ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, Dragomir R. Radev
Department of Computer Science, Yale University
{michihiro.yasunaga,r.zhang,alexander.fabbri,irene.li,dan.friedman,dragomir.radev}@yale.edu
jkasai@cs.washington.edu

Abstract

Scientific article summarization is challenging: large, annotated corpora are not available, and the summary should ideally include the article's impacts on research community. This paper provides novel solutions to these two challenges. We 1) develop and release the first large-scale manually-annotated corpus for scientific papers (on computational linguistics) by enabling faster annotation, and 2) propose summarization methods that integrate the authors' original highlights (abstract) and the article's actual impacts on the community (citations), to create comprehensive, hybrid summaries. We conduct experiments to demonstrate the efficacy of our corpus in training data-driven models for scientific paper summarization and the advantage of our hybrid summaries over abstracts and traditional citation-based summaries. Our large annotated corpus and hybrid methods provide a new framework for scientific paper summarization research.

Introduction

Fast-paced publications in scientific domains motivate us to develop automatic summarizers for scientific articles. Recent work in automatic summarization has achieved remarkable performance for news articles: Single-Document Summarization (Parveen, Ramsl, and Strube 2015; Cheng and Lapata 2016; See, Liu, and Manning 2017; Narayan, Cohen, and Lapata 2018), Multi-Document Summarization (Hong and Nenkova 2014; Cao et al. 2015; 2017). Scientific article summarization, on the other hand, is less explored, and differs from news article or other general summarization. For example, scientific papers are typically longer and contain more complex concepts and technical terms. Moreover, they are structured by section and contain citations.

To encourage research in scientific article summarization, several shared tasks have been organized recently: TAC 2014 (biomedical domain), CL-SciSumm 2016 (Jaidka et al. 2016) (computational linguistics domain; consisting of ACL Anthogoly papers). While these shared tasks have established a foundation for scientific paper summarization, their datasets are small, with just 30-50 articles. As understanding and annotating a scientific paper require domain-specific expert knowledge, annotation does not scale to a large corpus as compared to news articles, preventing us from applying data-driven approaches such as neural networks shown powerful in news article summarization (Cheng and Lapata 2016; See, Liu, and Manning 2017). In news article summarization, on the other hand, prior work (Woodsend and Lapata 2010; Cheng and Lapata 2016) has manually created gold summaries for 9,000 documents and extended them to 200K documents by heuristics. This type of annotation or crowdsourcing is not realistic for scientific papers due to their length and technical content.

Another characteristic of scientific papers is that they may have impacts that are not expected at the time of publication. For instance (Figure 1), the abstract of Bergsma and Lin emphasizes their techniques and improved performance in pronoun resolution, but a citation analysis reveals that their contribution to subsequent work lies largely in the noun gender dataset they created. While the abstract of a paper provides a solid summary of the *content* from the authors' point of view, it may fail to convey the actual *impact* of the paper on the research community. Additionally, the significance of a paper may change over time due to the progress and

Paper ID: P06-1005
Paper Title: Bootstrapping Path-Based Pronoun Resolution

Abstract:
We present an approach to pronoun resolution based on syntactic paths. Through a simple bootstrapping procedure, we learn the likelihood of coreference between a pronoun and a candidate noun based on the path in the parse tree between the two entities. This path information enables us to handle previously challenging resolution instances, and also robustly addresses traditional syntactic coreference constraints. Highly coreferent paths also allow mining of precise probabilistic gender/number information. We combine statistical knowledge with well known features in a Support Vector Machine pronoun resolution classifier. Significant gains in performance are observed on several datasets. **(mostly about technique)**

Citation Sentences:
Bergsma and Lin (2006) determine the like-lihood of coreference along the syntactic path connecting a pronoun to a possible antecedent, by looking at the distribution of the path in text. **(about technique)**

We use the approach of Bergsma and Lin (2006), both because it achieves state-of-the-art gender classification performance, and because a database of the obtained noun genders is available online. **(about both technique and dataset)**

For the gender task that we study in our experiments, we acquire class instances by filtering the dataset of nouns and their genders created by Bergsma and Lin (2006). **(about dataset)**

Figure 1: Abstract and citations of (Bergsma and Lin 2006). The abstract emphasizes their pronoun resolution techniques and improved performance; the citation sentences reveal that their noun gender dataset is also a major contribution to the research community, but it is not covered in the abstract.

evolution of research (Mei and Zhai 2008). In such situations our summary should ideally accommodate not only the major points highlighted by the authors (abstract) but also the views offered by the scientific community (citations).

This paper presents a novel dataset and summarization method to tackle the aforementioned problems in scientific paper summarization. Our corpus, which contains the citation network of ACL Anthology papers and human-written summaries for the 1,000 most cited papers, expands the existing CL-SciSumm project (Jaidka et al. 2016) and provides the largest manually-annotated dataset for scientific paper summarization. For each of the 1,000 papers (we call *reference papers*, or *RPs*), experts in CL/NLP read its abstract and incoming citation sentences to create a gold summary. This way, annotators can grasp broad, major aspects of the RP without reading the whole text, enabling faster annotation. We also conduct studies to validate that summaries created in this method are actually as comprehensive as summaries created by reading the full papers. Our dataset (1,000 papers) is significantly larger than the prior CL-SciSumm corpus (30 papers) and serves as a useful resource for supervised scientific paper summarization.

Further, we propose two novel summarization models for scientific papers that capture both the papers' *content* highlighted by the authors and *impact* perceived by the research community (hybrid summarization). In both models, given a reference paper (RP) to summarize, we take its abstract as the authors' insight, and identify a set of text spans (*cited text spans*) in the RP that are referred to by incoming citation sentences (i.e., community's views). The first approach then summarizes the union of the abstract and cited text spans, to integrate both components. The second approach, motivated by the fact that we already have the abstract as a clean self-summary of the paper, augments the abstract by adding salient texts extracted from the cited text spans (i.e., the community's views not covered in the abstract). For both approaches we also exploit the citation counts of the RP and its citing papers as an additional feature, to better reflect the authority of each work in the research community. To experiment with these two methods, we implement two neural network-based summarization models, which are also motivated by the architecture of Yasunaga et al. (2017)'s neural multi-document summarizer.

In evaluation, we use the CL-SciSumm shared task (Jaidka et al. 2016), an established benchmark for scientific paper summarization. This benchmark dataset contains gold summaries that are created by experts who read papers and their citation sentences. First, we find that our large training corpus enables neural summarizers to boost their performance and outperform all prior participants in the shared task. This confirms the usefulness of the proposed dataset. Second, we demonstrate that the proposed hybrid summarization methods can indeed incorporate both the authors' and research community's views, thereby producing more comprehensive summaries than abstracts. In summary, our contributions are as follows.

- A large manually-annotated corpus (1,000 examples) for scientific article summarization that facilitates research on supervised approaches.
- Novel scientific paper summarization methods that integrate both the authors' and research community's insights (hybrid summarization)

Background & Motivation

Text Summarization

Many existing summarization systems employ extractive methods to produce a summary, typically by ranking the salience of each sentence in a given document and then selecting sentences to be included in the summary (Erkan and Radev 2004; Parveen, Ramsl, and Strube 2015). Recently, in news article summarization, neural network-based approaches have proven successful (Cao et al. 2015; Cheng and Lapata 2016; Nallapati, Zhai, and Zhou 2017; See, Liu, and Manning 2017). This work presents neural network-based extractive models for scientific paper summarization.

Scientific Paper Summarization

Scientific paper summarization has been studied for decades (Paice 1981; Elkiss et al. 2008; Lloret, Romá-Ferri, and Palomar 2013; Jaidka et al. 2016; Parveen, Mesgar, and Strube 2016). While early work (Luhn 1958; Paice 1981; Paice and Jones 1993) focused on producing **content-based summaries** of target papers, the use of citations was later proposed to summarize target papers' contributions and lasting influence on the research community.

Citation-based summarization. Early work in citation-based summarization (Nakov, Schwartz, and Hearst 2004; Elkiss et al. 2008; Qazvinian and Radev 2008; Abu-Jbara and Radev 2011) aimed to summarize the contribution of a target paper (often called *reference paper*, or *RP*, in this context) by extracting a set of sentences from the citation sentences. We call a sentence that cites the RP a *citation sentence* (or *citing* sentence). A citation sentence can be viewed as a short summary of the RP written from the citing authors' perspective. Hence, a collection of citation sentences reflects the impact of the RP on the research community (Elkiss et al. 2008). While citation sentences provide the community's views of the RP, prior work (Siddharthan and Teufel 2007; Mei and Zhai 2008) pointed out issues in using citation sentences directly for summarization. In citing sentences, the discussion of the RP is often mixed with the content of the citing paper or with the discussion of other papers cited jointly, containing much irrelevant information.

To address such issues, recent work (Mei and Zhai 2008; Cohan and Goharian 2015; Jaidka et al. 2016; Li et al. 2017; Cohan and Goharian 2017a; 2017b) considers *cited text spans*-based summarization, where they identify a set of text spans (*cited text spans*; often a set of sentences) in the RP that its citing sentences refer to, and perform summarization on the identified text spans. This way, while the summary consists of words in the RP, it reflects the research community's insights. Experimental results in Mei and Zhai show that their cited text span-based model outperforms direct summarization of citing sentences. Cited text span-based summarization is also adopted as the default approach in two recent shared

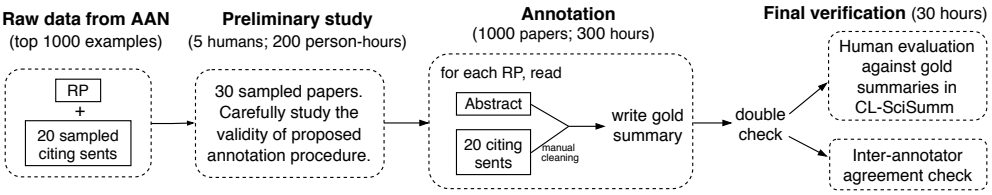

Figure 2: Overview of the dataset construction process.

tasks on scientific paper summarization: TAC 2014 and CL-SciSumm (Jaidka et al. 2016). Their datasets provide each RP and its incoming citation sentences, cited text spans, and a gold summary written by experts; the participants are asked to produce summaries using the RP and its citation sentences.

Our hybrid models. While the aforementioned citation-based summarization techniques inform us of the impact of an RP, they may overlook the authors' original message. For example, citation sentences (and consequently, cited text spans) often focus on the conclusion of the RP and may not cover other important aspects such as the motivation of the work. Moreover, in our preliminary study conducted on the CL-SciSumm shared task, we found that the quality of cited text span-based summaries produced by participants, often falls short of the abstracts in ROUGE evaluation against gold summaries. Conroy and Davis (2017) also find that the terms from abstracts in scientific documents often cover a large portion of human summaries. Our motivation in this work is therefore to integrate both the authors' original highlights (abstract) and research community's views (citations), and ultimately to improve upon the abstract.

Datasets. Previous datasets for scientific document summarization are small (Teufel and Moens; Jaidka et al.; TAC 2014), with only several dozen articles. Consequently, most of the existing summarizers for scientific papers are unsupervised or tuned on small data (Abu-Jbara and Radev 2011; Cohan and Goharian 2015; Li et al. 2016). In fact, in the previous CL-SciSumm shared task (30 data examples), no data-driven approaches like neural networks saw great success. The new dataset we introduce here (1,000 examples) is much larger than the prior CL-SciSumm corpus, enabling data-driven approaches to scientific paper summarization. In our experiments, we show that our dataset indeed allows neural network-based summarization models to outperform all prior participants in the shared task.

Recent work by Collins, Augenstein, and Riedel (2017) and Cohan et al. (2018) is related to ours in that they also introduce large-scale datasets and neural summarization models for scientific papers. Yet, while they focus on content-based summarization with automatically created gold summaries, our work constructs manually-annotated gold summaries as well as citation information to study the research community's view on each reference paper.

Dataset Construction

To overcome data scarcity in scientific paper summarization, we develop and release a manually-annotated, large-scale corpus for research papers in computational linguistics (CL). Our corpus contains the 1,000 most cited papers in the ACL Anthology Network (AAN) (Radev et al. 2013), their citation information, and gold summaries annotated by experts in the field. We follow the format of two prior datasets of scientific paper summarization, CL-SciSumm (Jaidka et al. 2016) and TAC 2014 (biomedical domain), so that systems trained / tested on our corpus can also be applied to or evaluated on those established datasets. Figure 2 depicts our data construction process.

Data Processing

We extract the 1,000 most cited papers and their citation sentences from AAN. The 1,000 papers have 21 - 928 citations in the anthology. For each of the RPs, we sample and clean 20 citation sentences, which are usually sufficient to study the research community's views of the RP (Mei and Zhai 2008). Specifically, following the prior datasets, we keep the oldest and latest citations and randomly sample the rest so that the 20 citations cover an extended period of time. We then remove inappropriate citation sentences (i.e., list citations, tables, those with bugs) and clean the rest, resulting in 15 citation sentences on average for each RP.

Annotation

We aim to develop gold summaries for the 1,000 papers in CL. In the prior datasets (CL-SciSumm and TAC 2014, containing $\sim$30 papers), gold summaries were prepared by humans with domain expertise, in the following manner (i.e., *expert* summaries): given a RP to summarize, the annotators read all the text of the RP and its citation sentences to grasp its content and impact, and wrote a comprehensive summary. Yet, due to the length and technical content of scientific papers, such annotation requires a significant amount of time as well as expertise, hindering the construction of a large-scale corpus for scientific article summarization.

To scale our annotation to the 1,000 papers, we develop a faster annotation procedure in this work. Five PhD students in NLP or people with equivalent expertise divide the 1,000 RPs, and read each paper's abstract and incoming citation sentences. Then, he or she identifies a few salient citation sentences that convey the RP's specific contributions not covered in the abstract and write a gold summary based on the abstract and selected citation sentences. This way, the annotators can save the time of reading the whole text of the

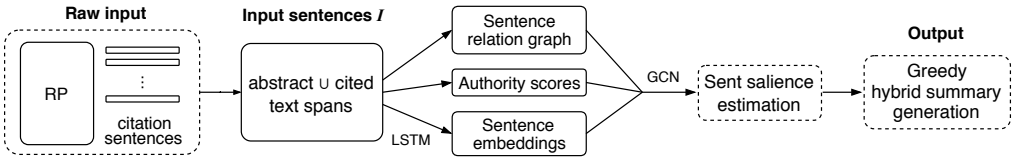

Figure 3: Overview of our summarization models.

RP, but can still grasp broad aspects of the paper to create a comprehensive summary. To take Bergsma and Lin (Figure 1) as an example again, while its abstract elaborates on the pronoun resolution techniques, its citation sentences reveal other major aspects of the RP such as the contribution of a noun gender dataset, which is discussed in some part of the RP but is not highlighted in the abstract. An expert summary would include both of these aspects to describe the RP. By reading the citation sentences in addition to the abstract, we can comprehend much of the content and impact of the RP without reading all its text.

Lastly, the citation sentence cleaning / selection process is double-checked to prevent mistakes.

Validation of our annotation procedure. Prior to annotation, we conducted preliminary studies on 30 sample papers. For each, we had the annotators list out the summary-worthy points they found 1) by reading the abstract + citing sentences, and 2) by reading the full paper; we observed that on average the former (our annotation method) covered over 90% of the major points found by reading the full papers, while just requiring 30% annotation time. This study suggests that by reading the abstract + citing sentences, annotators can create summaries comparable in quality to the ground truth in an inexpensive way.

Statistics. Our annotated summaries resulted in 151 words on average (similar to the gold summaries in the CL-SciSumm corpus, 150 words). To study the inter-annotator agreement on determining salient citations, we randomly picked 40 RPs and assigned another annotator; the Kappa coefficient (Cohen 1960) for inter-annotator agreement was 0.75, *substantial agreement*, on Landis and Koch scale. The high inter-annotator agreement further supports the efficacy of using abstract + citing sentences to create summaries.

Human evaluation. We also conducted human evaluation of our gold summaries against the gold summaries in the CL-SciSumm corpus, which were created by reading full papers. We studied the 15 papers in our corpus that already exist in CL-SciSumm. For each paper, we asked 5 computer science students who took an NLP course to evaluate which gold summary (ours or CL-SciSumm's) is more comprehensive, on an integer scale -2 to 2: 2 if the former (ours) is more comprehensive; -2 if the latter; 0 if they are similar; and 1, -1 are in between. The evaluated scores were 54% zero, 22% positive, and 20% negative, with average +0.02. This result indicates that our annotated summaries are comprehensive, and comparable to or slightly better than the summaries created by reading full papers.

The dataset construction took 600+ person-hours. This large corpus can be used to train scientific paper summarization models that utilize citations, facilitating research in supervised methods. In the next sections, we introduce data-driven hybrid summarization models and experiment on the proposed corpus.

Hybrid Summarization Models

Given a reference paper (RP) and its incoming citation sentences, our hybrid summarization models aim to reflect both the authors' and research community's voices on the RP. Specifically, we regard the abstract of the RP as the authors' original perspective, and obtain the research community's insights by identifying cited text spans in the RP (i.e., sentences in the PR that are referred to by the citation sentences). In this work, we consider the following two versions of hybrid summarization.

Hybrid 1: Summarizing the combination of the abstract and cited text spans

Hybrid 2: Augmenting the abstract with salient texts extracted from cited text spans

The motivation of Hybrid 2 is to build upon the clean self-summary provided by the authors and to add the community's views not covered in it. In both models, we take the union of the abstract and cited text spans as input I for summarization. Note that the input sentences in I (in particular, cited text spans) are not necessarily contiguous in the RP. This situation is analogous to multi-document summarization (MDS), which aims to produce a summary for a set of separate documents. Motivated by graph-based MDS methods (Erkan and Radev 2004; Yasunaga et al. 2017), we build a graph capturing the relations among the input sentences in I, and apply a Graph Convolutional Network (GCN) (Kipf and Welling 2017) on top to perform summarization.

Pre-processing

Given a reference paper (RP) and its incoming citation sentences, we first prepare the input sentences for summarization (i.e., abstract ∪ cited text spans), and build their sentence relation graph.

Cited text spans. We extract cited text spans in the RP for each incoming citation sentence, and then compile them for all the given citations. To identify cited text spans for a given citation sentence, we choose top two sentences in the RP

that are most similar in terms of the tf-idf cosine similarity measure (stop words excluded).

We repeat the extraction for all the given citation sentences, and take the union to construct the complete cited text spans of the RP. The union of the abstract and cited text spans of the RP will be the input I for summarization. In our experiments I contained about 40 sentences on average.

Sentence relation graph. We build a graph that takes the input sentences as nodes and captures their relationships via edges. We adopt the widely-used cosine similarity graph (Erkan and Radev 2004), where every pair of sentences has an edge with a weight equal to their tf-idf cosine similarity.

Authority feature. While cited text spans provide insights by the research community, they do not necessarily reflect the authority of each citation. Mei and Zhai (2008) argue that a citation made by a highly authoritative paper should be weighted more than that made by a less authoritative paper. To better reflect the authority in the research community, we consider an extra feature (authority score) for each cited text span, which is the sum of its citing papers' citation counts. Sentences in the abstract are given the citation count of the RP. We obtain citation counts from the ACL Anthology Network (Radev et al. 2013).

Main Architecture

Given the input sentences and their relation graph, we apply a GCN (Kipf and Welling 2017) to encode the whole input text together with the graph and to estimate the salience of each sentence in the global context. Based on the salience scores, Hybrid 1 and 2 employ two greedy heuristics to select sentences to be included in the summaries.

Graph convolutional network (GCN). GCNs are neural networks that operate on graphs to induce node features based on graph structure. GCNs have been shown effective not only in node classification tasks (Kipf and Welling 2017), but also in NLP applications such as syntactic tree-based sentence encoding (Marcheggiani and Titov 2017).

Given a graph G with N nodes, a GCN takes

- $\tilde{A} \in \mathbb{R}^{N \times N}$, the adjacency matrix of graph G with added self-connections.
- $X \in \mathbb{R}^{N \times D}$, the input node features (D is the dimension of the feature vector for each node).

and outputs high-level node features, $Z \in \mathbb{R}^{N \times D}$, which encode the graph structure. The function takes a form of layer-wise propagation. Specifically, in an L-layer GCN, the propagation from the l-th layer to the $(l+1)$-th layer is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (1)$$

where $H^{(l)} \in \mathbb{R}^{N \times D}$ denotes the l-th hidden layer, with $H^{(0)} = X, H^{(L)} = Z$. The adjacency matrix $\tilde{A}$ is normalized via the degree matrix $\tilde{D}$. σ is an activation function such as $\tanh$. $W^{(l)}$ is the learnable parameter in the l-th layer.

Sentence encoding. Given the input sentences $\{s_1, s_2, \ldots, s_N\}$ in I and their relation graph G, we first encode each sentence s_i by applying a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) on its word embeddings, and taking the final state of the LSTM as its initial sentence embedding, $\mathbf{x}_i \in \mathbb{R}^{D-1}$. The authority score of sentence s_i can be appended to $\mathbf{x}_i$ as an additional feature. The sentence embeddings $\mathbf{x}_i \in \mathbb{R}^D$ $(i=1,2,\ldots,N)$ are then grouped as a node feature matrix $X \in \mathbb{R}^{N \times D}$, and fed into a GCN with the adjacency matrix $\tilde{A}$ of the sentence relation graph G. Through multiple layers of propagation, the GCN encodes the whole input text and induces higher-level sentence embeddings based on the structure of G. The output of the GCN, $Z \in \mathbb{R}^{N \times D}$, gives updated sentence embeddings $\mathbf{s}_i \in \mathbb{R}^D$ that incorporate the global context.

Salience estimation. For each sentence s_i in our input, we estimate its salience score as follows:

$$\hat{R}(s_i) = \frac{\exp(\mathbf{v}^T\mathbf{s}_i)}{\sum_{s_j \in I}\exp(\mathbf{v}^T\mathbf{s}_j)} \quad (2)$$

where $\mathbf{s}_i$ is the updated embedding of sentence s_i, and $\mathbf{v}$ is a learnable parameter for projecting embeddings to be scalar scores. Note that the salience scores are normalized via softmax to be a probability distribution over all the input sentences.

Training

The model parameters include the weights in the LSTM and GCN, and $\mathbf{v}$. The model is trained to minimize the cross-entropy loss between the target salience scores (true labels) R and the estimated salience scores $\hat{R}$ of the input sentences:

$$L = -\sum_{s_i \in I} R(s_i)\log(\hat{R}(s_i)) \quad (3)$$

To construct the target scores R, we first take the average of ROUGE-1 & 2 scores for each sentence s_i evaluated with the gold summary (Cao et al. 2015), and then rescale the scores as a probability distribution over all the input sentences.

Summary Generation

Based on the salience scores estimated for the input sentences, Hybrid 1 and 2 employ two greedy heuristics to select sentences for the summaries.

Hybrid 1 (*extractive summarization of abstract* ∪ *cited text spans*). First, we sort all sentences in I in descending order of the salience score. We dequeue one sentence from the list and append it to the current summary if the sentence is of a reasonable length (more than 8 words, as in (Erkan and Radev 2004)) and is non-redundant. A sentence is redundant if it is similar to any sentence already in the summary, with tf-idf cosine similarity above 0.5 (Hong and Nenkova 2014). We keep adding sentences to the summary in this way until we reach the length limit. Finally, sentences in the summary are sorted in the original order in the RP.

Hybrid 2 (*augmentation of abstract with salient cited text spans*). We take all the cited text spans from I and sort them in descending order of the salience scores. Starting from the full abstract as the initial summary, we deque one sentence from the list of cited text spans and add to the current summary if it is of a reasonable length and is non-redundant. We repeat until the length limit, and finally sort the summary sentences in the original order in the RP.

Experiments

We experiment the hybrid summarization models on our training corpus to study the efficacy of the proposed dataset and models. We aim to show that our large-scale corpus allows the data-driven neural models to outperform prior work. We also analyze the outputs of hybrid summarization and illustrate their advantage over abstracts and traditional citation-based summaries.

Datasets & Evaluation

We train the GCN summarization models on our proposed corpus with 1,000 examples of RPs, citation sentences, and gold summaries. All models are validated and tested on established benchmarks, CL-SciSumm 2016 dev/test, where the gold summaries were created by experts reading full papers. In training, we exclude the few RPs in our corpus that also appear in the validation or test set.

We evaluate system summaries against the gold summaries by ROUGE (Lin 2004), which serves as a good metric for this work, as we aim to measure the comprehensiveness of summaries. To ensure comparability with the CL-SciSumm shared task, we measure ROUGE-2 Recall, F1 (2-R, 2-F) and -SU4 F1 (SU4-F), with the same configurations: -n 4 -2 -4 -u -m -s -f A.

Experimental Design

We conduct the following two experiments to study the proposed 1) corpus and 2) hybrid methods.

Exp 1. First, we study the usefulness of our dataset for data-driven models, by comparing the model performance after training on our corpus and after training on the existing CL-SciSumm corpus. For data-driven systems, we experiment with our GCN model. As the participants in the CL-SciSumm shared task adopted cited text span-based summarization, to ensure a fair comparison, we also let these models just summarize cited text spans. Specifically, we just select cited text spans, given the predicted salience scores (we call this *GCN Cited text spans*). We follow the same protocol as the shared task (no authority feature; summary length 250 words).

Exp 2. Next, we study the efficacy of the hybrid summarization models. As our goal is to learn to produce the gold summaries (average length 150 words) and compare them with abstracts[1] or traditional citation-based summaries, we experiment with the GCN Hybrid models with summary length 150 words (with/without authority feature), and analyze the output hybrid summaries against those baselines.

[1]The average length of abstracts is 110 words.

Summarizer	2-R	2-F	3-F	SU4-F
Trained on Our Corpus (size: 1000)				
GCN Hybrid 2 (Ours)	41.69*	29.30*	24.65*	18.56*
GCN Hybrid 1 (Ours)	36.47*	26.31*	21.33*	16.18*
GCN Cited text spans (Ours)	33.03*	23.49*	17.86*	14.15*
Trained on CL-SciSumm (size: 30)				
GCN Cited text spans (Ours)	24.93	18.46	12.77	12.21
Best participant 1	32.36	21.94	16.79	13.63
Best participant 2	26.67	18.85	12.83	12.45

*: higher than all models trained on the CL-SciSumm corpus.

Table 1: Results of **Exp 1**, showing ROUGE evaluations on the CL-SciSumm Test benchmark. Models trained on our corpus outperform all the models trained on the existing CL-SciSumm Train set.

Summarizer	2-R	2-F	3-F	SU4-F
Abstract	29.52	29.40	23.16	23.34
GCN Hybrid 2 w/ auth	**33.88**	**31.54**	**24.32**	**24.36**
GCN Hybrid 2	32.44	30.08	23.43	23.77
GCN Hybrid 1 w/ auth	29.65	28.05	21.83	20.22
GCN Hybrid 1	29.64	27.96	21.81	19.41
GCN Cited text spans w/ auth	26.30	24.39	18.85	17.31
GCN Cited text spans	25.16	24.26	18.79	17.67

w/ auth: using authority feature.

Table 2: Results of **Exp 2**, showing ROUGE evaluations on the CL-SciSumm Test benchmark. All models are trained on our corpus. The hybrid models outperform abstracts and pure citation summaries.

Training Details

We use 100-dimensional word embeddings for the input to the LSTM sentence encoder. The word embeddings are initialized with GloVe (Pennington, Socher, and Manning 2014). We set the dimension of the LSTM/GCN hidden states to be 200, 201 (i.e., $D=201$), and use two hidden layers for the GCN (i.e., $L=2$). We apply dropout (Srivastava et al. 2014) to the input word embeddings as well as the outputs of the LSTM and GCN, with dropout rate 0.5.

The model parameters and word embeddings are trained by the Adam optimizer (Kingma and Ba 2015), with batch size 5, learning rate 0.001, and a gradient clipping of 2.0 (Pascanu, Mikolov, and Bengio 2012). We employ early stopping (Caruana, Lawrence, and Giles 2001) based on the validation loss to prevent overfitting.

Results & Discussion

Exp 1. Table 1 shows the result of Exp 1, along with the top two participants in the CL-SciSumm shared task (Li et al.; Conroy and Davis). The upper part shows the model performance after training on our proposed corpus (1000 examples), and the lower part the existing CL-SciSumm corpus (30 examples). We find that the neural model, *GCN Cited text spans*, performs on par with the participants when

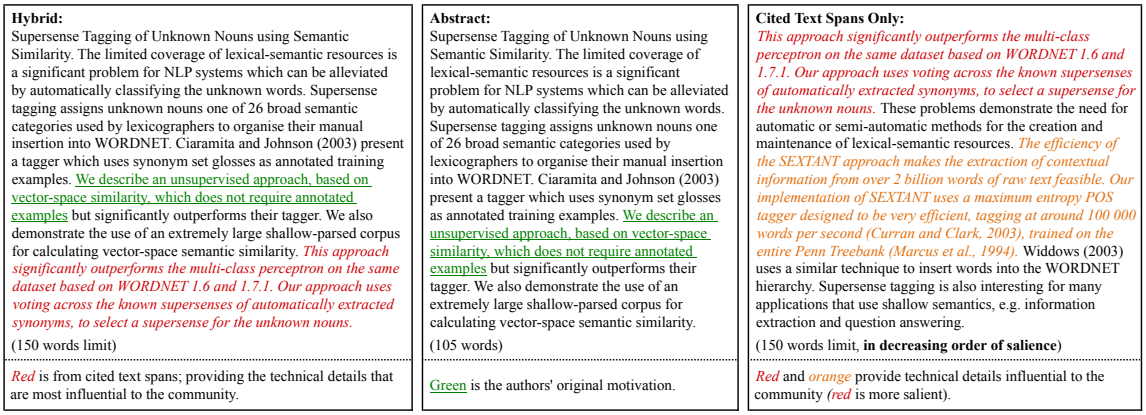

Figure 4: Comparison of our hybrid summary with the abstract and pure cited text spans summary, for paper P05-1004 in the CL-SciSumm 2016 test set. Our hybrid summary covers both the authors' original motivations (green) and the technical details influential to the research community (*red*).

trained on CL-SciSumm, but when trained on our corpus, it gains significant boosts in all the ROUGE metrics (e.g., +5 in ROUGE-3-F) and greatly outperform all the models trained on CL-SciSumm. With orders of magnitude more training examples than prior datasets, our corpus actually enables the data-driven neural network-based models to perform well on scientific paper summarization. This result suggests both the usefulness of the proposed corpus for training, and the feasibility of neural models in summarization given sufficient data.

Exp 2. Table 2 shows the result of Exp 2, along with the baselines (Abstract and GCN Cited text spans). We observe that both of the hybrid models perform clearly better than pure cited text span summaries. Moreover, Hybrid 1 surpasses abstracts in Recall, and Hybrid 2 outperforms abstracts in all ROUGE metrics, including the F1 of R-2 and -3, which have the highest correlation with human judgments (Cohan and Goharian 2016). Hybrid 2 performs better than Hybrid 1, most likely because Hybrid 2 builds on existing summaries (abstracts) and can ensure higher quality. In this experiment, Hybrid 2 added two sentences on average to the original abstract.

To qualitatively study the advantage of the hybrid summarization, we also compare and analyze the output summaries. As an example, Figure 4 shows the output summary of Hybrid 2 together with the abstract and pure cited text span summary for paper P05-1004 in the CL-SciSumm 2016 test set. The hybrid summary, which augments the abstract by taking in the most salient cited text spans, includes the technical contributions that are most influential to the community but are not covered in the abstract (*red*). The cited text span-based summary, on the other hand, provides more technical details, but lacks some of the author's original messages such as the motivation and objective of their work (green). Thus, the hybrid summary is indeed more comprehensive than the abstract and cited text span summary because it incorporates both the authors' original insights and the community's views on the paper.

Human evaluation. The above evaluation and analysis show the advantage of the hybrid models over the baselines. Here we conduct human evaluation of the hybrid summaries against gold summaries to study their utility. We asked 5 computer science students who took an NLP course to evaluate the coverage and coherence of the output summaries by our hybrid model, in a scale 1-5 (5 is the level of gold summaries). The model achieved 4.5 and 4.2 on average for these two metrics. While there is room for improving coherence, these scores suggest that the model can generate comprehensive and readable summaries.

Finally, we observe in Table 2 that all our models obtain moderate improvements by introducing the authority feature to reflect the authority of each citation made by the research community, suggesting the usefulness of this feature.

Conclusion

We proposed a novel dataset and hybrid models for scientific paper summarization. Our corpus, which contains 1,000 examples of papers, citation information and human summaries, is orders of magnitude larger than prior datasets and facilitates future research in supervised scientific paper summarization. We also presented hybrid summarization methods that integrate both authors' and community's insights, to overcome the limitations of abstracts (may not convey actual impacts) and traditional citation-based summaries (may overlook authors' original messages). Our experiments demonstrated that 1) the proposed dataset is indeed effective in training data-driven neural models, and that 2) the hybrid models produce more comprehensive summaries than abstracts and traditional citation-based summaries. We hope that our large annotated corpus and hybrid methods would open up new avenues for scientific paper summarization.

Acknowledgements

We thank Kokil Jaidka, Muthu Kumar Chandrasekaran, Min-Yen Kan, Yavuz Nuzumlali, Arman Cohan, as well as all the anonymous reviewers for their helpful feedback. We also thank everyone who helped the evaluation in this work.

References

Abu-Jbara, A., and Radev, D. R. 2011. Coherent citation-based summarization of scientific papers. In *ACL*.

Bergsma, S., and Lin, D. 2006. Bootstrapping path-based pronoun resolution. In *ACL*.

Cao, Z.; Wei, F.; Dong, L.; Li, S.; and Zhou, M. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*.

Cao, Z.; Li, W.; Li, S.; and Wei, F. 2017. Improving multi-document summarization via text classification. In *AAAI*.

Caruana, R.; Lawrence, S.; and Giles, C. L. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *NIPS*.

Cheng, J., and Lapata, M. 2016. Neural summarization by extracting sentences and words. In *ACL*.

Cohan, A., and Goharian, N. 2015. Scientific article summarization using citation-context and article's discourse structure. In *EMNLP*.

Cohan, A., and Goharian, N. 2016. Revisiting summarization evaluation for scientific articles. In *LREC*.

Cohan, A., and Goharian, N. 2017a. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In *SIGIR*.

Cohan, A., and Goharian, N. 2017b. Scientific document summarization via citation contextualization and scientific discourse. *IJDL*.

Cohan, A.; Dernoncourt, F.; Kim, D. S.; Bui, T.; Kim, S.; Chang, W.; and Goharian, N. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL-HLT*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*.

Collins, E.; Augenstein, I.; and Riedel, S. 2017. A supervised approach to extractive summarisation of scientific papers. In *CoNLL*.

Conroy, J., and Davis, S. 2015. Vector space and language models for scientific document summarization. In *NAACL-HLT*.

Conroy, J. M., and Davis, S. T. 2017. Section mixture models for scientific document summarization. *IJDL*.

Elkiss, A.; Shen, S.; Fader, A.; Erkan, G.; States, D.; and Radev, D. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *JASIST*.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Hong, K., and Nenkova, A. 2014. Improving the estimation of word importance for news multi-document summarization. In *EACL*.

Jaidka, K.; Chandrasekaran, M. K.; Rustagi, S.; and Kan, M.-Y. 2016. Overview of the cl-scisumm 2016 shared task. In *BIRNDL*.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kipf, T., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Li, L.; Mao, L.; Zhang, Y.; Chi, J.; Huang, T.; Cong, X.; and Peng, H. 2016. Cist system for cl-scisumm 2016 shared task. In *BIRNDL*.

Li, L.; Mao, L.; Zhang, Y.; Chi, J.; Huang, T.; Cong, X.; and Peng, H. 2017. Computational linguistics literature and citations oriented citation linkage, classification and summarization. *IJDL*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*.

Lloret, E.; Romá-Ferri, M. T.; and Palomar, M. 2013. Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*.

Marcheggiani, D., and Titov, I. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP*.

Mei, Q., and Zhai, C. 2008. Generating impact-based summaries for scientific literature. In *ACL-08: HLT*.

Nakov, P. I.; Schwartz, A. S.; and Hearst, M. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *SIGIR*.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL*.

Paice, C. D., and Jones, P. A. 1993. The identification of important concepts in highly structured technical papers. In *SIGIR*.

Paice, C. D. 1981. The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases. In *SIGIR*.

Parveen, D.; Mesgar, M.; and Strube, M. 2016. Generating coherent summaries of scientific articles using coherence patterns. In *EMNLP*.

Parveen, D.; Ramsl, H.-M.; and Strube, M. 2015. Topical coherence for graph-based extractive summarization. In *EMNLP*.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2012. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Qazvinian, V., and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. In *COLING*.

Radev, D. R.; Muthukrishnan, P.; Qazvinian, V.; and Abu-Jbara, A. 2013. The ACL anthology network corpus. *Language Resources and Evaluation* 1–26.

See, A.; Liu, P.; and Manning, C. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Siddharthan, A., and Teufel, S. 2007. Whose idea was this, and why does it matter? attributing scientific work to citations. In *HLT-NAACL*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR* 15:1929–1958.

Teufel, S., and Moens, M. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational linguistics* 28(4):409–445.

Woodsend, K., and Lapata, M. 2010. Automatic generation of story highlights. In *ACL*.

Yasunaga, M.; Zhang, R.; Meelu, K.; Pareek, A.; Srinivasan, K.; and Radev, D. R. 2017. Graph-based neural multi-document summarization. In *CoNLL-2017*.