SDP 2020: Shared Task Track

Muthu Kumar Chandrasekaran, Guy Feigenblat,

Ed Hovy, Abhilasha Ravichander,

Michal Shmueli-Scheuer, Anita de Waard

Now → Teaser to Shared Task Track

Next → Oral Sessions: Zoom Link 2

Overview of Evaluation and Results

Poster Session - Gather.Town – Room N

Shared Task Track
The Computational Linguistics Scientific Summarization Task (CL-SciSumm) @ SDP 2020

Muthu Kumar Chandrasekaran

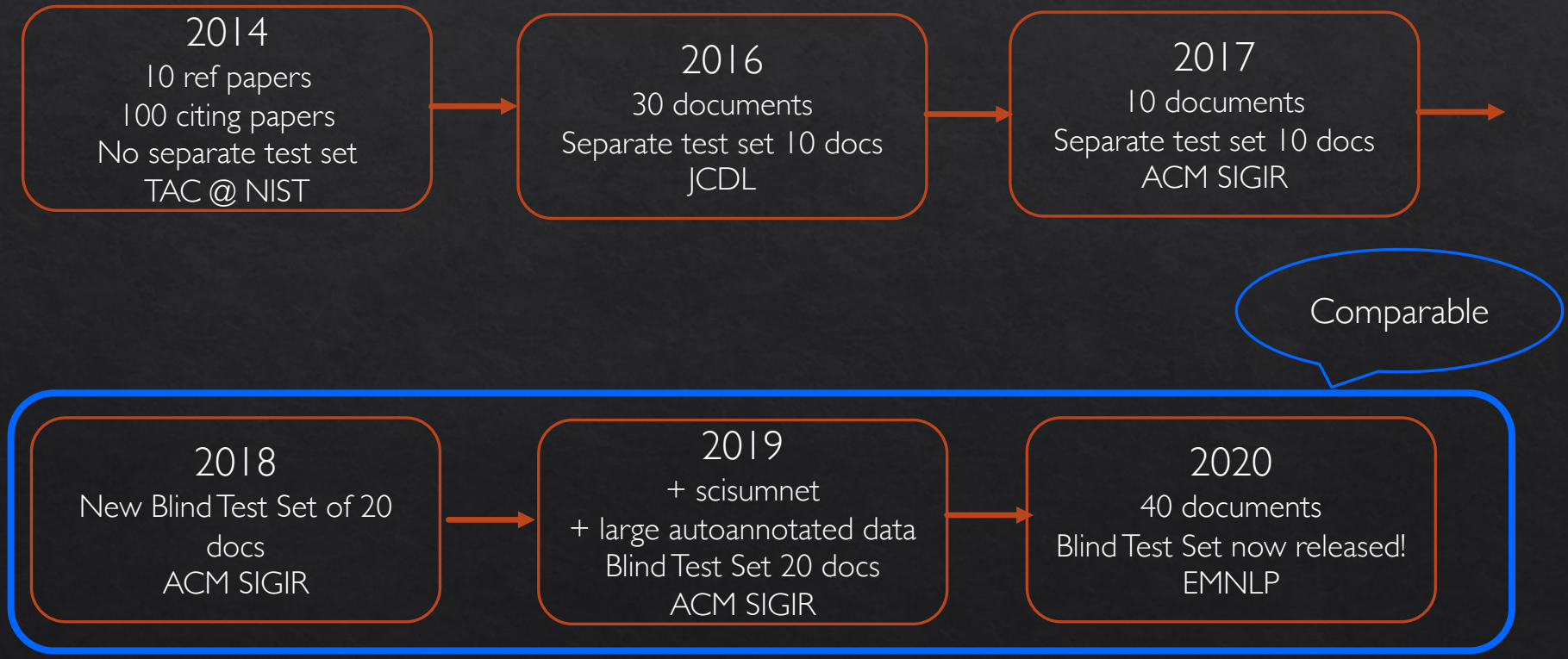
Amazon, Seattle

Outline

- Now
  - Background
  - Corpus
  - Tasks
    - Task 1A – Identify text span in the RP
    - Task 1B – Discourse Facet of the RP text
    - Task 2 – 250 words or less of summary
- Later (14:50 ET – After poser presentation)
  - Evaluation and Results

Background: CL-SciSumm

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

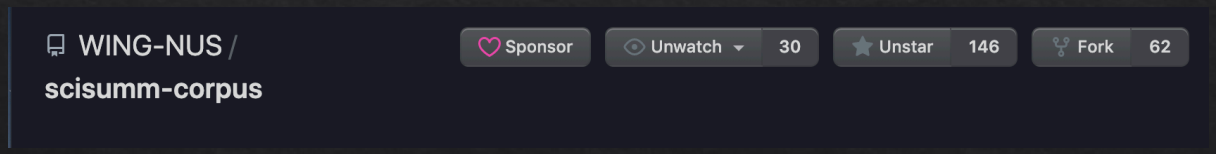


https://github.com/WING-NUS/scisumm-corpus/

CL-SciSumm over half a decade!

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

- Has inspired, spawned other resources in the community
  - SciSummnet, TalkSumm
  - Several state of the art papers/ systems published at EMNLP, ACL, AAAI that benchmark and improve on CL-SciSumm
  - Evolved into a de facto standard corpora for scientific document summarization with human written summaries



Introducing two new tasks this year:

LaySumm. – Scientific Summaries for the lay person

LongSumm – Longer Scientific Summaries (e.g., a paper blog)

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

State of the Corpus in 2020

- Task 1A and B: 40 target articles and 500+ citing articles
- Task 2 (Summarization): 40 + 1000 from SciSummnet = 1040 human written summaries and abstracts
- All tasks annotated by 6 paid and trained annotators from U-Hyderabad, India from 2016-2018
- Corpus development from 2016-2018 was sponsored by MSRA
- Details about Annotation and Corpus Construction are in the GitHub repo and in our previous overview papers:

Chandrasekaran, M. K., Yasunaga, M., Radev, D., Freitag, D., & Kan, M. Y. (2019). Overview and results: Cl-scisumm shared task 2019. arXiv preprint arXiv:1907.09854.

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., & Kan, M. Y. (2018). Insights from CL-SciSumm 2016: the faceted scientific document summarization Shared Task. International Journal on Digital Libraries, 19(2-3), 163-171.

Teaser to Shared Task Track

Oral Sessions (Zoom Link 2)

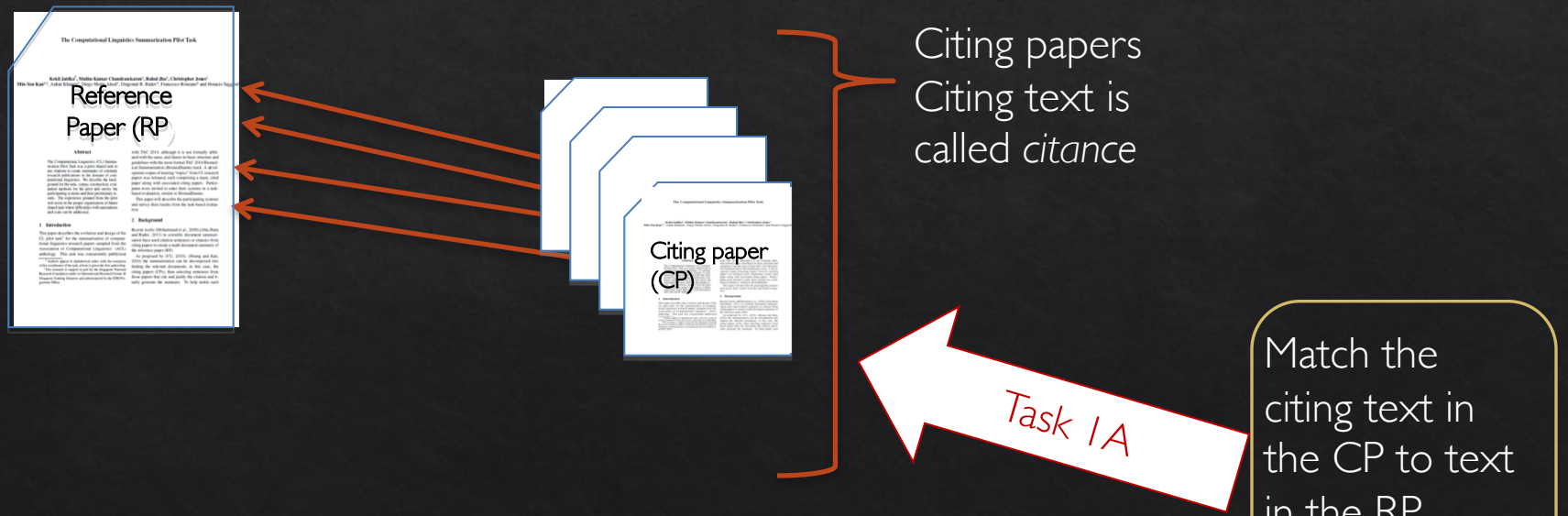
09:15 – 09:27	CL-SciSumm (Task 1a): Chai et al., NLP-PINGAN-TECH @ CL-SciSumm 2020
09:55 –10:10	Li et al., CIST@CL-SciSumm 2020, LongSumm 2020: Automatic Scientific Document Summarization
10:10-10:25	LongSumm 3 & CL-SciSumm (Task 2): Reddy et al., IIITBH-IITP@CL-SciSumm20, CL-LaySumm20, LongSumm20
11:50 – 12:05	CL-SciSumm (Task 1a): Aumiller et al., UniHD@CL-SciSumm 2020: Citation Extraction as Search
12:30 - 12:42	Umapathy et al., CiteQA@CLSciSumm 2020

Poster Session – 3:45 to 14:50 ET

Gather.Town – Room N

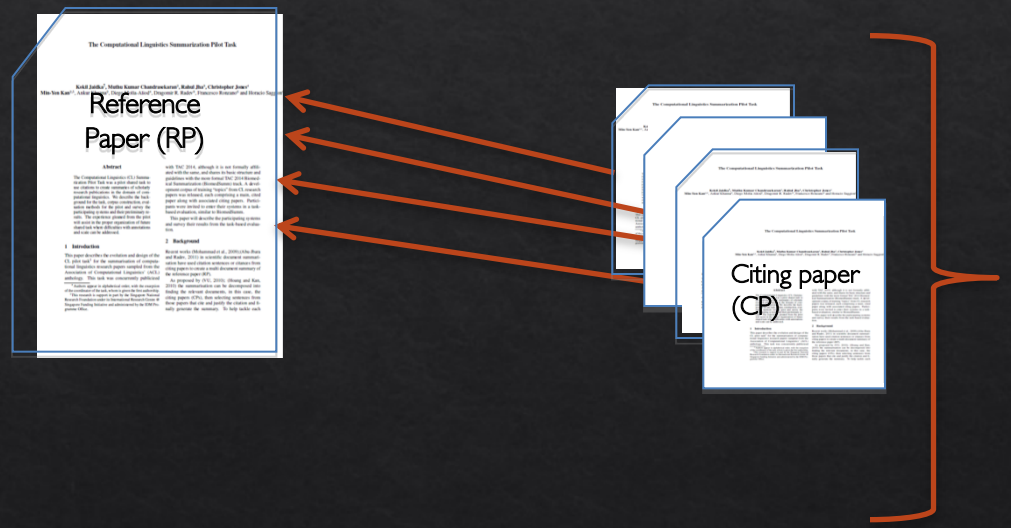
Tasks: Task 1A

Identify the text span in the RP which corresponds to the *citances* from the CP.



Tasks: Task 1B

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

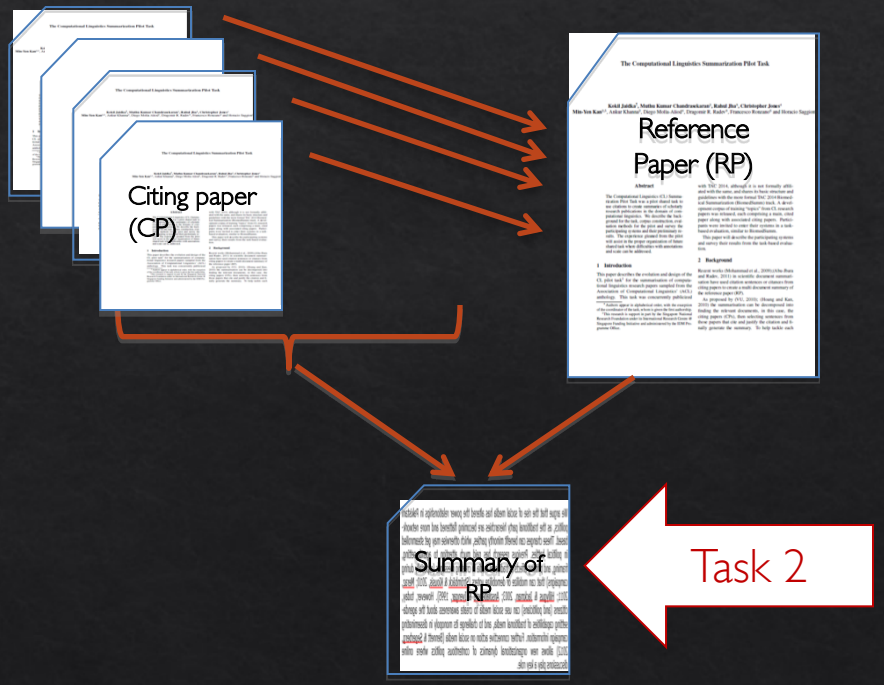


Classify the cited text in RP into one of several facets

Task 1B

Tasks: Task 2

Task 2: Generate a faceted summary of up to 250 words, of the reference paper, using itself and the citing papers.

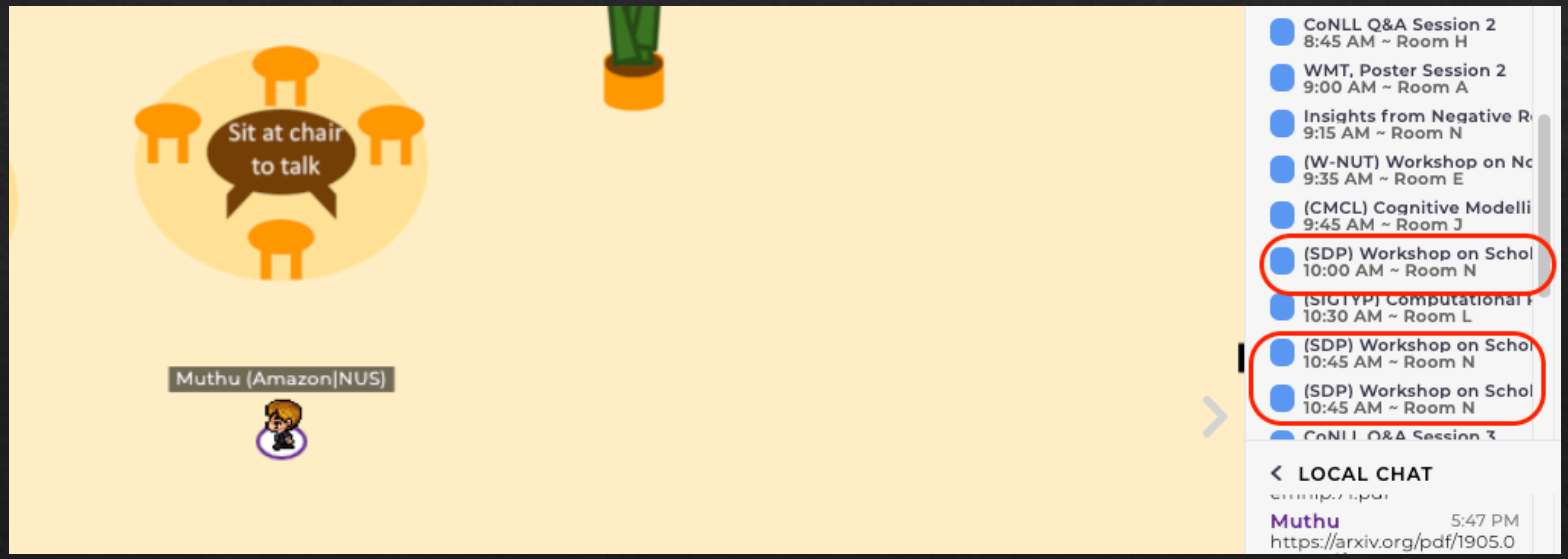


Use citances and the RP to create a summary

Teaser to Shared Task Track

Oral Sessions (Zoom Link 2)

Shared Task Poster Sessions 13:45 to 14:50 ET
Gather.Town – Room N and Room A



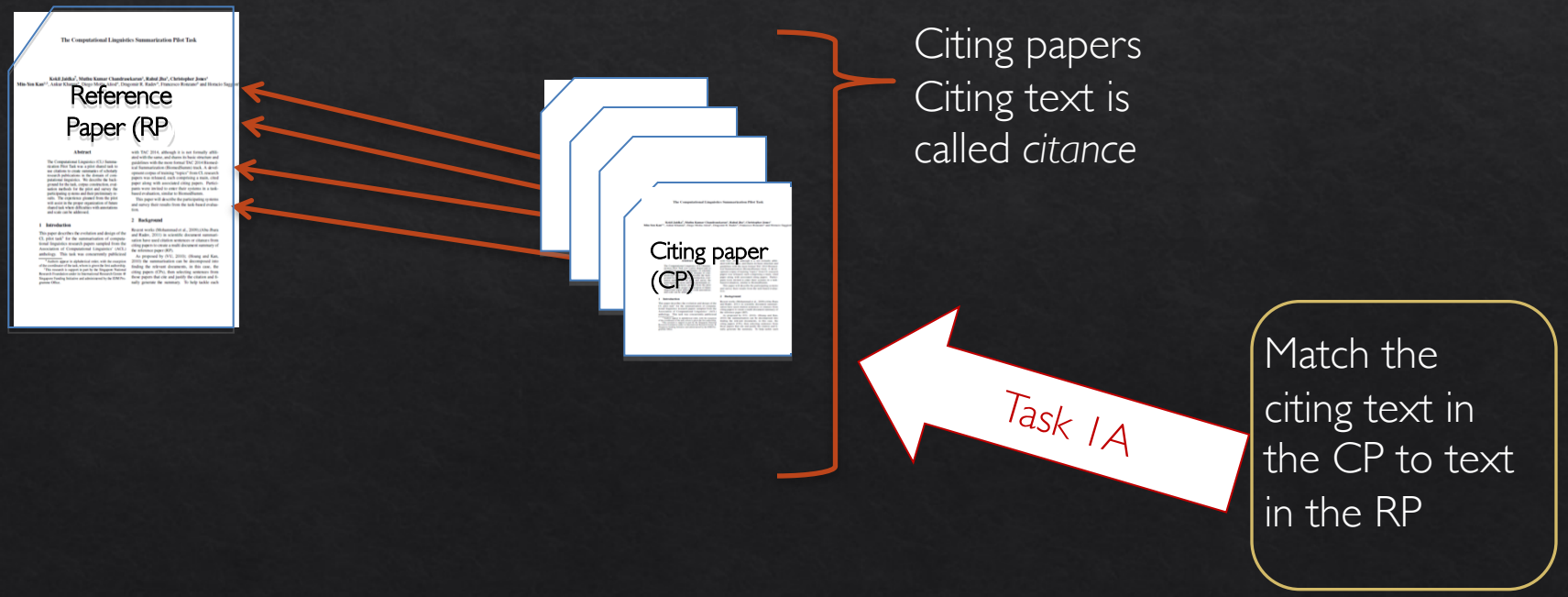
Shared Task Track
The Computational Linguistics Summarization Pilot Task @ SDP 2020

Muthu Kumar Chandrasekaran

Amazon, Seattle

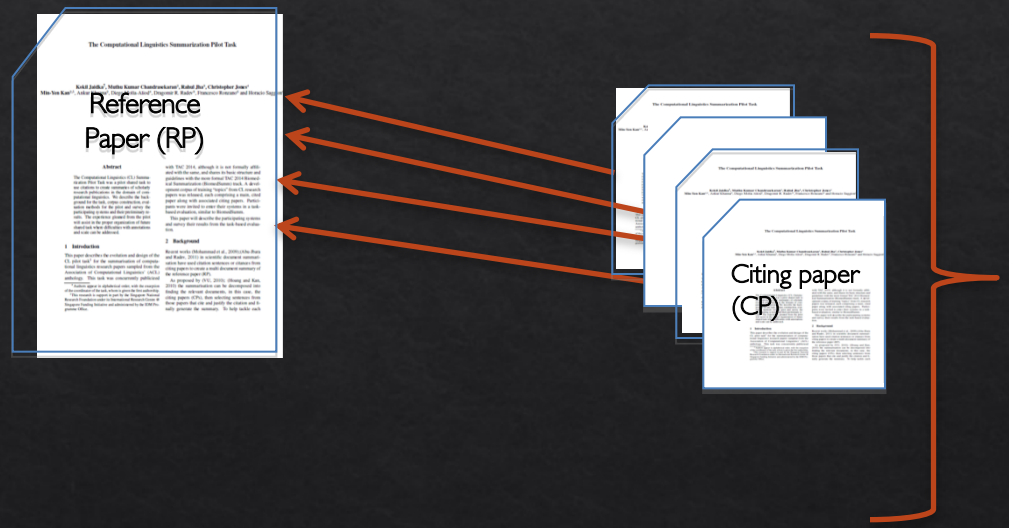
RECAP Tasks: Task 1A

Identify the text span in the RP which corresponds to the *citances* from the CP.



RECAP Tasks: Task 1B

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

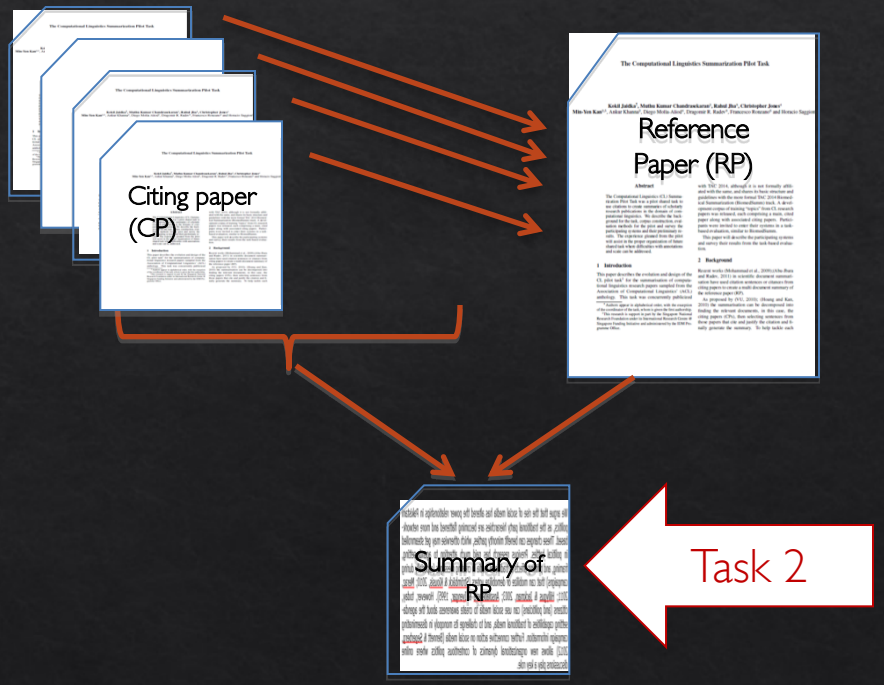


Classify the cited text in RP into one of several facets

Task 1B

RECAP Tasks: Task 2

Task 2: Generate a faceted summary of up to 250 words, of the reference paper, using itself and the citing papers.



Use citances and the RP to create a summary

Submissions

- 11 Systems Submitted for Evaluation
- Task 1A and B: All 11 systems
- Task 2 (Summarization): 4 out of the 11 systems
- 5 teams presented their systems in the forenoon session

Evaluation

- Run centrally from the CL-SciSumm repository. Replicable by anyone and everyone at anytime
- Task 1A –
  - Exact sentence id match
  - ROUGE 1, 2
- Task 1B –
  - conditional on Task 1A
  - Precision, Recall and F_1 over four discourse facets (classes)
- Task 2 - ROUGE-2

Best Performing System (Task 1A)

Team / System	Run	Exact Match			ROUGE-2		
		Precision	Recall	F_1	Precision	Recall	F_1
NLP_PINGAN TECH	sembert_scibert_all_top2	0.13	0.25	0.17	0.30	0.11	0.15
uniHD	intersection_2_field	0.26	0.12	0.16	0.32	0.08	0.11
CMU	run110	0.09	0.25	0.13	0.31	0.05	0.08
CIST	runs 22-42	0.07	0.26	0.11	0.34	0.03	0.05
AUTH	run 2	0.08	0.13	0.10	0.17	0.07	0.09
IITBH-IITP	variantU	0.05	0.22	0.08	0.30	0.02	0.03
IITP-AI-NLP-ML	runs 1-10	0.02	0.10	0.04	0.19	0.01	0.01
Martin-Luther-Universität Halle-Wittenberg		0.01	0.01	0.01	0.02	0.03	0.02

Best Performing System (Task1B)

Team / System	Run	Precision	Recall	F_1
CMU	run26	0.58	0.21	0.31
CIST	run40	0.41	0.24	0.30
uniHD	intersection_3_field	0.48	0.48	0.29
NLP_PINGAN TECH	run_sembert_scibert_all_top3	0.19	0.29	0.23
IITBH-IITP	variantU	0.49	0.15	0.23
AUTH	run 1	0.48	0.10	0.17
IITP-AI-NLP-ML		0.02	0.01	0.02
Martin-Luther-Universität Halle-Wittenberg	Martin-Luther-Universität Halle-Wittenberg	0.09	0.01	0.01

Best Performing System – Task 2

Team / System	Run	Vs. Abstract		
		P	R	F_1
AUTH	run 2 2	0.36	0.48	0.41
CIST	run43, 46, 49, 52, 55, 58, 61	0.14	0.37	0.21
IIT-NLP-AI-ML	run4	0.13	0.36	0.20
IITBH-IITP	variantA2, E2, F2, S2, U2, X2	0.11	0.25	0.15

Team / System	Run	Vs. Human		
		P	R	F_1
AUTH	run 2 2	0.23	0.22	0.22
CIST	run22, 25, 28, 31, 34, 37, 40	0.17	0.25	0.20
IIT-NLP-AI-ML	run4	0.14	0.25	0.17
IITBH-IITP	variantA2, E2, F2, S2, U2, X2	0.12	0.20	0.15

Best Performing System – Task 2

Team / System	Run	Vs. Community Summary		
		P	R	F_1
IITBH-IITP	variantU	0.20	0.43	0.27
CIST	run22, 25, 28, 31, 34, 37, 40	0.34	0.19	0.25
IIT-NLP-AI-ML	run4	0.27	0.15	0.19
AUTH	run 2 2	0.25	0.07	0.11

What do we see in CL-SciSumm 2020?

- Use of pretrained models, mostly transformer based models which have become ubiquitous in NLP
- Out of domain data e.g., arXiv, pubmed to train
- Graph Attention Network (CIST) models appears to be generating the holy grail summary that evaluates well on all human, abstract and community summaries
- AUTH's system (transformers / Pegasus) tops human and abstract summaries but ends up at the bottom of the list when evaluated on community summaries
- Neural models on Task B appear to saturate at 30% F1. Previous best is around ~38% on logistic regression / ensemble with rule based features

Limitations

- Supporting SoTA neural model training from scratch
- Preprocessing: OCR + Parsing from ACL Anthology PDFs
- Task 1A Evaluation
- Task 1B: limited number of samples for most (e.g., hypothesis) discourse facets, inconsistent labeling
- Scaling the corpus was difficult: key bottleneck in the corpus development
- Participant feedback?
  - Guidelines
  - The Task
  - The Corpus – size, #citing papers
  - Evaluation metrics

Acknowledgements

- Min-Yen Kan, National University of Singapore
- Dragomir Radev (Yale), Michihiro Yasunaga (Stanford), Rahul Jha (Microsoft)
- Chin-Yew Lin (MSRA)
- NIST and Hoa Dang
- Lucy Vanderwende, MSR
- Anita de Ward, Elsevier Data Services
- Kevin B. Cohen and colleagues (U. Colorado, Boulder)

Annotation Recruitment and Support: Vasudeva Varma, IIIT-H, India and group

U-Hyderabad Annotators:

Aakansha Gehlot, Ankita Patel, Fathima Vardha, Swastika Bhattacharya and Sweta Kumari

CL-SciSumm corpus was developed with funding from **Microsoft Research** over 2016-2018

Conclusions

- Successfully established a human annotator reference corpora for evaluating scientific document summarization
- Garnered interest and support of communities across fields (CS, NLP, Information Science, Bibliometrics, Linguistics)
- De facto standard for benchmarking Scientific Document Summarization
- Future: Summaries towards more depth – for researchers and practitioners; more breadth – wider reach through public understanding of science
- We invite teams to examine the detailed results available with the GitHub repo: https://github.com/WING-NUS/scisumm-corpus/tree/master/CLSciSumm_2020_Evaluation
- Thanks to all teams' participation and various co-organizers for the success of CL-SciSumm shared task series from 2014 through 2020!

Additional Slides

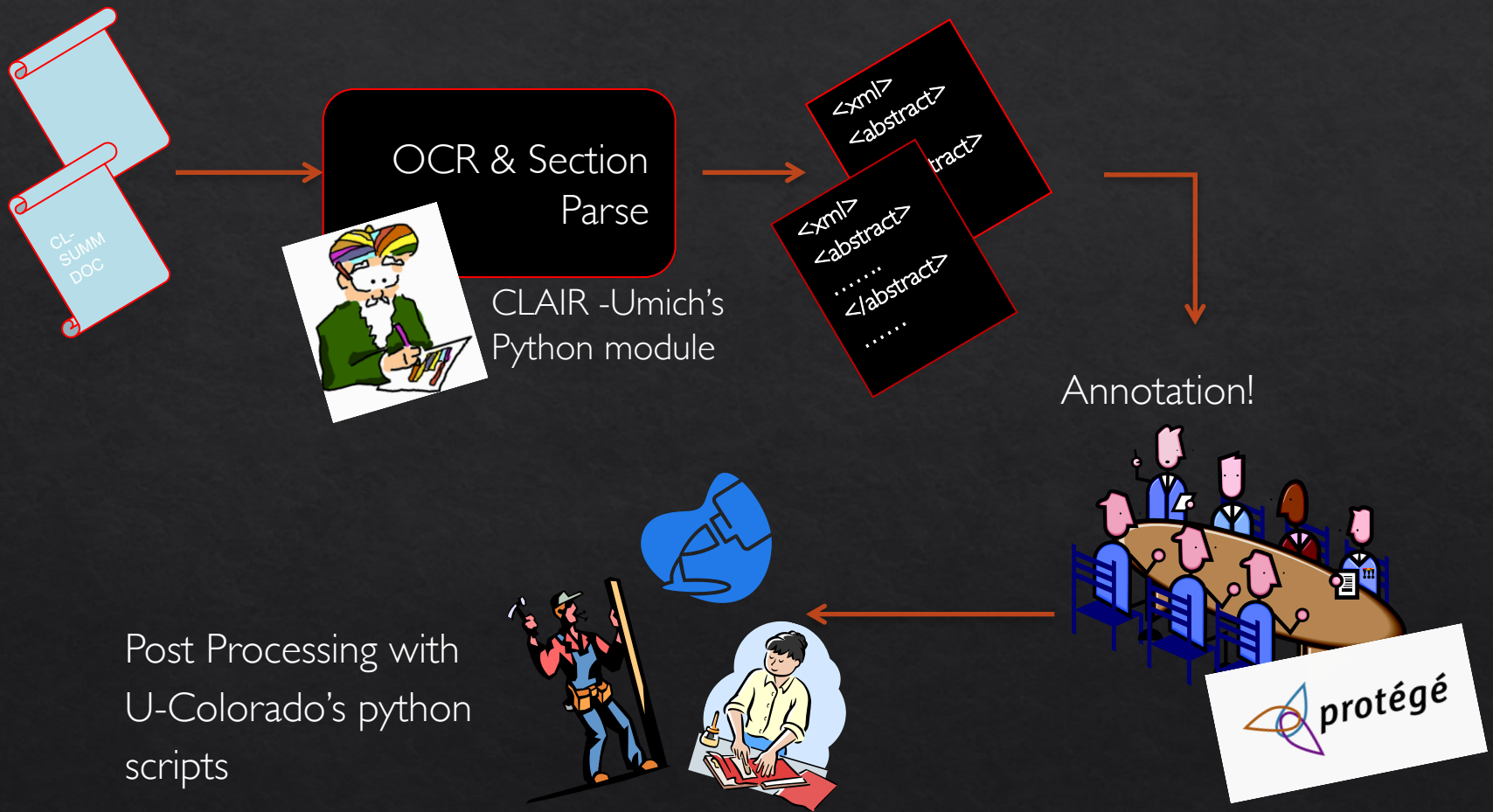
Scientific Document Summarization

- 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

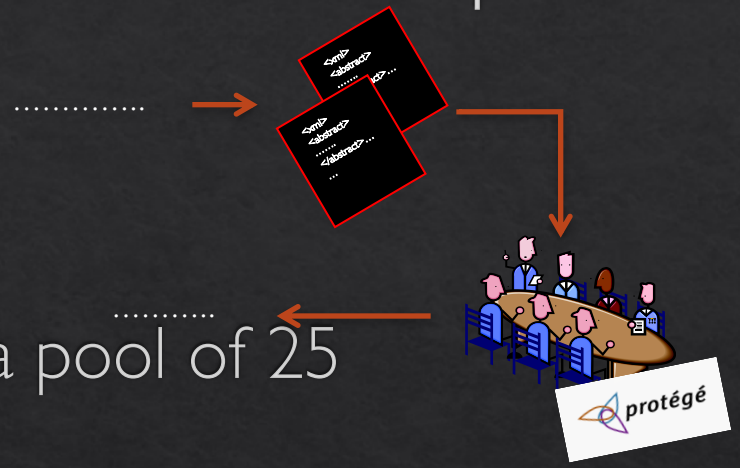
In summary

- 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



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