## Microsoft Research Funded Project Summary

Please submit your project summary to winnie.cui@microsoft.com before **4/15, 2017**.

This template is used for collecting the information and results of the funded projects in MSRA Call for Collaborative Research Program (CCRP). As stated in the project agreement, the results can be shared among the academic community through program web site and materials (poster, booklet, flyer, paper collection, etc.) to benefit the research community.

Please fill the following summary form in English.

Submission Date：4-20-2017 （Month-Day-Year, e.g. 4-10-2017）

Submitted By：Min-Yen KAN Phone：+65 6516 1885 E-mail：kanmy@comp.nus.edu.sg

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Project Information** | | | | | | |
| **Project Name** | | | | Scientific Document Summarization Standardized Reference Corpus and Shared Task | | |
| **Project Website** | | | | [*https://github.com/WING-NUS/scisumm-corpus*](https://github.com/WING-NUS/scisumm-corpus) | | |
| **Institution/University** | | | | National University of Singapore | | |
| **School/Department** | | | | School of Computing, Department of Computer Science | | |
| **Project Team**  *Note: Please attach one photo of principal investigator. Please list-up* ***all members*** *including graduate students* | | | | | | |
| **Name** | | | **E-mail** | | **Phone** | **Role** |
| Min-Yen Kan | | | kanmy@comp.nus.edu.sg | | +65 6516 1885 | Principal Investigator |
| Muthu Kumar Chandrasekaran | | | cmkumar087@gmail.com | |  | Collaborator |
| Kokil Jaidka | | | jaidka@sas.upenn.edu | |  | Collaborator |
|  | | |  | |  |  |
|  | | |  | |  |  |
| **Project Summary** | | | | | | |
| MSRA expects a result-oriented project summary (around 1200 words) that describes the solved problem, unique methods, applications, benefits to users and academia.  **Please fill the summary template with the following structure**   1. **Project Goal**   We developed resources and technologies that solve problems for scientific summarization. Current scientific summaries are written manually by scholars, synthesizing the goals and contributions of a study. Advances in automated document summarization, while significant, are not adapted to summarize the specialized scientific document format, typified with conventional argumentation patterns and use of technical terminology. Furthermore, automatic summarization systems do not support a researcher in the actual task of a literature survey – which may involve tracking a research over time, and following developments since a seminal publication, which could amass hundreds to thousands of citations per year. It is also difficult to quantitatively evaluate these summaries, because there is no single rubric of what comprises an ideal scientific summary. Importantly, the key resource of a standardised reference corpus is missing -- this is needed to interest the research community in dedicating resources and manpower, as comparative objective benchmarking is critical to reproducibility and assessment.  To address these issues, we built a corpus for scientific summarization, and led the scientific community with a shared task based upon the dataset. As a result of the project, we delivered two objectives: the collection and open-sourcing of a sizeable scientific document summarization corpus with articles culled from the ACL Anthology, a community workshop to gather partners of similar interests. Our corpus annotation work builds on our current and previous efforts in the area, with a co-proposed community workshop and pilot corpus and evaluation project led in the previous year (2014). Our workshop organization efforts synergize with MSRA’s interest in academic search and natural language processing and also is based on previous successful workshop in 2009.   1. **Technical breakthrough**   Big science is now a focal area in many scientific disciplines: in addressing concerns with the volume, velocity and variety of scientific data (Schatz, 2014). However, while many researchers focus on big data, we note that science is most profitably propagated through the scientific literature. The scientific literature has recently become its own information overload problem with the large scale growth of scientific publications. For researchers to stay relevant and keep abreast of research requires greater automation and management, even for this signal-rich big data problem.  One promising direction to address this problem is automatic summarization. Automated processes to ingest, digest and synthesise new results would help researchers focus on their core tasks and make long range connections among disparate fields of research. While the scientific article format has evolved to include an abstract, which is a type of summary, through various studies scientists have shown that the abstract only represents one perspective of the importance and contribution of a scientific work. Argumentative analyses of scientific articles have shown that contributions are scattered in various sections of the document (Teufel and Moens, 2002), and that literature surveys that categorise between integrative and descriptive select for different types of information usually prevalent in different sections (Jaidka et al., 2013). Furthermore, the contribution of a work as assessed by the research community at large is embodied elsewhere: namely, the citations to the work (Teufel et al., 2006). In others’ work, the sentences that reference a target paper also describe a “summary” of the target paper -- a community assessed importance of the work (Nakov et al., 2004). In creating a suitable machine algorithm for scientific document summarisation, the analyses of three sources of text must come together: the author-created abstract, the full text of the document, as well as the contexts of the citations from referencing documents.  The automated summarisation of individual papers is an important milestone towards the better semantic understanding of our own sciences. Tackling this problem represents progress towards greater contributions, such as automated survey paper generation and automated research problem and methodology formulation, which require multidocument summarization techniques (e.g., Lin and Hovy, 2002). The problem encapsulates a target application (scholarly documents) for specific natural language and information retrieval tasks (discourse understanding, information extraction, importance and relevance assessment), making the problem widely applicable to many researchers in the field.  Despite these important criteria, progress towards automated summarisation of science has been slow, due to the difficulties in drawing together the community on this problem. A key visible problem has been a lack of a standardised dataset from which the community can experiment and reproduce results for scientific document summarisation. We believe that building a standardised summarisation reference corpus and organising a shared task with the resultant materials would greatly spearhead research in this direction.  Our project address this key bottleneck in scientific summarisation through the execution of two complementary work packages: first, the compilation of a reference scientific document summarisation corpus; and second, the organisation of a shared task for building automated summarisation systems utilising the corpus. To garner more attention around the shared task, the PI and collaborators are proposing a scientific workshop on Text and Citation Analysis for Digital Libraries in parallel with this proposal submission.  Building on our teams’ previous experience in 2014 with the coordination of the pilot SciSumm task, we enlarged the pilot SciSumm corpus for the creation of the standardised summarisation reference corpus. Our methodology called for constructing document sets where a target document to be summarized is accompanied by later, citing documents such that the community’s viewpoint can also be imparted into the resultant summary. In this way, annotators have access to both the author’s point of view as well as the community’s assessment, which often varies over time (Abu-Jbara et al. 2013). The SciSumm corpus was built from sampled documents from the ACL Anthology, an open-source archive of scholarly documents that form the single largest source of natural language processing and computational linguistics research globally. We retained our limitation of the scope to computational linguistics articles (though the use of the ACL Anthology), as these works are immediately understandable and of import to the research community that would participate in the shared task, and the summaries themselves would thus of direct interest to the prospective participants. We note that this although any resultant study will be broadly applicable to many sciences that employ the convention of the scientific method (hypotheses, experiment, validation), and not limited to NLP and CL specifically.  In our preliminary preparation, we have identified a few potential sources for the dataset annotation. Dr Jaidka and the PI have been in contact with the identified parties and directly oversee the annotation training, collection and cleaning phases of the corpus building. The lessons learned from the pilot task as well as the revised annotation guidelines serve an important role in the training and annotation phases of the corpus annotation.  Our second objective was synergistic with the first. A shared task was organized as part of the workshop. Following previous successful shared task organisation, we asked competition entrants to also help with the annotation process, in a Cranfield-styled evaluation pool. This method broadens the interest of the community in maintaining the corpus and further trains manpower in the annotation methodology needed to broaden the task beyond the annotated sets developed in the first work package.   1. **Innovative Applications**   Our CL-SciSumm project enables future research that is on the horizon, on using natural language processing technologies such as information extraction from scientific documents and cross-document scientific document summarization. We built a community around the largely natural language processing oriented research faculty and students, in organizing CL-SciSumm, but also helped to network these scholars with the larger community of scholars working in scholarly data processing for IR by aligning our work with the Digital Libraries and Information Retrieval communities (JCDL and SIGIR). The CL-SciSumm task brought a unifying theme to the NLPIR4DL and BIR workshops that have been longstanding activities in those communities with the interests of the NLP community in performing text summarization.  Another reference shared task that was similar and related in scope is the ScienceIE 2017 SemEval task which aimed at a similar scope of bringing NLP scholars together for complementary purposes. We’re pleased to report that our CL-SciSumm work set the benchmark for work in this area and with the creation of independent scholars at University College London (UCL) and University of Massachusetts, Amherst, has already created synergistic work and energy that can be claimed under MSR and CL-SciSumm’s integral contributions.    Fig 1. Dr Kokil Jaidka presenting our CL-SciSumm efforts at BIRNDL 2016, co-located with JCDL 2016 in New Jersey, USA.    Fig 2. All of the workshop participants for BIRNDL 2016 pose for a group photo.   1. **Academic Achievement**  Our CL-SciSumm project created both a reference scholarly dataset of use in academic work, as well as three non-refereed reports on the project and its accompanying workshop: the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). These works have already started to garner citations from scholars outside of our organizing committee, as others have recognized the importance of this dataset and its active support and revision to enlarge the quality and quantity of the dataset.We note that the published work only references the 2016 required deliverables and are sure of the long-term impact our work will have in shaping the future work in natural language processing of scholarly articles. **5. Achievement in Talent Fostering**  Our work has helped develop talent on several levels. At the lowest level, our project helped over ten applied linguistics students at the University of Hyderabad apply their scholarship to the creation of the dataset and annotations. Our work helped our collaborators improve the portfolio of research and leadership, in rallying the community around the shared task of summarization of scientific articles. Dr Kokil Jaidka and Mr Muthu Chandrasekaran now have gained experience in the entire lifecycle of a shared task: the sourcing, cleaning and annotation of a dataset, and the work towards garnering participation and organizing a final venue for reporting, and generating the reports for archival reflection by the community.  **6. Collaboration with Microsoft Research**  We have updated our MSRA collaborator, Chin Yew Lin, on the developments of the second and third runs on our CL-SciSumm shared task and corpus development. Our team has worked on outreach with MSR and MSRA by contacting Lucy Vanderwende (MSR), Alex Wade (former MSR) and Ming Zhou (MSRA), and informing them of the developments in our shared task.  **7. Project Development**  The original timeframe for the project development ended in 2016. However, there were surplus funds (due to the use of less expensive but equally rigorous annotation efforts) and the community and our recruited manpower were interested in continuing the work. We are making good use of the surplus funding to fund the additional year’s (2017) work to give better value and visibility to both our recruited research staff and Microsoft. | | | | | | |
| **Publications**  *List of publications to date or planned in next 2 years related to the project (paper, patent, report, presentation, book, book section)* | | | | | | |
| **Paper publication** | *N/A* | | | | | |
| **Other Publication** | 1)Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi and Min-Yen Kan (2016) *Overview of the CL-SciSumm 2016 Shared Task*. In Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016). June. Newark, New Jersey, USA. pp. 93-1022) Guillaume Cabanac, Muthu Kumar Chandrasekaran, Ingo Frommholz, Kokil Jaidka, Min-Yen Kan, Philipp Mayr and Dietmar Wolfram (2016) *Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016)*. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16). ACM, New York, NY, USA, 299-300. 3) Guillaume Cabanac, Muthu Kumar Chandrasekaran, Ingo Frommholz, Kokil Jaidka, Min-Yen Kan, Philipp Mayr and Dietmar Wolfram (2016) *Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016)*. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16). ACM, New York, NY, USA, 299-300. | | | | | |
| **Project Demonstration** | | | | | | |
| **Demo list** | | *N/A.* | | | | |