Divide and Conquer: From Complexity to Simplicity for Lay Summarization

Rochana Chaturvedi[1], Saachi[2*], Jaspreet Singh Dhani[2*], Anurag Joshi[2*], Ankush Khanna[2*], Neha Tomar[2*], Swagata Duari[2], Alka Khurana[2] and Vasudha Bhatnagar[2]

[1]Keshav Mahavidyalaya, University of Delhi
[2]Department of Computer Science, University of Delhi

{*rochana.chaturvedi, saachi.mcs19.du, jaspreet.mcs19.du, anuragjoshi.mca19.du, ankush.mcs19.du, neha.mcs19.du*}@gmail.com, {*sduari, akhurana, vbhatnagar*}@*cs.du.ac.in*

Abstract

We describe our approach for the 1st Computational Linguistics Lay Summary Shared Task CL-LaySumm20. The task is to produce non-technical summaries of scholarly documents. The summary should be within easy grasp of a layman who may not be well versed with the domain of the research article. We propose a two step *divide-and-conquer* approach. First, we judiciously select segments of the documents that are not overly pedantic and are likely to be of interest to the laity, and over-extract sentences from each segment using an unsupervised network based method. Next, we perform abstractive summarization on these extractions and systematically merge the abstractions. We run ablation studies to establish that each step in our pipeline is critical for improvement in the quality of lay summary. Our approach leverages state-of-the-art pre-trained deep neural network based models as zero-shot learners to achieve high scores on the task.

1 Introduction

Acceptance of science by society is accelerated by sharing scientific knowledge and engaging with the public at large. Scientifically backed information, when suitably summarized and conveyed to the common man, spurs empowerment to combat the spread of misinformation. Lay summary of a scientific scholarly text, targeted for the general public, captures the broad scientific idea and its potential impact with minimal technical jargon. Funding agencies, scientists within and outside the field, and science journalists also benefit from lay summaries (Kuehne and Olden, 2015).

CL-LaySumm20 shared task aims to develop NLP methods to bridge the gap between advances made by the scientific community and non-specialist audience, by summarizing scholarly scientific articles in language understandable by lay persons. Evaluation for the task is done on the basis of Recall and F1-scores of ROUGE-1, -2, and -L metrics (Lin, 2004). Additionally, selective summaries are evaluated by science journalists and communicators for ease of comprehension as well as for interestingness. Chandrasekaran et al. (Forthcoming) document the results and insights from the shared task.

1.1 Abstractive Vs. Extractive Summarization

Automatic summarization of generic documents is accomplished by either using *Extractive* or *Abstractive* approach. *Extractive* summarization algorithms rank salient sentences in the input text, and subsequently select top ranked sentences for inclusion in summary. These algorithms effectively identify sentences containing important facts, but often suffer from weak coherence. An extractive summary, which is more like bullet points, does not compare favourably with human written summary, which is a cohesive piece of text generally written after paraphrasing and fusing different sentences or phrases from the text. Overall coherence between sentences in an extractive summary depreciates because of severe loss of context and several dangling anaphora (Antunes et al., 2018).

With rapid and remarkable developments in neu-

[*]These authors have equal contribution to this work.

Proceedings of the First Workshop on Scholarly Document Processing, pages 344–355

https://doi.org/10.18653/v1/P17

ral language models, *Abstractive* summarization algorithms have gained traction. These models are trained on sequence to sequence text generation (Sutskever et al., 2014) and are able to generate high quality natural language texts. They are competent to abstract long sentences into short and meaningful sentences, and are germane enough to introduce novel expressions and paraphrases while maintaining almost human-like quality. The current state-of-the-art neural abstractive summarizers are based on transformers (Vaswani et al., 2017), which use self-attention mechanism to allow contextual encoding of input sequence.

A major shortcoming of transformer based models is that their memory requirement and computational cost depends quadratically on the length of input sequence. A two stage extractive-abstractive pipeline is usually proposed to alleviate this shortcoming including in the prominent works of Chen and Bansal (2018), Gehrmann et al. (2018) and Zhao et al. (2020). Extractive step before abstraction has also been deemed important to improve the content selection in abstractive summaries (Liu and Liu, 2009; Mehdad et al., 2014).

1.2 Lay Summarization

Dubé and Lapane (2014) provide a checklist for manually writing lay summary for specified audience, which serves as a desiderata for designing algorithms for lay summarization. Manually translating complex research ideas into lay language incurs extensive patience, time, subject knowledge and effort. This has motivated research in the area of automatic *lay summarization*, which aims at condensing core ideas of scientific research and transforming them in accessible language for lay audiences, while remaining true to science.

Extractive summarization is insufficient for the task of lay summarization because of two reasons. First, when the sentences are selected for inclusion in summary they carry the burden of scientific jargon along with them, which degrades readability and comprehension for lay audience. Second, loss of contextual information and the consequent lack of coherence seriously detriments the purpose of lay summary.

As discussed earlier, state-of-the-art transformer based neural abstractive summarizers do not scale well for documents exceeding 1000 sequence tokens (Zhao et al., 2020). As scholarly articles are usually much longer, abstractive summarizers cannot be effectively used standalone for the CL-Laysum20 task.

Lay Summarization can benefit from tactfully exploiting the strengths of extractive and abstractive summarization, while renouncing their respective caveats. Distilling important sentences conveying core scientific ideas from the paper using extractive summarization, and feeding it to state-of-the-art abstractive summarizer has potential to yield desired non-technical summary of the scientific article in simple and understandable language.

1.3 Our Approach

We propose a two step approach that *divides* the scientific scholarly text into segments to *conquer* the *complexity* before generating *simple* lay summary. Following the heuristic advanced by Collins et al. (2017a) that certain sections of the document are more pertinent from the summarization viewpoint, we exploit the structure of scientific scholarly text to select information rich segments. We discerningly combine state-of-the-art extractive and abstractive summarization methods, to first extract important sentences from the selected segments, and then compress and paraphrase these sentences via abstractive summarizer. Subsequently, we combine the summaries in a rule based manner to obtain the final lay summary. We report systematic ablation studies to demonstrate the benefit of (i) using abstraction after extraction, and (ii) focusing on specific sections for lay summarization.

2 Background and Related Work

Earlier works on summarization of scientific articles aim to automatically produce the summary for researchers from multiple perspectives that complement each other. These cover automatic creation of abstract (Luhn, 1958; Lloret et al., 2013), extraction of keywords (Duari and Bhatnagar, 2019; Campos et al., 2020), title generation (Putra and Khodra, 2017), extraction of highlights (Collins et al., 2017b; Cagliero and La Quatra, 2020), query-focused summarization (Erera et al., 2019) and citation based summarization of articles (Cohan and Goharian, 2018; Yasunaga et al., 2019).

Various supervised and unsupervised techniques have been used so far for accomplishing distinctive tasks pertinent to scientific articles (Altmami and Menai, 2020). Recently, Miller (2019) propose to leverage the state-of-the-art BERT model (Devlin et al., 2018) for extractive summarization

of lectures. In this approach, K-means clustering is performed on sentence embeddings obtained from BERT, and the sentences that are closest to cluster centroids are extracted to create the summary. Among non-neural models, a popular approach is to capture relations between sentences or word phrases via a weighted graph. Gupta et al. (2014, 2019) model the sentences of the document as nodes of a weighted directed graph and compute idf based entailment scores between sentence pairs. They use weighted minimum vertex cover to extract most salient sentences.

Most recent neural abstractive summarizers are trained on masked language modeling task where random sequences of inputs are masked and the model learns to reproduce the masked portions of text. One such model that has achieved state-of-the-art results on abstractive summarization datasets is BART (Lewis et al., 2019). BART is an autoencoder which is pretrained to reproduce the original input after it has been corrupted with arbitrary noise. BART uses transformer (Vaswani et al., 2017) based architecture that employs self-attention mechanism to allow contextual encoding of input sequence.

3 Data

The organizers provide training and validation corpora for CL-LaySumm20 task, named Laysumm2 (215 documents) and Batch3 (357 documents). These documents comprise abstracts and full texts of scholarly articles from epilepsy, archaeology, and materials engineering domains. Each document in the two corpora is accompanied with a gold-standard lay summary. The test set contains 37 documents (abstracts and full texts). Table 1 presents basic statistics for the training, validation, and test datasets.

Stats	Dataset				
	Fulltext + Abstract			Gold-Summary	
	Laysumm2	Batch3	Test	Laysumm2	Batch3
$\mathbf{N_{avg}}$	5493	4803	6125	116	93
$\mathbf{NS_{avg}}$	230	109	272	5	3
$\mathbf{S_{avg}}$	24	46	23	23	31

Table 1: Descriptive statistics for complete text and gold-standard summaries of training and test corpora. N_{avg}: average document length in words, NS_{avg}: average number of sentences in documents, and S_{avg}: average sentence length in words.

4 Methodology

Our approach is based on the premise that not all sections of scientific scholarly text are equally comprehensible to non-experts. Gist of the scientific ideas and the important findings are concentrated in *Abstract* and *Conclusion* sections, while most of the technical details of the research are liberally spread in sections describing methodology and experimentation. *Introduction* and *Discussion* sections lie somewhere in between the spectrum.

Based on the intuition that *Abstract, Introduction and Conclusion* sections in scientific scholarly text are information rich, Kavila and Radhika (2015) construct summaries sourced from these sections. Collins et al. (2017a) argue that the *Abstract*, being an author generated summary is most important section in a paper. Using corpus of 10K computer science research papers, they empirically compare the overlap between different sections and paper highlights. It is reported that among *Abstract, Conclusion, Discussion* and *Introduction* sections (ACDI), *Introduction* section shows least overlap. The authors attribute low importance of *Introduction* section to its longer length.

We empirically test this conjecture for lay summaries. We divide the scientific document in two parts - (i) combined ACDI text, and (ii) rest of the document and compute the ROUGE scores of the two parts[1] with respect to the gold standard summary. Table 2 shows the result of the experiment for both corpora, affirming the observations documented by Collins et al. (2017a). For both corpora the combined ACDI sections, despite being shorter, boast of higher ROUGE scores compared to the remaining text. The results confirm that ACDI sections are apposite for generating summaries from layman perspective.

	Section	$\mathbf{N_{avg}}$	1F	1R	2F	2R	LF	LR
L	ACDI	1581	**12.92**	**90.57**	**7.98**	**57.32**	**9.07**	**65.08**
	Rest	3720	9.85	82.02	3.52	36.19	5.75	53.04
B	ACDI	1901	**8.35**	**91.66**	**4.63**	**54.13**	**5.76**	**65.23**
	Rest	2406	6.30	83.17	2.49	35.55	4.11	56.44

Table 2: Average ROUGE scores of combined ACDI sections vs rest of the text for Laysumm2 (L) and Batch3 (B) datasets.

4.1 Section-wise Analysis

We further study the relative importance of each of these four sections from lay summary perspective

[1]We use pre-processed text for this analysis.

and present our results in Table 3. Note that all research papers in the corpora are not structured uniformly and there is a variation in the sections present in a paper. Column N_{Doc} shows the number of documents that contain the particular section. All ROUGE scores are computed over the *existing* sections in the documents.

For Laysumm2 corpus, *Abstract* consistently exhibits high ROUGE scores, except for slightly better ROUGE-recall of *Introduction*. Length of the *Introduction* section, which is almost four times that of abstract, possibly begets this advantage. *Discussion* and *Introduction* sections, which have similar average lengths score comparably. *Conclusion*, the shortest section, displays relatively higher F-score for its length. Its low recall score is clearly due to its short length.

Documents in Batch3 corpus evince different trend in ROUGE scores due to difference in the lengths of the sections. *Discussion* section is strikingly longer compared to others, gaining higher recall scores. Interestingly, the gain due to length is annulled by F scores, which are the lowest among the four sections. *Abstract* consistently earns second highest score, despite short length.

	Section	N_{Doc}	N_{avg}	1F	1R	2F	2R	LF	LR
L	A	205	210	**49.33**	69.91	**29.20**	**41.71**	**36.74**	**52.12**
	C	175	326	31.78	56.94	10.10	18.66	17.53	31.40
	D	129	816	19.89	66.91	5.93	21.79	10.98	37.99
	I	206	**852**	22.24	**71.44**	8.48	28.90	13.06	42.74
B	A	357	302	30.55	68.89	**13.68**	31.57	19.62	44.78
	C	24	113	**33.89**	42.21	11.25	12.80	**20.13**	24.24
	D	356	**1298**	10.79	**81.16**	4.50	**35.50**	6.97	**54.12**
	I	355	505	21.81	73.24	7.95	27.24	12.45	42.55

Table 3: Average ROUGE scores for *Abstract* (A), *Conclusion*(C), *Discussion*(D), *Introduction*(I) sections wrt gold standard lay summaries for Laysumm2 (L) and Batch3 (B) datasets. The averages are taken by considering only the cases where these sections are present in the document. N_{Doc}: is the number of documents in which a particular section appears.

The experiment leads to conclusion not different from (Collins et al., 2017a), and forms the basis of rules we use for generation of lay summary (described in the following subsection).

4.2 Lay Summarization Framework

The complete pipeline of our system is shown in Figure 1. We reconstruct the input for summarization by extricating *Abstract, Conclusion, Discussion* and *Introduction* sections from the preprocessed text. Recognizing the richness and simplicity of the information contained in *Abstract*, supported by high ROUGE scores, we choose not to perform extractive summarization over it. We over-determine important sentences from each of the remaining three sections using a common extractive summarization method. The four segments, viz. *Abstract* and *Conclusion, Discussion* and *Introduction*, are further condensed using an abstractive summarizer to obtain corresponding simplified texts. Finally, the four abstractive summaries are concatenated one by one in the ACDI order until the desired *Lay summary* length is achieved. We observe from Table 3 that some sections might be missing in some documents. We simply move on to the next most important section (as per ACDI order) in such cases.

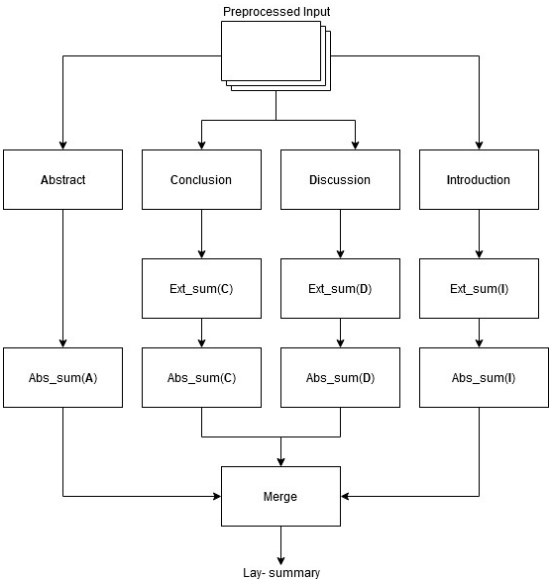

Figure 1: Pipeline of our Methodology. Ext-sum: Extractive summarization step, Abs-sum: Abstractive summarization step.

5 Experimental Setting

In this section we describe the choices we made for implementing the lay summarization framework. All the source code is made publicly available on Github [2].

5.1 Data Pre-processing

We pre-process the input text by removing redundant whitespaces, hyperlinks, and references. Based on the intuition that sentences containing relatively more mathematical symbols and non-English characters might not be comprehensible to lay readers, we completely remove the sentences which are comprised of more than one fifth special

[2] github.com/anuragjoshi3519/laysumm20

characters. We also remove any single character or numeral preceded by a period, since this character will constitute beginning of a valid sentence only if it was upper case and was preceded by a period and single whitespace- for instance, 'ab.c' is replaced with 'ab c'. We also replace common acronyms with their full forms. Finally, we remove all punctuation symbols except periods, question marks and exclamation marks which constitute end of sentence markers for effective sentence tokenization.

5.2 Extractive Summarization

We experiment with two extractive summarization methods belonging to different genres with the objective of comparing the cost and benefit. We choose a pretrained supervised neural model BioBERT with k-means clustering (BioBERT_SUM), and a frugal, unsupervised network based summarization algorithm. The two methods are briefly described below.

(i) Supervised: BioBERT_SUM - Motivated by the approach proposed in Miller (2019), we apply k-means clustering on BioBERT embeddings. BioBERT (Lee et al., 2020) is initialized with weights from BERT (Devlin et al., 2018) model pretrained on general domain corpora followed by further training on scholarly text specific to Biomedical domain, which is one of the specified domains of our input corpora. Thus, we expect it to perform well with both general domain as well as biological domain inputs. We use the fine tuned version (BioBERT-NLI[3]) of BioBERT with the bert-extractive-summarizer package[4] for extractive summarization.

(ii) Unsupervised: Entailment based Weighted Minimum Vertex Cover (wMVC) is an unsupervised network based approach proposed by Gupta et al. (2014). The sentences are modelled as vertices of the graph, and Inverse Document Frequency (IDF) based entailment is employed to link sentences Gupta et al. (2019). The algorithm considers those sentences important, which entail many sentences. The extent to which a sentence A entails another sentence B is captured by the weight of directed edge (A, B) defined as:

$$E_{A,B} = \frac{\sum_{w \in A \cap B} idf_w}{\sum_{w \in B} idf_w}$$

where, the idf score of a word w is computed as:

$$idf_w = \log \frac{N}{n_i}$$

n_i = number of sentences containing w, N = Total number of sentences in a document

The connectivity score $Conn_u$ of the vertex determines the importance of the corresponding sentence:

$$Conn_u = \sum_{u \neq v} E_{u,v}$$

In a vertex pruning step, all vertices having connectivity score below a threshold are removed. Finally, minimum number of sentences that encapsulate the essence of document are identified using weighted minimum vertex cover. The aim is to prefer vertices with high connectivity score therefore the vertex weights are inverted for reduction to weighted minimum vertex cover. Highest scoring k sentences are extracted from the solution. These sentences are then re-ordered as per original document ordering. We implement wMVC using Python 3.8 and NetworkX (Hagberg et al., 2008) package.

5.3 Abstractive Summarization

We noticed through manual checks that the gold summaries provided for the task are abstractive in nature. Therefore we extract a longer than required length summary using extractive summarizer and compress them using BART abstractive summarizer. We use the transformers library provided by Wolf et al. (2019) and weights from pretrained model facebook/bart-large-cnn[5] for experiments. We run the BART and BioBERT_SUM on Google Colaboratory with GPU setting while wMVC experiments are run on a CPU.

5.4 Experimental Design

We design experiments to answer three research questions.

I. How does unsupervised wMVC method compare with BioBERT_SUM for extractive summarization?

II. Does staging of extractive and abstractive summarization bring in improvement in the quality of lay summaries?

III. Does *divide-and-conquer* approach for generating lay summaries pay-off?

[3] https://huggingface.co/gsarti/biobert-nli

[4] https://pypi.org/project/bert-extractive-summarizer/0.4.2/

[5] https://huggingface.co/facebook/bart-large-cnn

To answer questions (I) and (II), we extract summaries from the full text using BioBERT_SUM and wMVC. Next we feed the extracted summaries to BART for comparison. The findings are described in Section 6.1. To answer question (III), we compare lay summaries generated by combining ACDI as single unit and those generated by the framework. The observations are discussed in Section 6.2.

6 Results

6.1 Experiment I and II

We compare the performances of wMVC and BioBERT based summarizers by extracting 100 word summaries from the full text of the given documents and computing their respective ROUGE scores (Section (i) of Table 4). Macro-averaged scores for both datasets are higher for wMVC summaries, indicating that wMVC yields better quality summary for the two corpora.

In order to test effectiveness of staging extractive and abstractive summarization, we extract 200 word (twice the length of stipulated summaries) summaries from the full text using the two extractive summarizers, and feed these to BART abstractive summarizer to obtain two sets of lay summaries. Final average summary length is 103 words for Laysumm2 and 93 for Batch3 documents, meeting the stipulated length restriction. ROUGE scores of the summaries are recorded in Section (ii) of Table 4.

It is observed that abstraction distinctly improves ROUGE scores in all cases for all metrics. Interestingly, the quantum of improvement is apparently more for BioBERT based summaries, which makes it winner for lay summaries of Batch3 documents. The conclusion is not confirmatory, however. We plan to investigate deeper using statistical tests.

This experiment indicates that performance of abstractive summarizers for generating lay summaries can be leveraged by feeding them focused and quality content obtained by extractive summarizer. However, the quantum of boost is not predictable and depends on the input. It is noteworthy that staging of pretrained extractive and abstractive summarizers for inference is less data and resource intensive than training a model end to end.

6.2 Experiment III

Next we describe our experiment to test our conjecture of *divide-and-conquer*. We generate lay summaries using the *Abstract, Conclusion, Discussion* and *Introduction* sections as single unit, and compare with those obtained by the proposed framework.

	Data	Model	1F	1R	2F	2R	LF	LR
(i)	L	BioBERT_SUM	35.88	36.77	8.50	8.67	18.88	19.37
		wMVC	**37.52**	**37.96**	**10.88**	**10.97**	**20.70**	**20.93**
	B	BioBERT_SUM	29.87	35.01	6.04	7.09	16.48	19.41
		wMVC	**32.22**	**38.70**	**8.69**	**10.52**	**18.51**	**22.43**
(ii)	L	BART_BioBERT_SUM	37.09	37.95	10.92	11.04	20.85	21.29
		BART_wMVC	**38.54**	**39.60**	**11.68**	**11.97**	**21.29**	**21.88**
	B	BART_BioBERT_SUM	**34.72**	**39.68**	**10.16**	**11.64**	**20.20**	**23.19**
		BART_wMVC	33.13	38.21	8.90	10.30	18.46	21.38

Table 4: Evaluation scores on training (L) and validation (B) datasets for complete text (i) Extractive only (ii) Extractive + Abstractive: Abstracting after extracting 200 word summaries using BART.

We extract 200 word summaries after combining *Abstract, Introduction, Discussion* and *Conclusion* sections in document order and abstract using BART which delivers lay summaries of average length 113. Next, based on the framework (Figure 1), we generate 150-170 length extractive summaries from the *Conclusion, Discussion* and *Introduction* sections. If any section has length less than 250 words, it is not subjected to extractive summarization. Section-wise lay summaries for each of the four sections are obtained, which are finally assembled by appending in order of ACDI till desired length of lay summary is achieved (90-110 words). Note that in case any section is missing from the paper, the framework quietly ignores it. We present the results in Table 5.

All ROUGE scores for both data sets show significant improvement over the scores obtained for lay summaries of full text (Part (ii) of Table 4 and part (i) of Table 5). It is abundantly clear that *Abstract, Introduction, Discussion* and *Conclusion* sections are most useful for generating lay summaries. Part (ii) of Table 5, further reveals that summary generation from individual sections and their subsequent merging in ACDI order results in higher scoring lay summaries than those generated from combined text of ACDI.

This validates the *divide-and-conquer* approach of focussing on limited segments of scholarly scientific documents, extracting the gist and abstracting it to make it comprehensible. The insight available from this result may help in designing better lay summarizers.

We report the evaluation results of summaries produced from our final experiment ACDI Incremental on the test set in Table 6. In the first variant,

	Data	Model	1F	1R	2F	2R	LF	LR
(i)	L	BART_BioBERT_SUM	41.93	45.07	15.56	16.70	25.01	26.85
		BART_wMVC	**43.29**	**46.82**	**16.34**	**17.68**	**24.91**	**26.97**
	B	BART_BioBERT_SUM	**36.92**	**46.70**	**12.10**	**15.47**	**21.56**	**27.45**
		BART_wMVC	36.56	46.48	11.73	15.06	20.63	26.45
(ii)	L	BART_BioBERT_SUM	**47.04**	52.97	**21.66**	**24.37**	**28.74**	**32.33**
		BART_wMVC	46.93	**52.99**	21.00	23.66	28.45	32.08
	B	BART_BioBERT_SUM	**38.33**	**51.93**	**13.90**	**18.98**	**22.28**	**30.46**
		BART_wMVC	36.74	49.95	12.77	17.45	21.11	28.87

Table 5: Evaluation scores for experiments on partial input text for both training (L) and validation (B) sets. (i) ACDI Combined. (ii) ACDI Incremental.

BioBERT_SUM is used for extractive summarization and summary lengths are 90 words on an average while in the next two rows, wMVC is used for extractive step and summary lengths are 100 and 110 words respectively. Our final results submitted towards shared task are from BART_wMVC_110 variant. [6]

System Variant	1F	1R	2F	2R	LF	LR
BART_BioBERT_SUM_90	42.43	49.53	17.30	20.19	24.84	29.03
BART_wMVC_100	**42.76**	50.32	17.23	20.13	**25.28**	29.68
BART_wMVC_110*	42.53	**51.59**	**17.48**	**21.02**	25.26	**30.55**

Table 6: Results on the Test corpus. BART_wMVC_110 is submitted towards evaluation for the competition.

It is noteworthy that the quality of generated lay summary is sensitive to the order of the sections. In case the abstract is simple and long enough, there is a possibility that the lay summary might be a condensed form of abstract only. Lay summary of this paper is shown in A.

7 Discussion

We present two sample system summaries along with the gold standard summaries in appendix B in Tables 8 and 9. We can observe that Table 8 has remarkably high overlap (highlighted) with the gold summary. In Table 9, we observe that some technical terms (highlighted) do find their way into lay summary. Use of appropriate ontologies and substituting these terms with their synonyms or entity classes can possibly make the meaning clearer to laity. At times, some sentences (example highlighted in Table 9) end up being extracted that are loosely coupled with the rest of the summary and are not even important from the viewpoint of a layman. Such sentences may increase the ROUGE score, but deteriorate the overall readability.

[6]The evaluation scores for test set are retrieved from codalab- https://competitions.codalab.org/competitions/25516#results.

Manual inspection by authors for few other system summaries indicates that we need to improve pre-processing and sentence tokenization. For example, one of the summaries contains confidence interval values which may not be comprehensible to laity. Certain inconsistencies in the input format, confuse our parsing algorithms leading to inaccurate segmentation of sections in a few cases. Moreover, we notice that in few scenarios, BART leaves out incomplete sentences towards the end, which degrades the quality of lay summary. An astute post-processing check is desirable to address this problem.

Dependence on abstractive summarizer for the quality of lay summary is the main caveat of the proposed framework. Anticipating constant improvement in the state-of-the-art in NLG, we expect the framework to yield high quality lay summaries.

8 Conclusion

We propose a framework for generating *Lay Summaries* of scientific scholarly documents. The framework is based on the core idea of extractive-abstractive pipeline to generate lay summaries. We *divide* the text into segments and focus on information rich segments to extract important sentences. These extracts are fed to the state-of-the-art abstractive summarizer for further compression which improves readability of the summary. This strategy improves the quality of lay summary, while cutting down on the training data requirement as well as computational resources. The proposed framework is frugal in terms of both types of resources. We show that reusing pre-trained publicly available models can be favoured over devising new training architectures. Thereby, reaping advantages of transfer learning for specialized tasks.

Acknowledgments

We thank the anonymous reviewers for helpful feedback. We also thank Google Colaboratory for GPU access.

References

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2020. Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*.

Jamilson Antunes, Rafael Dueire Lins, Rinaldo Lima, Hilario Oliveira, Marcelo Riss, and Steven J Simske. 2018. Automatic cohesive summarization with pronominal anaphora resolution. *Computer Speech & Language*, 52:141–164.

Luca Cagliero and Moreno La Quatra. 2020. Extracting highlights of scientific articles: A supervised summarization approach. *Expert Systems with Applications*, 160:113659.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

M. K. Chandrasekaran, G. Feigenblat, Ravichander A. Hovy. E., Shmueli-Scheuerand M., and A. De Waard. Forthcoming. Overview and insights from scientific document summarization shared tasks 2020: Cl-scisumm, laysumm and longsumm. In *Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020)*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Arman Cohan and Nazli Goharian. 2018. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, 19(2-3):287–303.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017a. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017b. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Swagata Duari and Vasudha Bhatnagar. 2019. scake: semantic connectivity aware keyword extraction. *Information Sciences*, 477:100–117.

Catherine E Dubé and Kate L Lapane. 2014. Lay abstracts and summaries: Writing advice for scientists. *Journal of Cancer Education*, 29(3):577–579.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, et al. 2019. A summarization system for scientific documents. *arXiv preprint arXiv:1908.11152*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Anand Gupta, Manpreet Kaur, Ahsaas Bajaj, and Ansh Khanna. 2019. Entailment and spectral clustering based single and multiple document summarization. *International Journal of Intelligent Systems and Applications*, 11(4):39.

Anand Gupta, Manpreet Kaur, Shachar Mirkin, Adarsh Singh, and Aseem Goyal. 2014. Text summarization through entailment-based minimum vertex cover. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 75–80.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Selvani Deepthi Kavila and Y Radhika. 2015. Extractive text summarization using modified weighing and sentence symmetric feature methods. *International Journal of Modern Education and Computer Science*, 7(10):33.

Lauren M Kuehne and Julian D Olden. 2015. Opinion: Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences*, 112(12):3585–3586.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fei Liu and Yang Liu. 2009. From extractive to abstractive meeting summaries: Can it be done by sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 261–264.

Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. 2013. Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88:164–175.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Jan Wira Gotama Putra and Masayu Leylia Khodra. 2017. Automatic title generation in scientific articles for authorship assistance: a summarization approach. *Journal of ICT Research and Applications*, 11(3):253–267.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.

Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Seal: Segment-wise extractive-abstractive long-form text summarization.

A Lay Summary of Present Paper

Title: Divide and Conquer: From Complexity to Simplicity for LaySummarization
Summary
The task is to produce non technical summaries of scholarly documents. The summary should be within easy grasp of a layman who may not be well versed with the domain of the research article. We propose a two step divide and conquer approach. We judiciously select segments of the documents that are not overly pedantic and are likely to be of interest to the laity. We over extract sentences from each segment using an unsupervised network based method. We perform abstractive summarization on these extractions and systematically merge the abstractions. We run ablation studies to establish that each step in our pipeline is critical for improvement in the quality of lay summary.

Table 7: Lay Summary of the paper. The length of the *Abstract* is about 140 words, and it contains few technical terms. Due to the length restriction of 100 words the lay summary is an abstraction of *Abstract* only.

B Sample System Summaries vs Gold Summaries

Summaries Corresponding to Paper ID : S2352409X18305169	
Title	GIS modeling of agricultural suitability in the highlands of the Jornada branch of the Mogollon culture of southcentral New Mexico
Gold Standard	Measures of site location in relation to agricultural potential are an important tool for identifying relative shifts in the importance of agriculture in prehistoric economies over time. We examine GIS modeling of agricultural potential based on soil characteristics, topography, and proximity to drainage in the highlands of the Mogollon culture of southcentral New Mexico. We describe methods, limitations, and advantages of this approach. Preliminary results support other evidence of strong agricultural reliance in the pithouse period, substantially greater than in the Archaic; the pueblo period may be slightly more linked to optimal agricultural land, though the latter conclusion is uncertain.
BART_wMVC	Measures of site location in relation to agricultural potential are an important tool for identifying relative shifts in the importance of agriculture over time within a given region. We examine the application of GIS modeling of agricultural potential based on soil characteristics, topography, and proximity to drainage in the highlands of the Jornada branch of the Mogollon culture of southcentral New Mexico. Our results support other evidence of strong agricultural reliance in the pithouse period, substantially greater than in the Archaic the pueblo period occupation may be slightly more tightly linked to optimal agricultural land, though the latter conclusion is uncertain. Our results have potential implications for both the interpretation of Formative period settlement in the Sierra Blanca Capitan Mountain highlands, and for further methodological approaches to settlement analysis.

Table 8: Gold standard and System Summaries for document scoring highest in ROUGE-1F (79.48). Highlighted sentences appear in gold standard lay summary.

Summaries Corresponding to Paper ID:S2352409X18303663	
Title	An evaluation of classical morphologic and morphometric parameters reported to distinguish wolves and dogs
Gold Standard	Visual traits and measurements that support distinguishing dog and wolf skeletal remains have been long-used, but insufficiently researched. We evaluated 14 of these, including dental abnormalities; mandible shape; orbital angle; hard palate; snout dimensions; and skull dimensions. We found only a few reliable measures, including skull height, very small or large orbital angle, snout width index, and specific measures of the 1st molar and 4th premolar teeth. Thus, much earlier research now must be re-considered toward use of combined visual, measured, and genetic traits for accurate archaeological identifications.
BART_wMVC	Morphological and morphometric differences between wolves and dogs are often overlooked. This article shows how these differences can be used to better understand the history of wolf-dog relations. The study also shows that the differences between the two species are not as large as previously thought. The results of the study were published in the Journal of Archaeology and Ethnology, a journal of the American Museum of Natural History and the American Academy of Arts and Sciences. Traditional morphometric identification of potential early domesticated dogs largely has been based on low numbers of specimens, as well as unverified diagnostic methods and variables. We propose the use of much larger canid reference groups to explore whether variation identified as signs of domestication in these specimens actually reflects natural variation that will be seen more easily within larger sample groups.

Table 9: Gold standard and System Summaries for document scoring lowest in ROUGE-1F (20.96). Highlighted terms and sentences are poor choices for lay audience.