The Computational Linguistics Summarization Task @ BIRNDL 2019

Muthu Kumar Chandrasekaran, Michihiro Yasunaga,

Dragomir Radev, Dayne Freitag, Min-Yen Kan

Corpus Highlights

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

- Annotated Corpus is same as CL-SciSumm 2018.
- 40 articles with 500 plus cited articles
- Annotation by 6 paid and trained annotators from U-Hyderabad
- CL-SciSumm 2019 traing data augmented in SciSummnet corpus (Yasunaga et al 2018) – 1000 articles
- Auto annotation 1000 datapoints for Task 1a (nomoto 2018)

https://github.com/WING-NUS/scisumm-corpus/

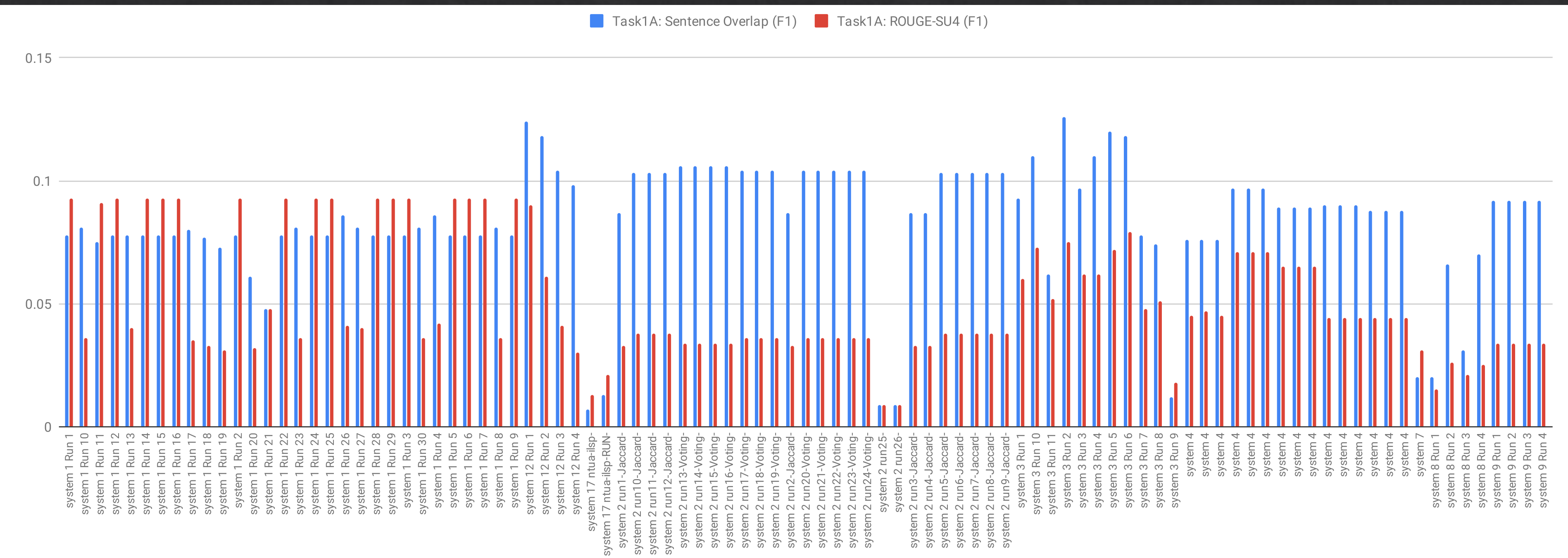
Outline

- Highlights
  - Results and Analyses of Runs
  - Task 1A – Identify text span in the RP
  - Task 1B – Discourse Facet of the RP text
  - Task 2 – 250 words or less of summary
- Conclusion

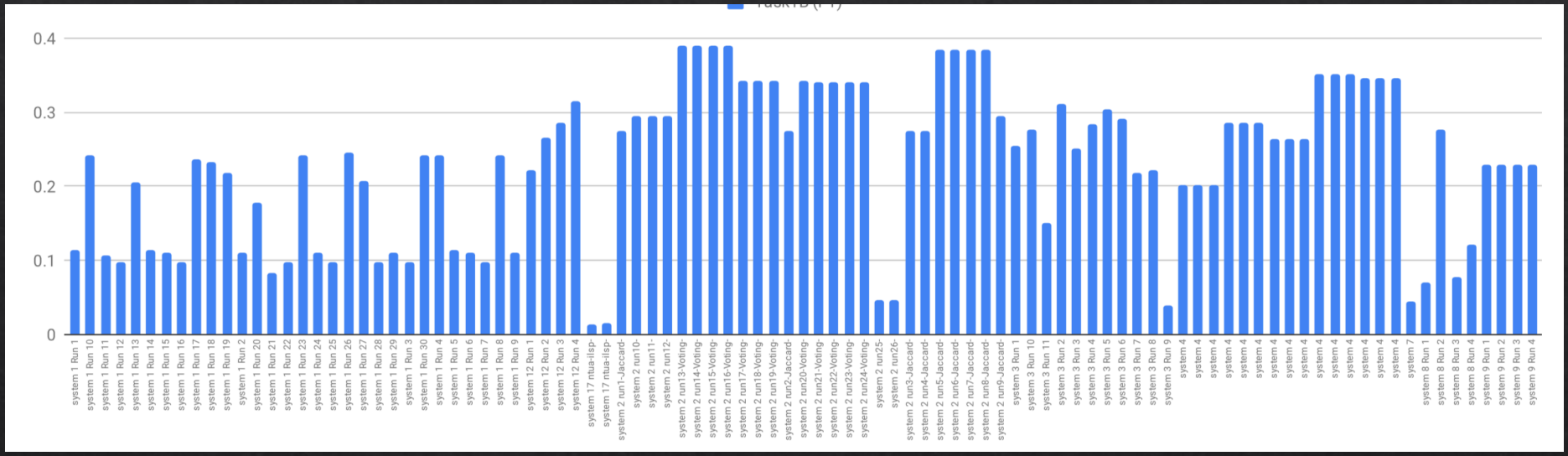
Evaluation

- Task 1A – Exact sentence id match, ROUGE-2 overlap
- Task 1B –
  - conditional on Task 1A
- Task 2 - ROUGE-2 and ROUGE-4

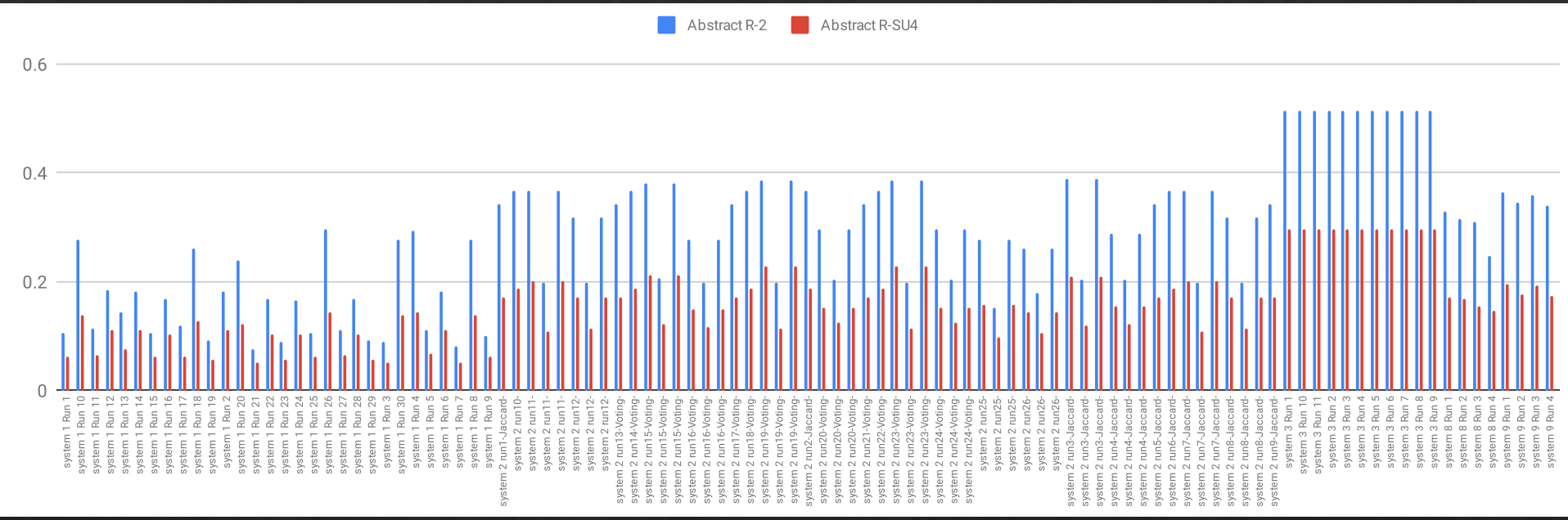
Results (Task 1A)



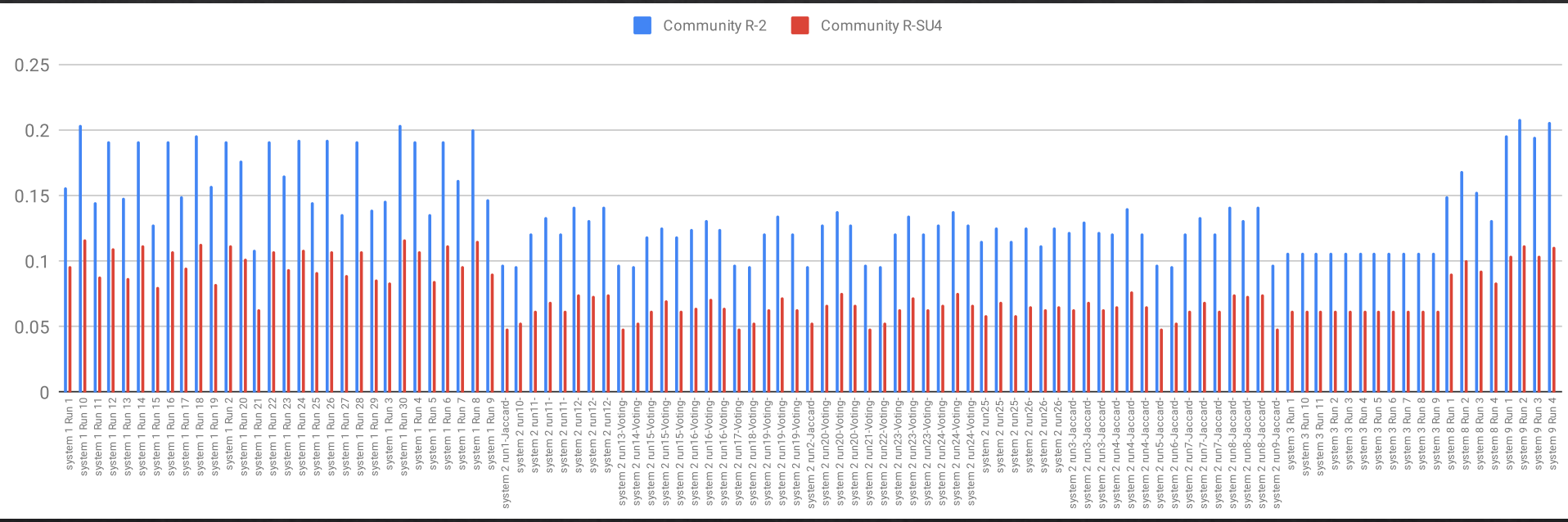
Results (Task 1B)



Results (Task 2) - Abstract



Results (Task 2) - Community



Results (Task 2) - Human

Key Research Questions

- Did data augmentation help?
  - Yes, for deep learning based models. No, for traditional machine learning models

- Did deep learning papers do better than traditional methods?
  - Yes, since they are robust and can leverage on noisy training data

Limitations

- Task 1B: limited number of samples for most (e.g., hypothesis) discourse facets, inconsistent labeling
- Preprocessing: OCR + Parsing
- Auto annotation is too noisy than expected
- Participant feedback?
  - Guidelines
  - The Task
  - The Corpus – size, #citing papers
  - Evaluation metrics

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

Additional Slides

Scientific Document Summarization

- Abstractive summary
  - Authors' own summary
- Extractive summary
  - Surface, lexical, semantic or rhetorical features of the paper
- Citation summary
  - Community creates a summary when citing
- Faceted summary
  - Capture all aspects of a paper

Scientific Document Summarization
Citation-based extractive summaries

Scope of Citation

- Qazvinian, V., and Radev, D. R. "Identifying non-explicit citing sentences for citation-based summarization" (ACL, 2010)
- Abu-Jbara, Amjad, and Dragomir Radev. "Reference scope identification in citing sentences." (ACL, 2012)

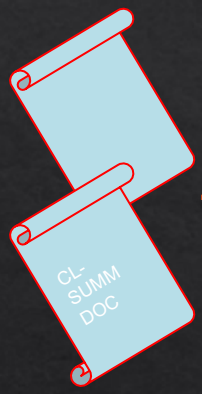
Coherence

- Abu-Jbara, Amjad, and Dragomir Radev. "Coherent citation-based summarization of scientific papers." (ACL 2011)

In summary

- Community concurs that a citation-based summary of a scientific document is important.
- Citing papers cite different aspects of the same reference paper.
- Assigning facets to these citances may help create coherent summaries.

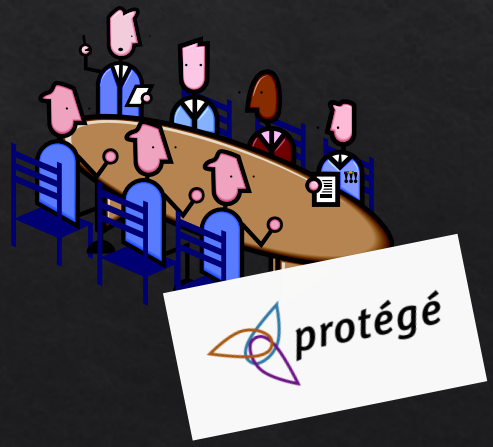
Annotation Pipeline

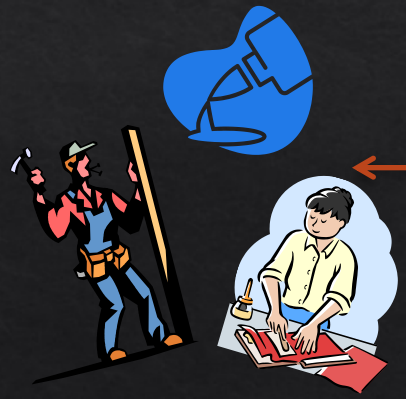


OCR & Section Parse

CLAIR -Umich's Python module

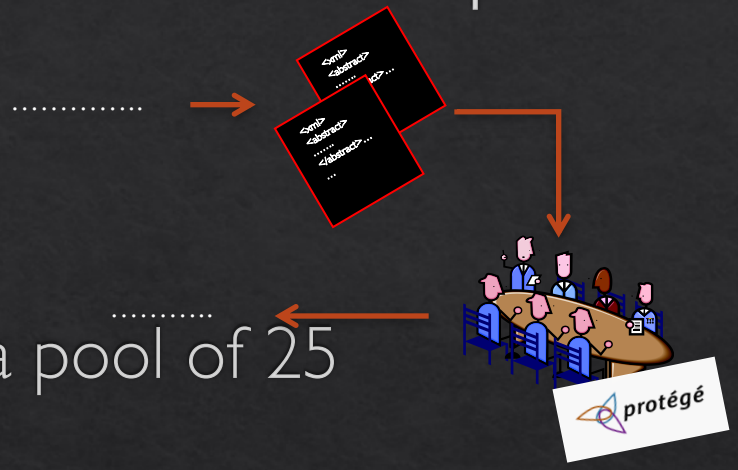
Annotation!





Post Processing with U-Colorado's python scripts

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