

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

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

Kokil Jaidka<sup>1</sup>, Muthu Kumar Chandrasekaran<sup>2</sup>, Min-Yen Kan<sup>2</sup>,

*<sup>1</sup>University of Pennsylvania*

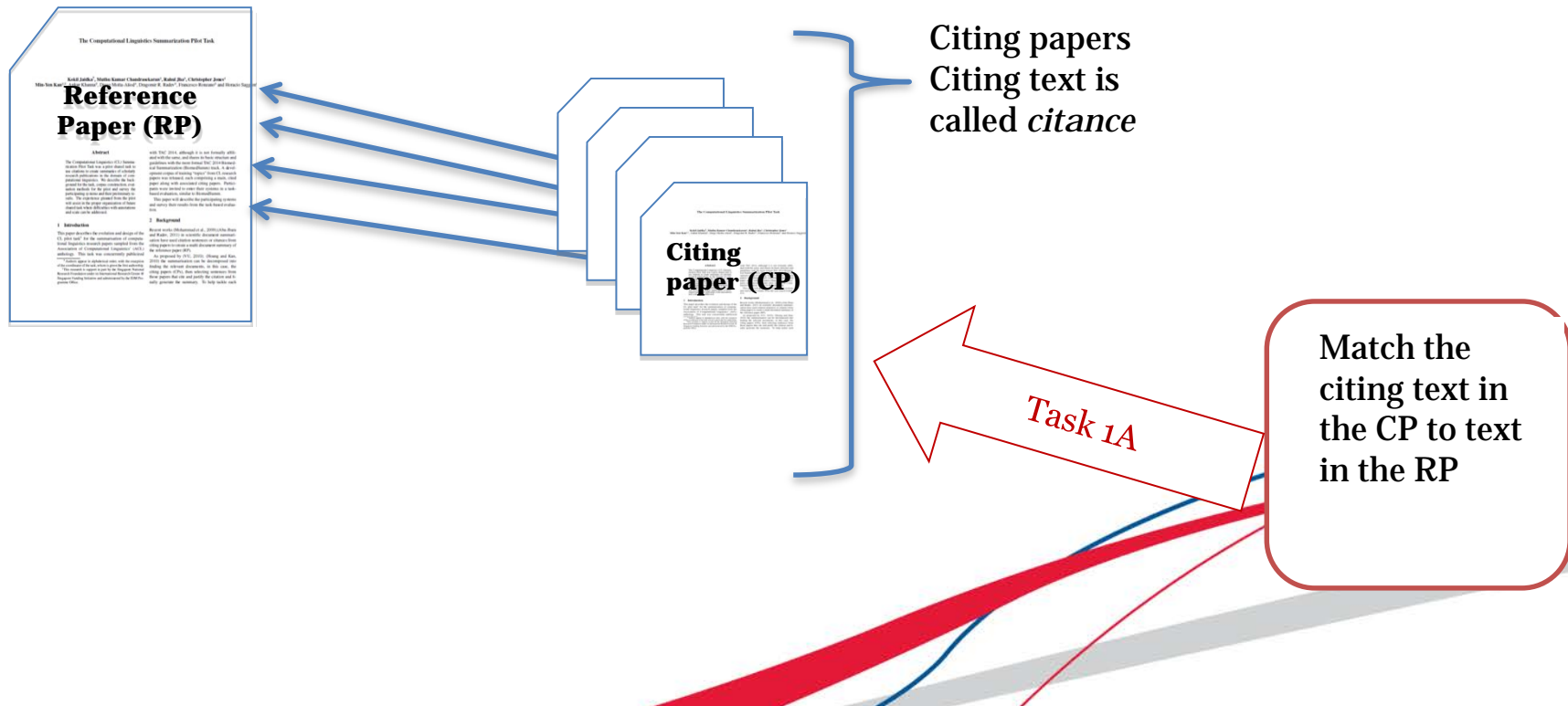
*<sup>2</sup>National University of Singapore*

# 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.

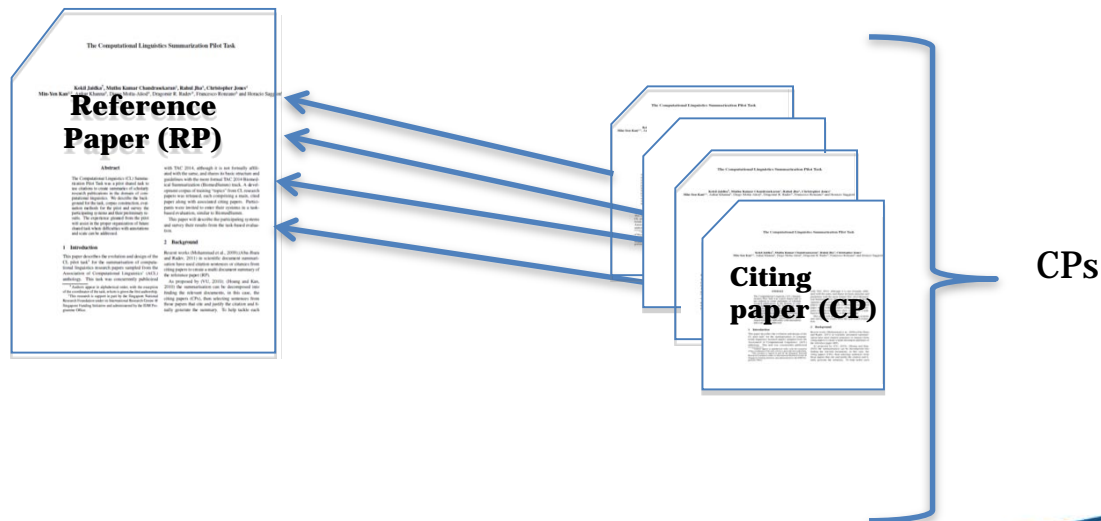


# The CL-SciSumm Shared Task

**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

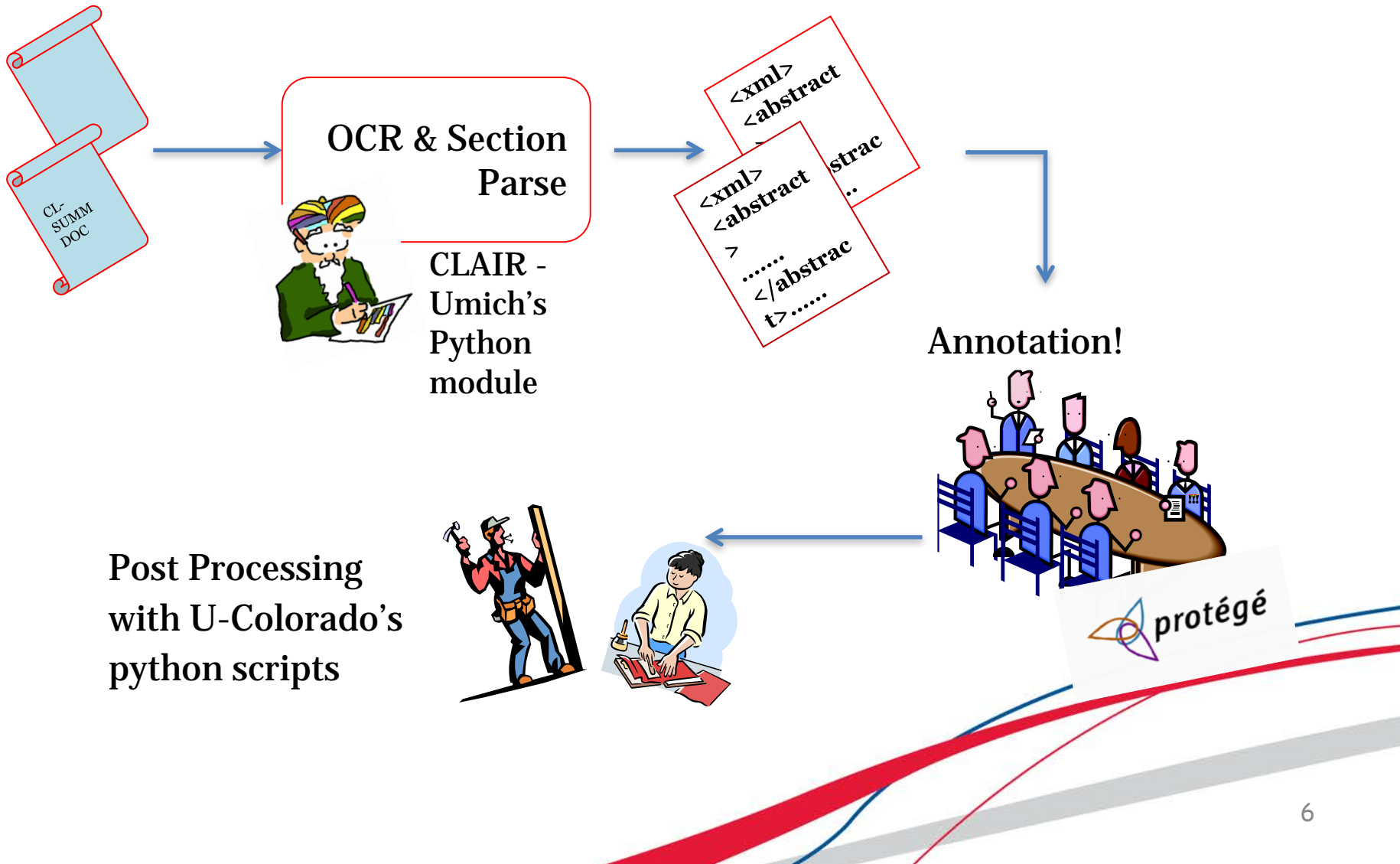
Task 1B



# 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



# 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



**Table 3** System ids mapped to system descriptions

System id	Reference Paper	System Description
[4]	<i>sys3</i>	Vector space model, with non- negative matrix factorization (NNMF)
[14]	<i>sys5</i>	Transdisciplinary Scientific Lexicon (TSL) and Maximal Marginal Relevance
[19]	<i>sys6</i>	TF-IDF and a tripartite neural network
[11]	<i>sys8</i>	SVM classifiers, voting methods
[10]	<i>sys9</i>	TextSentenceRank with similarity functions
[23]	<i>sys10</i>	Linear regression
[13]	<i>sys12</i>	Learning to Rank approach
[1]	<i>sys13</i>	Heuristic approach
[17]	<i>sys15</i>	SVM with convolution kernel
[2]	<i>sys16</i>	SVMRank, decision tree classifier, Manifold Ranking method

1. Aggarwal, P., Sharma, R.: Lexical and Syntactic cues to identify Reference Scope of Citation. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 103{112. Newark, NJ, USA (2016)

2. Cao, Z., Li, W., Wu, D.: PolyU at CL-SciSumm 2016. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 132{138. Newark, NJ, USA (2016)

4. Conroy, J., Davis, S.: Vector space and language models for scientific document summarization. In: NAACL-HLT, pp. 186{191. Association of Computational Linguistics, Newark, NJ, USA (2015)

10. Klamp, S., Rexha, A., Kern, R.: Identifying Referenced Text in Scientific Publications by Summarisation and Classification Techniques. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 122{131. Newark, NJ, USA (2016)

11. Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., Cong, X., Peng, H.: CIST System for CL-SciSumm 2016 Shared Task. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 156{167. Newark, NJ, USA (2016)

11. Lu, K., Mao, J., Li, G., Xu, J.: Recognizing reference spans and classifying their discourse facets. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 139{145. Newark, NJ, USA (2016)

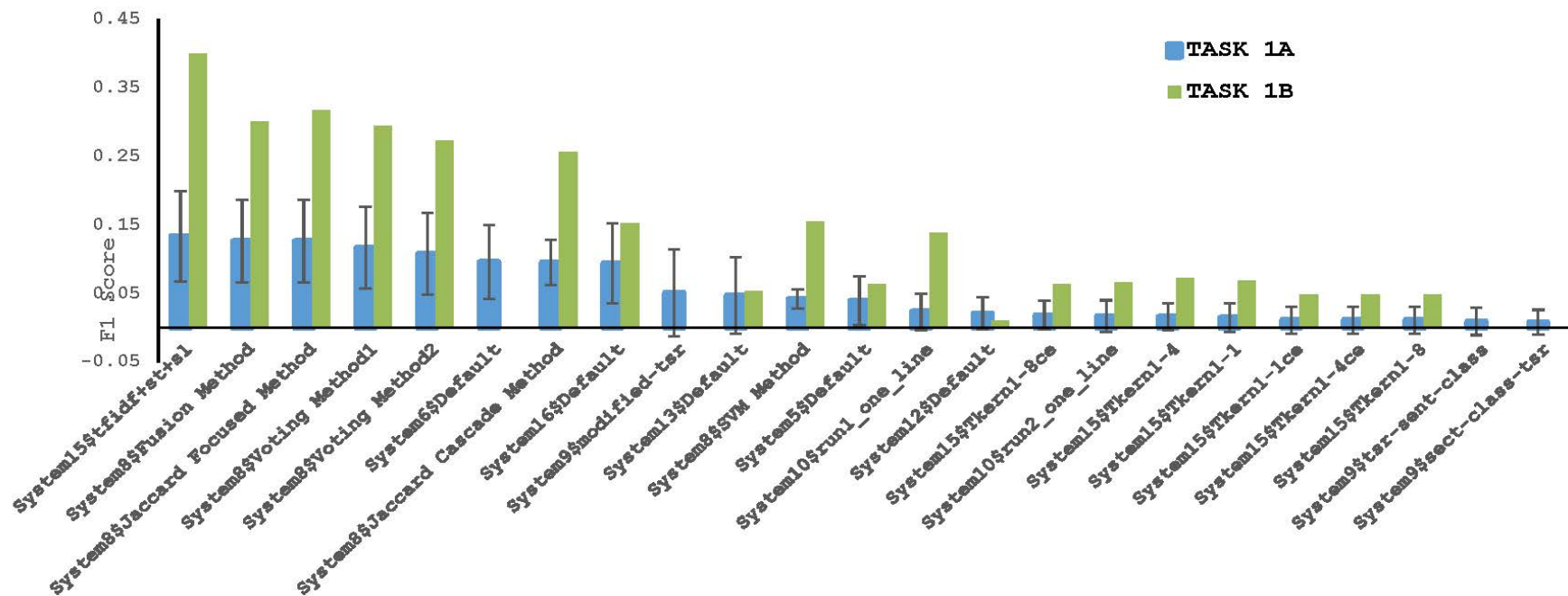
14. Malenfant, B., Lapalme, G.: RALI System Description for CL-SciSumm 2016 Shared Task. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 146{155. Newark, NJ, USA (2016)

17. Moraes, L., Baki, S., Verma, R., Lee, D.: University of Houston at CL-SciSumm 2016: SVMs with tree kernels and Sentence Similarity. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 113{121. Newark, NJ, USA (2016)

19. Nomoto, T.: NEAL: A neurally enhanced approach to linking citation and reference. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 168{174. Newark, NJ, USA (2016)

23. Saggion, H., AbuRa'Ed, A., Ronzano, F.: Trainable Citation-enhanced Summarization of Scientific Articles. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 175{186. Newark, NJ, USA (2016)

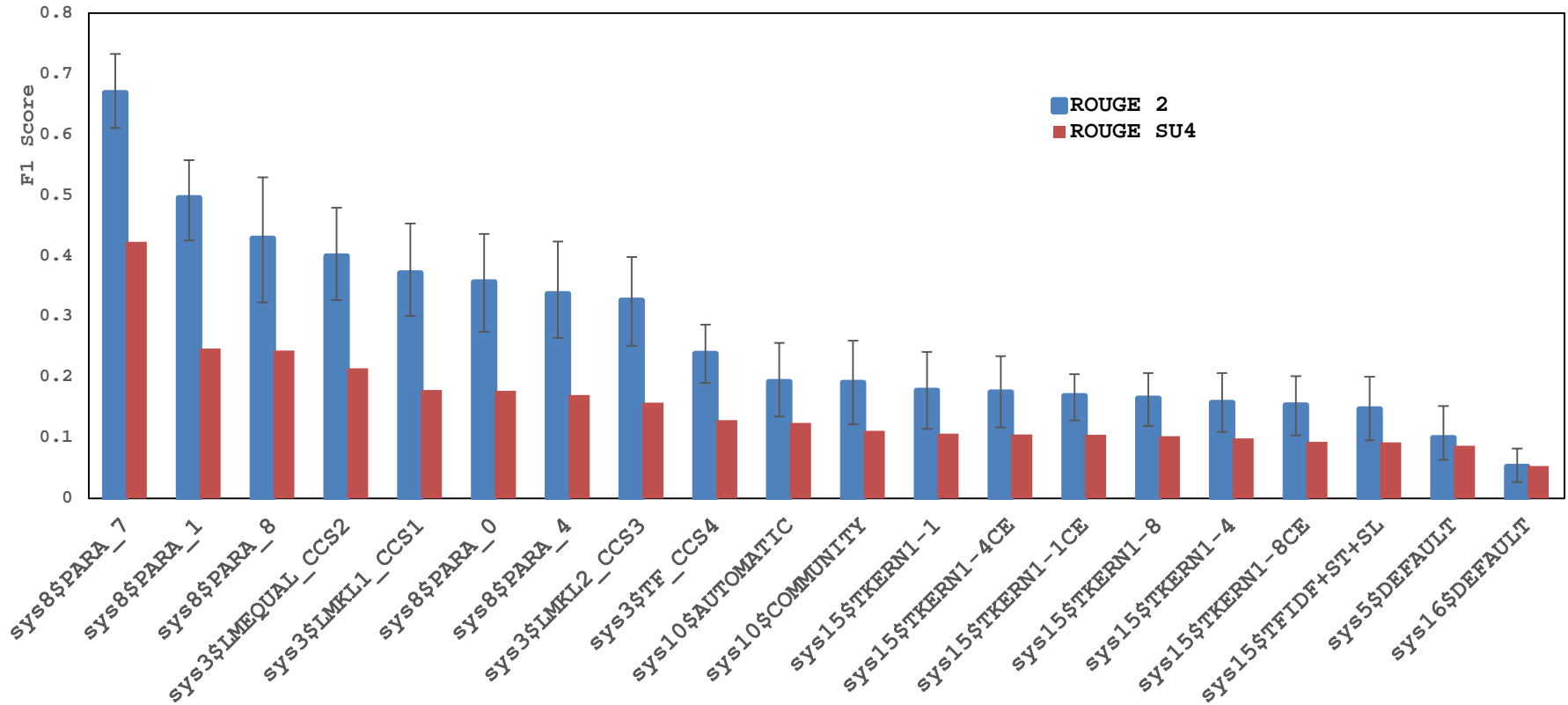
# Best Performing Approaches: Task 1



- System 15 – Tfidf
- System 8 - Combinations of SVM Classifier + term frequencies + surface features
- System 6 - Tfidf + embeddings-based neural network
- System 16 – SVMRank, decision tree classifier

# Best Performing Approaches (Task 2)

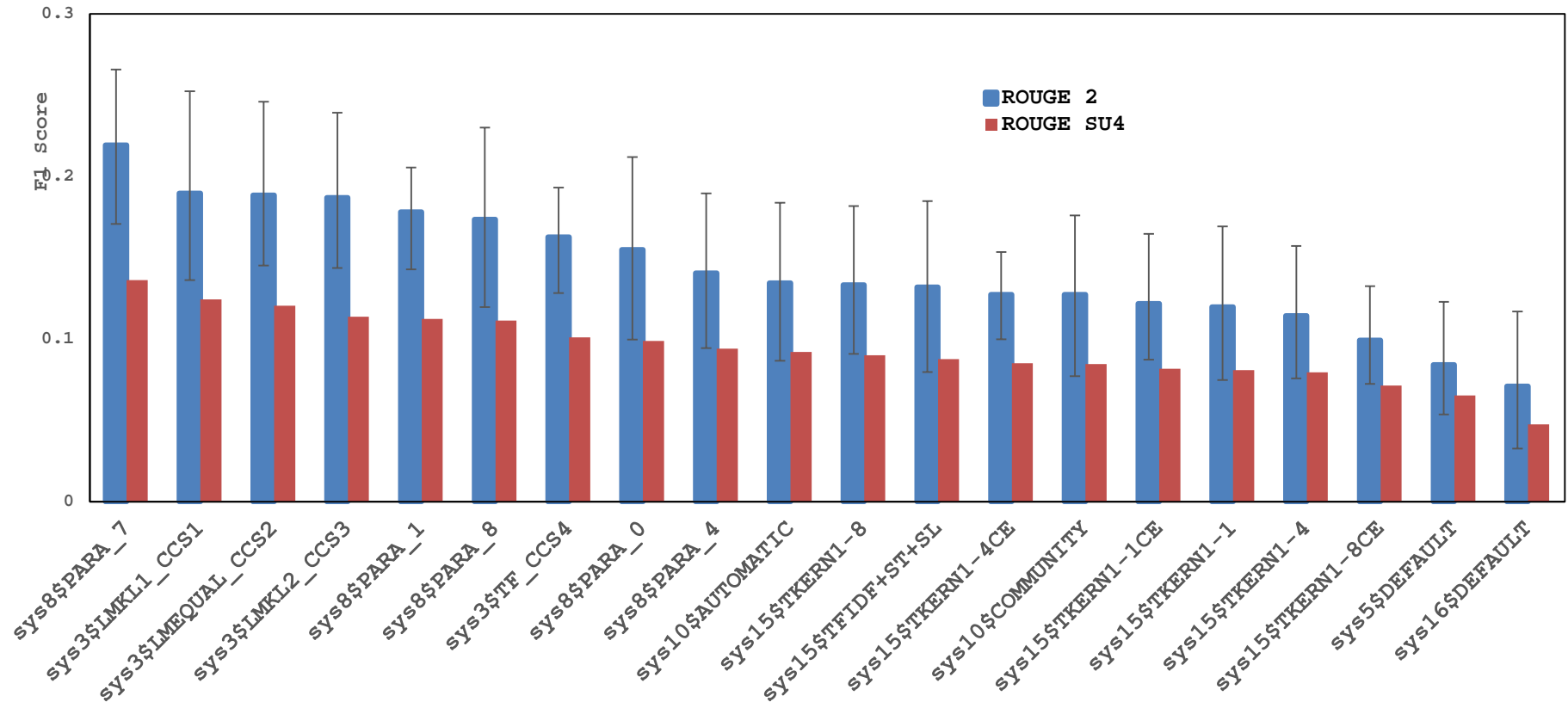
System summaries vs Paper Abstracts



- System 8 - SVM classifiers, voting methods
- System 3 - Term frequency + NNMF
- System 10 - WEKA + feature relevance scores

# Best Performing Approaches (Task 2)

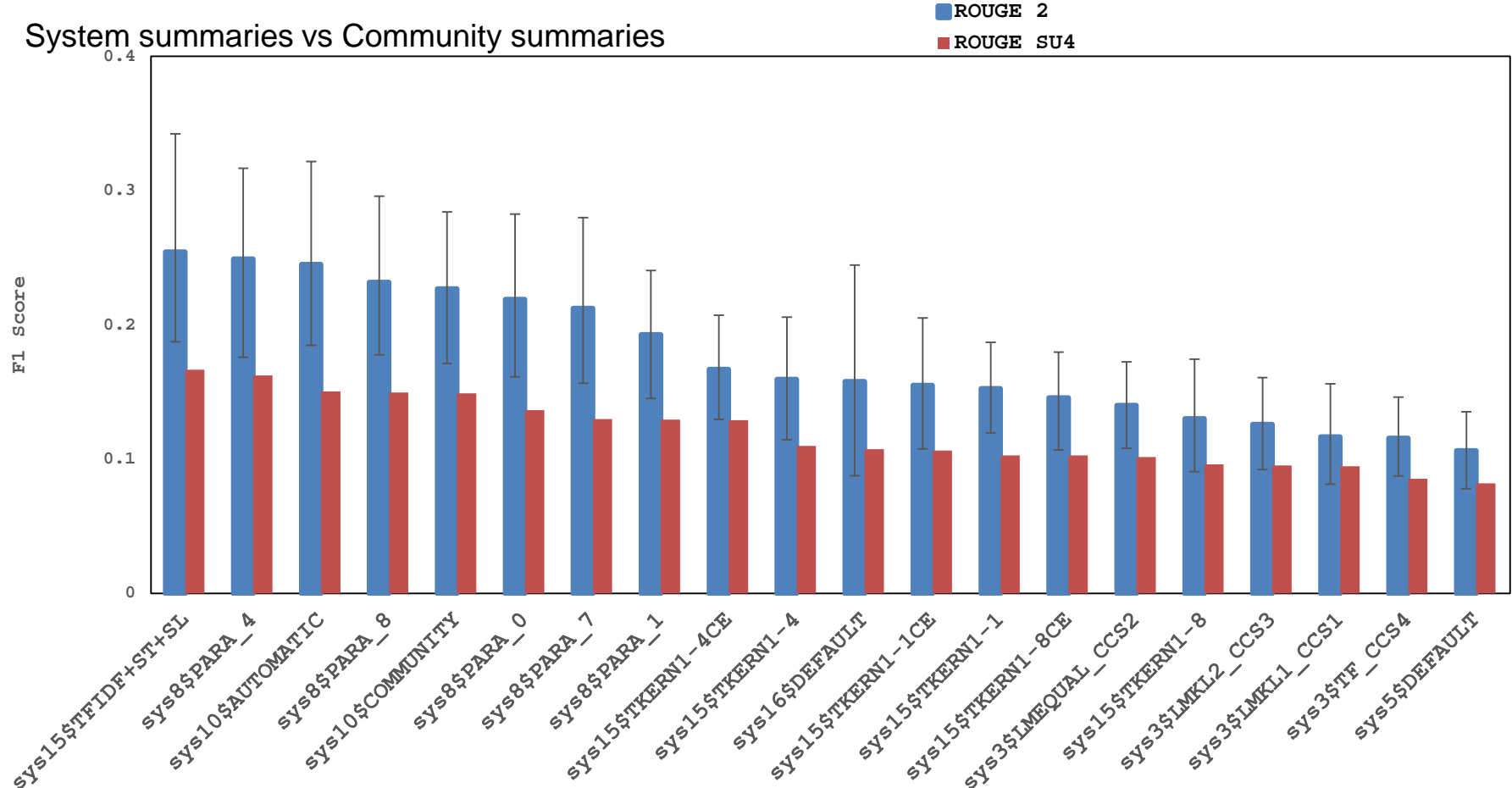
System summaries vs. Human summaries



- System 8 - SVM classifiers, voting methods
- System 3 - Term frequency + NNMF
- System 10 - WEKA + feature relevance scores

# Best Performing Approaches (Task 2)

System summaries vs Community summaries



- System 15 - SVM with convolution kernel
- System 8 – SVM classifiers, voting methods
- System 10 - WEKA + feature relevance scores

# Dataset Limitations

- Task 1B: limited number of samples for most (e.g., hypothesis) discourse facets, inconsistent labeling
- Preprocessing: OCR + Parsing **Rolf Kümmerli,<sup>1,2</sup> Andy**

↓  
Rolf K<sup>..</sup>ummerli,<sup>1,2</sup>

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

# Acknowledgements

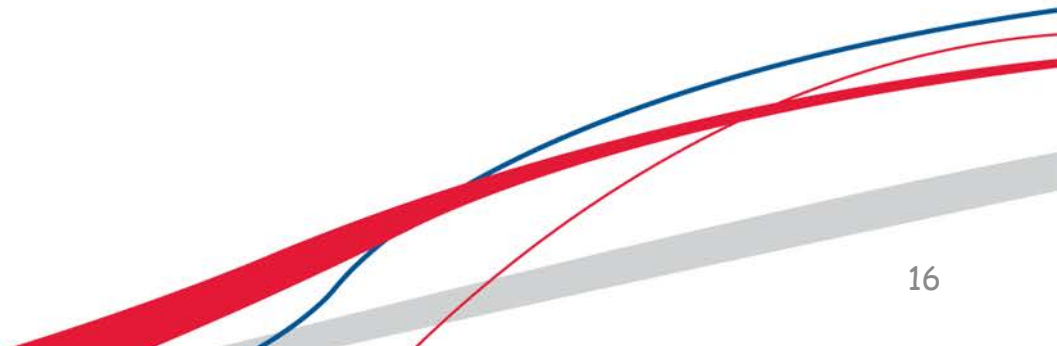
- Chin-Yew Lin (MSRA)
- NIST and Hoa Dang
- Lucy Vanderwende, MSR
- Anita de Ward, Elsevier Data Services
- Kevin B. Cohen, Prabha Yadav (U. Colorado, Boulder)
- Rahul Jha (Google)
- U-Hyderabad Annotators:
  - Aakansha Gehlot, Ankita Patel, Fathima Vardha, Swastika Bhattacharya and Sweta Kumari
- System 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