Scientific Document Summarization
Standardized Reference Corpus and Shared Task
Min-Yen Kan*, Kokil Jaidka+, Muthu Kumar Chandrasekaran*
*National University of Singapore
+University of Pennsylvania
Chin-Yew Lin, Microsoft Research

Where we are after 3 years?

2014
- 10 target papers; 100 citing papers; annotated by three of the organisers at NUS
- CL-SciSumm Pilot Shared Task @ TAC, MD, USA
- Self-reported evaluation cross-validated on a single corpus

2016
- 30 target papers; 500 citing papers; annotated by a 5+ applied linguistics students from U. Hyderabad, India
- Largest corpus in the world
- CLSciSumm – 16 @ JCDL, New Jersey
- Centralised evaluation on a test set of 10 papers

2017
- Upcoming! Expanding to 50 target papers
- Centralised, automated, transparent evaluation on codalab.org
- CLSciSumm – 17 @ SIGIR, Tokyo
- Sign up! http://wing.comp.nus.edu.sg/cl-scisumm2017

Corpus annotation and shared task funded by MSR

Tasks
- Task 1A: Identify the text span in the RP which corresponds to the citing sentences from the cited paper (CP).
- Task 1B: Identify the discourse facet for every cited text span from a predefined set of facets.
- Task 2: Output a summary of the reference paper (RP).

Result Overview from CLSciSumm – 16

Annotated Corpus
- Gold standard annotations for Task 1A and 1B, per topic or reference paper
- Community and hand-written summaries for Task 2, per topic

Download the corpus form GitHub repo, accessible using this QR code. https://github.com/WING-NUS/scisumm-corpus

Acknowledgements: Chin-Yew Lin (MSRA), Hoa Dang and NIST, Lucy Vanderwende (MSR), Anita de Ward (Elsevier Data Services), Kevin B Cohen and team (U. Colorado, Boulder), Rahul Jha (Google), CLAIR, (U.Michigan, Ann Arbor) and University of Hyderabad (India)