Insights from CL-SciSumm 2016: the faceted scientific document summarization Shared Task

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi & Min-Yen Kan

International Journal on Digital Libraries

ISSN 1432-5012

Int J Digit Libr
DOI 10.1007/s00799-017-0221-y

Springer

Int J Digit Libr
DOI 10.1007/s00799-017-0221-y

Insights from CL-SciSumm 2016: the faceted scientific document summarization Shared Task

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

Received: 14 November 2016 / Revised: 31 May 2017 / Accepted: 6 June 2017

Abstract We describe the participation and the official results of the 2nd Computational Linguistics Scientific Summarization Shared Task (CL-SciSumm), held as a part of the BIRNDL workshop at the Joint Conference for Digital Libraries 2016 in Newark, New Jersey. CL-SciSumm is the first medium-scale Shared Task on scientific document summarization in the computational linguistics (CL) domain. Participants were provided a training corpus of 30 topics, each comprising of a reference paper (RP) and 10 or more citing papers, all of which cite the RP. For each citation, the text spans (i.e., citances) that pertain to the RP have been identified. Participants solved three sub-tasks in automatic research paper summarization using this text corpus. Fifteen teams from six countries registered for the Shared Task, of which ten teams ultimately submitted and presented their results. The annotated corpus comprised 30 target papers—currently the largest available corpora of its kind. The corpus is available for free download and use at https://github.com/WING-NUS/scisumm-corpus.

Keywords Summarization · Automated literature review · Scientific document summarization · Computational linguistics

✉ Kokil Jaidka
jaidka@sas.upenn.edu

1 School of Arts and Sciences, University of Pennsylvania, Pennsylvania, USA

2 School of Computing, National University of Singapore, Singapore, Singapore

3 Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee, India

4 Smart Systems Institute, National University of Singapore, Singapore, Singapore

1 Introduction

The CL-SciSumm task provides resources and benchmark tasks to encourage research on 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 comprise a synopsis of its key points and key contributions, and can be a surrogate measure of importance within an academic community [19]. An advantage of using citances is that they are embedded with meta-commentary and offer a contextual, interpretative layer to the cited text. In contrast, traditional automatic summarization approaches, which focus on salience and offer a gist of the paper, do not consider its academic context [9,25], or the context of the reader. They do not exploit community feedback from other citing papers, which serve to verify the claim of a paper and highlight certain facets over others [8].

CL-SciSumm explores the summarization of scientific research in the domain of computational linguistics. 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 response to this need, we introduced the CL-SciSumm pilot task in 2014, as a part of the larger BioMedSumm task at TAC,[1] to encourage the incorporation of new kinds of information in automatic scientific paper summarization, such as the facets of research information being summarized in the research paper. CL-SciSumm encourages the use of citing mini-summaries embedded in

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

Published online: 14 June 2017

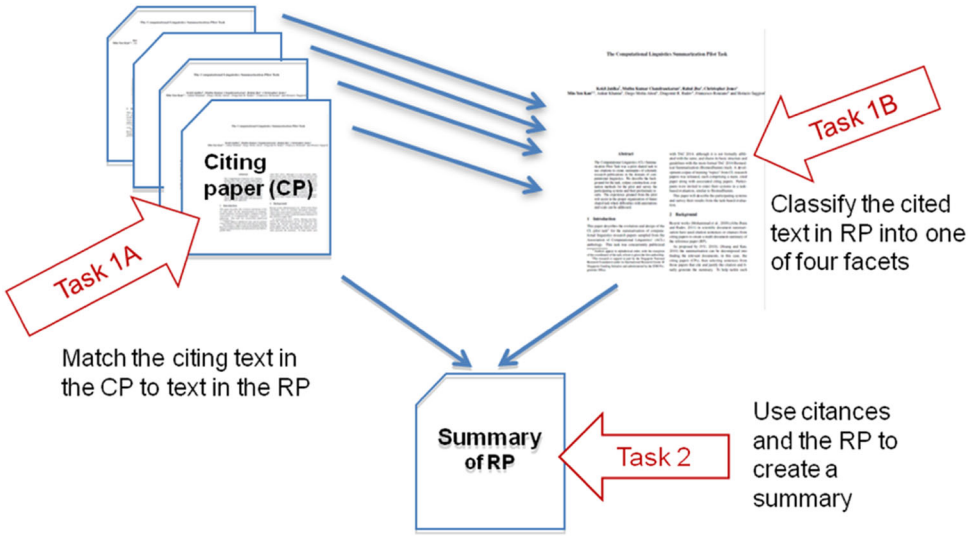

Fig. 1 Illustration of tasks: Task 1A, Task 1B and Task 2

other papers, written by other scholars, when they refer to the paper, which are expected to reflect the most important contributions and applications of the paper. Few studies have explored citation-based approaches in summarization (e.g., [21]). Through the CL-SciSumm task, we aim to spur the creation of new resources and tools to automate the synthesis and updating of automatic summaries of CL research papers, and therefore help to advance the overall state of the art in scientific summarization.

2 Task

The CL-SciSumm Shared Task comprised three sub-tasks (Fig. 1) in automatic research paper summarization on a text corpus. The development corpus was an extended version of the dataset used by the CL pilot task at the Text Analysis Conference 2014 (TAC 2014) [7]. Participants were required to develop approaches to solve some or all of the three sub-tasks, and submit the system outputs from the test set to the task organizers.

Given: A topic consisting of a reference paper (RP) and up to ten citing papers (CPs) that all contain citations to the RP. Citations in the CP are preidentified as the text spans (i.e., citances) that cite the RP.

Task 1A: For each citance, identify the spans of text (reference 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 reference text span, identify what discourse facet of the paper it belongs to, from a predefined set of 4 discourse facets: aim, method, results and implication. Discourse facets describe the type of information in the reference text span.

Task 2: Generate a structured summary of the RP from the reference text spans of the RP. The length of the summary should not exceed 250 words. This was an optional bonus task.

Evaluation. We used an automatic evaluation script to measure system performance for Task 1A, in terms of the overlaps in sentence ID marked by those identified in system output against the gold standard created by human annotators. We evaluated Task 1B as a proportion of the 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. Recall-Oriented Understudy for Gisting Evaluation is a set of metrics to automatically evaluate summarization systems [12]. ROUGE measures the overlap between computer-generated summaries and multiple human-written reference summaries. In previous studies, ROUGE scores have significantly correlated with human judgments on summary quality [13]. Different variants of ROUGE differ according to the granularity at which overlap is calculated. For instance, ROUGE-2 measures the bigram overlap between the candidate computer-generated summary and the reference summaries. More generally, ROUGE-N measures the n-gram overlap. ROUGE-L measures the overlap in longest common subsequence (LCS). ROUGE-S measures overlaps in skip bigrams or bigrams with arbitrary gaps in between. ROUGE-SU uses skip bigram plus unigram overlaps. CL-SciSumm 2016 uses ROUGE-2 and ROUGE-SU4 for evaluating Task 1A and Task 2. CL-SciSumm pilot used only ROUGE-L for evaluation.

Data: The CL-SciSumm 2016 dataset comprised ten topic pairs each, in the training set, development and test set. Each topic pair comprises a reference paper and the citing

information which pairs with it—i.e., the citances, discourse facets and summaries of the reference paper. Inadvertently, we included the gold standard annotations for Tasks 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 in training their systems.

3 CL-SciSumm pilot 2014

In a previous iteration of this task, we conducted the CL summarization pilot task [7] as a part of the BioMedSumm task at the Text Analysis Conference 2014 (TAC 2014).[2] The dataset comprised only ten topic pairs. Participants 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; and `TALN.UPF` used various $tf.idf$ scores. These systems are described in detail in Jaidka et al. [7].

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 with the OCR export to XML.[3] These issues hindered system building and evaluation in the pilot. These informed our current iteration, and accordingly, we changed the annotation file format and the XML transformation process in the current task.

4 Corpus development

The CL-SciSumm 2016 task included the original training dataset of the pilot task as the development corpus, with the aim of encouraging teams from the previous edition to participate. We augmented the development corpus with ten additional sets for system training and a separate, new 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 summaries, by the annotators of the CL-SciSumm annotation effort

We followed the general procedure of the CL-SciSumm pilot task to construct the enlarged CL-SciSumm corpus (for details 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 number of CPs provided in the former corpus, which was 8, was increased to 10 in the construction of the latter corpus. Secondly, the maximum number of CPs provided in the former was 10, but this limit was later raised to 40 CPs, in order to have many more citances which could a. capture diverse facts about the reference paper and b. avoid overfitting in supervised models.

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`[4] annotation package of the `Protege` editing environment.[5]

We followed the previous annotation scheme unchanged from the previous pilot task, which was: 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.

4.2 Dataset

The final CL-SciSumm dataset comprised 30 RPs separated into three sets of 10 documents each: training, development and test sets. The dataset comprises approximately 6700 reference sentences and 750 citances (Table 1). We refer to the sparsity statistics calculated by Moraes et al. [18] who identified that the total size of the dataset comprises 23,356 unique words among the reference documents in the dataset and 5520 unique words in the citances. Table 2 identifies the sections of the reference papers, which were being cited, in the order of their popularity. These were identified by traversing the XML structure of the reference paper and identifying the parent section for each referenced sentence.

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

[3] The text of the documents was extracted from the original PDF documents; an optical character recognition (OCR) system was applied.

[4] http://knowtator.sourceforge.net/.

[5] http://protege.stanford.edu/about.php.

Table 1 Statistics of the CL-SciSumm corpus

Description	Count
Number of documents	506
Number of reference documents	30
Number of citation documents	702
Average citing documents for each reference	15.9
Average gold summary length (words)	134.2
Stdev gold summary length	27.9
Sentence length in citances (words)	34
Sentence length in reference spans (words)	22
Results	2.8

Table 2 Proportion of sections cited in the CL-SciSumm dataset

Paper section	Proportion of citations
Method	32.5
Introduction	28.7
Conclusion	8.5
Evaluation	5.3
Abstract	5.2
Discussion	5.2
Research objectives	4.1
Related work	3.9
Experiments	3.5
Results	2.8

5 Overview of approaches

We discuss the approaches followed by the participating systems, in no particular order. Except for the top performing systems in each sub-task, we do not provide detailed relative performance information for each system in this paper. The evaluation scripts are available for free download and use in CL-SciSumm Github repository,[6] to enable interested parties inclusive of participants to run their own evaluation.

Malenfant et al. [15] 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, to ensure new information was actively being added to the summary (Table 3). Nomoto [20] proposed a hybrid model for Task 2, comprising of $tf.idf$ and a tripartite neural network. His system performed stochastic gradient descent 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 maximal marginal Reference [3].

Li et al. [11] used an SVM classifier with a topical lexicon to identify the best-matching reference spans for a citance, using inverse df (document frequency) similarity, Jaccard similarity and context similarity. They finally submitted six system runs, each following a variant of similarity measures and approaches: fusion, Jaccard Cascade, Jaccard Focused, SVM and two other 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 *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. [24] 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 reference text spans and the relevance to a community-based summary. Two runs were submitted, using `SUMMA` [23] to score and extract all matched sentences and only the top sentences, respectively.

Lu et al. [14] cast Task 1A as a ranking problem, applying Learning to Rank strategies. In contrast, they treat Task 1B as a standard text classification problem and focussed on novel feature engineering. Along this vein, they considered features of both citation contexts and cited spans.

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

Moraes et al. [18] used SVM with the 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 setup scored better than their $tf.idf$ baseline. They submitted three system runs with this approach.

Cao et al. [2], for Task 1A, use SVM rank with lexical and document structural features to rank reference text sentences for every citance. Task 1B was tackled using a decision tree classifier. They modeled summarization as a query-focused summarization task with citances as queries. They generate summaries (Task 2) by improvising on a manifold ranking method.

Conroy and Davis [4] attempted to solve Task 2 with an adaptation of a system developed for the TAC 2014

[6] https://github.com/WING-NUS/scisumm-corpus/tree/master/evaluation_scripts.

Table 3 System ID prefixes mapped to system description papers

System id	*sys3*	*sys5*	*sys6*	*sys8*	*sys9*	*sys10*	*sys12*	*sys13*	*sys15*	*sys16*
System paper	[4]	[15]	[20]	[11]	[10]	[24]	[14]	[1]	[18]	[2]

Table 4 All systems from each submission ranked by their performance in Tasks 1A and 1B. NS indicates that no submission was received from the system on the task

Task 1A		Task 1B	
System id	F_1 score	System id	F_1 score
sys15$tfidf+st+sl	0.134	*sys15$tfidf+st+sl*	0.399
sys8$Fusion	0.126	*sys8$Jaccard Focused*	0.317
sys8$Jaccard Focused	0.126	*sys8$Jaccard Fusion*	0.300
sys8$Voting1	0.116	*sys8$Voting1*	0.295
sys8$Voting2	0.108	*sys8$Voting2*	0.274
sys6$Default	0.096	*sys8$Jaccard Cascade*	0.257
sys8$Jaccard Cascade	0.095	*sys8$SVM*	0.155
sys16$Default	0.094	*sys16$Default*	0.153
sys9$Modified-tsr	0.051	*sys10$run1_one_line*	0.139
sys13$Default	0.047	*sys15$Tkern1-4*	0.073
sys8$SVM	0.042	*sys15$Tkern1-1*	0.069
sys5$Default	0.039	*sys10$run2_one_line*	0.066
sys10$run_1_line	0.023	*sys15$Tkern1-8ce*	0.064
sys12$Default	0.021	*sys5$Default*	0.064
sys15$Tkern1-8ce	0.018	*sys13$Default*	0.053
sys10$run_2_line	0.017	*sys15$Tkern1-1ce*	0.049
sys15$Tkern1-4	0.016	*sys15$Tkern1-4ce*	0.049
sys15$Tkern1-1	0.015	*sys15$Tkern1-8*	0.049
sys15$Tkern1-1ce	0.011	*sys12$Default*	0.011
sys15$Tkern1-4ce	0.011	*sys6*	NS
sys15$Tkern1-8	0.011	*sys9*	NS
sys9$tar-sent-class	0.009		
sys9$sect-class-tsr	0.008		

BioMedSumm task.[7] They provided the results from a simple vector space model, wherein they used a TF representation of the text and nonnegative 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

The 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 (Sect. 4.1). This was used to calculate precision, recall and F_1 score for each system. As Task 1B is a multi-label classification, this task was also scored by the same metrics of precision, recall and F_1 score.

Nine systems submitted outputs for Task 1A, and of these, seven submitted their results for Task 1B. Table 3 maps each output to the actual system description. Table 4 ranks all submitted system runs for Task 1A and Task 1B by their average $F1$ scores. All the system runs have been identified by a concatenated string of their system identification number and run ID (used by the authors in the systems' description paper) for the sake of convenience.

For the summarization task (Task 2), the ROUGE package [12] was used to compare the three types of gold summaries against the system generated summaries. We list the system results in Table 5 and plot them graphically (in Figs. 2, 3, 4), presenting ROUGE-2 and ROUGE-SU4 F_1 scores for the six systems that attempted Task 2 (ROUGE-1 and ROUGE-3 results showed similarly, and have omitted for succinctness.

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

Table 5 Systems' performance measured in ROUGE-SU4 on Task 2 (summarization), evaluated against the target paper's abstract, human summaries and community summaries. Systems' rank appears in parentheses

System id	Versus human summary	Versus community summary	Versus abstract
sys8$PARA_7	0.136 (1)	0.130 (7)	0.423 (1)
sys3$LMKL1_CCS1	0.124 (2)	0.095 (18)	0.179 (5)
sys3$LMEQUAL_CCS2	0.121 (3)	0.102 (15)	0.214 (4)
sys3$LMKL2_CCS3	0.114 (4)	0.095 (17)	0.158 (8)
sys8$PARA_1	0.112 (5)	0.129 (8)	0.247 (2)
sys8$PARA_8	0.111 (6)	0.150 (3)	0.244 (3)
sys3$TF_CCS4	0.101 (7)	0.085 (19)	0.129 (9)
sys8$PARA_0	0.099 (8)	0.137 (6)	0.177(6)
sys8$PARA_4	0.094 (9)	0.162 (2)	0.170(7)
sys10$AUTOMATIC	0.092 (10)	0.150 (3)	0.124 (10)
sys15$TKERN18	0.090 (11)	0.096 (16)	0.102 (15)
sys15$TFIDF+ST+SL	0.088 (12)	0.167 (1)	0.092 (18)
sys15$TKERN14CE	0.085 (13)	0.129 (8)	0.105 (13)
sys10$COMMUNITY	0.085 (14)	0.149 (5)	0.111 (11)
sys15$TKERN11CE	0.082 (15)	0.106 (12)	0.105 (14)
sys15$TKERN11	0.081 (16)	0.103 (13)	0.107 (12)
sys15$TKERN14	0.080 (17)	0.110 (10)	0.099 (16)
sys15$TKERN18CE	0.071 (18)	0.103 (14)	0.093 (17)
sys5$DEFAULT	0.065 (19)	0.082 (20)	0.087 (19)
sys16$DEFAULT	0.048 (20)	0.107 (11)	0.053 (20)

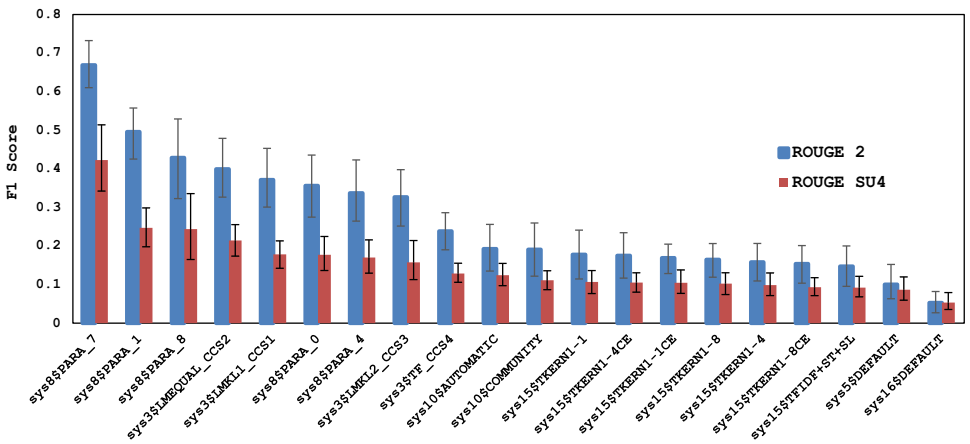

Fig. 2 Systems' performance on Task 2 versus *abstract* summaries sorted by their average ROUGE score as compared against the abstracts of the test set documents. *Bars* show standard deviation

In Task 1A, the $tf.idf$ baseline by *sys15* [18] performed best. Several runs of *sys8* [11] yielded the following top performance followed by *sys6* [20].
The best-performing system for Task 1B was *sys15* [18], followed by *sys8* [11], *sys16* [2] and *sys10* [24].

The systems ranked in the top 3 places did significantly better than systems ranked in the bottom 3 places. However, given the skewed nature of data for Task 1B, where most of the discourse facets were annotated as "methods," we were not able to establish any statistically significant differences in systems' performance from the Task 1B results.

For Task 2, the ROUGE-SU4 scores follow similar trends as the ROUGE-2 scores. The multiple runs submitted by *sys8* [11] performed the best on abstract summaries, with the highest ROUGE scores. Note that the names of the *sys8* runs for Task 2 are different from those in Task 1 as different approaches were applied for summary generation, as detailed in the authors' original paper. The next best performers were *sys3* [4] and *sys10* [24].

Performance comparison against community summaries identified *sys15* [18] as the best-performing system, and *sys8* [11] and *sys10* [24] were close behind.

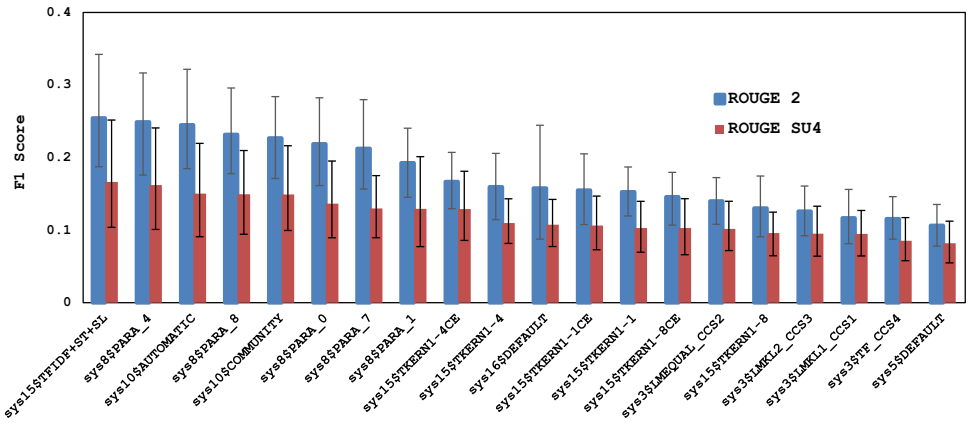

Fig. 3 Systems' performance on Task 2 versus *community* summaries sorted by their average ROUGE score over target summaries in the test set. *Bars* show standard deviation

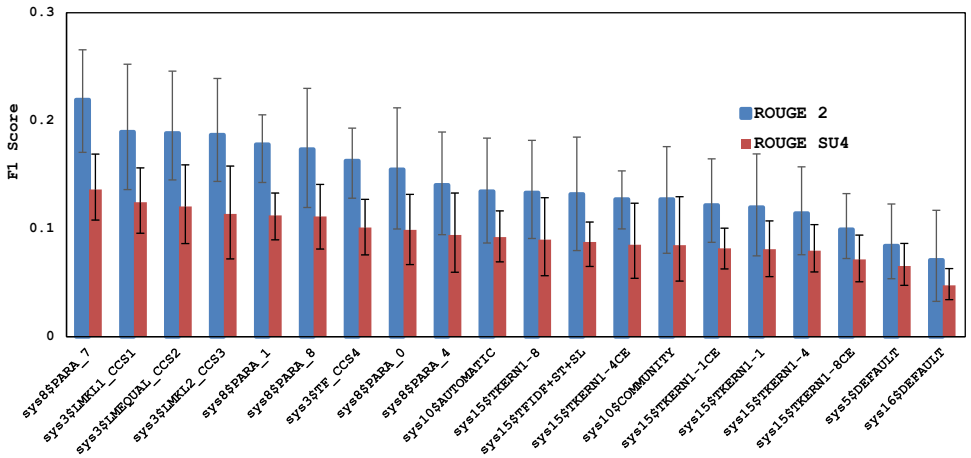

Fig. 4 Systems' performance on Task 2 versus *human* summaries, sorted by their average ROUGE score over target summaries in the test set. *Bars* show standard deviation

On human summaries, *sys8* [11] performed the best, followed closely by multiple runs from *sys3* [4].

Considering the implications of these findings, it is not surprising that the systems which performed well against abstract summaries are those which incorporated surface features, such as sentence position (*sys8* and *sys10*) and lexical features, such as term frequency (*sys3*). Both *sys8* and *sys3* performed well when compared against the abstract and the human summaries, while *sys15* and *sys10* did better on community summaries. When evaluated against abstracts, there is a prominent gap in the performance of the best-performing run of *sys8* and the next best performance.

Besides the overall success of lexical approaches, the successful runs from *sys3* suggest the potential of a probabilistic approach for scientific summarization and also indicate the generalizability of a summarization approach developed for one domain on another—in this case, the *sys3*'s language model was trained on biomedical research papers, but was applied to summarize computational linguistics research papers. The $tf.idf$-based approach of *sys15* performed better on the citation classification task; however, the kernel-based approach did moderately better in summary generation. The performance of *sys8*—against both abstract and human summaries—reiterates the importance of sentence position in the summarization task, followed by the sentence length and the span of cited text.

As a part of the development and maintenance of the CL-SciSumm Shared Task, we have considered the feedback from the participants in the task in order to improve the quality and usability of our dataset. In the following paragraphs, we provide an error analysis which will be used to improve the quality of the CL-SciSumm dataset in future iterations:

- Annotated citances included instances where citation markers were mapped to the paper title of the reference paper as the reference span, if no matching text or ideas were found. Although this was an annotation rule followed uniformly across all topic sets, it appears to lead to a drop in accuracy and will not be continued in subsequent tasks.
- As is often the case in abstractive summarization, citances often paraphrase facts from the reference paper; as a result, there may be no overlap of salient words between the citance and reference text span. In the future, we may consider incorporating new attributes to easily identify

such sentences, in order to facilitate sub-tasks related to paraphrasing.
- In the future, we may consider dropping ambiguous citances from our dataset. This refers to the large proportion of citances referencing general information about the paper, which may be mapped to its first mention, i.e., Introduction section of the reference paper (ref Table 2), but is actually mentioned several times.
- On observing all the system runs taken together, it was apparent that in two of the ten test topics (C00-2123 and P98-1046) that the average F_1-score was one standard deviation or more away from the mean average F_1-score. This highlights the challenging nature of the task, and suggests that more general systems and methods should be devised to perform well different instances of scientific summarization.
- The fine-grained analysis highlighted that kernel-based approaches provided by *sys15* were the most inconsistent in their performance, often obtaining no matches for their Task 1A responses. Possibly, more experiments would help to adjust the parameters in such approaches to be suitable for the scientific summarization task.

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 Tasks 1A and 1B were also participants in the CL-SciSumm pilot task. In general, those systems which implemented weights based on term and document frequency tended to perform better than those which did not. 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 other systems. We also note that the systems performing well on Tasks 1A and 1B also do well in generating summaries—this supports our expectations on the Shared Task, and validates the need to push the state of the art in scientific summarization.

The participants have provided us with valuable insights about our dataset quality and provided feedback for further development. Besides enriching the quality of the SciSumm dataset, we are again planning to extend the dataset itself, with a larger corpus enriched with metadata compiled by the AAN [22].

Based on the interest of the community and the participants' feedback, we believe that the CL-SciSumm Shared Task and its associated corpus has broad applicability to related problems in computational linguistics and natural language processing, especially in the sub-disciplines of text summarization, discourse structure in scholarly discourse, paraphrasing, textual entailment and text simplification. 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 sub-tasks which could be added into our purview. We are also scouting for other related research problems of relevance to the scientific summarization community.

Acknowledgements 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. 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. We also thank Rahul Jha and Dragomir Radev for sharing their software to prepare the XML versions of papers, and Kevin B. Cohen and colleagues for sharing their annotation schema, export scripts and the Knowtator package implementation on the Protege software. These parties have all made indispensable contributions in realizing this Shared Task.

References

1. Aggarwal, P., Sharma, R.: Lexical and Syntactic cues to identify Reference Scope of Citance. In: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 103–112. Newark, NJ, USA (2016)
2. Cao, Z., Li, W., Wu, D.: PolyU at CL-SciSumm 2016. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 132–138. Newark, NJ, USA (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. Presses Universitaires de Louvain, Louvain-la-Neuve (2010)
6. Hoang, C., Kan, M.: Towards automated related work summarization. In: Proceedings 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: Proceedings of ENLG, pp. 125–135 (2013)
9. Jones, K.S.: Automatic summarising: the state of the art. Inf. Process. Manag. **43**(6), 1449–1481 (2007)
10. Klampfl, S., Rexha, A., Kern, R.: Identifying referenced text in scientific publications by summarisation and classification techniques. In: Proceedings of the Joint Workshop on Bibliometric-

Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 122–131. Newark, NJ, USA (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: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 156–167. Newark, NJ, USA (2016)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text summarization branches out. In: Proceedings of the ACL-04 workshop **8** (2004)
13. Liu, F., Liu, Y.: Correlation between rouge and human evaluation of extractive meeting summaries. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pp. 201–204. Association for Computational Linguistics (2008)
14. Lu, K., Mao, J., Li, G., Xu, J.: Recognizing reference spans and classifying their discourse facets. In: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 139–145. Newark, NJ, USA (2016)
15. Malenfant, B., Lapalme, G.: RALI system description for CL-SciSumm 2016 shared task. In: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 146–155. Newark, NJ, USA (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: Proceedings 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: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 113–121. Newark, NJ, USA (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: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 168–174. Newark, NJ, USA (2016)
21. Qazvinian, V., Radev, D.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, pp. 689–696. ACL (2008)
22. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL anthology network corpus. Lang. Resour. Eval. (2013). doi:10.1007/s10579-012-9211-2
23. Saggion, H.: SUMMA: a robust and adaptable summarization tool. Traitement Autom. des Lang. **49**(2), 103–125 (2002)
24. Saggion, H., AbuRa'Ed, A., Ronzano, F.: Trainable citation-enhanced summarization of scientific articles. In: Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2016), pp. 175–186. Newark, NJ, USA (2016)
25. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Comput. Linguist. **28**(4), 409–445 (2002)