Overview of the CL-SciSumm 2016 Shared Task

Kokil Jaidka[1], Muthu Kumar Chandrasekaran[2], Sajal Rustagi[3], and Min-Yen Kan[2,4]

[1] Big Data Experience Lab, Adobe Research India
[2] School of Computing, National University of Singapore, Singapore
[3] Dept. of Computer Science and Engineering, Indian Institute of Technology, Roorkee, India
[4] Interactive and Digital Media Institute, National University of Singapore, Singapore
kokil@pmail.ntu.edu.sg

Abstract. The CL-SciSumm 2016 Shared Task is the first medium-scale shared task on scientific document summarization in the computational linguistics (CL) domain. The task built off of the experience and training data set created in its namesake pilot task, which was conducted in 2014 by the same organizing committee. The track included three tasks involving: (1A) identifying relationships between citing documents and the referred document, (1B) classifying the discourse facets, and (2) generating the abstractive summary. The dataset comprised 30 annotated sets of citing and reference papers from the open access research papers in the CL domain. This overview paper describes the participation and the official results of the second CL-SciSumm Shared Task, organized as a part of the Joint Workshop onBibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016), held in New Jersey,USA in June, 2016. The annotated dataset used for this shared task and the scripts used for evaluation can be accessed and used by the community at: https://github.com/WING-NUS/scisumm-corpus.

1 Introduction

The CL-SciSumm task provides resources to encourage research in a promising direction of scientific paper summarization, which considers the set of citation sentences (i.e., "citances") that reference a specific paper as a (community created) summary of a topic or paper [21]. Citances for a reference paper are considered a synopses of its key points and also its key contributions and importance within an academic community [19]. The advantage of using citances is that they are embedded with meta-commentary and offer a contextual, interpretative layer to the cited text. The drawback, however, is that though a collection of citances offers a view of the cited paper, it does not consider the context of the target user [9] [24], verify the claim of the citation or provide context from the reference paper, in terms the type of information cited or where it is in the referenced paper [8].

CL-SciSumm explores summarization of scientific research, for the computational linguistics research domain. It encourages the incorporation of new kinds

of information in automatic scientific paper summarization, such as the facets of research information being summarized the research paper. Our previous task suggested that scholars in CL typically cite methods information from other papers. CL-SciSumm also encourages the use of citing mini-summaries written in other papers, by other scholars, when they refer to the paper. It is anticipated that these selected facts would closely reflect the most important contributions and applications of the paper. These insights have been explored in a smaller scope by previous work. We propose that further explorations can help to advance the state of the art. Furthermore, we expect that the CL-SciSumm Task could spur the creation of new resources and tools, to automate the synthesis and updating of automatic summaries of CL research papers.

Previous work in scientific summarization has attempted to automatically generate multi-document summaries by instantiating a hierarchical topic tree[6], generating model citation sentences[17] or implementing a literature review framework[8]. However, the limited availability of evaluation resources and human-created summaries constrains research in this area. In 2014, the CL-SciSumm Pilot task was conducted as a part of the larger BioMedSumm Task at TAC [5]. In 2016, our proposal was not successful with ACL; fortunately it was accepted as a part of the BIRNDL workshop [15] at JCDL-2016[6].

The development and dissemination of the CL-SciSumm dataset and the related Shared Task has been generously supported by the Microsoft Research Asia (MSRA) Research Grant 2016.

2 Task

Given: A topic consisting of a Reference Paper (RP) and up to 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.

Task 2: Finally, 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. This was an optional bonus task.

Evaluation: Participants were required to submit their system outputs from the test set to the task organizers. An automatic evaluation script was used to measure system performance for Task 1a, in terms of the sentence id overlaps between the sentences identified in system output, versus the gold standard created by human annotators. Task 1b was evaluated as a proportion of the

[5] http://www.nist.gov/tac/2014
[6] http://www.jcdl.org

correctly classified discourse facets by the system, contingent on the expected response of Task 1a. Task 2 was optional, and evaluated using the ROUGE-N [12] scores between the system output and three types of gold standard summaries of the research paper.
Data: The dataset comprises ten pairs of training sets, development and test sets. Each pair comprises the annotated citing sentences for a research paper and the discourse facets being referenced, and summaries of the research paper.

3 CL-SciSumm Pilot 2014

The CL Summarization Pilot Task [7] was conducted as a part of the BiomedSumm Track at the Text Analysis Conference 2014 (TAC 2014) [7]. Ten pairs of annotated citing sentences and summaries were made available to the participants, who reported their performance on the same Tasks described above, as a cross-validation over the same dataset. System outputs for Task 1a were scored using word overlaps with the gold standard measured by the ROUGE–L score. Task 1b was scored using precision, recall and F_1. Task 2 was an optional task where system summaries were evaluated against the abstract using ROUGE–L. No centralized evaluation was performed. All scores were self-reported.

Three teams submitted their system outputs. clair_umich was a supervised system using lexical, syntactic and WordNet based features; MQ system used information retrieval inspired ranking methods; TALN.UPF used various TF-IDF scores.

During this task, the participants reported several errors in the dataset including text encoding and inconsistencies in the text offsets. The annotators also reported flaws in the xml encoding, and problems in the OCR export to XML. These issues hindered system building and evaluation. Accordingly, changes were made to the annotation file format and the XML transformation process in the current task.

4 Development

The CL-SciSumm 2016 task included the original training dataset of the Pilot Task, to encourage teams from the previous edition to participate. It also incorporated a new development corpus of ten sets for system training, and a separate test corpus of ten sets for evaluation. Additionally, it provided three types of summaries for each set in each corpus -

- the abstract, written by the authors of the research paper
- the community summary, collated from the reference spans of its citances
- human-written summary written by the annotators of the CL-SciSumm annotation effort

[7] http://www.nist.gov/tac/2014

For the general procedure followed to construct the CL-SciSumm corpus, please see [7]. There are two differences in the selection of citing papers (CP) for the training corpus, as compared to the development and test corpora. Firstly, the minimum numbers of CP provided in the former, which was 3, was increased to 8 in the construction of the latter. Secondly, the maximum number of CPs provided in the former was 10, but this limit was removed in the construction of the latter, so that up to 60 CPs have been provided for a single RP. This was done to have more citances of which potentially more would mention the RP in greater detail. This would also produce a wider perspective in the community summary.

4.1 Annotation

The annotators of the development and test corpora were five postgraduate students in Applied Linguistics, from University of Hyderabad, India. They were selected out of a larger pool of over twenty-five participants, who were all trained to annotate an RP and its CPs on their personal laptops, using the Knowtator[8] annotation package of the Protege editing environment[9].

The annotation scheme was unchanged from what was followed by [7]: Given each RP and its associated CPs, the annotation group was instructed to find citations to the RP in each CP. Specifically, the citation text, citation marker, reference text, and discourse facet were identified for each citation of the RP found in the CP. Inadvertently, we included the gold standard annotations for Task 1a and 1b when we released the test corpus. We alerted the participating teams to this mistake and requested them not to use that information for training their systems.

5 Overview of Approaches

The following paragraphs discuss the approaches followed by the participating systems, in no particular order. Except for the top performing systems in each of the sub-tasks, we do not provide detailed relative performance information for each system, in this paper. The evaluation scripts have been provided at the CL- SciSumm Github respository [10] where the participants may run their own evaluation and report the results.

The approach by [14] used the Transdisciplinary Scientific Lexicon (TSL) developed by [5] to build a profile for each discourse facet in citances and reference spans. Then a similarity function developed by [16] was used to select the best-matching reference span with the same facet as the citance. For Task 2, the authors used Maximal Marginal Relevance [3] to choose sentences so that they brought new information to the summary.

[8] http://knowtator.sourceforge.net/
[9] http://protege.stanford.edu/about.php
[10] github.com/WING-NUS/scisumm-corpus

Nomoto [20] proposed a hybrid model for Task 2, comprising TFIDF and a tripartite neural network. Stochastic gradient descent was performed on a training data comprising of triples of citance, the true reference and the set of false references for the citance. Sentence selection was based on a dissimilarity score similar to MMR.

Mao *et al.* [11] used an SVM classifier with a topical lexicon to identify the best matching reference spans for a citance, using ifd similarity, Jaccard similarity and context similarity. They finally submitted six system runs, each following a variant of similarities and approaches - the fusion method, the Jaccard Cascade method, the Jaccard Focused method, the SVM method and two voting methods.

Klampfl *et al.* [10] developed three different approaches based on summarization and classification techniques. They applied a modified version of an unsupervised summarization technique, termed it TextSentenceRank, to the reference document. Their second method incorporates similarities of sentences to the citation on a textual level, and employed classification to select from candidates previously extracted through the original TextSentenceRank algorithm. Their third method used unsupervised summarization of the relevant sub-part of the document that was previously selected in a supervised manner.

Saggion *et al.* [23] reported their results for the linear regression implementation of WEKA used together with the GATE system. They trained their model to learn the weights of different features with respect to the relevance of cited text spans and the relevance to a community-based summary. Two runs were submitted, using SUMMA [22] to score and extract all matched sentences and only the top sentences respectively.

Lu *et al.* [13] regarded Task 1a as a ranking problem, applying Learning to Rank strategies. In contrast, the group cast Task 1b as a standard text classification problem, where novel feature engineering was the team's focus. Along this vein, the group considered features of both citation contexts and cited spans.

Aggarwal and Sharma [1] propose several heuristics derived from bigram overlap counts between citances and reference text to identify the reference text span for each citance. This score is used to rank and select sentences from the reference text as output.

Baki *et al.* [18] used SVM with subset tree kernel, a type of convolution kernel. Computed similarities between three tree representations of the citance and reference text formed the convolution kernel. Their set-up scored better than their TF-IDF baseline method. They submitted three system runs with this approach.

The PolyU system [2], for Task 1a, use SVM-rank with lexical and document structural features to rank reference text sentences for every citance. Task 1b is solved using a decision tree classifier. Finally, they model summarization as a query–focussed summarization with citances as queries. They generate summaries (Task 2) by improvising on a Manifold Ranking method (see [2] for details).

Finally, the system submitted by Conroy and Davis [4] attempted to solve Task 2 with an adaptation of a system developed for the TAC 2014 BioMedSumm Task [11]. They provided the results from a simple vector space model, wherein they used a TF representation of the text and non- negative matrix factorization (NNMF) to estimate the latent weights of the terms for scientific document summarization. They also provide the results from two language models based on the distribution of words in human-written summaries.

6 System Runs

Performance of systems for Task 1a was measured by the number of sentences output by the system that overlap with the sentences in the human annotated reference text span (see section 4.1). These numbers were then used to calculate the precision, recall and F_1 score for each system. As Task 1b is a multi-label classification, this task was also scored by metrics - precision, recall and F_1 score.

Nine systems submitted outputs for Task 1. The following plots rank the systems for Task 1 by their F_1 scores. In the figures, all the systems have been identified by their participant number. Only the top performing systems for Tasks 1a, 1b and 2 have been identified by name in sections 6 and 7.

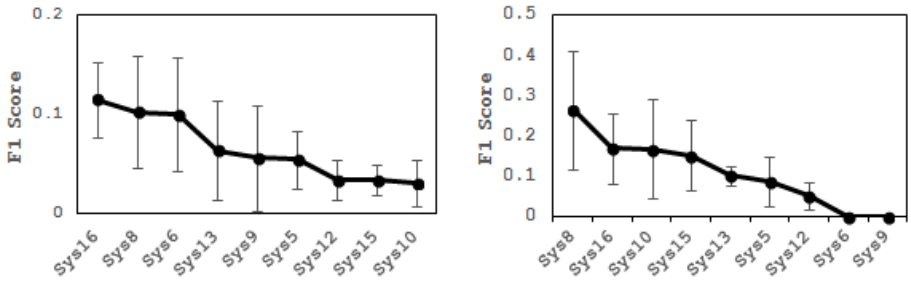

Fig. 1. System performances on Task 1a(left) and Task 1b(right).

Task 2, to create a summary of the reference paper from citances and the reference paper text, was evaluated against 3 types of gold standard summaries: the reference paper's abstract, a community summary and a human summary. A Java Implementation of ROUGE[12] was used to compare the gold summaries against summaries generated by systems. We calculated ROUGE–2 and ROUGE–4 F_1 scores for the system summaries against each of the 3 summary types. ROUGE–1 and ROUGE–3, which showed similar results have been omitted from this paper.

[11] http://www.nist.gov/tac/2014/BiomedSumm
[12] http://kavita-ganesan.com/content/rouge-2.0

Four of the nine system that did Task 1 also did the bonus Task 2. Following are the plots with their performance measured by ROUGE–2 and ROUGE–4 against the 3 gold standard summary types.

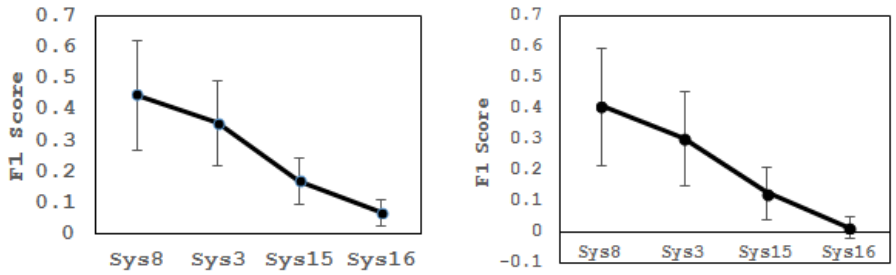

Fig. 2. Task 2 system performances on abstract summaries measured by ROUGE–2 (left) and ROUGE–4 (right)

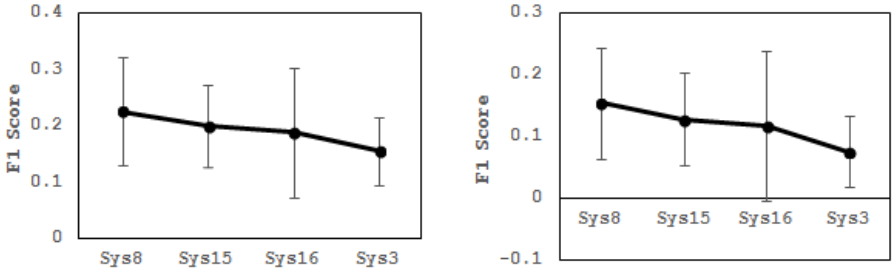

Fig. 3. Task 2 system performances on community summaries measured by ROUGE–2 (left) and ROUGE–4 (right)

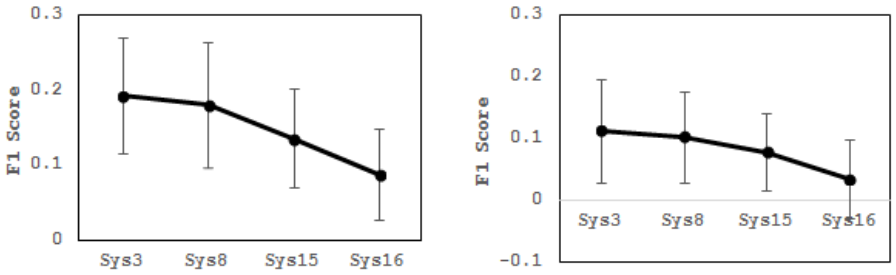

Fig. 4. Task 2 system performances on human summaries measured by ROUGE 2 (left) and 4 (right)

For Task 1a, the best performance was shown by *sys16*, developed by [2]. The next best performance was shown by *sys8* [11] and *sys6* [20].

For Task 1b, the best performance was shown by *sys8* [11], followed by the systems *sys16*[2] and *sys10*[23].

For Task 2, the system by [11], *sys8*, performed the best against abstract and community summaries, while *sys16* [2] performed well on community summaries. The system by *sys15* [1], was also a strong performer on these tasks. On human summaries, the best performance was seen from *sys3* [4].

The F_1 scores of all systems on Tasks 1a and 1b were generally low. However, the systems ranked in the first 3 places, did significantly better than systems ranked in the last 3 places.

On Task 2, all systems except *sys16* performed better when evaluated against abstracts, than against other summary types. Furthermore, system performances did not differ significantly from one another when evaluated against human and community summaries. However, when evaluated against abstracts, the best performing system significantly outperforms systems ranked in the lower half.

7 Conclusion

Ten systems participated in the CL-SciSumm Task 2016. A variety of heuristical, lexical and supervised approaches were used. Two of the best performing systems in Task 1a and 1b were also participants in the CL-SciSumm Pilot Task. The results from Task 2 suggest that automatic summarization systems may be adaptable to different domains, as we observed that the system by [4], which had originally been developed for biomedical human summaries, outperformed the others. We also note that systems performing well on Tasks 1a and 1b also do well in generating community summaries - this supports our expectations about the Shared Task, and validates the need to push the state-of-the-art in scientific summarization. In future work, other methods of evaluation can be used for comparing the performance of the different approaches, and a deeper analysis can lead to new insights about which approaches work well with certain kinds of data. However, such an inquiry was beyond the scope of this overview paper. We deem our Task a success, as it has spurred the interest of the community and the development of tools and approaches for scientific summarization. We are investigating other potential subtasks which could be added into our purview. We are also scouting for other related research problems, of relevance to the scientific summarization community.

Acknowledgement. The organizers of the CL-SciSumm16 shared task would like to thank Microsoft Research Asia, for their generous funding. We would also like to thank Vasudeva Varma and colleagues at IIIT-Hyderabad, India and University of Hyderabad, India for their efforts in convening and organizing our annotation workshops. We acknowledge the continued advice of Hoa Dang, Lucy Vanderwende and Anita de Waard from the pilot stage of this task and thank them for the same. We thank Rahul Jha and Dragomir Radev for sharing their software to prepare the XML versions of papers. We are grateful to Kevin B. Cohen and colleagues for their support, and for sharing their annotation schema, export scripts and the Knowtator package implementation on the Protege software - all of which have been indispensable for this Shared Task.

References

1. Aggarwal, P., Sharma, R.: Lexical and Syntactic cues to identify Reference Scope of Citance. 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 (June 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 (June 2016)
3. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 335–336. Association of Computational Linguistics (1998)
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)
5. Drouin, P.: Extracting a bilingual transdisciplinary scientific lexicon. In: eLexicography in the 21st century: new challenges, new applications. pp. 43–53. Louvain-la-Neuve: Presses Universitaires de Louvain (2010)
6. Hoang, C., Kan, M.: Towards automated related work summarization. In: Proc. of COLING: Posters. pp. 427–435. ACL (2010)
7. Jaidka, K., Chandrasekaran, M.K., Elizalde, B.F., Jha, R., Jones, C., Kan, M.Y., Khanna, A., Molla-Aliod, D., Radev, D.R., Ronzano, F., et al.: The Computational Linguistics Summarization Pilot Task. In: Proceedings of Text Analysis Conference. Gaithersburg, USA (2014)
8. Jaidka, K., Khoo, C.S., Na, J.C.: Deconstructing human literature reviews–a framework for multi-document summarization. In: Proc. of ENLG. pp. 125–135 (2013)
9. Jones, K.S.: Automatic summarising: The state of the art. Information Processing and Management 43(6), 1449–1481 (2007)
10. Klampfl, 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 (June 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 (June 2016)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL-04 workshop 8 (2004)
13. 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 (June 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 (June 2016)
15. Mayr, P., Frommholz, I., Cabanac, G., Wolfram, D.: Editorial for the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) at JCDL 2016. In: Proc. of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016). pp. 1–5. Newark, NJ, USA (June 2016)
16. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: 21st national conference on Artificial Intelligence. pp. 775–780. AAAI (2006)
17. Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D.R., Zajic, D.: Using citations to generate surveys of scientific paradigms. In: Proc. of NAACL. pp. 584–592. ACL (2009)
18. 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 (June 2016)
19. Nakov, P.I., Schwartz, A.S., Hearst, M.: Citances: Citation sentences for semantic analysis of bioscience text. In: Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics. pp. 81–88 (2004)
20. 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 (June 2016)
21. Qazvinian, V., Radev, D.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 689–696. ACL (2008)
22. Saggion, H.: SUMMA: A Robust and Adaptable Summarization Tool. Traitement Automatique des Langues 49(2), 103–125 (2002)
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 (June 2016)
24. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics 28(4), 4099–445 (2002)