The Computational Linguistics Summarization Pilot Task

Kokil Jaidka[1*], **Muthu Kumar Chandrasekaran**[2], **Rahul Jha**[3], **Christopher Jones**[4]
Min-Yen Kan[2,5], **Ankur Khanna**[2], **Diego Molla-Aliod**[4], **Dragomir R. Radev**[3],
Francesco Ronzano[6] and **Horacio Saggion**[6]

[1] Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore
[2] Web, IR NLP Group, School of Computing, National University of Singapore, Singapore
[3] School of Information, University of Michigan, USA
[4] Division of Information & Communication Sciences, Computing Department, Macquarie University, Australia
[2] Interactive and Digital Media Institute, National University of Singapore, Singapore
[6] Universitat Pompeu Fabra, Barcelona, Spain

Abstract

The Computational Linguistics (CL) Summarization Pilot Task was a pilot shared task to use citations to create summaries of scholarly research publications in the domain of computational linguistics. We describe the background for the task, corpus construction, evaluation methods for the pilot and survey the participating systems and their preliminary results. The experience gleaned from the pilot will assist in the proper organization of future shared task where difficulties with annotations and scale can be addressed.

1 Introduction

This paper describes the evolution and design of the Computational Linguistics (CL) pilot task[1] for the summarization of computational linguistics research papers sampled from the Association of Computational Linguistics' (ACL) anthology. This task was run concurrently with the Text Analysis Conference 2014 (TAC '14), although not formally affiliated with it. This shared task shares the same basic structure and guidelines with the formal TAC 2014 Biomedical Summarization (BiomedSumm) track. A training corpus "topics" from CL research papers was released, each comprising a reference paper along with some sampled papers that cited the reference paper. Participants were invited to enter their systems in a task-based evaluation, similar to BiomedSumm.

2 Background

Recent work (Mohammad et al., 2009; Abu-Jbara and Radev, 2011) in scientific document summarisation have used citation sentences (also known as *citances*) from citing papers to create a multi document summary of the reference paper (RP).

As proposed by (Vu, 2010; Hoang and Kan, 2010) the summarization can be decomposed into finding the relevant documents; in this case, the citing papers (CPs), then selecting sentences from those papers that cite and justify the citation and finally gen-

[*] Authors appear in alphabetical order, with the exception of the task coordinator, who was given first authorship.

[1]This research is supported in part by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, and in part made possible thanks to a summer internship granted to Christopher Jones by the Department of Computing, Macquarie University.

erating the summary. To tackle each subproblem, we created a gold standard dataset where human annotators identified the citances in each of (up to) ten randomly sampled citing papers for the RP.

Jaidka and Khoo (2013)'s work on summarizing information science articles indicated that most citations clearly refer to one or more specific discourse facets of the cited paper. In the CL domain, during our corpus construction, we identified that the discourse facets being cited were usually the aim of the paper, methods followed, and the results or implications of the work. Accordingly, we used a different set of discourse facets than BiomedSumm which suit our target domain of CL papers better. The resultant corpus should be viewed as a development corpus only, such that the community can enlarge it to a proper shared task with training, development and testing set divisions in the near future.

3 Corpus Construction

A large and important portion of scholarly communication in the domain of computational linguistics is publicly accessible and archived at the ACL Anthology[2]. The texts from this archive are also under a Creative Commons license, which allows unfettered access to the published works for any purposes, including downstream research on summarization of its contents. We thus view the ACL Anthology as a corpus and randomly sampled published research papers as a base for building our annotated corpus. In selecting materials for resultant corpus from the Anthology, we wanted to enable citation-based summarization. To this end, with consultation from the BiomedSumm organizers, we needed to ensure that the reference paper was cited with appropriate diversity.

As of the corpus construction date (18 September 2014), the live Anthology contained approximately 25K publications, exclusive of the third-party papers hosted (i.e., with metadata but without the actual .PDF of the paper) and extraneous files (i.e., front matter and full volumes). To ensure sufficient opportunity to use citation based summarization, we further removed papers published after and including 2006, leaving 13.8K publications. We randomized this list to remove any ordering effects. Starting from the top of the list, we used a combination of Google Web and Scholar searches to approximate the number of citations (i.e., citing papers (CP)). We retained any paper with over 10 citations. We vetted the citations to ensure that the citation spread was at least a window of three years, as previous work had indicated that citations over different time periods (with respect to the publication date of the RP) exhibit different tendencies (**?**).

We then used the title search facility of the ACL Anthology Network[3] (AAN, February 2013 version), to locate the paper. We inspected and listed all citing papers' Anthology ID, title and year of publication. We note the citation count from Google / Google Scholar and AAN differ substantially.

To report the final list of citing papers, we strived to provide at least three CP for each RP. We defined the following criteria (in order or priority):

1. Non-list citation (i.e., at least one citation in the body of the CP for the RP not of the form [RP,a,b,c]);
2. The oldest and newest citations within AAN; and,
3. Citations from different years.

We included the oldest and newest citation regardless of criteria 1) and 3) and included a randomized sample of up to 8 additional citing paper IDs that met either criteria 1) and 3).

The resulting final list was divided among the annotator group, whom are a subset of the authors of this paper from NUS and NTU. We used the same scheme used by annotators of the BiomedSumm track's corpus. Given each RP and up to 10 associated CPs, the annotation group was instructed to find citations to the RP in each CP. Annotators followed instructions used for BiomedSumm task annotation, to re-use the resources created for BiomedSumm and reduce necessary effort. Specifically, the citation text, citation marker, reference text, and discourse facet were marked for each citation of the RP found in the CP.

4 The CL-Summ Task

This shared task proposes to solve same problems posed of the BioMedSumm track, but in the domain

[2]`http://aclweb.org/anthology/`

[3]`http://clair.eecs.umich.edu/aan/index.php`

of Computational Linguistics. This task calls for summarization frameworks to build a structured summary of a research paper – which incorporates facet information (such as Aims, Methods, Results and Implications) from the text of the paper, and "community summaries" from its citing papers.

We define the *CL-Summ Task* as follows:

Given: a topic, comprising of the PDF and extracted text of an RP and up to 10 CPs. In each provided CP, the citations to the RP (or citances) have been identified. The information referenced in the RP is also annotated. Note that both the text, and the citations may be noisy, and that there could be additional citing papers that were not provided (due to sampling).

Output systems to perform the following tasks, where the numbering of the task corresponds to those used in the BiomedSumm task.

- Task 1A: Identify the text span in the RP which corresponds to the citances from the CP. These may be of the granularity of a full sentence or several sentences, and may be contiguous or not. It may also be a sentence fragment (no more than 5).

- Task 1B: Identify the discourse facet for every cited text span from a predefined set of facets.

Evaluation: Assess Task 1 performance by using the ROUGE (Lin, 2004) score to compare the overlap of text spans in the system output versus the gold standard created by human annotators.

5 Participating teams

Nine teams expressed an interest in participating in the shared task which are listed below in alphabetical order.

1. **CCS2014**, from the IDA Center for Computing Sciences, USA. They proposed to employ a language model based on the sections of the document to find referring text and related sentences in the cited document.
2. **clair_umich**$^{*\$}$ from University of Michigan, Ann Arbor, USA.
3. **IHMC**, A team from IHMC, USA.
4. **IITKGP_sum**, from Indian Institute of Technology, Kharagpur, India. They planned to use citation network structure and citation context analysis to summarize the scientific articles.
5. **MQ**$^{*\$}$, from Macquarie University, Australia. They plan to use the same system that was used for the BiomedSumm track, with the exception that they will not incorporate domain knowledge (UMLS). For Task 1A they planned to use similarity metrics to extract the top n sentences from the documents. For Task 1B they used a logistic regression classifier.
6. **PolyAF**, from The Hong Kong Polytechnic University.
7. **TabiBoun14**, from the Boazii University, Turkey. They planned to modify an existing system for CL papers, which uses LIBSVM as a classification tool for facet classification, and plan to use cosine similarity to compare text spans.
8. **Taln.UPF***, from Universitat Pompeu Fabra, Spain. They have proposed to adapt available summarization tools to scientific texts.
9. **TXSUMM**, from University of Houston, Texas. Their system consists of applying similarity kernels in an attempt to better discriminate between candidate text spans (with sentence granularity). Their system uses an extractive, ranking method.

Three teams submitted system descriptions. A further two (of the three) submitted their findings. The system descriptions and self-reported task results are reported in the next sections (denoted with '*' and '$', respectively in the above text).

6 The clair_umich System — Comparing Overlap of Word Synsets

6.1 Data Preprocessing

For each RP, citing sentences were extracted from all its CP. Each citing sentence was then matched to a text segment in the original paper creating the final annotated dataset. The original source text for the papers in the CL-Summ corpus was not sentence-segmented, which made it difficult to compute evaluation metrics.

Data preprocessing of the CL-Summ corpus was done in the following way – First, sentences from the reference papers were segmented and then matched to each of these source sentences to the CL-Summ annotation files. This yielded a fixed set of source sentences from the original files, a subset of which were matched to each citing sentence. In this way, given a citing sentence, matching sentences from the source paper were compared to the gold standard sentences matched from the source paper and compute precision / recall.

The average number of source sentences matched for each citing sentence was 1.28 (with standard deviation 1.92). The maximum number of source sentences matched for a citing sentence was 7. Given that the total number of source sentences for papers ranged from between 100 to 600, this made it a very challenging classification problem.

6.2 Baseline System

The team first created a baseline system based on TF.IDF cosine similarity. For any citing sentence, the system computed the TF.IDF cosine similarity with all the sentences in the RP, thus the IDF values differed across each of the 10 RPs.

6.3 Supervised System

The supervised system used knowledge-based features derived from WordNet, syntactic dependency based features, and distributional features in addition to the simple lexical features like cosine similarity. These features are described below.

1. **Lexical Features:** Two lexical features were used – TF.IDF and the LCS (Longest Common Subsequence) between the citing sentence (C) and source sentence S, which is computed as:

$$\frac{|LCS|}{min(|C|, |S|)}$$

2. **Knowledge Based Features:** The system also used set of features based on Wordnet similarity. Six wordnet based word similarity measures were combined to obtain six knowledge based sentence similarity features using the method proposed in (Banea et al., 2012). The wordnet based word similarity measures used are path similarity, WUP similarity (Wu and Palmer, 1994), LCH similarity (Leacock and Chodorow, 1998), Resnik similarity (Resnik, 1995), Jiang-Conrath similarity (Jiang and Conrath, 1997), and Lin similarity (Lin, 1998).

   Given each of these similarity measures, the similarities between two sentences was computed by first creating a set of senses for each of the words in each of the sentences. Given these two sets of senses, the similarity score between citing sentence C and source sentence S was calculated as follows:

$$sim_{wn}(C, S) = \frac{(\omega + \sum_{i=1}^{|\phi|} \phi_i) * (2|C||S|)}{|C| + |S|}$$

   Here ω is the number of shared senses between C and S. The list ϕ contains the similarities of non-shared words in the shorter text, ϕ_i is the highest similarity score of the ith word among all the words of the lower text (Zhu and Lan, 2013).

3. **Syntactic Features:** An additional feature based on similarity of dependency structures was used, by applying the method described in (Zhu and Lan, 2013). The Stanford parser was used to obtain dependency parse all the citing sentences and source sentences. Given a candidate sentence pair, two syntactic dependencies were considered equal if they have the same dependency type, govering lemma, and dependent lemma. If R_c and R_s are the set of all dependency relations in C and S, the dependency overlap score was computed using the formula:

$$sim_{dep}(C, S) = \frac{2 * |R_c \cap R_s| * |R_c||R_s|}{|R_c| + |R_s|}$$

7 The MQ System — Finding the Best Fit to a Citance

Given the text of a citance, the MQ system ranks the sentences of the reference paper according to its similarity to the citance. Every sentence and its citance was modeled as a vector and compared using cosine similarity. The team experimented with different forms of representing the information in the vectors, and different forms of using the similarity scores to perform the final sentence ranking.

$$\text{MMR} = \arg\max_{D_i \in R \setminus S} \left[\lambda(\text{sim}(D_i, Q)) - (1 - \lambda) \max_{D_j \in S} \text{sim}(D_i, D_j) \right]$$

Where:

- Q is the citance text.
- R is the set of sentences in the document.
- S is the set of sentences that haven been chosen in the summary so far.

Figure 1: Maximal Marginal Relevance (MMR)

7.1 Baseline – Using TF.IDF

For the baseline system (similar to the clair_umich team), the TF.IDF of all lowercased words was used, without removing stop words. Separate TF.IDF statistics were computed for each reference paper, using the set of sentences in the paper and the citance text of all citing papers.

7.2 Adding texts of the same topic

Since the amount of text used to compute the TF.IDF in Section 7.1 was relatively little, the complete text of all citing papers was added, under the presumption that citing papers are presumably of the same topic as the reference paper. By adding this text we hope to include complementary information that can be useful for extending and computing the IDF component.

7.3 Adding context

In order to extend the information of each sentence in the reference paper and further add to the approach in Section 7.2, the text from the reference papers was added within a context window of 20 sentences by including the neighouring sentences, centered in the target sentence.

7.4 Re-ranking using MMR

The last experiment used Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to rank the sentences. All sentences were represented as TF.IDF vectors of extended information as described in Section 7.3. Then, the final score of a sentence was the combination of the similarity with the citance and similarity of the other sentences of the summary according to the formula shown in Figure 1. A value of $\lambda = 0.97$ was chosen.

8 The Taln.UPF System

8.1 Pre-processing / documents preparation:

The UPF system carried out the following set of preprocessing steps on the papers of each topic:

- Sentence segmentation: To identify candidate sentences that will be validated or rejected in the following pre-processing steps;
- Tokenizer and POS tagger: Using the open-source GATE software
- Sentence sanitizer: To remove incorrectly annotated sentences, relying on a set of rules and heuristics;
- Document structural analyzer: To classify each sentence as belonging to one of the following document structural categories: Abstract, Introduction, Result_Discussion, Experimental_Procedure, Supplemental_Data, Material_Method, Conclusion, Acknowledgement_Funding, and Reference;
- Sentence TF.IDF vector computation: To associate to each sentence a TF.IDF vector where the IDF values are computed over all the papers of the related topic (up to 10 CP and one RP).

8.2 Task 1A: Algorithm for identifying reference paper text spans for each citance

For each citance, its global citance context span was considered as the union of the citance context spans marked by human annotators (this was required for BiomedSumm; but in the case of CL-summ, there was only one available human annotation, so no union was required).

Then we select the sentences of the citing paper that overlap totally or partially the global citatance context span; these sentences are referred to as the citance context sentences ($CtxSent_1$,..., $CtxSent_N$), We characterize the citance by the document structural category associated with most of its citance context sentences ($CtxSent_1$,...,$CtxSent_N$).

In case of a tie in the number of occurrences of document structural categories among all the citance context sentences, we choose the document structural category that is most frequent in the citing paper. In case of a further tie, we select the document structural category that is most frequent in the whole set of citing and reference papers. We associate to each reference paper sentence (RefSent) a score equal to the sum of its TF.IDF vector cosine similarity with each citance context sentence ($CtxSent_1$,..., $CtxSent_N$).

Then, we weight the score of RefSent by the relative relevance in the whole training corpus of this kind of link between document structural categories. In order to do so, we consider the document structural category previously associated to the citation together with the document structural category of the RefSent. For instance, if there is a citance associated to the Introduction that references a RefSent belonging to the Abstract and we know that in the whole training corpus this situation occurs in 6.5% of citance-referenced sentence pairs, we multiply the RefSent score for 0.065, obtaining the final RefSent score.

We choose the 3 reference paper sentences (RefSents) with the highest final RefSent score as the reference paper text spans.

8.3 Task 1B: Algorithm for identifying the discourse facet of the cited text spans

A linear-kernel SVM classifier was trained to associate each citance with one of the five text facets considered in Task 1B. Each citance was characterized by lexical and semantic features extracted from the sentences belonging to the citance context together with the sentences of the reference paper selected as outcome of Task 1A. Some of the features exploited were:

- relative number of sentences belonging to each document structural category;
- relative number of sentences belonging to the citance context or reference paper;
- relative number of POS;
- presence of key lexical patterns.

9 Evaluation and Results

Two teams have submitted their results so far, as self-assessed using ROUGE (Lin, 2004). ROUGE (in specific, the ROUGE-L variant) is a popular evaluation method for summarisation systems that compares the text output of the system against a set of target summaries. Since ROUGE uses the actual contents words, and not the offset information of the setences chosen by the annotation team, we expect non-zero results for cases when a system chooses a sentence that is somewhat similar to (but not identical) to one chosen by annotators.

The MQ system was an unsupervised system while clair_umich system was supervised. clair_umich reports cross validated performance over the 10 topics while MQ evaluated their system over all 10 topics in a single run. For the MQ system, the output is the set of selected sentences, and the target summaries are the sentences given by the annotators. For the clair_umich system, the ROUGE-L scores were computed for each citing sentence in each annotation file separately and then averaged for a topic. While the results are not strictly comparable, they do allow for cautious conclusions to be drawn.

The following paragraphs describe the results for Tasks 1A, 1B, and the bonus Task 2 which was attempted by the MQ system.

9.1 Task 1A: For each citance, identify the spans of text (cited text spans) in the RP

Table 2 shows the ROUGE-L F_1 scores of each individual reference document from the CL-Summ dataset.

9.2 Task 2: Generate a structured summary of the RP and all of the community discussion of the paper represented in the citances

The MQ team performed an additional test to see whether information from the citances were useful for building an extractive summary, as is the case

MQ			clair_umich		
P	R	F_1	P	R	F_1
0.212	0.335	0.223	n/a	n/a	0.738

Table 1: Task 1A performance for the participating systems expressed as ROUGE-L score averaged over all topics.

Paper ID	MQ	clair_umich	Paper ID	MQ System	clair_umich
C90-2039	0.235	0.801	J00-3003	0.196	0.644
C94-2154	0.288	0.842	J98-2005	0.101	0.658
E03-1020	0.239	0.846	N01-1011	0.221	0.689
H05-1115	0.350	0.736	P98-1081	0.200	0.776
H89-2014	0.332	0.770	X96-1048	0.248	0.622

Table 2: Task 1A ROUGE-L F1 scores for individual topics.

with the BiomedSumm data (Mollá et al., 2014). They implemented extractive summarization systems with and without information from the citances. The summarizers without information from the citances scored each sentence as the sum of the TF.IDF values of the sentence elements. They tried the TF.IDF approach described in Section refsec:tfidf.

The summarizers with information from the citances scored each candidate sentence i on the basis of rank(i,c) obtained in Task 1A, which has values between 0 (first sentence) and n (last sentence), and represents the rank of sentence i in citance c:

$$\text{score}(i) = \sum_{c \in \text{citances}} 1 - \frac{\text{rank}(i, c)}{n}$$

The summaries were evaluated using ROUGE-L, where the model summaries are the abstracts of the corresponding papers. Since paper X96-1048 of the CL-Summ data did not have an abstract, it was omitted from this experiment. Table 3 shows the breakdown of ROUGE-L F_1 scores per document.

10 Discussion

10.1 Comparison with the BioMedSumm task

Table 4 compares the results of the MQ system's experiments with the CL-Summ data, against the results from the BiomedSumm data. In all results the systems were designed to return 3 sentences, as specified in the shared task. All short sentences (under 50 characters) were ignored, to avoid including headings or mistakes made by the sentence segmentation algorithm.

The results show an improvement in both domains, with the exception that MMR does not improve over the run that uses TF.IDF over context in CL-Summ, whereas there is an improvement in BiomedSumm. The absolute values are better in the BiomedSumm data, and looking at the confidence intervals it can be presumed that the difference between the best and the worst run is statistically significant in the BiomedSumm data. The results in the CL-Summ data are poorer in general and there are no statistically significant differences. However, this may be an artifact of the small size of the corpus. Overall, the improvement of results in CL-Summ mirrors that of the BiomedSumm data, so it can be suggested that on adding more information to the models that compute TF.IDF, the results improve. It is expected that alternative approaches, which gather related information to be added for computing the vector models will produce even better results. The results with MMR appears to be contradictory across the two domains, but the difference is small and may not be statistically significant.

10.2 Tweaking the Parameters — the clair_umich Baseline

For any citing sentence, the TF.IDF cosine similarity was computed with all the sentences in the source paper, and any sentences that had a cosine similarity higher than a given threshold were added to the matched sentences. Table 5 shows the precision/recall for different values of the cosine thresh-

Paper ID	TF.IDF	Task 1A TF.IDF	Task 1A MMR	Paper ID	TF.IDF	Task 1A TF.IDF	Task 1A MMR
C90-2039_TRAIN	0.347	0.315	0.293	J00-3003_TRAIN	0.221	0.382	0.367
C94-2154_TRAIN	0.095	0.123	0.120	J98-2005_TRAIN	0.221	0.216	0.233
E03-1020_TRAIN	0.189	0.189	0.196	N01-1011_TRAIN	0.187	0.268	0.284
H05-1115_TRAIN	0.134	0.306	0.321	P98-1081_TRAIN	0.241	0.210	0.206
H89-2014_TRAIN	0.294	0.319	0.320	Average	0.214	0.259	0.260

Table 3: ROUGE-L F_1 results for summaries generated by the MQ system.

	CL-Summ				BiomedSumm			
Run	P	R	F_1	CI	P	R	F_1	CI
TF.IDF	0.198	0.316	0.211	0.185–0.240	0.326	0.273	0.279	0.265–0.293
topics	0.201	0.324	0.217	0.191–0.245	0.357	0.288	0.300	0.285–0.316
context	0.214	0.339	0.225	0.197–0.255	0.372	0.291	0.308	0.293–0.323
MMR	0.212	0.335	0.223	0.195–0.251	0.375	0.290	0.308	0.293–0.323

Table 4: ROUGE-L results of the MQ system runs for Task 1A.

Similarity Threshold	Precision	Recall	F_1
0.01	0.027	0.641	0.051
0.05	0.048	0.426	0.087
0.1	0.060	0.235	0.095
0.2	0.079	0.081	0.080
0.3	0.062	0.032	0.042
0.4	0.022	0.085	0.012
0.5	0.007	0.002	0.003

Table 5: Precision/Recall for different values of the cosine threshold for the baseline clair_umich system.

old:

The F_1 scores seems to reach a maxima at a similarity threshold of about 0.1. The recall at the threshold of 0.1 is about 0.23, while the precision is only 0.06. This suggests that initial progress can be made on this problem by first removing these spurious matches that have high lexical similarity. We present some error analysis in the next section.

Error Analysis for the Baseline System. A number of errors made by the baseline system are due to source sentences that match the words but differ slightly in their information content. An example is shown on the right:

Citing text: "use the BNC to build a co-occurrence graph for nouns, based on a co-occurrence frequency threshold"

True positives:

- "Following the method in (Widdows and Dorow, 2002), we build a graph in which each node represents a noun and two nodes have an edge between them if they co-occur in lists more than a given number of times."

False positives:

- "Based on the intuition that nouns which co-occur in a list are often semantically related, we extract contexts of the form Noun, Noun,... and/or Noun, e.g. "genomic DNA from rat, mouse and dog"."
- "To detect the different areas of meaning in our local graphs, we use a cluster algorithm for graphs (Markov clustering, MCL) developed by van Dongen (2000)."
- "The algorithm is based on a graph model representing words and relationships between them."

Even though the false positive sentences contain the same lexical items (nouns, co-occurrence, graph), they differ slightly in the facts presented. Detection of such subtle differences in meaning might be challenging for an automated system.

Another set of difficult sentences is when the citing sentence says something that is implied by the sentence in the source paper. For example:

Citing text: "The line of our argument below follows a proof provided in ... for the maximum likelihood estimator based on nite tree distributions"
False negatives:

- "We will show that in both cases the estimated probability is tight."

Here, the citing text mentions a proof from the RP, but to match the sentence in the RP, the system needs to understand that the act of showing something in a scientific paper constitutes a proof.

11 Shortcomings and Limitations

There were several errors and shortcomings of the dataset which were identified in the process of annotating and parsing the corpus for use by the participating systems.

- The use of "..." where text spans are snippets: The use of "..." follows the BioMedSumm standard practice of indicating discontiguous texts. In Citation Text and Reference Text fields, the "..." means that there is a gap between two text spans (citation spans or reference spans). They may be on different pages, so the gap might be a text. There might be a formula or a figure there, or some text encoding which is not a part of the annotation. However, this notation caused mismatches for sentences which used text from different parts of the same sentence.

- Small size of the training corpus: The corpus comprised only a set of 10 topics, each with upto 10 citing documents. In this small dataset, participants were asked to conduct a 10-fold cross validation. The small size of the data set meant that there were no statistically significant results, but significance could only be guessed at, from the overall trend of the data.

- Errors in parsing the file: Some of the older PDF files, when parsed to text or XML, had such as misspelled words, spaces within words, sentences in the wrong place and so on. Unfortunately these errors were OCR parsing errors, and not within our control. We recommended that participants configure their string matching to be lenient enough to alleviate such problems.

- Errors in citation/reference offset numbers: In the original annotations, citation/reference offset numbers were character-based, and relative to an XML encoding which was not shared in the task, and did not match with the offset numbers on the text-only, cleaned version of the document. Although the text versions of the source documents were shared with the intention to help the participants, this often made their tasks more difficult if their system was geared towards numerical and not system matching. A solution was found for reference offsets by revising them to sentence id numbers based on available XML files from the clair_umich system's pre-processing stage; however, the citation offsets remain character-based.

- Text encoding: Often, the text was not in UTF-8 format as expected. Some participating teams, like the UPF, solved this by running the universal charset tool provided by Google Code over all the text and annotations in order to determine the right file encoding to use. It was found that some of the files were also in WINDOWS-1252 and GB18030 formats.

- Errors in file construction: An automatic, open-source software was used to map the citation annotations from a software, Protege, to a text file. However, participants identified several errors in the output - especially in cases where there was one-to-many mapping between citations and references. Besides this, several annotation texts had no annotation ID (Citance Number field).

12 Conclusions

This paper describes the computational linguistics pilot task for the faceted summarization of scholarly papers. We describe the three systems participated in the shared task, and describe the evaluation of two submitted runs. The teams used versions of TF.IDF as baselines. The MQ system followed an unsupervised algorithm while clair_umich followed a supervised algorithm. For identifying referenced text spans in reference papers, the best performance was obtained by clair_umich's supervised algorithm using lexical, syntactic and knowledge-based features to calculate the overlap between sentences in the citation span and the reference paper. Although no system submitted results for Task 1B, the task involving identifying the discourse facets of reference text, TALN.UPF submitted an algorithm which they aim to implement. Finally, an added experiment by the MQ system sought to compare baseline summaries of reference papers, based on a TF*IDF calculation, against gold standard summaries, comprising the reference paper's abstracts.

The clair_umich system incorporated WordNet synsets for expanding and comparing cited text with reference papers, and the use of syntactic features further enriched the calculation of overlap. On the other hand, the MQ system relied exclusively on reading and comparing texts. Furthermore, their system was originally built for the BioMedSumm task – however, they had to discard some domain-specific features for this task. It is possible that the lack of domain knowledge, coupled with OCR-related and PDF parsing errors, affected the performance of their system in the CL domain.

This task is an initiative for encouraging the development of tools and approaches for scientific summarisation. It helped us identify existing tools and resources to leverage on for this purpose and also the hindrances which needed to be overcome in order to have a systematic and well-coordinated evaluation. However, with results of only for two systems, it is not possible to conjecture at what may be the better methods for summarizing CL research papers. The resources from this task, and its corpus, are freely available for interested research groups to experiment with; the corpus is first-of-its-kind summarization corpus for compuational linguistics.

The results of the pilot are encouraging: there seems to be ample interest from the community and it seems possible to answer more detailed methodological questions with a more detailed analysis and a larger datasets. We encourage the community to support a future proposal to enlarge the pilot to a full scale shared task. We plan a systematic annotation of a training, development as well as test sets, and the availability of more than one gold standard annotation, and open-sourced tools and resources to support the efforts of participating teams. We invite the community to join us in this endeavour with any resources and time they can spare.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 500–509. Association for Computational Linguistics.

Carmen Banea, Samer Hassan, Michael Mohler, and Rada Mihalcea. 2012. Unt: A supervised synergistic approach to semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 635–642. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 335–336, New York, New York, USA. ACM Press.

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

Kokil Jaidka, Christopher SG Khoo, and Jin-Cheon Na. 2013. Literature review writing: how information is selected and transformed. *Aslib Proceedings*, 65(3):303–325.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International. Conference on Research in Computational Linguistics*, pages 19–33.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL Workshop on Tech Summarisation Branches Out*.

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

Diego Mollá, Christopher Jones, and Abeed Sarker. 2014. Impact of Citing Papers for Summarisation of Clinical Documents. In *Proceedings of the Australasian Language Technology Workshop 2014 (ALTA '14)*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hoang Cong Duy Vu. 2010. Towards automated related work summarization. Master's thesis, National University of Singapore.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tiantian Zhu and Man Lan. 2013. ECNUCS: Measuring short text semantic equivalence using multiple similarity measurements. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 124–131. Association for Computational Linguistics.