The Computational Linguistics Summarization Pilot Task

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

[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] Universitat Pompeu Fabra, Barcelona, Spain
[4] School of Information, University of Michigan, USA
[5] Division of Information & Communication Sciences, Computing Department, Macquarie University, Australia
[6] Interactive and Digital Media Institute, National University of Singapore, Singapore

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. The annotated development corpus used for this pilot task is publicly available at here:

`https://github.com/WING-NUS/scisumm-corpus`

1 Introduction

This paper describes the evolution and design of the Computational Linguistics (CL) pilot task 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. We released a training corpus of "topics" from CL research papers, each comprising a reference paper along with sample papers that cited the reference paper. Participants were invited to enter their systems in a task-based evaluation, similar to BiomedSumm.

This paper describes the participating systems and surveys the results from the task-based evaluation.

2 Background

Work (Mohammad et al., 2009; Abu-Jbara and Radev, 2011) in scientific document summarization have used citation sentences (also known as *citances*) from citing papers (hereafter, *CPs*) to create a multi document summary of a reference paper (hereafter, *RP*). Their approach followed a three-part process: finding the relevant documents, or the CPs, then selecting sentences which justify the citation in the RP and, finally, generating the summary. In our task, we have created a training corpus comprising human annotations for each of these sub-problems.

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

Human annotators identified the citances in each of (up to) ten randomly sampled CPs 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. Discourse facets indicate the type of information described in the reference span. E.g., "Aim" indicates that the citation concerns the aims of the reference paper. From our exploration of the computational linguistics domain, we observed that the discourse facets being cited were usually the aim of the paper, its methods and the results or implications of the work. We applied these observations in annotating discourse facets in our training corpus.

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[1]. 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 the resultant corpus from the Anthology, we aimed to enable citation-based summarization. To this end, with consultation from the main BiomedSumm task 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 version 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 Google 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 (Abu-Jbara et al., 2013).

We then used the title search facility of the ACL Anthology Network[2] (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 criterion 1) and 3). The final list was divided among the annotator group, who are a subset of the authors of this paper, from National University of Singapore and Nanyang Technological University, Singapore. Annotators re-used the resources created for BiomedSumm, which reduced the efforts required; however, a different set of discourse facets were used, which better represented the content of research papers in computational linguistics. 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.

3.1 Corpus Preprocessing

The original source text for the papers in the CL-Summ corpus was not sentence-segmented, which made it difficult to compute evaluation metrics. Both the clair_umich system and the TALN.UPF system

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

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

performed significant corpus pre-processing to create a sanitized, annotated dataset for their systems.

In the clair_umich system, for each RP, citing sentences were extracted from all its CPs. Each CP sentence was matched to the RP to create the final annotated dataset. Given a citing sentence, matching sentences from the RP were compared to the gold standard RP sentences to compute precision / recall. The average number of RP sentences matched for each CP sentence was 1.28 (with standard deviation 1.92). The maximum number of RP sentences matched for a CP sentence was 7. Given that the total number of RP sentences ranged from between 100 to 600, this made it a very challenging classification problem.

The UPF system performed a heavy sanitization process to overcome encoding issues in the corpus:

1. **A**utomatic PDF-to-text conversion: Conversion of PDF versions of the paper into text, by means of Poppler[3], a robust PDF-to-text convertor;
2. **M**anual verification of output: Manual validation of the PDF-to-text conversion errors in order to get a clean textual version of each paper;
3. **Sentence splitter and Sentence Sanitizer**: A rule-based sentence splitter and sanitizer was used to identify candidate sentences, and to remove incorrectly annotated sentences
4. **M**apping annotations to clean textual versions: Inspection of the textual contents of each of the annotation files, and manual mapping of the annotations to the clean textual version of each paper.

This resulted in two sanitized versions of the initial corpus that was shared as a part of this task. Both versions are shared, along with the original, in the official repository of the CL Corpus with the consent of the participants, and comprise a valuable contribution to the task.

4 The CL-Summ Task

This shared task proposes to solve the same problems posed in the BioMedSumm track, but in the domain of Computational Linguistics. It poses the research problem of building 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 reference paper (RP) and up to 10 citing papers (CPs). In each provided CP, the citations to the RP (or citances) have been identified and manually annotated. The information referenced in the RP is also annotated.
The Challenge: Output systems are required to perform the following tasks, where the numbering of the task corresponds to those used in the Biomed-Summ 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 (up to five sentences), and may be contiguous or not. It may also be a sentence fragment.
- Task 1B: Identify the discourse facet for every cited text span from a predefined set of facets. Discourse facets categorize the type of information described in the reference span. A maximum of three reference spans can be marked for every citance. In case these spans describe different different discourse facets, the most prevalent discourse facet is annotated.

Evaluation: Evaluate Task 1A 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.

An additional task in BioMedSumm, which was not advertised with this shared task, was:

Task 2: Generate a faceted summary of up to 250 words, of the reference paper, using itself and the citing papers.

5 Participating teams

Nine teams expressed an interest in participating in the shared task, and the three teams reported back with system descriptions and self-assessed results are listed below:

[3]http://poppler.freedesktop.org/

1. **clair_umich** from University of Michigan, Ann Arbor, USA. They proposed a supervised system based on lexical, syntactic and knowledge-based features to calculate similarity scores between sentences in the CPs and the RP.
2. **MQ**, from Macquarie University, Australia. They applied their system developed for the BiomedSumm Task, with the exception that they did not incorporate domain knowledge (UMLS). For Task 1A they used similarity metrics to extract the top n sentences from the documents. For Task 1B they used a logistic regression classifier. For the bonus Task 2 they incorporated the distances from Task 1A to rank the sentences.
3. **Taln.UPF**, from Universitat Pompeu Fabra, Spain. They adapted available summarization tools to scientific texts.

The system descriptions and task results for clair_umich, MQ and TALN.UPF are provided in the following sections.

6 The clair_umich System — Comparing Overlap of Word Synsets

6.1 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.

6.2 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:** Six wordnet-based similarity measures were combined to obtain six sentence similarity features(Banea et al., 2012):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). Using these measures, the similarities between two sentences was computed by 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:** Given a candidate sentence pair, two syntactic dependencies were considered equal if they had the same dependency type, govering lemma, and dependent lemma (Zhu and Lan, 2013). The Stanford parser was used to obtain dependency parses of all the citing sentences and source sentences. Then, 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 ranked 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.

Baseline – Using TF.IDF For the baseline system, the TF.IDF of all lowercased words was used, without removing stop words (similar to the clair_umich team). Separate TF.IDF statistics were

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

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)

computed for each reference paper, using the set of sentences in the paper and the citance text of all citing papers.

Adding texts of the same topic: Since the amount of text used to compute the TF.IDF was relatively little, it was presumed that citing papers are of the same topic. Accordingly the complete text of all citing papers was added in calculations for the IDF component.

Adding context: In order to extend the information on each sentence in the reference paper the text from the reference papers was added within a context window of 20 neighboring sentences to the target sentence.

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 previous paragraph. Then, the final score of a sentence was the combination of the similarity with the citance, and similarity with 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

In the TALN.UPF system, the following text analysis tools were first used to pre-process text in the sanitized document corpus:

1. **Tokenizer** and **POS tagger**: The GATE[4] tool and its ANNIE NLP tools for English were used to tokenize and tag the corpus.
2. **Sentence TF.IDF vector calculator**: A TF.IDF vector was generated for each sentence. The IDF values of the terms of each document were computed by considering the CPs and RP as a complete corpus.

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

For Task 1A, the TALN.U PF system implemented an algorithm to map every citation in the CP to one or more (up to three) **reference text spans** from the RP. Sentences from the CP that overlapped totally or partially with the citation text were selected and referred to as the *citation context* (CtxSent1,..., CtxSentN). For every sentence in the RP, the system associated a *score* equal to the sum of the TF.IDF vector cosine similarities computed between that sentence and each sentence belonging to the citation context (CtxSent1,..., CtxSentN). Finally, the top N sentences from the RP with the highest *score* were chosen as the reference span. Conflicts in choice were resolved by preferring sentences that occurred in the same document section in the RP. Furthermore, scores for referencing sentences were weighted by the prominence of document sections selected reference spans; for instance, if there 6.5% of all reference spans were sourced from the Abstract section, then the score for a sentence from the Abstract of the RP is multiplied by 0.065. The system evaluated the performance of the algorithm with varying values of N, the number of sentences included in the reference span.

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

Task 1B is cast as a sentence classification problem. From the corpus, the system selected sentences from the CPs that overlapped totally or partially with a manually annotated reference text span and they

[4] https://gate.ac.uk/ie/annie.html

were classified as the discourse facet of the manually annotated reference text span. This resulted in a set of 266 cited papers' sentences, each characterized by a discourse facet (see Table 1).

Docset	Citing papers
Aim	46
Hypothesis	1
Implication	25
Results	29
Method	165
TOTAL:	266

Table 1: Discourse facet of the sentences of cited papers belonging to a manually annotated reference text span.

Every sentence was modelled as a word vector of unigrams, bigrams, trigrams and lemmatized versions of all three. Order of stopwords was preserved as well. Sentence classification performance was compared across 3 classifiers – *Naive Bayes (NB)*, *SVM* with linear kernel and *Logistic Regression (LR)*. Results from a 10-fold cross validation over the set of CP sentences (see Table 1) are shown in Table 4. Clearly, LR performed the best with an averaged F_1 of 0.719.

9 Evaluation and Results

Three teams submitted their self-assessed results, using ROUGE (Lin, 2004) for task-1A. ROUGE (in specific, the ROUGE-L variant) is a popular evaluation method for summarization 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 sentences 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.

For Task 1a, the MQ and TALN.UPF systems were unsupervised while clair_umich system was supervised. The former two systems were evaluated over all 10 topics in a single run, while clair_umich reports cross validated performance over the 10 topics. For Task 1b, the TALN.UPF system also followed a supervised approach with 10-fold cross validation. The ROUGE-L scores have been calculated using the system output of a set of selected sentences as the system summary, and comparing their overlap against the target summaries are the sentences given by the annotators.

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

Results for Task 1: Table 3 shows the ROUGE-L F_1 scores of each individual reference document from the CL-Summ dataset.

Disc. facet	NB	SVM	**LR**
Aim	0.725	0.734	**0.732**
Method	0.706	0.826	**0.828**
Implication	0.049	0.000	**0.200**
Results	0.509	0,533	**0.533**
Hypothesis	0.024	0.000	**0.000**
WEIGHED AVG. F_1	0.623	0.698	**0.719**

Table 4: Task 1B TALN.UPF: F1-Score comparison of classification algorithms

Results for Task 2: 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 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 7.

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:

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

The summaries were evaluated using ROUGE-L, where the model summaries are the abstracts of the corresponding papers. Since paper X96-1048 of the SciSumm data did not have an abstract, it was omitted from this experiment.

An example excerpt from a target summary (Abstract) for the reference paper J03-3003 is:

MQ			clair_umich			TALN.UPF		
P	R	F_1	P	R	F_1	P	R	F_1
0.212	0.335	0.223	0.444	.574	0.487	0.194	0.344	0.225

Table 2: Task 1A performance for the participating systems expressed as ROUGE-L score averaged over all topics.

Paper ID	MQ	clair_umich	TALN.UPF	Paper ID	MQ System	clair_umich	TALN.UPF
C90-2039	0.235	0.635	0.180	J00-3003	0.196	0.559	0.263
C94-2154	0.288	0.536	0.200	J98-2005	0.101	0.344	0.196
E03-1020	0.239	0.478	0.198	N01-1011	0.221	0.498	0.254
H05-1115	0.350	0.375	0.233	P98-1081	0.200	0.367	0.211
H89-2014	0.332	0.546	0.275	X96-1048	0.248	0.535	0.240

Table 3: Task 1A ROUGE-L F1 scores for individual topics.

We describe a statistical approach for modeling dialogue acts in conversational speech, i.e., speech- act-like units such as STATEMENT, QUESTION, BACKCHANNEL, AGREEMENT, DISAGREEMENT, and APOLOGY. Our model detects and predicts dialogue acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialogue act sequence. The dialogue model is based on treating the discourse structure of a conversation as a hidden Markov model and the individual dialogue acts as observations emanating from the model states. Constraints on the likely sequence of dialogue acts are modeled via a dialogue act n-gram... We achieved good dialogue act labeling accuracy (65% based on errorful, automatically recognized words and prosody, and 71% based on word transcripts, compared to a chance baseline accuracy of 35% and human accuracy of 84%) and a small reduction in word recognition error.

The MQ System's output baseline summary for the same reference paper is 20 sentences long; below is an excerpt:

Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. In all these cases, DA labels would enrich the available input for higher-level processing of the spoken words. The relation between utterances and speaker turns is not one-to-one: a single turn can contain multiple utterances, and utterances can span more than one turn (e.g., in the case of backchanneling by the other speaker in midutterance). The most common of these are the AGREEMENT/ACCEPTS. One frequent example in our corpus was the distinction between BACKCHANNELS and AGREEMENTS (see Table 2), which share terms such as "right" and "yeah". Networks compare to decision trees for the type of data studied here. Neural networks are worth investigating since they offer potential advantages over decision trees.

Table 5 shows the breakdown of ROUGE-L F_1 scores per document.

10 Discussion

10.1 MQ System Performance: BioMedSumm Vs. CL-Summ

Table 2 compares the results of the MQ system's experiments with the SciSumm 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.

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 5: 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 6: ROUGE-L results of the MQ system runs for Task 1A.

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 7 shows the precision / recall for different values of the cosine threshold.

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

The F_1 scores seems to reach a maximum at the similarity threshold of 0.1. The recall at the threshold of 0.1 is 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.

10.3 Error Analysis for the Participating Systems

Some drawbacks were observed in the approach and evaluation for the MQ system. The example below illustrates the MQ system's output for Task 1a, for the reference paper H89-2014:

"The statistical methods can be described in terms of Markov models." "An alternative approach taken by Jelinek, (Jelinek, 1985) is to view the training problem in terms of a "hidden" Markov model: that is, only the words of the training text are available, their corresponding categories are not known." "In this regard, word equivalence classes were used (Kupiec, 1989)." The target sentence was: "The work described here also makes use of a hidden Markov model."

The first sentence of the sample output was very similar to the target sentence. It was not the best match, but it was a close match, and an evaluation metric such as ROUGE would reward it. On the other hand, the second sentence, even though it talked about HMMs, it was not strictly about the approach used by the paper and therefore it should not be rewarded with a good score. However, ROUGE would be too lenient here. This is one of the issues identified by the MQ system in following a purely lexical approach.

In the clair_umich 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 in Figure 2. Here, even

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.”

Figure 2: Lexically similar false positive sentences.

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.”

Figure 3: Implied example.

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 RP, as evident in Figure 3.

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.

- **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, thus making difficult the implementation of an automated homogeneous textual processing pipeline.
- **Content**: Some of the older PDF files, when parsed to text or XML, presented several text formatting issues: hyphenation problems, words not separated by blank spaces, page headers and footnotes included in the textual flow, 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 offsets:** 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. As a consequence, in order to retrieve the annotated texts, systems like the TALN-UPF manually searched through citing documents to identify the correct offset, and the clair_umich system created an automatic program to generate sentence offsets.
- **Discontiguous texts**: The use of "..." follows the BioMedSumm standard practice of indicating discontiguous texts, meaning that there was a gap between two text spans (citation spans or reference spans). The gap might be because text moves onto a new page. Sometimes there was a formula, page number or figure between two text spans 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 corpus:** The corpus comprised only a set of 10 topics, each with up to 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 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 Conclusion

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 implemented an unsupervised algorithm, while the clair_umich system decided on a supervised approach. For identifying referenced text spans in reference papers, clair_umich's supervised algorithm performed best, using lexical, syntactic and knowledge-based features to calculate the overlap between sentences in the citation span and the reference paper. Finally, an added experiment by the MQ team compared the baseline summaries of reference papers against gold standard summaries, based on TF.IDF calculations.

The clair_umich system incorporated WordNet synsets for expanding and comparing cited text with reference papers, and used syntactic features to further enrich overlap calculations.

In contrast, the MQ system relied exclusively on reading and comparing texts. Furthermore, the MQ system was originally built for the BioMedSumm task – however, they had to discard some domain-specific features for this task. We believe that the lack of domain knowledge, coupled with OCR-related and PDF parsing errors, affected the performance for our CL task.

This task is an initiative for encouraging the development of tools and approaches for scientific summarization. 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 more detailed analyses over 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.

13 Acknowledgments

This shared task 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. The authors also acknowledge and thank the BiomedSumm organizers – especially Lucy Vanderwende, Kevin B. Cohen, Prabha Yadav, and Hoa Trang Dang – for lending their expertise in organizing this pilot.

The **MQ system** was made possible thanks to a winter internship granted to Christopher Jones by the Department of Computing, Macquarie University.

The **clair_umich system** wishes to acknowledge the helpful suggestions of Ben King, Mohamed Abouelenien and Reed Coke.

The **TALN.UPF** system is supported by the project Dr. Inventor (FP7-ICT-2013.8.1 611383), programa Ramón y Cajal 2009 (RYC-2009-04291), and the project TIN2012-38584-C06-03 Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain.

References

Amjad Abu-Jbara and Dragomir R. Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proc of ACL: HLT-Vol. 1*, pages 500–509. ACL.

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R. Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proc. of NAACL*, pages 596–606. ACL.

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

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

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 *Proc. of 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. *WordNet: An electronic lexical database*, 49(2):265–283.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proc. of ICML*, pages 296–304.

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 R. Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proc. of NAACL*, pages 584–592. ACL.

Diego Mollá, Christopher Jones, and Abeed Sarker. 2014. Impact of citing papers for summarisation of clinical documents. In *Proc. of the Australasian Language Technology Workshop 2014 (ALTA '14)*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc of IJCAI - Vol. 1*, pages 448–453.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proc.of ACL*, pages 133–138. ACL.

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: Proc. of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 124–131. ACL.