

The Computational Linguistics Summarization Task @ BIRNDL 2016

Kokil Jaidka¹, Muthu Kumar Chandrasekaran², Sajal Rustagi¹, Min-Yen Kan^{2,3}

¹Adobe Research India

²School of Computing, National University of Singapore, Singapore

³Interactive and Digital Media Institute, National University of Singapore, Singapore

Microsoft Research



Corpus Highlights

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

- ◆ Corpus enlarged from 10 (pilot) to 30 CL articles
- ◆ Annotation by 6 paid and trained annotators from U-Hyderabad
- ◆ Sponsorship from Microsoft Research Asia

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

Oral Sessions

14:35-14:55	<ul style="list-style-type: none">• System 8 – Top in Task 1B, among top performers for Task 1A and Task 2• Remote presentation from China
14:55-15:15	<ul style="list-style-type: none">• System 6 – Among top performers for Task 1A• Local presentation

Outline

- ◆ Highlights

now > 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

Still a work in progress:



Old
results

Will present results based on the CEUR paper (“old”), stacked average of all runs...



New
results

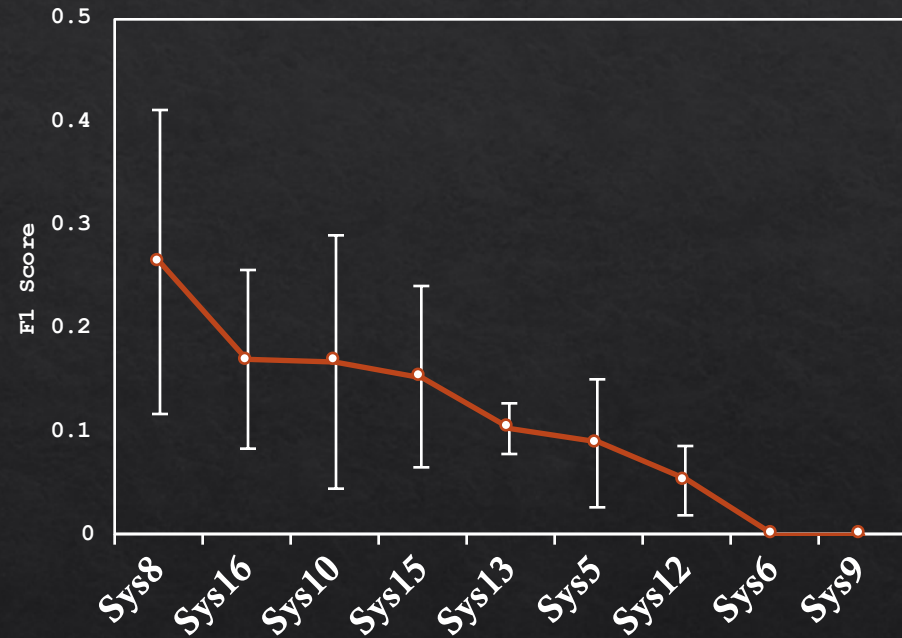
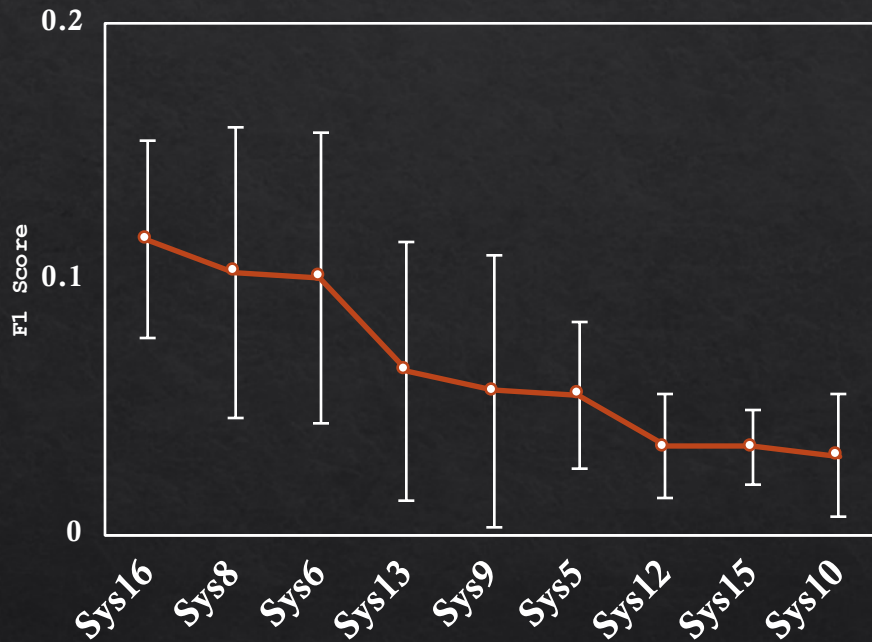
... and contrast with newer (still preliminary) results (“new”), individual runs separated

Evaluation

- Task 1A – Exact sentence ID match
- Task 1B –
 - conditional on Task 1A
 - Bag of Words (BOW) overlap between discourse facets
- Task 2 - ROUGE-2 and ROUGE-4



System Results (Task 1A & 1B)

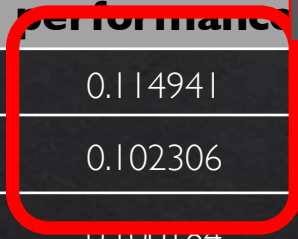




Best Performing System (Task 1A)

System ID	Task 1a	
	Avg performance	
16	0.114941	0.038295
8	0.102306	0.056893
6	0.100164	0.056926
13	0.063622	0.050519
9	0.056172	0.053044
5	0.054283	0.028954
12	0.034219	0.020178
15	0.034122	0.014837
10	0.03073	0.023688

Best performing Systems





Old results

Best Performing System (Task I B)

System ID	Task I b	
	Avg performance	Std
16	0.1696516	0.000000
8	0.264754	0.1473109
13	0.10294	0.0236852
5	0.088737	0.0617396
12	0.052747	0.0341898
15	0.152984	
10	0.168061	0.122391

Best performing System

Best performing System



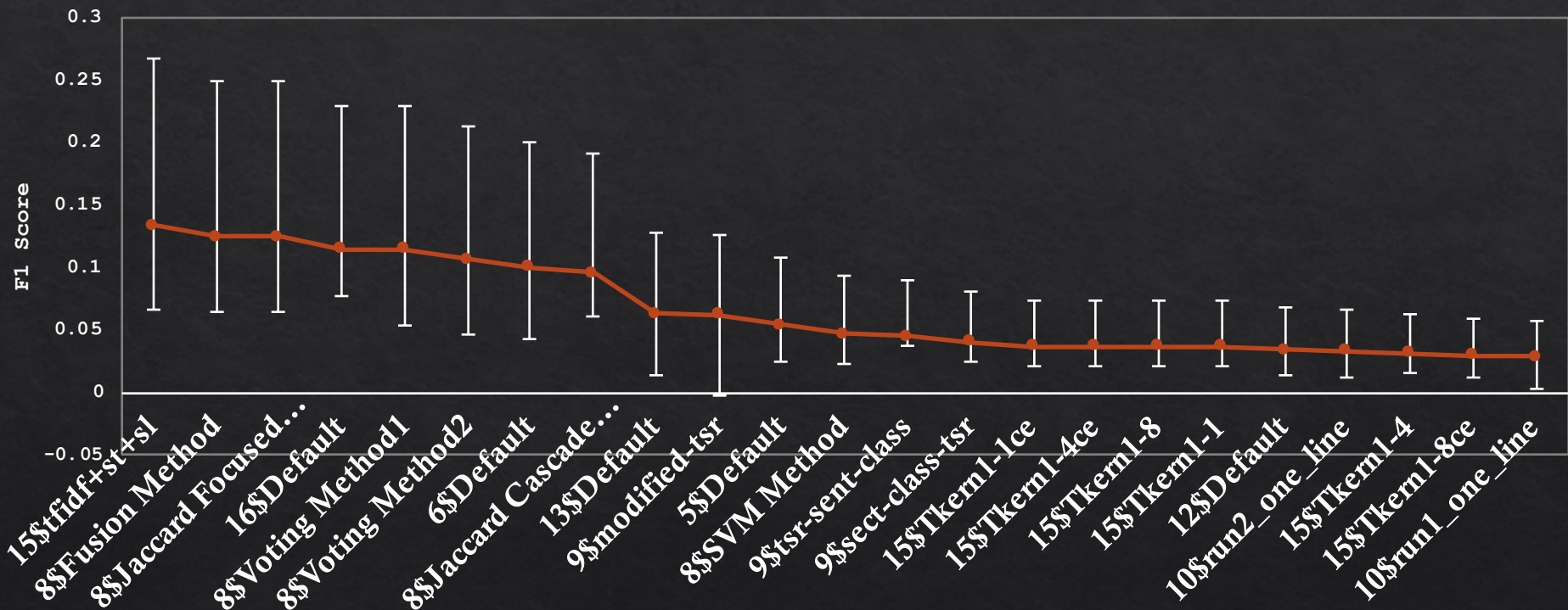
Best Performing System – Task 2

Best performing Systems

System ID	Approaches	Comments
3	<ul style="list-style-type: none"> • NNMF for BioMedSumm 	The best for human summaries
8	<ul style="list-style-type: none"> • hLDA topic modeling • Sentence length/position • Cited text spans • RST 	The best for abstract and community summaries
15	<ul style="list-style-type: none"> • Tkeml-1 • Tkeml-1ce • Tkeml-4 • Tkeml-4ce • Tkeml-8 • Tkeml-8ce 	Kemel-based approaches are worthy of exploration
16	<ul style="list-style-type: none"> • Manifold Ranking System 	Ranking approaches do not seem to work

New results

New Results (Task 1A)



New results

New Results (Task 1A)

System ID	Approach	Task 1a	Comments
5	<ul style="list-style-type: none">Discourse profiling, similarity function	<ul style="list-style-type: none">0.03	Some assumptions might be misplaced
6	<ul style="list-style-type: none">Tfidf + neural network, dissimilarity score	<ul style="list-style-type: none">0.10	Tfidf approach performed among the best, like last year
8	<ul style="list-style-type: none">Sentence fusionJaccard CascadeJaccard FocusedSVM methodVoting Method 1Voting Method 2	<ul style="list-style-type: none">0.120.090.120.040.110.10	Second best performance, second highest deviation
9	<ul style="list-style-type: none">Sect-class TSRModified TSRTSR-sent-class	<ul style="list-style-type: none">0.000.050.00	Ranking methods have not worked well

Best performing Systems

Best performing Systems

New results

New Results (Task 1A)

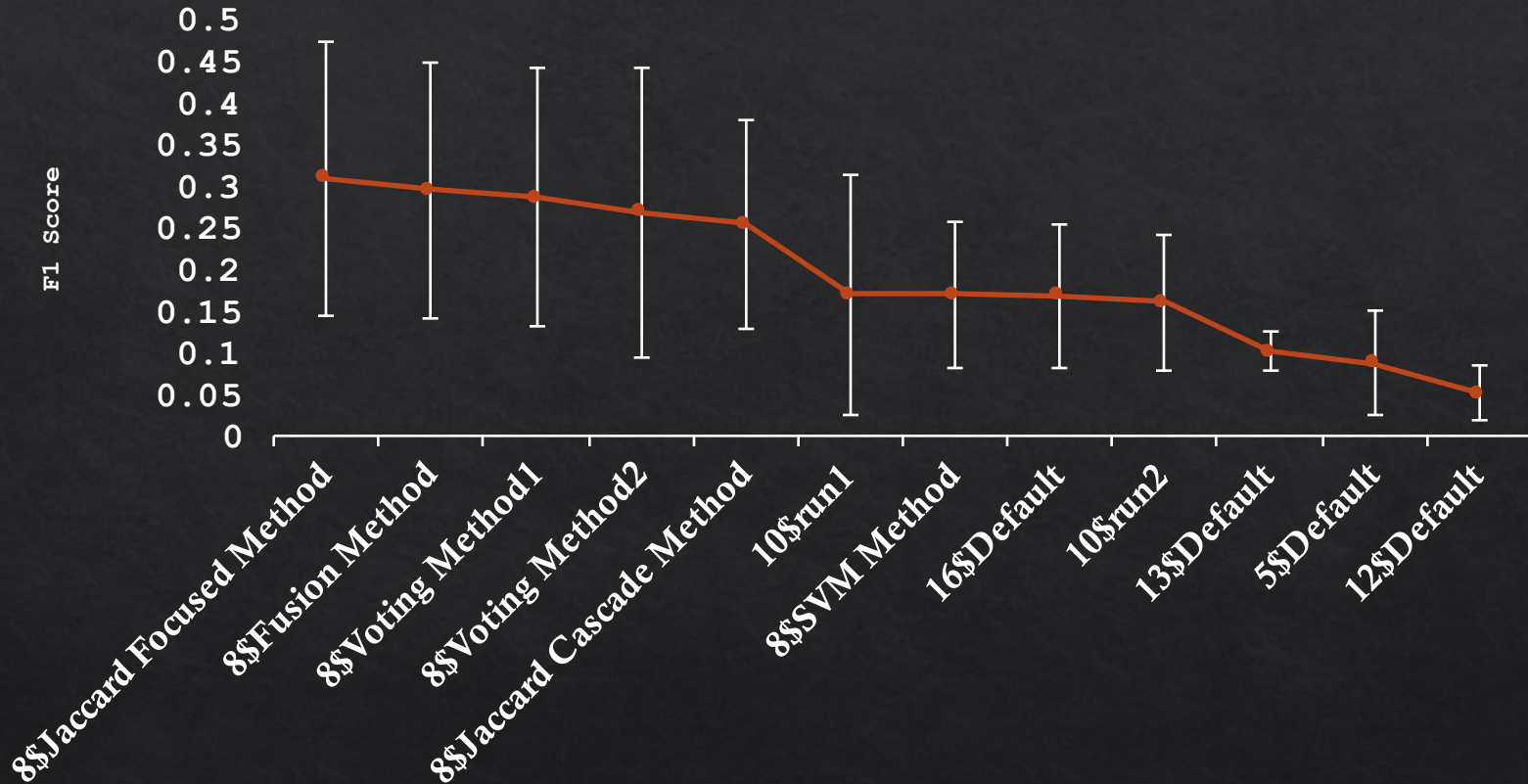
ID	Approach	Task 1A	Comments
10	WEKA + SUMMA <ul style="list-style-type: none"> Method 1 Method 2 	<ul style="list-style-type: none"> 0.02 0.01 	<ul style="list-style-type: none"> Regression did not perform well
12	<ul style="list-style-type: none"> Ranking problem, Text classification problem 	<ul style="list-style-type: none"> 0.02 	<ul style="list-style-type: none"> Suggests that Task 1a is not IR
13	<ul style="list-style-type: none"> Unsupervised bigram overlap method 	<ul style="list-style-type: none"> 0.04 	<ul style="list-style-type: none"> Middle order performance
15	<ul style="list-style-type: none"> Tfidf+st+sl Tkernl-1 Tkernl-1ce Tkernl-4 Tkernl-4ce Tkernl-8 Tkernl-8ce 	<ul style="list-style-type: none"> 0.13 0.01 0.01 0.01 0.01 0.01 0.01 	<ul style="list-style-type: none"> Best performance, most deviation
16	<ul style="list-style-type: none"> SVMRank, Manifold Ranking System 	<ul style="list-style-type: none"> 0.10 	<ul style="list-style-type: none"> Most consistent out of top performing systems

Best performing Systems

Best performing Systems

New results

New Results (Task 1 B)



New
results

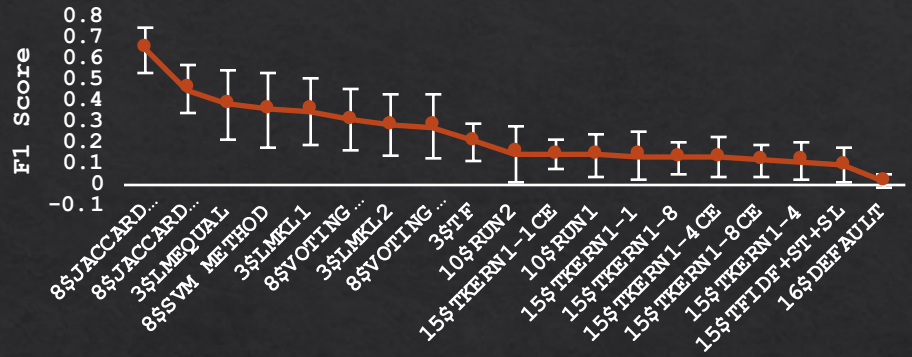
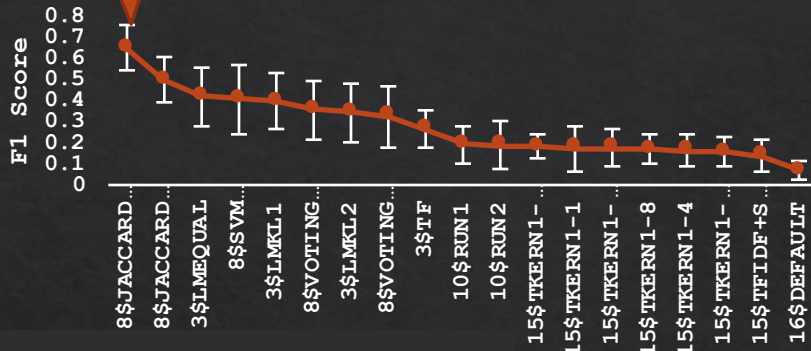
New Results (Task 1B)

ID	Approach	Task 1B	Comments
5	<ul style="list-style-type: none">Transdisciplinary Scientific Lexicon	0.06	Dependency on Task 1A hurts performance
8	<ul style="list-style-type: none">Sentence fusionJaccard CascadeJaccard FocusedSVM methodVoting Method 1Voting Method 2	<ul style="list-style-type: none">0.290.250.310.170.280.26	Combinations of Voting methods with Task 1A approaches worked well
10	<p>WEKA + SUMMA</p> <ul style="list-style-type: none">Text classification 1Text classification 2	<ul style="list-style-type: none">0.130.06	Domain knowledge improves classification
12	<ul style="list-style-type: none">Text classification	<ul style="list-style-type: none">0.01	Citation context is not enough; More features need to be explored
13	<ul style="list-style-type: none">Rule-based approach	<ul style="list-style-type: none">0.05	Dependency on Task 1A and paper structure
16	<ul style="list-style-type: none">Manifold Ranking System	<ul style="list-style-type: none">0.15	Ranking did not perform well

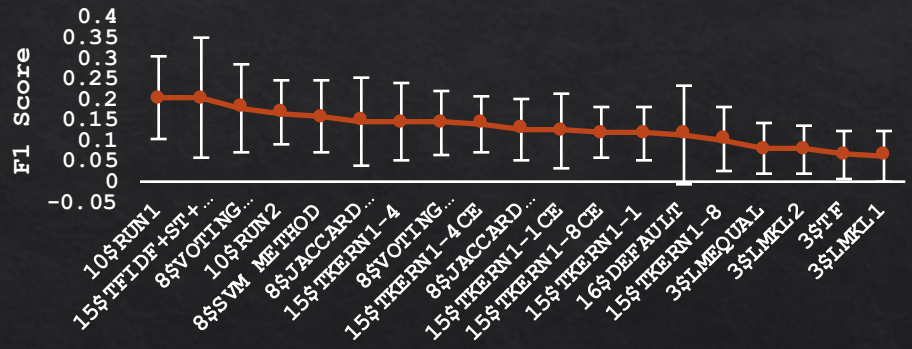
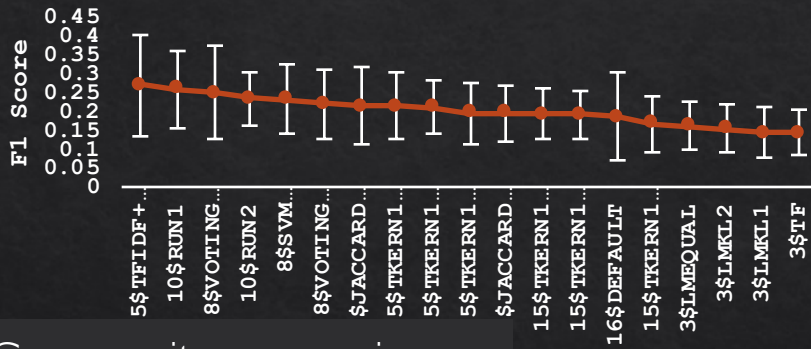
Best
performing
Systems

New results

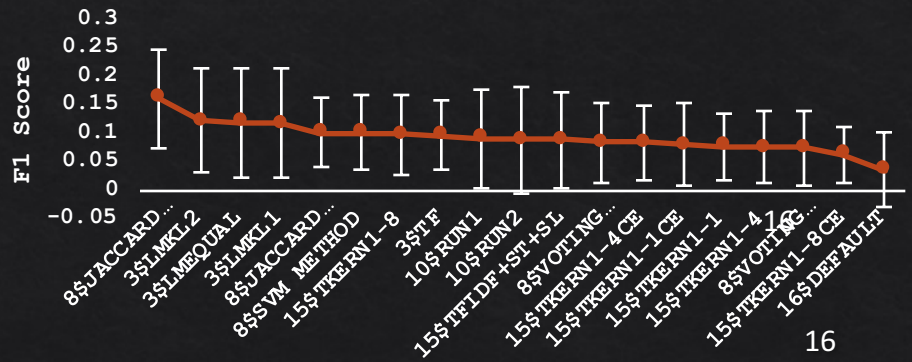
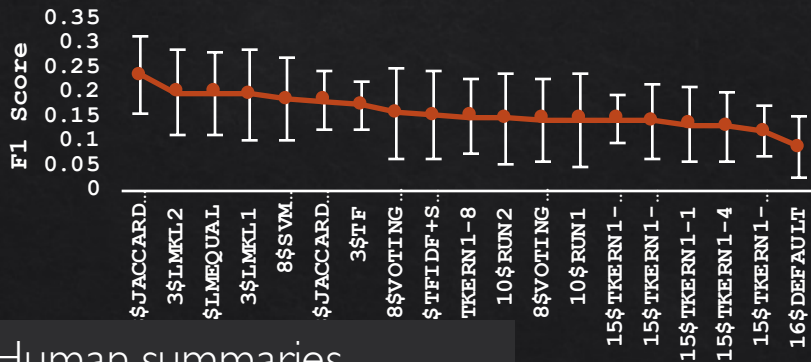
New Results – Task 2



Abstract summaries



Community summaries



Human summaries

Supplemental Analyses

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

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
- ◇ Participant feedback?
 - ◇ Guidelines
 - ◇ The Task
 - ◇ The Corpus – size, #citing papers
 - ◇ Evaluation metrics

Next Steps: IJDL Special Issue



Other shared tasks have a notebook version of the proceedings.

Authors wishing to revise should submit a revised version of their paper to the ACL Anthology.

We also encourage extended versions (e.g., with more detailed analyses) to the IJDL special issue:

<http://bit.ly/birndl-ijdl>

First submission deadline: 30 September

Notification: 15 November

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

Conclusions

- ◆ Successful enlargement of the 2014 pilot task, albeit with some clarification issues
- ◆ We invite teams to examine the detailed results available with the GitHub repo: <https://github.com/WING-NUS/scisumm-corpus/>
- ◆ Results and finalized analyses still in development; CEUR version should be deemed preliminary notebook version of paper
- ◆ Look forward to your discussion for the planning and coordination of the next iteration!
- ◆ Thanks to all teams' participation for the success of CL-SciSumm 2016!

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)

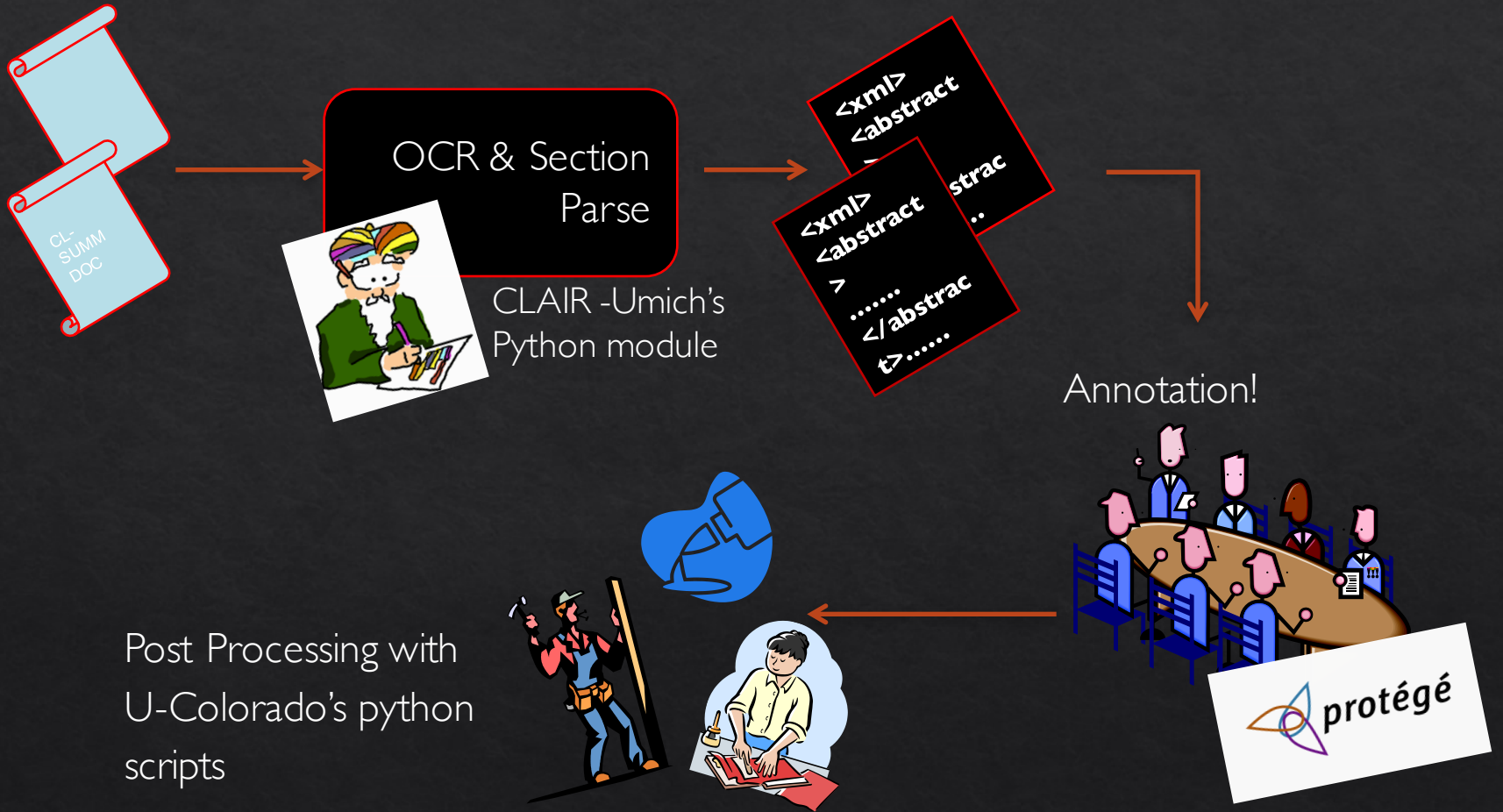
Coherence

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

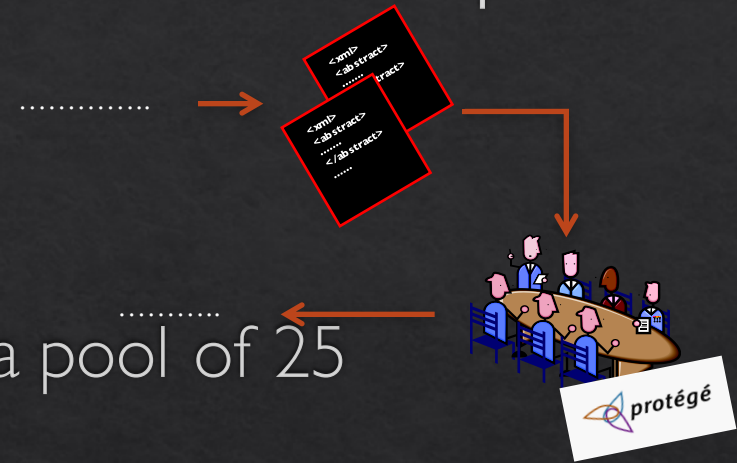
In summary

- ◇ 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 citations 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