Introduction to the Computation Linguistics Shared Task

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

We describe the Computational Linguistics Shared Task: citation-based summarisation of computational linguistics research papers. We give background information on the data set (ACL research papers) and the evaluation method, present a general overview of the systems that have taken part in the task and discuss their preliminary results.

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 [1][2] 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 [3] 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 Hoang et al. [4] 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 sub problems, 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 [5]. 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$_{UMICH from University of Michigan, Ann Arbor, USA}$ (?) previously showed ...
- 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.
- A team from IHMC, USA

4.1 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on US-letter paper are:

Citations: Citations within the text appear in parentheses as (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved. Collapse multiple citations as in (**?**; **?**). Also refrain from using full citations as sentence constituents. We suggest that instead of

"(**?**) showed that ..."

you use

"Gusfield (**?**) showed that ..."

If you are using the provided LaTeX and BibTeX style files, you can use the command `\newcite` to get "author (year)" citations.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g.,

"We previously showed (**?**) ..."

should be avoided. Instead, use citations such as

"Gusfield (**?**) previously showed ..."

Please do not use anonymous citations and do not include acknowledgements when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (**?**). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (**?**).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

5 Length of Submission

Long papers may consist of up to nine (9) pages of content and an unlimited number of reference pages, and short papers may consists of up to five (5) pages of content and an unlimited number of reference pages. Papers that do not conform to the specified length and formatting requirements are subject to resubmission.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.