
- Scaling the corpus was difficult: key bottleneck in the corpus development

Acknowledgements

- Chin-Yew Lin (MSRA)
- NIST and Hoa Dang
- Lucy Vanderwende, MSR
- Anita de Ward, Elsevier Data Services
- Kevin B. Cohen, Prabha Yadav (U. Colorado, Boulder)
- Rahul Jha (Google)
- U-Hyderabad Annotators:
  - Aakansha Gehlot, Ankita Patel, Fathima Vardha, Swastika Bhattacharya and Sweta Kumari
- System Paper Reviewers:
  - Akiko Aizawa, Dain Kaplan, John Lawrence, Lucy Vanderwende, Philipp Mayr, Vasudeva Verma and John Conroy

This task was possible through the generous support of Microsoft Research

Supplemental Analysis

- We investigated whether high deviations could be because of the topic sets themselves
- Topics with both high and low number of citances have mixed results
- No significant patterns of performance against:
  - Number of citances of the topic set
  - Age of the paper

Thank you

jaidka@sas.upenn.edu