Slides available at http://bit.ly/cl-scisumm16-slides and will be filed in GitHub.

The Computational Linguistics Summarization Pilot Task @ BIRNDL 2016

Kokil Jaidka[1], Muthu Kumar Chandrasekaran[2], Sajal Rustagi[1], Min-Yen Kan[2,3]

[1]Adobe Research India

[2]School of Computing, National University of Singapore, Singapore

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

Microsoft Research

Slides available at http://bit.ly/cl-scisumm16-slides and will be filed in GitHub.

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/

Slides available at http://bit.ly/cl-scisumm16-slides and will be filed in GitHub.

Oral Sessions

14:35-14:55	• System 8 Top in Task 1B, among top performers for Task 1A and Task 2 • Remote presentation from China
14:55-15:15	• 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:

Will present results based on the CEUR paper (™old"), stacked average of all runs

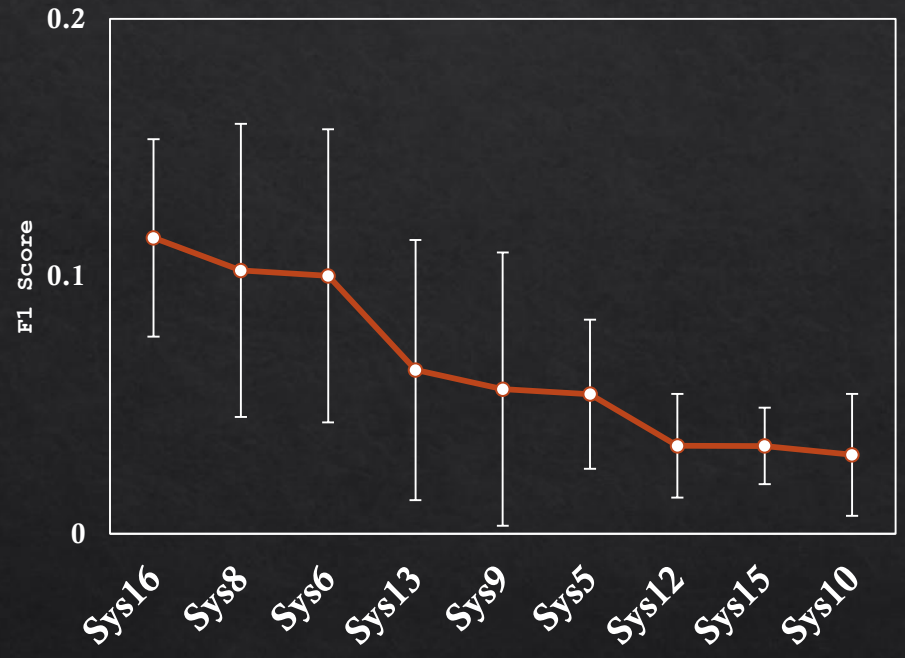
and contrast with newer (still preliminary) results (™new"), individual runs separated

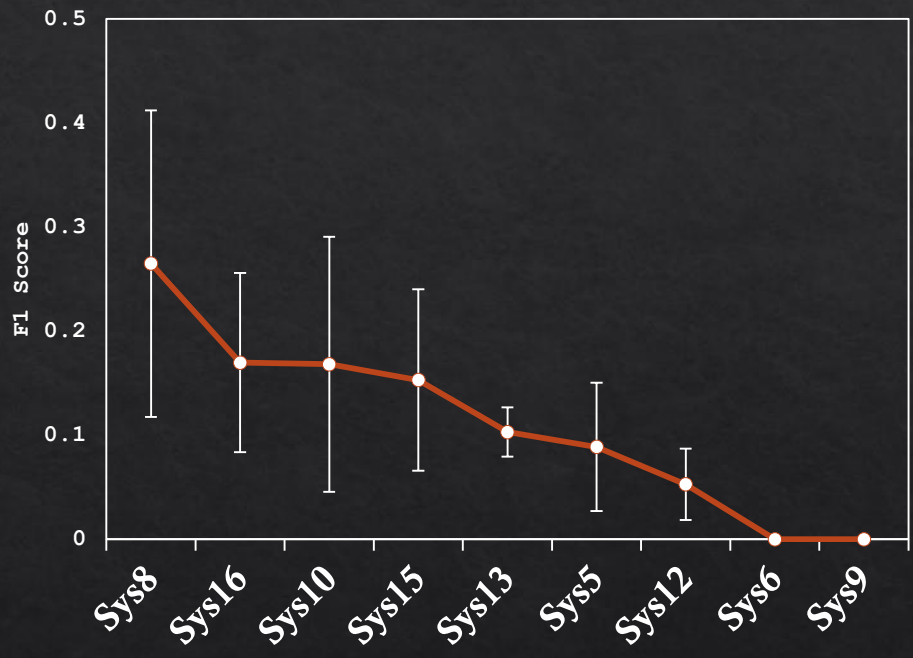
Evaluation

- Task 1A Exact sentence id match
- Task 1B
  - conditional on Task 1A
  - BoW overlap between discourse facets
- Task 2 - ROUGE-2 and ROUGE-4

Old results

System Results (Task 1A & 1B)





CEUR version (all system runs averaged)

Old results

Best Performing System (Task 1A)

System ID	Task 1a	
	Avg performance	
16	0.114941	0.038295
8	0.102306	0.056893
6	0.100184	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

Best Performing System (Task 1B)

System ID	Task 1b	
	Avg performance	
16	0.1696516	
8	0.264754	01473109
13	0.10294	0.0236852
5	0.088737	0.0617396
12	0.052747	0.0341898
15	0.152984	
10	0.168061	

Best performing System

Best performing System

Old results

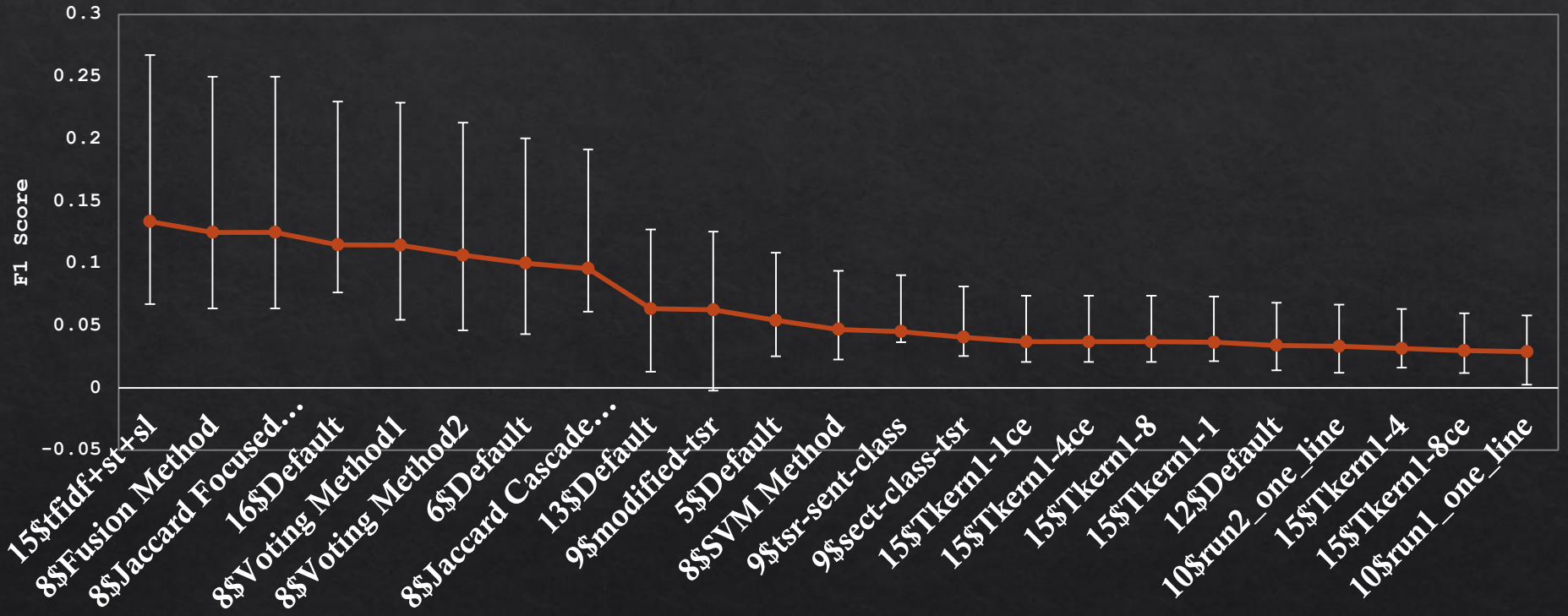
Best Performing System Task 2

Best performing Systems

System ID	Approaches	Comments
3	• NNMF for BioMedSumm	The best for human summaries
8	• Topic modeling	The best for abstract and community summaries
15	• Tkern1-1 • Tkern1-1ce • Tkern1-4 • Tkern1-4ce • Tkern1-8 • Tkern1-8ce	Kernel-based approaches are worthy of exploration
16	• 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	• Discourse profiling, similarity function	• 0.03	Some assumptions might be misplaced
6	• Tfidf + neural network, dissimilarity score	• 0.10	Tfidf approach performed among the best, like last year
8	• Sentence fusion • Jaccard Cascade • Jaccard Focused • SVM method • Voting Method 1 • Voting Method 2	• 0.12 • 0.09 • 0.12 • 0.04 • 0.11 • 0.10	Second best performance, second highest deviation
9	• Sect-class TSR • Modified TSR • TSR-sent-class	• 0.00 • 0.05 • 0.00	Ranking methods have not worked well

Best performing Systems

Best performing Systems

New results

New Results (Task 1A)

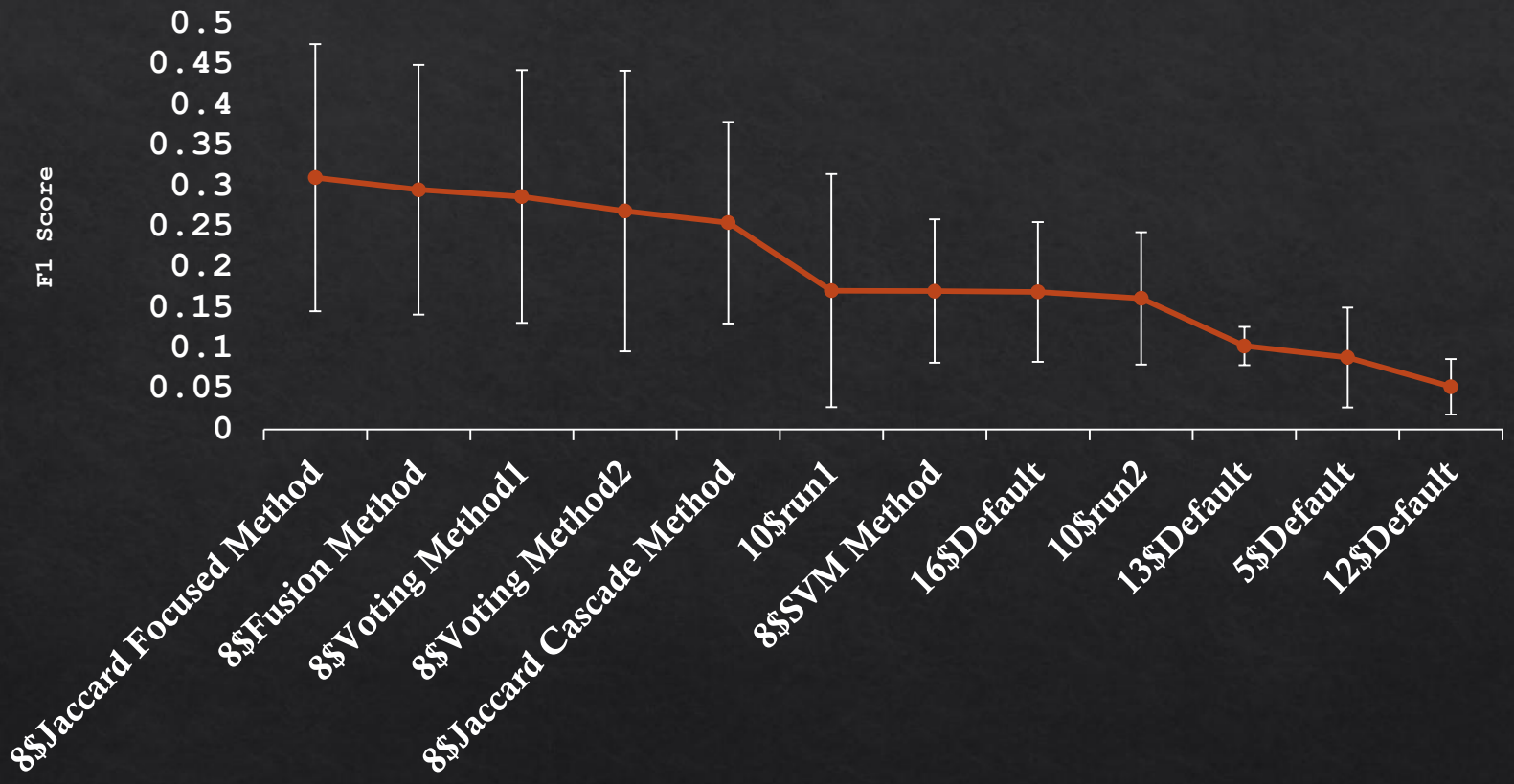
ID	Approach	Task 1A	Comments
10	WEKA + SUMMA • Method 1 • Method 2	• 0.02 • 0.01	• Regression did not perform well
12	• Ranking problem, Text classification problem	• 0.02	• Suggests that Task 1a is not IR
13	• Unsupervised bigram overlap method	• 0.04	• Middle order performance in
15	• Tfidf+st+sl • Tkern1-1 • Tkern1-1ce • Tkern1-4 • Tkern1-4ce • Tkern1-8 • Tkern1-8ce	• 0.13 • 0.01 • 0.01 • 0.01 • 0.01 • 0.01 • 0.01	• Best performance, most deviation
16	• SVMRank, Manifold Ranking System	• 0.10	• Most consistent out of top performing systems

Best performing Systems

Best performing Systems

New results

New Results (Task 1B)



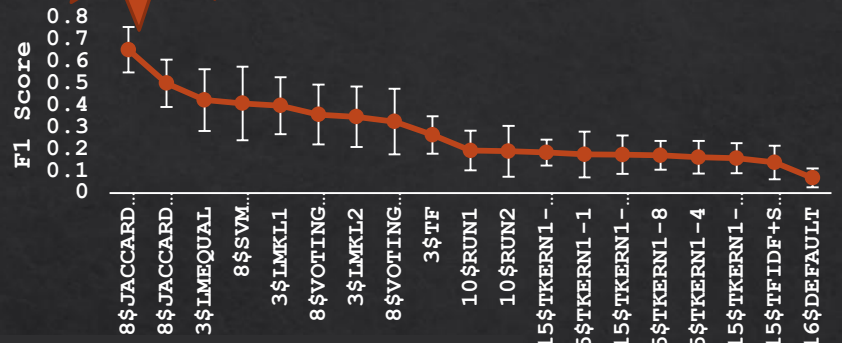
New Results (Task 1B)

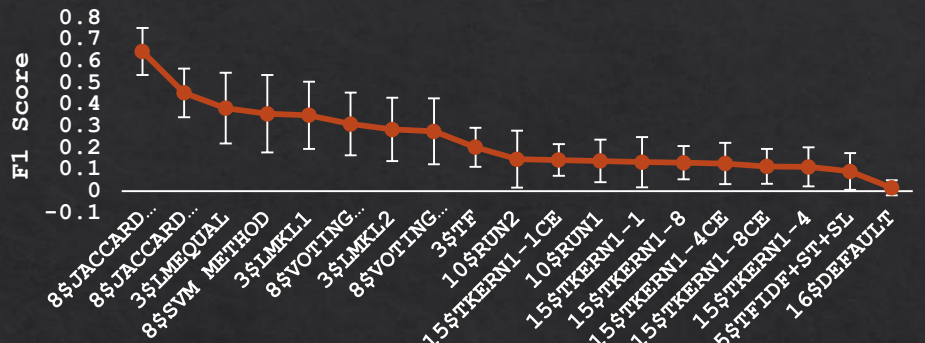
ID	Approach	Task 1B	Comments
5	• Transdisciplinary Scientific Lexicon	0.06	Dependency on Task 1A hurts performance
8	• Sentence fusion • Jaccard Cascade • Jaccard Focused • SVM method • Voting Method 1 • Voting Method 2	• 0.29 • 0.25 • 0.31 • 0.17 • 0.28 • 0.26	Combinations of Voting methods with Task 1A approaches worked well
10	WEKA + SUMMA • Text classification 1 • Text classification 2	• 0.13 • 0.06	Domain knowledge improves classification
12	• Text classification	• 0.01	Citation context is not enough; More features need to be explored
13	• Rule-based approach	• 0.05	Dependency on Task 1A and paper structure
16	• Manifold Ranking System	• 0.15	Ranking did not perform well

Best performing Systems

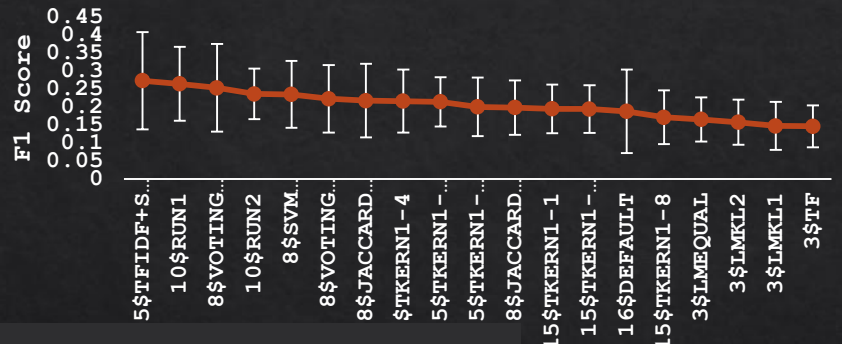
New Results Task 2

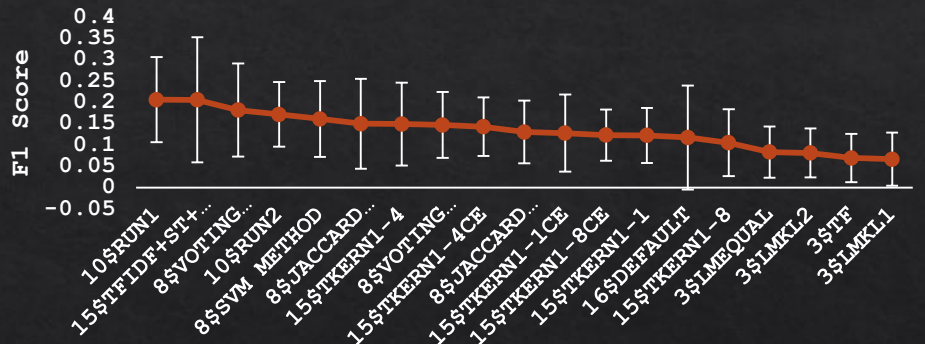
New results



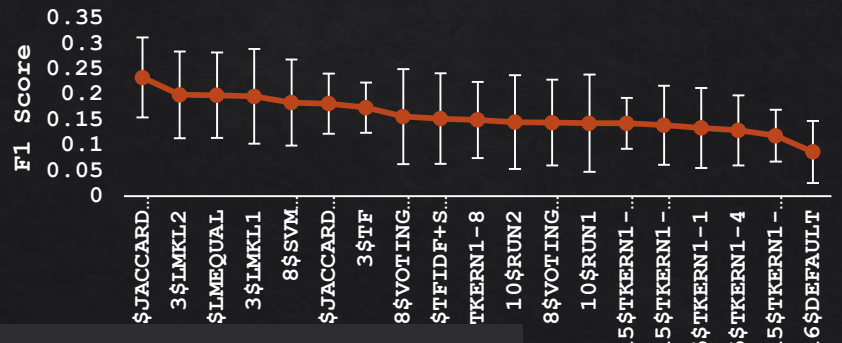


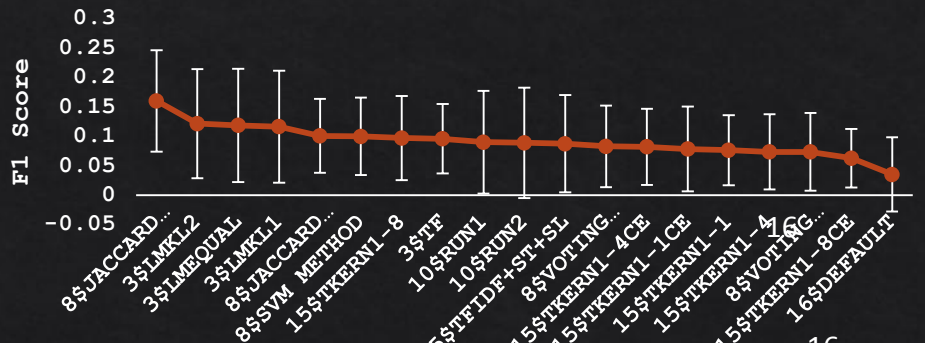
Abstract summaries





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 citances have mixed results
- No significant patterns of performance against:
  - Number of citances 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

Slides available at http://bit.ly/cl-scisumm16-slides and will be filed in GitHub.

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)

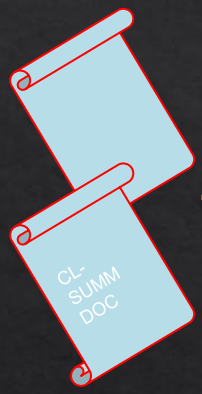
Coherence

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

In summary

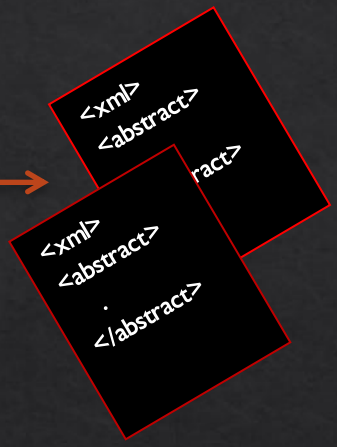
- Community concurs that a citation-based summary of a scientific document is important.
- Citing papers cite different aspects of the same reference paper.
- Assigning facets to these citances may help create coherent summaries.

Annotation Pipeline

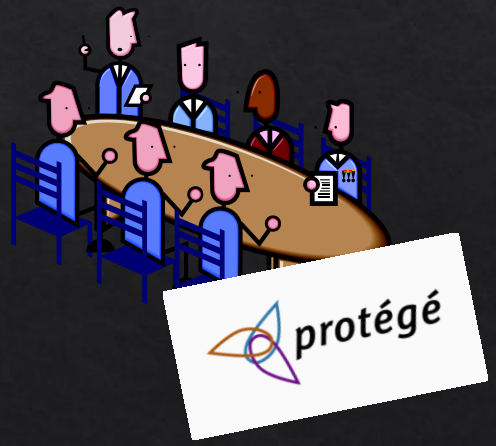


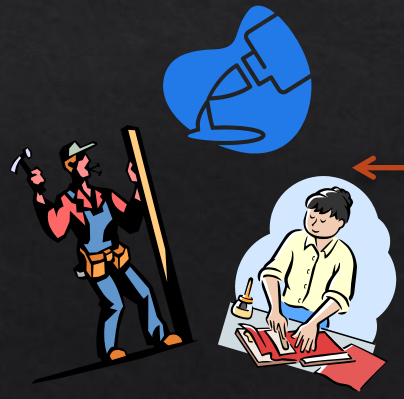
OCR & Section Parse

CLAIR -Umich‰s Python module



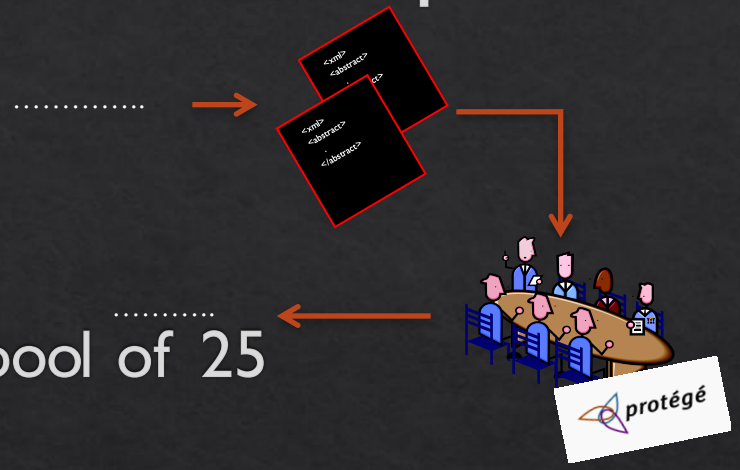
Annotation!





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