The Computational Linguistics Summarization Pilot Task

Kokil Jaidka
Wee Kim Wee School of
Communication & Information
Nanyang Technological University
koki0001@e.ntu.edu.sg

Muthu Kumar Chandrasekaran
Dept. of Computer Science
School of Computing
National University of Singapore
muthu.chandra@comp.nus.edu.sg

Min-Yen Kan
Dept. of Computer Science
School of Computing
National University of Singapore
kanmy@comp.nus.edu.sg

Ankur Khanna
Web, IR NLP Group
School of Computing
National University of Singapore
khanna89ankur@gmail.com

Abstract

1 Introduction

This paper describes the evolution and design of the SciSumm Shared Task for the scientific summarisation of computational linguistics research papers. It was concurrently publicized with the upcoming TAC 2014, although it is not formally affiliated with the same, and shares its basic structure and guidelines with the more formal BiomedSumm track of TAC 2014. A development corpus of training "topics" from computational linguistics (CL) research papers was released, each comprising a main, cited paper alongwith associated citing papers. Participants were invited to enter their systems in a task-based evaluation, similar to the one announced by BioMedSumm.

This paper will describe the participating systems and survey their results from the task-based evaluation.

2 Background

Recent works (Mohammad et al., 2009)(Abu-Jbara and Radev, 2011) in scientific document summarisation have used citation sentences or citances from citing papers to create a multi document summary of the reference paper (RP). The computational linguistics (CL) community uses the ACL Anthology Reference Corpus (Bird et al., 2008) to evaluate and report performance of such systems. To support further research in this direction we built a manually annotated corpus of 10 randomly sampled documents from the ACL anthology reference corpus.

As proposed by (VU, 2010), (Hoang and Kan, 2010) the summarisation can be decomposed into finding the relevant documents, in this case, the citing papers (CPs), then selecting sentences from those papers that cite and justify the citation and finally generate the summary. To help tackle each of these subproblems, we created gold standard datasets where human annotators identify the citances in each of about 10 randomly sampled citing papers for the RP.

Given a reference paper and up to 10 citing papers, annotators from National University of Singapore and Nanyang Technological University were instructed to find citations to the RP in the 10 CPs. Annotators followed instructions used for annotation of corpus for the TAC 2014 Biomedical Summarisation task (BiomedSumm) to encourage cross participation across the two tasks. Specifically, the citation text, citation marker, reference text, and discourse facet were marked for each citation of the RP found in the CP.

A pilot study conducted in the information science domain indicated that most citations clearly refer to one or more specific aspects of the cited paper (Jaidka et al., 2013). For computational linguistics, we identified that the discourse facets being cited were usually the aim of the paper, methods followed and the results or implications of the work. Accordingly, we used a different set of discourse facets than BiomedSumm which suit CL papers better.

Please note that this is a development corpus and only a training set is available for use now. Although, we plan to release a test set of documents for next years evaluation, we plan to report k fold cross-validated performance over the 10 documents for the two systems registered for participation.

3 The Task

In this task, we explore a new form of structured summary: a faceted summary of the traditional self-summary (the abstract) and the community summary (the collection of citances). As a third component, we propose to group the citances by the facets of the text that they refer to. We propose that by identifying first, the cited text span, and second, the facet of the paper (Aim, Method, Result or Implication), we can create a faceted summary of the paper by clustering all cited/citing sentences together by facet.

The SciSumm Shared Task is defined as follows:

Given: A topic consisting of a Reference Paper (RP) and upto ten Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) have been identified that pertain to a particular citation to the RP.

Task 1a: For each citance, identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These are of the granularity of a sentence fragment, a full sentence, or several consecutive sentences (no more than 5).

Task 1b: For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets.

Evaluation: Task 1 will be scored by overlap of text spans in the system output vs the gold standard created by human annotators.

4 Participating teams

The following teams have expressed an interest in participating, and may be submitting their findings in this paper:

- Taln.UPF, from Universitat Pompeu Fabra, Spain. They have proposed to adapt available summarisation tools to scientific texts.

- Clair_UMICH from University of Michigan, Ann Arbor, USA.

- IITKGP_sum, from Indian Institute of Technology, Kharagpur, India. They plan to use citation network structure and citation context analysis to summarise the scientific articles.

- CCS2014, from the IDA Center for Computing Sciences, USA. They will employ a language model based on the sections of the document to find referring text and related sentences in the cited document.

- TabiBoun14, from the Boazii University, Turkey. They plan to modify an existing system for CL papers, wherein they use LIBSVM as a classification tool for face classification. They also plan to use the cosine similarity metric to compare text spans.

- PolyAF, from The Hong Kong Polytechnic University.

- TXSUMM, from University of Houston, Texas. Their system consists of applying similarity kernels in an attempt to better discriminate between candidate text spans (with sentence granularity). They are using an extractive procedure with ranking algorithms.

- MQ, from Macquarie University, Australia. They plan to use the same system that was used for the BiomedSumm track, with the exception that they will not incorporate domain knowledge (UMLS). For task 1a they will use similarity metrics to extract the top n sentences from the documents. For task 1b we will use a logistic regression classifier. For task 2 we will incorporate the distances from task 1 to rank the sentences.

- A team from IHMC, USA

Acknowledgments

Do not number the acknowledgment section. Do not include this section when submitting your paper for review.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 500–509. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 427–435. Association for Computational Linguistics.

Kokil Jaidka, Christopher SG Khoo, and Jin-Cheon Na. 2013. Literature review writing: how information is selected and transformed. In *Aslib Proceedings*, volume 65, pages 303–325. Emerald Group Publishing Limited.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 584–592. Association for Computational Linguistics.

HOANG CONG DUY VU. 2010. Towards automated related work summarization.