The CL-SciSumm Shared Task 2017: Results and Key Insights

Kokil Jaidka[1], Muthu Kumar Chandrasekaran[2], Devanshu Jain[1], and Min-Yen Kan[2,3]

[1] University of Pennsylvania, USA
[2] School of Computing, National University of Singapore, Singapore
[3] Smart Systems Institute, National University of Singapore, Singapore
jaidka@sas.upenn.edu

Abstract. The CL-SciSumm Shared Task is the first medium-scale shared task on scientific document summarization in the computational linguistics (CL) domain. In 2017, it comprised three tasks: (1A) identifying relationships between citing documents and the referred document, (1B) classifying the discourse facets, and (2) generating the abstractive summary. The dataset comprised 40 annotated sets of citing and reference papers from the open access research papers in the CL domain. This overview describes the participation and the official results of the CL-SciSumm 2017 Shared Task, organized as a part of the 40^{th} Annual Conference of the Special Interest Group in Information Retrieval (SIGIR), held in Tokyo, Japan in August 2017. We compare the participating systems in terms of two evaluation metrics and discuss the use of ROUGE as an evaluation metric. 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

CL-SciSumm explores summarization of scientific research in the domain of computational linguistics research. It encourages 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 also encourages the use of citing mini-summaries written in other papers, by other scholars, when they refer to the paper. The Shared Task dataset comprises the set of citation sentences (i.e., "citances") that reference a specific paper as a (community-created) summary of a topic or paper [19]. Citances for a reference paper are considered a synopses of its key points and also its key contributions and importance within an academic community [17]. The advantage of using citances is that they are embedded with meta-commentary and offer a contextual, interpretative layer to the cited text. Citances offer a view of the cited paper which could complement the reader's context, possibly as a scholar [7] or a writer of a literature review [6].

The CL-SciSumm Shared Task is aimed at bringing together the summarization community to address challenges in scientific communication summarization. Over time, we anticipate that the Shared Task will spur the creation of new resources, tools and evaluation frameworks.

A pilot CL-SciSumm task was conducted at TAC 2014, as part of the larger BioMedSumm Task[4]. In 2016, a second CL-Scisumm Shared Task [5] was held as part of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) workshop [16] at the Joint Conference on Digital Libraries (JCDL[5]). This paper provides the results and insights from CL-SciSumm 2017, which was held as part of subsequent BIRNDL 2017 workshop[15] at the annual ACM Conference on Research and Development in Information Retrieval (SIGIR[6]).

2 Task

CL-SciSumm defined two serially dependent tasks that participants could attempt, given a canonical training and testing set of papers.

Given: A topic consists of a Reference Paper (RP) and ten or more 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. Additionally, the dataset provides three types of summaries for each RP:

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

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.

3 Development

We built the CL-SciSumm corpus by randomly sampling research papers (Reference papers, RPs) from the ACL Anthology corpus and then downloading the

[4] http://www.nist.gov/tac/2014
[5] http://www.jcdl2016.org/
[6] http://sigir.org/sigir2017/

citing papers (CPs) for those which had at least ten citations. The prepared dataset then comprised annotated citing sentences for a research paper, mapped to the sentences in the RP which they referenced. Summaries of the RP were also included.

The CL-SciSumm 2017 corpus included a refined version of the CL-SciSumm 2016 corpus of 30 RPs as a training set, in order to encourage teams from the previous edition to participate. The test set was an additional corpus of 10 RPs.

Based on feedback from CL-SciSumm 2016 task participants, we refined the training set as follows:

- In cases where the annotators could not place the citance to a sentence in the referred paper, the citance was discarded. In prior versions of the task, annotators were required to reference the title (Reference Offset: ['0']) but the participants complained that this resulted in a drop in system performance.
- Citances were deleted if they mentioned the referred paper in a clause as a part of multiple references and did not cite specific information about it.

For details of the general procedure followed to construct the CL-SciSumm corpus, and changes made to the procedure in CL-SciSumm-2016, please see [5].

3.1 Annotation

The annotation scheme was unchanged from what was followed in previous editions of the task and the original BiomedSumm task developed by Cohen et. al[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.

4 Overview of Approaches

Nine systems participated in Task 1 and a subset of five also participated in Task 2. The following paragraphs discuss the approaches followed by the participating systems, in lexicographic order by team name.

The Beijing University of Posts and Telecommunications team from their Center for Intelligence Science and Technology (CIST, [11]) followed an approach similar to their 2016 system submission [10]. They calculated a set of similarity metrics between reference spans and citance – *idf* similarity, Jaccard similarity, and context similarity. They submitted six system runs which combined similarity scores using a fusion method, a Jaccard Cascade method, a Jaccard Focused method, an SVM method and two ensemble methods using voting.

The Jadavpur University team (Jadavpur, [3]) participated in all of the tasks. For Task 1A, they defined a cosine similarity between texts. The reference paper's

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

sentence with the highest score is selected as the reference span. For Task 1B, they represent each discourse facet as a bag of words of all the sentences having that facet. Only words with the highest *tf.idf* values are chosen. To identify the facet of a sentence, they calculated the cosine similarity between a candidate sentence vector and each bag's vector. The bag with the highest similarity is deemed the chosen facet. For Task 2, a similarity score was calculated between pairs of sentences belonging to the same facets. If the resultant score is high, only a single sentence of the two is added to the summary.

Nanjing University of Science and Technology team (NJUST, [14]) participated in all of the tasks (Tasks 1A, 1B and 2). For Task 1A, they used a weighted voting-based ensemble of classifiers (linear support vector machine (SVM), SVM using a radial basis function kernel, Decision Tree and Logistic Regression) to identify the reference span. For Task 1B, they created a dictionary for each discourse facet and labeled the reference span with the facet if its dictionary contained any of the words in the span. For Task 2, they used bisecting K-means to group sentences in different clusters and then used maximal marginal relevance to extract sentences from each cluster and combine into a summary.

National University of Singapore WING (NUS WING, [18]) participated in Tasks 1A and B. They followed a joint-scoring approach, weighting surface-level similarity using *tf.idf* and longest common subsequence (LCS), and semantic relatedness using a pairwise neural network ranking model. For Task 1B, they retrofitted their neural network approach, applying it to output of Task 1A.

The Peking University team (PKU, [21]) participated in Task 1A. They computed features based on sentence-level and character-level *tf.idf* scores and word2vec similarity and used logistic regression to classify sentences as being reference spans or not.

The Graz University of Technology team (TUGRAZ, [4]) participated in Tasks 1A and 1B. They followed an information retrieval style approach for Task 1A, creating an index of the reference papers and treating each citance as a query. Results were ranked according to a vector space model and BM25. For Task 1B, they created an index of cited text along with the discourse facet(s). To identify the discourse facet of the query, a majority vote was taken among the discourse facets found in the top 5 results.

The University of Houston team (UHouston, [8]) used a combination of lexical and syntactic features for Task 1A, based on the position of text and textual entailment They tackled Task 1B using WordNet expansion.

The University of Mannheim team (UniMA, [9]) also participated in all of the tasks. For Task 1A, they used supervised learning to rank paradigm to rank the sentences in the reference paper using features such as lexical similarity, semantic similarity, entity similarity and others. They formulated Task 1B, as a one-versus-all multi-class classification. They used an SVM and a trained convolutional neural network (CNN) for each of the five binary classification tasks. For Task 2, they clustered the sentences using single pass clustering algorithm using a Word Mover's similarity measure and sorted the sentences in each cluster according to their Text Rank score. Then they ranked the clusters according to the average Text Rank score. Top sentences were picked from the clusters and added to summary until the word limit of 250 words was reached.

Finally, the Universitat Pompeu Fabra team (UPF, [1]) participated in Tasks 1A, 1B and 2. For Task 1A, they used a weighted voting ensemble of systems that used word embedding distance, modified Jaccard distance and BabelNet embedding distance. They formulated Task 1B as a one-versus-all multi-class classification. For Task 2, they trained a linear regression model to learn the scoring function (approximated as cosine similarity between reference paper's sentence vector and summary vector) of each sentence.

5 Evaluation

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. The raw number of overlapping sentences were used to calculate the precision, recall and F_1 score for each system. We followed the approach in most SemEval tasks in reporting the overall system performance as its micro-averaged performance over all topics in the blind test set.

Additionally, we calculated lexical overlaps in terms of the ROUGE-2 and ROUGE-SU4 scores [12] between the system output and the human annotated gold standard reference spans.

ROUGE scoring was used for CL-SciSumm 17, for Tasks 1a and Task 2. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a set of metrics used to automatically evaluate summarization systems [12] by measuring 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 2017 uses ROUGE-2 and ROUGE-SU4 for its evaluation.

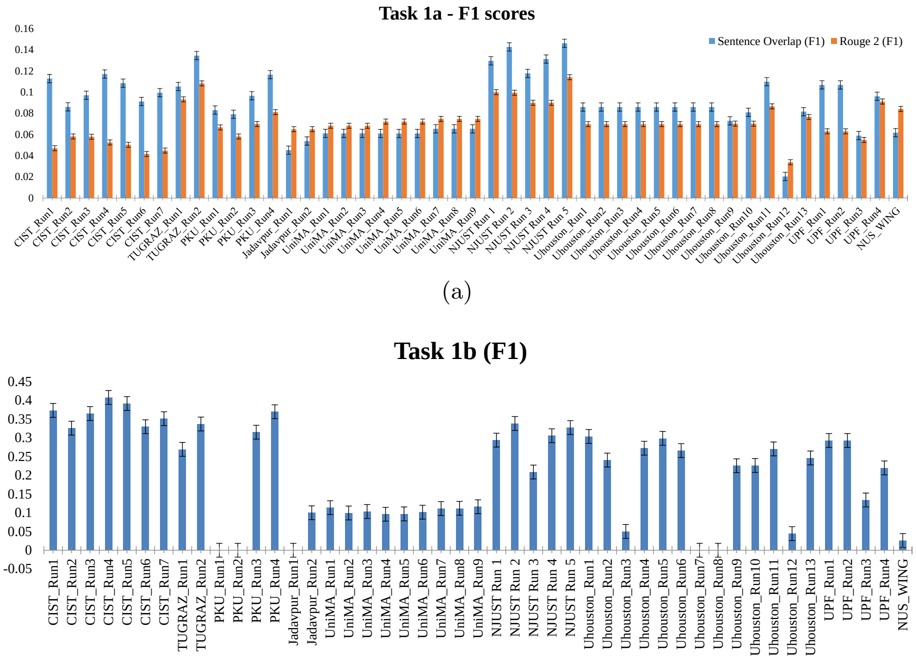
Task 1B was evaluated as a proportion of the correctly classified discourse facets by the system, contingent on the expected response of Task 1A. As it is a multi-label classification, this task was also scored based on the precision, recall and F_1 scores.

Task 2 was optional, and also evaluated using the ROUGE–2 and ROUGE–SU4 scores between the system output and three types of gold standard summaries of the research paper: the reference paper's abstract, a community summary, and a human summary.

The evaluation scripts have been provided at the CL-SciSumm Github repository[8] where the participants may run their own evaluation and report the results.

6 Results

This section compares the participating systems in terms of their performance. Five 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–SU4 against the 3 gold standard summary types. The results are provided in Table 1 and Figure 1. The detailed implementation of the individual runs are described in the system papers included in this proceedings volume.



(a)

(b)

Fig. 1: Performances on (a) Task 1A in terms of sentence overlap and ROUGE-2, and (b) Task 1B conditional on Task 1A

For Task 1A, the best performance was shown by three of the five runs from NJUST [14]. Their performance was closely followed by TUGRAZ [4]. The third best system was CIST [11] which was also the best performer for Task 1B. The

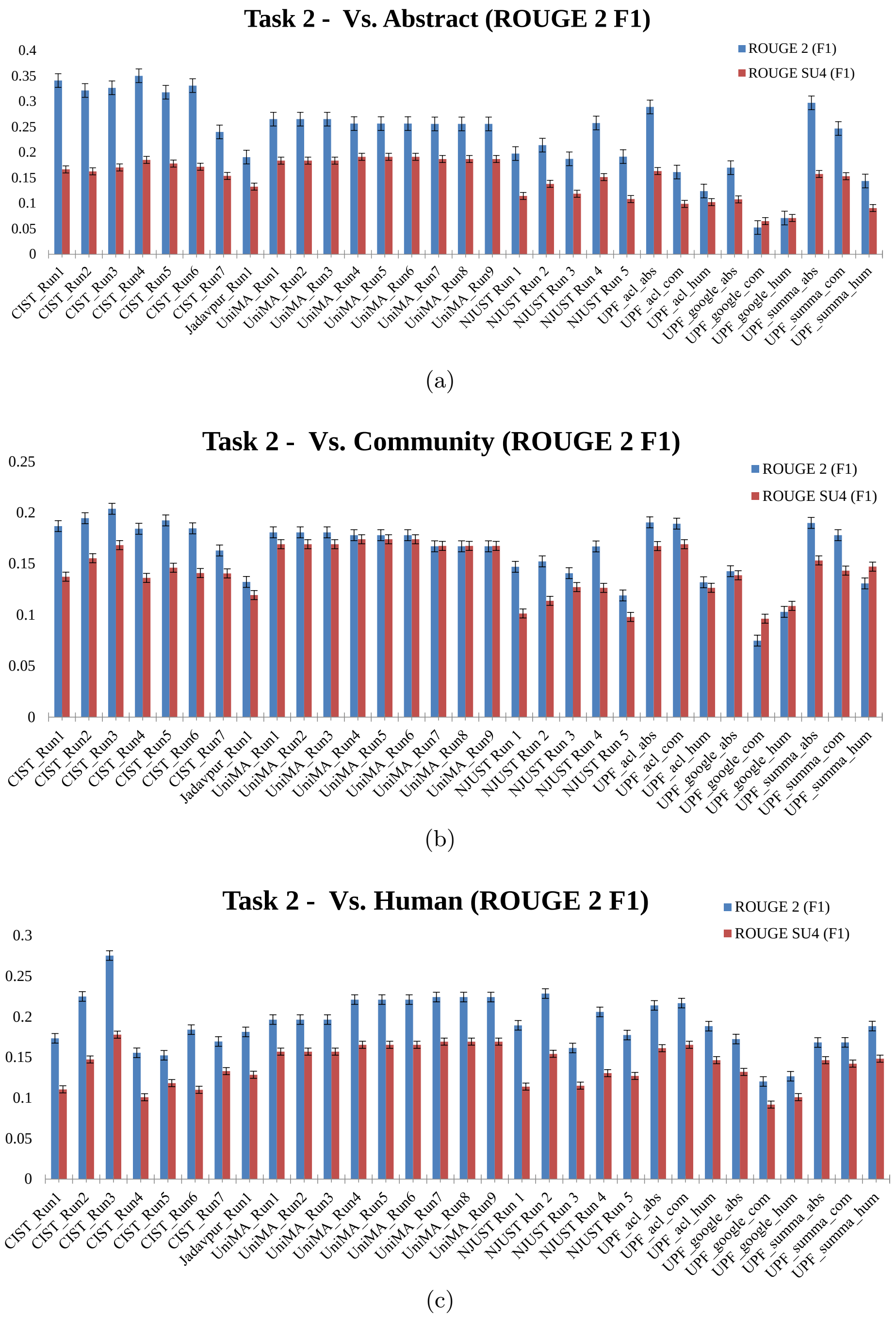
[8] github.com/WING-NUS/scisumm-corpus

System	**Task 1A: Sentence Overlap (F_1)**	**Task 1A: ROUGE F_1**	**Task 1B**
NJUST [14] Run 2	**0.124**	0.100 (3)	0.339 (7)
NJUST [14] Run 5	0.123	**0.114** (1)	0.328 (10)
NJUST [14] Run 4	0.114	0.090 (7)	0.306 (13)
TUGRAZ [4] Run 2	0.110	0.108 (2)	0.337 (8)
CIST [11] Run 1	0.107	0.047 (44)	0.373 (3)
CIST [11] Run 4	0.105	0.053 (42)	**0.408** (1)
PKU [21] Run 4	0.102	0.081 (11)	0.370 (4)
CIST [11] Run 5	0.100	0.050 (43)	0.392 (2)
NJUST [14] Run 1	0.097	0.100 (3)	0.294 (16)
NJUST [14] Run 3	0.094	0.090 (7)	0.209 (28)
UHouston [8] Run 1	0.091	0.087 (9)	0.271 (20)
TUGRAZ [4] Run1	0.088	0.093 (5)	0.269 (21)
UPF [1] Run 1	0.088	0.063 (36)	0.293 (17)
UPF [1] Run 2	0.088	0.063 (36)	0.293 (17)
CIST [11] Run 3	0.086	0.058 (38)	0.365 (5)
CIST [11] Run 7	0.084	0.045 (45)	0.351 (6)
PKU [21] Run 3	0.083	0.070 (19)	0.315 (12)
CIST [11] Run 6	0.077	0.042 (46)	0.330 (10)
UHouston [8] Run 1	0.074	0.070 (19)	0.304 (14)
UHouston [8] Run 2	0.074	0.070 (19)	0.241 (24)
UHouston [8] Run 3	0.074	0.070 (19)	0.050 (40)
UHouston [8] Run 4	0.074	0.070 (19)	0.272 (19)
UHouston [8] Run 5	0.074	0.070 (19)	0.299 (13)
UHouston [8] Run 6	0.074	0.070 (19)	0.266 (22)
UHouston [8] Run 7	0.074	0.070 (19)	0 (43)
UHouston [8] Run 8	0.074	0.070 (19)	0 (43)
PKU [21] Run 1	0.071	0.067 (33)	0 (43)
UPF [1] Run 4	0.071	0.091 (6)	0.220 (27)
UHouston [8] Run 9	0.068	0.070 (16)	0.226 (25)
UHouston [8] Run 10	0.068	0.070 (16)	0.226 (25)
UHouston [8] Run 13	0.068	0.077 (12)	0.246 (23)
CIST [11] Run 2	0.067	0.058 (38)	0.326 (11)
PKU [21] Run 2	0.066	0.058 (38)	0 (43)
NUS [18]	0.055	0.084 (10)	0.026 (42)
UniMA [9] Run 1	0.053	0.068 (30)	0.114 (31)
UniMA [9] Run 2	0.053	0.068 (30)	0.100 (36)
UniMA [9] Run 3	0.053	0.068 (30)	0.103 (34)
UPF [1] Run 3	0.052	0.055 (41)	0.134 (29)
UniMA [9] Run 7	0.051	0.075 (13)	0.111 (33)
UniMA [9] Run 8	0.051	0.075 (13)	0.112 (32)
UniMA [9] Run 9	0.051	0.075 (13)	0.116 (30)
UniMA [9] Run 4	0.049	0.072 (16)	0.096 (39)
UniMA [9] Run 5	0.049	0.072 (16)	0.097 (38)
UniMA [9] Run 6	0.049	0.072 (16)	0.102 (35)
Jadavpur [3] Run 2	0.042	0.065 (34)	0.100 (36)
Jadavpur [3] Run 1	0.037	0.065 (34)	0 (43)
UHouston [8] Run 12	0.014	0.034 (47)	0.044 (41)

Table 1: Systems' performance in Task 1A and 1B, ordered by their F_1-scores for sentence overlap on Task 1A. Each system's rank by their performance on ROUGE on Task 1A and 1B are shown in parentheses.

System	**Vs. Abstract**		**Vs. Human**		**Vs. Community**	
	R–2	**RSU–4**	**R–2**	**RSU–4**	**R–2**	**RSU–4**
CIST [11] Run 4	**0.351**	0.185(3)	0.156(22)	0.101(23)	0.184(9)	0.136(16)
CIST [11] Run 1	0.341	0.167(8)	0.173(16)	0.111(21)	0.187(7)	0.137(15)
CIST [11] Run 6	0.331	0.172(6)	0.184(13)	0.110(22)	0.185(8)	0.141(12)
CIST [11] Run 3	0.327	0.171(7)	**0.275**(1)	**0.178**(1)	**0.204**(1)	0.168(4)
CIST [11] Run 2	0.322	0.163(9)	0.225(3)	0.147(9)	0.195(2)	0.155(7)
CIST [11] Run 5	0.318	0.178(5)	0.153(23)	0.118(18)	0.192(3)	0.146(10)
UPF [1] summa_abs	0.297	0.158(11)	0.168(19)	0.147(9)	0.190(5)	0.153(8)
UPF [1] acl_abs	0.289	0.163(9)	0.214(7)	0.161(5)	0.191(4)	0.167(5)
UniMA [9] Runs 1,2,3	0.265	0.184(4)	0.197(9)	0.157(6)	0.181(10)	0.169(2)
NJUST [14] Run 4	0.258	0.152(14)	0.206(8)	0.131(15)	0.167(13)	0.126(19)
UniMA [9] Run 4,5,6	0.257	**0.191**(1)	0.221(5)	0.166(3)	0.178(11)	**0.174**(1)
UniMA [9] Run 7,8,9	0.256	0.187(2)	0.224(4)	0.169(2)	0.167(13)	0.167(5)
UPF [1] summa_com	0.247	0.153(13)	0.168(19)	0.142(12)	0.178(11)	0.143(11)
CIST [11] Run 7	0.240	0.154(12)	0.170(18)	0.133(13)	0.163(15)	0.141(12)
NJUST [14] Run 2	0.214	0.138(15)	0.229(2)	0.154(7)	0.152(16)	0.114(21)
NJUST [14] Run 1	0.198	0.114(18)	0.190(10)	0.114(20)	0.147(17)	0.101(23)
NJUST [14] Run 5	0.192	0.108(19)	0.178(15)	0.127(17)	0.119(23)	0.098(24)
Jadavpur [3] Run 1	0.191	0.133(16)	0.181(14)	0.129(16)	0.132(19)	0.119(20)
NJUST [14] Run 3	0.187	0.119(17)	0.162(21)	0.115(19)	0.141(19)	0.127(17)
UPF [1] google_abs	0.170	0.108(19)	0.173(16)	0.132(14)	0.143(18)	0.139(14)
UPF [1] acl_com	0.161	0.099(22)	0.217(6)	0.166(3)	0.189(6)	0.169(2)
UPF [1] summa_hum	0.144	0.091(23)	0.189(11)	0.148(8)	0.131(21)	0.147(9)
UPF [1] acl_hum	0.124	0.102(21)	0.188(12)	0.147(9)	0.132(19)	0.127(17)
UPF [1] google_hum	0.071	0.071(24)	0.127(24)	0.101(23)	0.103(24)	0.109(22)
UPF [1] google_com	0.052	0.065(25)	0.120(25)	0.092(25)	0.075(25)	0.096(25)
Mean Score	**0.237**	**0.150**	**0.193**	**0.141**	**0.164**	**0.145**

Table 2: Systems' performance for Task 2 ordered by their ROUGE–2(R–2) and ROUGE–SU4(R–SU4) F_1-scores. Each system's rank by their performance on the corresponding evaluation is shown in parentheses. Winning scores are bolded.



(a)

(b)

(c)

Fig. 2: Task 2 Performances on (a) Abstract, (b) Community and (c) Human summaries. Plots correspond to the numbers in Table 2.

next best performers of Task 1B were by PKU [21] and NJUST [14].

For Task 2, CIST had the best performance against the abstract, community and human summaries [11]. UPF [1] had the next best performances against the abstract and community summaries while NJUST [14] and UniMA [9] were close runner-ups against the human summaries.

In this edition of the task, we used ROUGE-1 as a more lenient way to evaluate Task 1A – however, as Figure 1 shows, many systems' performance on ROUGE scores was lower than on the exact match F_1. The reasons for this aberration are discussed in Section 7.

7 Error Analysis

We carefully considered participant feedback from CL-Scisumm 2016 Task [5] and made a few changes to the annotation rules and evaluation procedure. We discuss the key insights from Task 1A, followed by Task 2.

Task 1A: In 2017, we introduced the ROUGE metric to evaluate Task 1A, which we anticipated would be a more lenient way to score the system runs, especially since it would consider bigrams separated by over up to four words. However, we found that system performance on ROUGE was not always more lenient than sentence overlap F_1 scores. Table 3 provides some examples to demonstrate how the ROUGE score is biased to prefer shorter sentences over longer ones. ROUGE scores are calculated for candidate reference spans (RS) from system submissions against the gold standard (GS) reference span (Row 1 of Table 3). Here, we consider 3 examples, each with a pair of RS compared with one another. The RS of Submission 2 is shorter than that of Submission 1. Both systems retrieve one correct sentence (overlap with GS) and one incorrect sentence. Although F_1 score overlap for exact match of sentences for both will be the same, the ROUGE score for Submission 2 (shorter) is greater than that of Submission 1. In the next example, neither system retrieves a correct match. Submission 1 is shorter than that of Submission 2. The exact match for both systems are the same: 0. However the ROUGE scores for Submission 1 (shorter) is higher than that of Submission 2. In the last example, both the submissions correctly retrieve GS. However, they also retrieve an additional false positive sentence. Submission 1's RS is longer than Submission 2. Similar to the previous example, ROUGE score for Submission 1 (shorter) is less than that of Submission 2.

Evaluation on ROUGE recall instead of ROUGE F_1 will prevent longer candidate summaries from being penalized. However, there is a caveat – a system would retrieve the entire article (RP) as the reference span and achieve the highest ROUGE recall. On sorting all the system runs by their average recall measure, we find that the submission by [3] ranked the first. Considering the overall standing of this system, we infer that there were probably a lot of false positives due to there being a lack of a stringent word limit. In future tasks, we will impose a limit on the length of the reference span that can be retrieved.

Although our documentation advised participants to return reference spans of three sentences or under, we did not penalize longer outputs in our evaluation.

On the other hand, evaluation on ROUGE precision would encourage systems to return single-sentences with high information overlap. A large body of work in information retrieval and summarization has measured system performance in terms of task precision. In fact, as argued by Felber and Kern [4], Task 1A can be considered akin to an information retrieval or a question answering task. We can then use standard IR performance measures such as Mean Average Precision (MAP) over all the reference spans. We plan to pilot this measure in the next edition of the Shared Task.

Task 1A topic level meta-analysis: We conducted a meta-analysis of system performances for Task 1A over all the topics in the test set. We observed that for only one of the ten test topics (specifically, W09-0621), the average F_1 score was one standard deviation away from the mean average F_1 score of all the topics taken together. At the topic level, we observed that the largest variances in system performance were for W11-0815, W09-0621, D10-1058 and P07-1040, for which nearly two-thirds of all the submitted runs had an ROUGE or an overlap F_1 score that was more than one standard deviation from the average F_1 score for that topic. We note that since most of the participants submitted multiple runs, some of these variances are isolated to all the runs submitted by a couple of teams (specifically, NJUST and UniMA) and may not necessarily reflect an aberration with the topic itself. All participants were recommended to closely examine the outputs these and other topics during their error analysis. They can refer to the topic-level results posted in the Github repository of the CL-SciSumm dataset [9].

Task 1B: Systems reported difficulty in classifying discourse facets (classes) with few datapoints. The class distribution, in general, is skewed towards the 'Method' facet. Systems reported that the class imbalance could not be countered effectively by class weights. This suggests that the 'Method' facet is composed of other sub-facets which need to be identified and annotated as ground truth.

Task 2: While considering the results from Task 2, we observed that ensemble approaches were the most effective against all three sets of gold standard summaries. Some systems – for instance, the system by Abura'Ed et. al [1] – tailored their summary generation approach to improve on one type of summary at a time. We plan to discourage this approach in future tasks, as we envision that systems would converge towards a general, optimal method for generating salient scientific summaries. Based on the results from CL-SciSumm 2016, we had expected that approaches that did well against human summaries would also do well against community summaries. However, no such inferences could be made from the results of CL-SciSumm 2017. In the case of NJUST [14], one of their approaches (Run 2) was among the top approaches against abstract and human summaries, but was a poor performer against the community summaries. On the other hand, different runs by UPF [11] performed well against different

[9] https://github.com/WING-NUS/scisumm-corpus

Table 3: Error Analysis: Why ROUGE did not improve systems’ performances

Summary Type	**Sentence id**	**Text**	**Rouge-F**
Gold Standard	36,37	’Identifying semantic relations in a text can be a useful indicator of its conceptual structure.’, ’Lexical cohesion is expressed through the vocabulary used in text and the semantic relations between those words.	
Submission 1	36,45	’To automatically detect lexical cohesion tics between pairwise words, three linguistic features were considered: word repetition, collocation and relation weights.’, ’Lexical cohesion is expressed through the vocabulary used in text and the semantic relations between those words.’	0.20
Submission 2	36, 119	Each text was only 500 words in length and was related to a specific subject area.’, ’Lexical cohesion is expressed through the vocabulary used in text and the semantic relations between those words.	0.39
Submission 1	45,119	’To automatically detect lexical cohesion tics between pairwise words, three linguistic features were considered: word repetition, collocation and relation weights.’ ’Each text was only 500 words in length and was related to a specific subject area.’	0.07
Submission 2	45, 118	’To automatically detect lexical cohesion tics between pairwise words, three linguistic features were considered: word repetition, collocation and relation weights.’, ’In this investigation, recall rates tended to be lower than precision rates because the algorithm identified fewer segments (4.1 per text) than the test subjects (4.5).	0.03
Submission 1	36,37,118	’Identifying semantic relations in a text can be a useful indicator of its conceptual structure.’, ’Lexical cohesion is expressed through the vocabulary used in text and the semantic relations between those words.’, ’In this investigation, recall rates tended to be lower than precision rates because the algorithm identified fewer segments (4.1 per text) than the test subjects (4.5).’	0.33
Submission 2	36, 37, 119	’Each text was only 500 words in length and was related to a specific subject area.’, ’Identifying semantic relations in a text can be a useful indicator of its conceptual structure.’, ’Lexical cohesion is expressed through the vocabulary used in text and the semantic relations between those words.’	0.58

summaries. One of their runs (‘UPF_acl_com’) was among the top against human and community summaries but was near the bottom against abstract summaries.

8 Conclusion

Nine systems participated in CL-SciSumm 2017 shared tasks. The tasks provided a larger corpus with further refinements over 2016. Compared with 2016, the task attracted additional submissions that attempted neural network-based methods. Participants also experimented with the use of word embeddings trained on the shared task corpus, as well as on other domain corpora. We recommend that future approaches should go beyond off-the-shelf deep learning methods, and also exploit the structural and semantic characteristics that are unique to scientific documents; perhaps as an enrichment device for word embeddings. The results from 2016 suggest that the scientific summarization task lends itself as a suitable problem for transfer learning [2].

For CL-SciSumm 2018, we are planning to collaborate with Yale University and introduce semantic concepts from the ACL Anthology Network [20].

Acknowledgement. We 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 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 would also like to 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. Abura’Ed, A., Chiruzzo, L., Saggion, H., Accuosto, P., lex Bravo: LaSTUS/TALN @ CL-SciSumm-17: Cross-document Sentence Matching and Scientific Text Summarization Systems. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
2. 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)
3. Dipankar Das, S.M., Pramanick, A.: Employing Word Vectors for Identifying,Classifying and Summarizing Scientific Documents. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)

4. Felber, T., Kern, R.: Query Generation Strategies for CL-SciSumm 2017 Shared Task. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
5. Jaidka, K., Chandrasekaran, M.K., Rustagi, S., Kan, M.Y.: Insights from cl-scisumm 2016: the faceted scientific document summarization shared task. International Journal on Digital Libraries pp. 1–9 (2017)
6. 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)
7. Jones, K.S.: Automatic summarising: The state of the art. Information Processing and Management 43(6), 1449–1481 (2007)
8. Karimi, S., Verma, R., Moraes, L., Das, A.: University of Houston at CL-SciSumm 2017. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
9. Lauscher, A., Glavas, G., Eckert, K.: Citation-Based Summarization of Scientific Articles Using Semantic Textual Similarity. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
10. 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)
11. Li, L., Zhang, Y., Mao, L., Chi, J., Chen, M., Huang, Z.: CIST@CLSciSumm-17: Multiple Features Based Citation Linkage, Classification and Summarization. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
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. 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. Ma, S., Xu, J., Wang, J., Zhang, C.: NJUST@CLSciSumm-17. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
15. Mayr, P., Chandrasekaran, M.K., Jaidka, K.: Editorial for the 2nd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL) at SIGIR 2017. In: Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Tokyo, Japan, August 11, 2017. pp. 1–6 (2017), http://ceur-ws.org/Vol-1888/editorial.pdf
16. 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 Pro-

cessing for Digital Libraries (BIRNDL2016). pp. 1–5. Newark, NJ, USA (June 2016)

17. 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)
18. Prasad, A.: WING-NUS at CL-SciSumm 2017: Learning from Syntactic and Semantic Similarity for Citation Contextualization. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)
19. 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)
20. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The acl anthology network corpus. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries. pp. 54–61. Association for Computational Linguistics (2009)
21. Zhang, D.: PKU @ CLSciSumm-17: Citation Contextualization. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017)