A Corpus of Very Short Scientific Summaries

Yifan Chen
University of Cambridge, UK
yc462@cam.ac.uk

Tamara Polajnar
Royal Society of Chemistry
and University of Cambridge, UK
polajnart@rsc.org

Colin Batchelor
Royal Society of Chemistry, UK
batchelorc@rsc.org

Simone Teufel
University of Cambridge, UK
sht25@cam.ac.uk

Abstract

We present a new summarisation task, taking scientific articles and producing journal table-of-contents entries in the chemistry domain. These are one- or two-sentence author-written summaries that present the key findings of a paper. This is a first look at this summarisation task with an open access publication corpus consisting of titles and abstracts, as input texts, and short author-written advertising blurbs, as the ground truth. We introduce the dataset and evaluate it with state-of-the-art summarisation methods.

1 Introduction

Table-of-contents (TOC) entries are short summaries written by authors that are placed in the table of contents of journals, often with an eye-catching accompanying image, to advertise their paper to readers. They are meant to be a clear and concise summary of a paper's main contribution, but different from the title and abstract, which have a different communicative function. We take the titles and abstracts from chemistry papers published by the Royal Society of Chemistry as input in this initial study, as they are freely available and more numerous, but will also release full text for the smaller subset of open access publications. As such, this particular corpus is different from other scientific corpora, which normally take the abstract as the summary.

We include an analysis of the corpus properties, and experimental results with strong baselines and three different state-of-the-art deep learning models: an attention-based RNN, a reinforcement learning approach, and a BERT-based transformer method. We also perform a human evaluation study to compare the models and to validate the usefulness of the summarisation task.

The main aim of releasing this particular corpus is to see how useful deep learning models are in producing a latent semantic representation of chemical texts. Within this paper we constrain ourselves to the question: Can we decode this representation into a useful summary and perhaps aid in the journal editing process? But the real goal would be to transfer this representation into useful editing tasks like plagiarism detection or semantic search and discovery.

In addition, chemistry has a complex domain lexicon involving a potentially infinite set of molecules that can be described both by formulae and (often multi-word) terms, as well as a set of other techniques. While everything from tokenisation onward in an NLP pipeline would benefit from customisation, there is a lack of domain-specific resources. In this paper we start addressing this problem by introducing a summarisation corpus, but we hope this is the first of many resources that will aid researchers in this field. Ultimately, underlying all these tasks is an ability to produce representations that can accurately substitute multi-word terms for chemical formulae or an accurate hyponym like *ketone*, without ontological knowledge. This corpus is one in a set of tools that will help us compare models' abilities to do this.

2 Related Work

Supervised summarisation tasks are primarily evaluated on large news text corpora: CNN/Daily Mail (See et al., 2017), XSum (Narayan et al., 2018), and Newsroom (Grusky et al., 2018). Most of these use professionally written summaries consisting of one or more sentences provided by the publisher. There are also domain-specific datasets like arXiv/PubMed (Cohan et al., 2018) for mixed science summarisation and BIGPATENT (Sharma et al., 2019), both of which use abstracts as the ground truth. The number of documents in these datasets varies between 200,000 and 1.3 million.

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 153–164

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

While initial efforts in the field concentrated on unsupervised extractive summarisation methods (Mihalcea and Tarau, 2004; Vanderwende et al., 2007; Moratanch and Chitrakala, 2017), recent work has seen an explosion of deep learning-based models that leverage these large datasets to produce more abstractive and natural-sounding summaries. Out of these we choose three methods that each take a different approach. First is the pointer-generator method by See et al. (2017), which balances extractive and abstractive summarisation by substituting phrases according to a learned parameter. The second method from Chen and Bansal (2018) uses a reinforcement learning (RL) algorithm to extract informative sentences and then rewrites them using a sequence-2-sequence model with an additional re-ranking algorithm to avoid repetitive phrasings. The final method we test is a BERT-based model introduced by Liu and Lapata (2019) which uses BERT embeddings as the pre-trained encoder, stacked Transformer layers as a decoder and a fine-tuning process to produce more natural abstractive summaries. Although Cohan et al. (2018) designed a method for use with scientific text, it was specifically created for full text documents and our input text is much shorter.

There are a few historic examples of extractive summarisation in the chemistry domain (Boudin et al., 2008; Pollock and Zamora, 1975), but the research in this subfield was not as active as in other NLP applications. In their summarisation approach, Boudin et al. (2008) highlight the need for keeping the case in chemical names and for using character-based similarity measures for relevance ranking. We, likewise, employ customised tokenisation and named entity recognition while pre-processing the corpus (Corbett and Boyle, 2018) to enable future researchers to forgo the NER step. Further related applications of NLP include information retrieval (Sun et al., 2011; Hawizy et al., 2011) and literature mining (Zaslavsky et al., 2017; Öztürk et al., 2020), while lessons learned from deep learning in NLP have been applied to strings representing chemical structures to successfully discover new potential antibiotics (Stokes et al., 2020).

3 Corpus Description

The RSCSUM corpus contains 307,847 papers published in the chemistry domain between 2000 and 2019.[1] It is split into around 80% training set, 10% test set, and 10% validation set. The titles and abstracts comprise input documents, while the reference is the short table-of-contents summary.

Split	Training set examples	246268
	Test set examples	30776
	Validation set examples	30803
Training set stats	Total vocabulary size	487610
	Appearing over 10 times	52230
	Average summary length	29.07
	Average title length	18.37
	Average abstract length	169.08
	Compression ratio	21.01
	Named entities in title	11.11%
	Named entities in summary	10.40%
	Named entities in abstract	8.28%

Table 1: RSCSUM statistics.

Tab. 1 shows some vital statistics of the corpus. The compression ratio ($R_i = L_{D_i}/L_{S_i}$) indicates the ratio between the length of the input document and the gold standard summary. Our compression ratio is 21, substantially higher than that of CNN and DM (14). On the other hand, titles and abstracts are very information-dense and provide a large part of the relevant information useful for the summary. Our vocabulary size is similar to that of the CNN/DM and XSum corpora (Narayan et al., 2018), although many of the tokens occur very rarely. Of course, this is a count of lower case tokens, which in chemistry can lead to the undesired effect of collapsing significantly different chemical names into the same string. About 10% of the text is also taken up by chemical named entities, including chemical formulae. These are difficult to abstract, and this process may require knowledge from the full text documents or external sources. In fact about 1% of summaries have 3 or more chemical names that do not appear in the input text, while 26% of them have at least one. Tab. 2 shows an example document where automatically extracted chemical named entities (Corbett and Boyle, 2018) are highlighted.

Tab. 3 shows how our corpus compares to frequently used standard summarisation corpora. Although we have a relatively high number of training instances, both the summary and document lengths are at the short end of the spectrum. This will obviously make abstractive summarisation a more difficult task.

The concepts of *extractive fragment coverage*

[1] You can download the corpus at http://rsc.li/RSCsum

doc id: C2JM16014E
Summary: Highly luminescent Cd1-xZnxSe1-ySy quantum dot (QD)-encoded poly(styrene-co-ethylene glycol dimethacrylate-co-methacrylic acid) beads (PSEMBs) were prepared by a novel in situ synthesis method.
Title: Facile single step preparation of high-performance quantum dot barcodes.
Abstract: We demonstrate the facile preparation of highly luminescent Cd1-xZnxSe1-ySy quantum dot (QD)-encoded poly(styrene-co-ethylene glycol dimethacrylate-co-methacrylic acid) beads (PSEMBs) in a straightforward and reproducible manner. The monodisperse mesoporous PSEMBs are first swelled in chloroform. Afterwards , the reaction precursors, composed of Cd, Zn, Se and S, are impregnated into the microspheres. Subsequently, the Cd1-xZnxSe1-ySy QDs are synthesised directly within the polymer beads by thermal decomposition.....

Table 2: Example document "C2JM16014E" from RSCsum: summary, title and abstract.

dataset	documents	Avg. length (char)	
		article	summary
CNN	92K	656	43
DM	219K	693	53
XSum	227K	431	23
arXiv	215K	4938	220
PubMed	133K	3016	203
RSCsum	308K	187	29

Table 3: Comparison of corpora statistics.

and *extractive fragment density* were introduced by Grusky et al. (2018). The extractive fragment coverage is the percentage of unigrams in the summary that were directly copied from the input text, regardless of their order. The extractive fragment density is the average of the sum of squares of lengths of the extracted n-grams. The higher the coverage score is, the more individual words the summary extracts from the input text, but this may indicate re-wording using the same tokens and not necessarily direct copying of phrases. On the other hand, high density suggests a higher number of copied n-grams and therefore a more extractive dataset. Fig. 1 visualises the distribution of fragment density and coverage of the five compared corpora[2]. As we can see, RSCSUM has lower coverage and lower density. The graphs also indicate that the ground truth summaries in RSCSUM and XSum are more distinct from the reference content, compared to the bullet point summaries in CNN/DM or the abstracts in arXiv and PubMed.

If we take as reasonable that chemical names should be copied we can adapt the idea of fragment coverage to this dataset. We use the following formula to detect the percentage overlap between the tokens in each sentence n in the abstract A_i and the tokens in the summary S_i^t, disregarding the tokens that belong to chemical named entities t_c. We take the maximum value of all sentence overlaps.

$$\theta(A_i, S_i) = \max\left(\frac{|S_i^t \cap A_{i1}^{t-t_c}|}{|A_{i1}^{t-t_c}|}, \ldots \frac{|S_i^t \cap A_{in}^{t-t_c}|}{|A_{in}^{t-t_c}|}\right) \quad (1)$$

This allows us to examine the abstractiveness of the corpus in finer detail, as show in Tab. 2. In about 54k cases, authors appear to have taken the easy route of directly copying at least one sentence from the abstract.

4 Methods

We implement three extractive baselines and three abstractive deep models.

4.1 Baselines

The most basic extractive baseline is **Lead-2**, where we take the first two sentences of an abstract as the summary. We also use SumBasic[3] and a GloVe vector[4] enhanced TextRank algorithm.

4.2 Deep models

4.2.1 Pointer–Generator

We adapt the PyTorch re-implementation[5] of the original pointer–generator network (PGN) (See et al., 2017) by inserting ELMo embeddings (Peters et al., 2018) trained on PubMed texts,[6] hereafter refered to as PGN-E. The added ELMo embeddings were computed by a pre-trained two-layered bidirectional language model (biLM) resulting in 512-dimensional word vectors, which is more than twice of the original pointer generator embedding size. We also experimented with the original PGN, and while we found the difference in performance as evaluated by ROUGE metrics negligible, PGN-E produced a higher rate of novel n-grams indicating better abstractiveness.

[2]The data for CNN/DM and XSum is gathered from Hugging Face (https://huggingface.co/datasets), and PubMed and arXiv from Cohan et al. (2018). For all datasets, we consider the training set only.

[3]https://pypi.org/project/sumy/

[4]https://github.com/stanfordnlp/GloVe

[5]https://github.com/atulkum/pointer_summarizer

[6]https://allennlp.org/elmo/

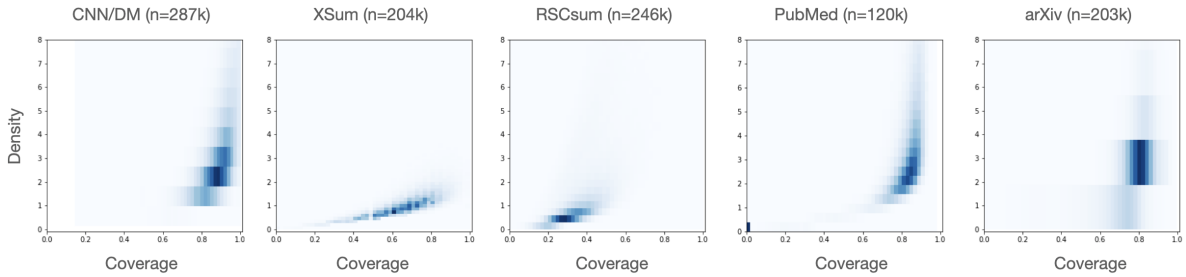

Figure 1: Extractive fragment coverage and density distributions across the compared datasets, where n indicates the number of documents.

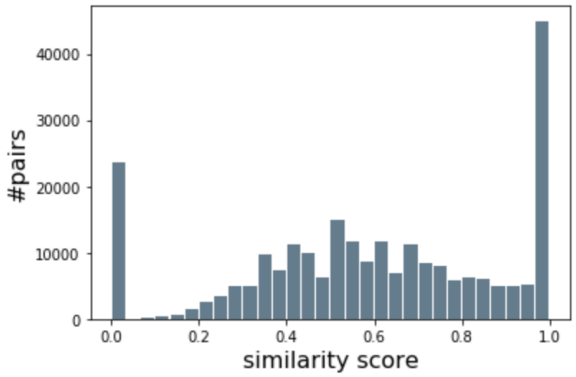

Figure 2: Distribution of the similarity scores between summary and abstract according to Eq. 1.

Likewise, we adjusted some of the hyperparameters: the vocabulary size is 100K for both the source and target text, and the Adagrad (Duchi et al., 2011) learning rate 0.05 was chosen from 0.15,[7] 0.1 and 0.05 with the initial accumulator value set to 0.1. We used gradient clipping with a maximum gradient norm of 2 and early stopping triggered by the loss on the validation set. The generated summaries were constrained to the range from 25 to 100 tokens. During training, the batch size was 8, and at test time the beam size of the beam search algorithm was 4. We trained the enhanced model for a maximum of 35K iterations due to computational restrictions. Finally, to speed up the training process, input content sentences were truncated to a maximum length of 100 tokens and summary sentences to 30 tokens.

4.2.2 Reinforcement Learning

We use the model introduced by Chen and Bansal (2018) (RL-EA) which has two separate learning mechanisms, maximum likelihood (ML) and reinforcement learning (RL). When training with Adam (Kingma and Ba, 2014), their respective learning rates were 10^{-3} and 10^{-4} respectively, and the discount factor was 0.95. The abstractor and extractor were trained separately until convergence with ML objectives, then RL was applied to the trained sub-modules. Each single-layered LSTM network includes 256 hidden units in all models. The final encoder states linearly map to the initial decoder states in the abstractor module. We also applied early stopping and the 2-norm of 2.0 gradient clipping (Pascanu et al., 2012) here.

To prime the embedding matrix of the ML model, we trained 128-dimensional word2vec (Mikolov et al., 2013) vectors with a constrained vocabulary size of 30K tokens. These embeddings were then updated in downstream training. A sentence ending token (v_{EOE}) was added as a trainable parameter for RL to learn when to stop extracting sentences from the input source, so the total summary length has no strict limitation compared with the other two abstractive models. At test time, input content was not limited, but the output summary was constrained to a maximum of 30 tokens per sentence for the abstractor.

We also tested the re-ranking mechanism of this algorithm, but decided not to include results as they produced similar ROUGE scores and lower percentages of novel n-grams.

4.2.3 Transformer

We adapt the open-source code provided by Liu and Lapata and replace the "BERT-base-uncased" pre-trained model with SciBERT (Beltagy et al., 2019). SciBERT follows the BERT model architecture, which is a multi-bidirectional transformer, by training an objective which predicts masked tokens and the next sentence, but is trained on scientific texts including PubMed. The input source content and target summaries were tokenised with BERT's subword tokeniser. We refer to this model as SciBERT Abstractive, SciBERTA for short.

[7] 0.15 is the best learning rate for the CNN/DM dataset.

The specific hyperparameter values of the abstractive component are shown in Appendix A.1. The maximum encoding text size is set to 512, because only 210 input texts in our corpus are longer than 512, and 512 is the maximum length in the original BERT model position embeddings. The model was trained for 100K steps with gradient accumulation for every five steps. Its intermediate models were saved every 2000 iterations and evaluated on the validation set every 2500 steps. The top-3 intermediate results which have the highest validation accuracy are chosen, and results on the test set are averaged to provide the final score.

The transformer decoder contains 768 hidden units and a hidden size of 2048 for all feed-forward layers. During decoding the beam search size is 5 and the output summary is limited the the range of 20 to 100 tokens. The decoding ends when an end-of-sequence token is generated. Trigram blocking (Paulus et al., 2017) is used to avoid repetition. Because of the sub-word tokeniser, OOV tokens are rarely observed.

5 Quantitative Evaluation

We perform a set of classic quantitative experiments by training and tuning on the train and validation portions of the data and testing once on the test portion. We evaluate the performance of the models using four standard variations of ROUGE (Lin, 2004) and report the F1 values and the confidence interval (CI) in Tab. 4. ROUGE-1 and ROUGE-2 measure the unigram and bigram overlap between the automatic and reference summaries, whereas ROUGE-L measures the longest extended matching sequence of words using the longest common subsequence (LCS). An advantage of using LCS is that it uses all in-sequence matches that reflect sentence-level word order, rather than requiring consecutive matches. Since it automatically includes the longest in-sequence common n-grams, a predefined n-gram length is not necessary. Finally, the skip-bigram/unigram variation ROUGE-SU measures overlap of word pairs that have a maximum of two gaps between words combined with unigram overlap.

5.1 Results

Results are presented in Tab. 4. We can observe from the CI overlap that SciBERTA is significantly better than the other two deep learning methods, whereas Lead-2 provides the most competitive baseline. In fact PGN-E barely outperforms the Lead-2 baseline. An examination of the summaries produced on the validation set confirms that this method most often adopts the Lead-2 strategy by copying first and/or second sentence, an issue that was also observed in prior work (Qiu et al., 2020; Gehrmann et al., 2018). This result pattern is also confirmed by prior work (Tab. 5) where we see that the original PGN does not even beat the Lead-3 baseline, while the BERT-based model outperforms the other two. The introduction of ELMo vectors slightly alleviates this issue over the original PGN-coverage model. The original PGN has 71.28% unigram overlap with Lead-2, while PGN-E is at 68.22%. RL-EA has unigram overlap of 61.35% and SciBERTA again shows highest diversity with only about 53.86% overlap with the Lead-2 baseline.

As an indication of abstractiveness, Fig. 4 shows the average percentage of novel n-grams for each of the deep learning methods. We can see that SciBERTA also outperforms the other methods with this measure, and although PGN-E and RL-EA were indistinguishable based on ROUGE, PGN-E is actually producing more novel n-grams.

5.2 Filtering Training Data

As we saw in Sec.3 we filter out the training examples which have at least one sentence with 80% overlap and higher and we are left with 174180 training examples. We call this training set RSCSUM-T80, and leave the test and validation sets untouched, with 30776 and 30803 examples respectively. The exclusion of the strong extractive signal changes the training data profile and also reduces the training data size, resulting in a decrease in ROUGE results by between 1 and 4 F1 points. On the other hand, this strategy does lead to a larger proportion of novel n-grams for PGN-E and RL-EA (Fig. 5). PGN-E's performance increases slightly over the original SciBERTA scores, which unfortunately drop slightly on the pruned training data. This indicates that SciBERTA is more resilient to biased signals in the training data, but benefits from more training data. The unigram overlap between the generated summaries and the Lead-2 baseline is also reduced across the board (PGN: 59.61%, PGN-E: 59.49%, RL-EA:49.44%, SciBERTA: 52.46%).

Models	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-SU	
	F1	CI	F1	CI	F1	CI	F1	CI
Lead-2	**45.4**	(-0.2, +0.2)	**30.5**	(-0.3, +0.3)	**38.1**	(-0.2, +0.2)	**21.8**	(-0.2, +0.2)
SumBasic	38.5	(-0.2, +0.2)	20.4	(-0.2, +0.2)	29.8	(-0.2, +0.2)	15.0	(-0.2, +0.2)
Textrank	38.2	(-0.2, +0.2)	21.8	(-0.3, +0.3)	30.7	(-0.3, +0.3)	16.6	(-0.2, +0.2)
PGN-E	46.2	(-0.2, +0.3)	31.0	(-0.4, +0.3)	41.2	(-0.3, +0.3)	26.5	(-0.3, +0.3)
RL-EA	47.4	(-0.3, +0.3)	31.7	(-0.3, +0.3)	40.8	(-0.3, +0.3)	25.8	(-0.3, +0.3)
SciBERTA	**48.3**	(-0.3, +0.3)	**32.4**	(-0.4, +0.3)	**42.4**	(-0.3, +0.4)	**28.1**	(-0.3, +0.3)

Table 4: Results of extractive and abstractive models on RSCSUM (best in each section bold-faced). Fig. 3 gives the bar chart for a better view.

Models	ROUGE-1	ROUGE-2	ROUGE-L
Baseline			
Lead-3 (See et al., 2017)	40.3	17.7	36.6
Abstractive models			
PGN (See et al., 2017)	36.4	15.7	33.4
PGN+coverage (See et al., 2017)	39.5	17.3	36.3
RL-EA+rerank (Chen and Bansal, 2018)	40.9	17.8	38.5
BERTSumAbs (Liu and Lapata, 2019)	**41.7**	**19.4**	**38.8**

Table 5: Performance of related algorithms from prior work on the non-anonymised CNN/Daily Mail dataset.

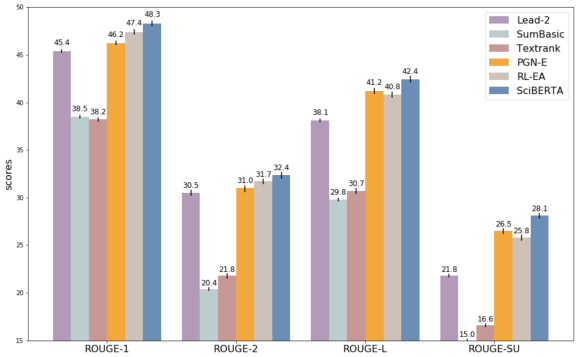

Figure 3: Bar chart for Tab. 4.

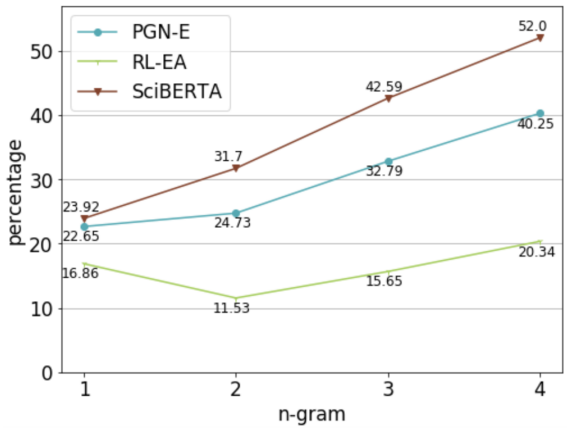

Figure 4: Average percentage of novel n-grams in the generated summaries.

6 Qualitative Evaluation

Reference-based automatic evaluation has accepted limitations such as the fact that any given reference is not the only possible summary or even necessarily the best one. We, therefore, perform a small-scale human evaluation using three participants with a background in the chemical sciences. They volunteered to help. We choose to compare SciBERTA as the best-performing system, in addition to RL-EA, which is less similar to the Lead-2 baseline than PGN-E and also has a different learning objective. In order to test the quality of the gold standard reference summaries objectively, we also evaluate them in our setup.

We perform a 3×3 Latin Square design with 42 items (documents) and three conditions (systems). The advantage of the Latin Square design (cf. Tab. 6) is that each participant sees each item in only one condition, thus avoiding repetition bias. The systems (α, β, γ) are run on different batches of 14 items (1, 2, 3) and then shown to judges (A_1, A_2, A_3), so that A_1 is seeing summaries from batch 1 generated by system α, batch 2 generated by system β, and batch 3 summaries by system γ. We then randomise the order of items shown. So that the conditions are not readily distinguishable, we aim to provide a distribution of summary lengths which is as even as possible. Fig. 6 shows the natural distribution of the summary lengths produced on the test set.

The participants were presented with the document title and abstract (concatenated together) and a summary, which they rated on scale of 0-5 according to the criteria in Tab. 7. Dimensions considered are grammaticality, informativeness, relevance and overall quality. We report the systems' mean scores

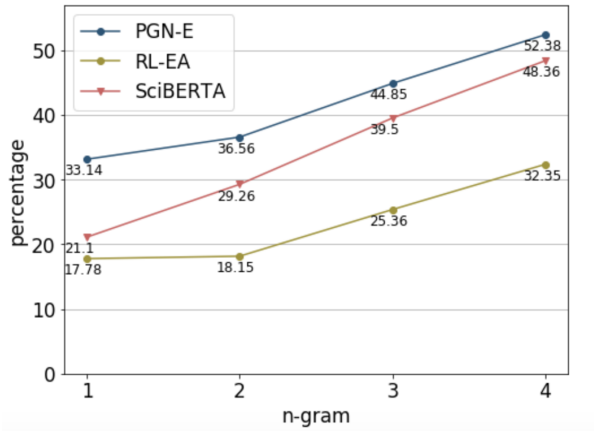

Figure 5: Average percentage of novel n-grams in the generated summaries with the filtered training dataset RSCSUM-80.

	batch 1	batch 2	batch 3
A_1	$\alpha 1$	$\beta 2$	$\gamma 3$
A_2	$\gamma 1$	$\alpha 2$	$\beta 3$
A_3	$\beta 1$	$\gamma 2$	$\alpha 3$

Table 6: 3×3 Latin square design with conditions (systems) α, β, and γ, participants A_1, A_2, A_3 and item batches 1, 2, 3.

for each dimension separately as well as the average across dimensions.

6.1 Results

Tab. 8 shows the results, with SciBERTA outperforming both RL-EA and the ground truth (GT) on all counts. SciBERTA achieves the highest mean score on all four criteria at 4.52, 4.81, 3.55 and 3.48 respectively, whereas the ground truth has only 3.33 overall quality and 3.86 on average, which is quite unexpected. One possible reason is that a few of the ground truth examples are sentences directly extracted from the abstracts and are thus less interesting for this task. Although we took care to distribute the summary lengths as evenly as possible, in general RL-EA favours longer summaries, yet the more succinct summaries of SciBERTA are preferred even in the informativeness dimension. RL-EA also sometimes produces unfinished sentence fragments and this could be one of the reasons for its relatively low grammaticality and quality values.

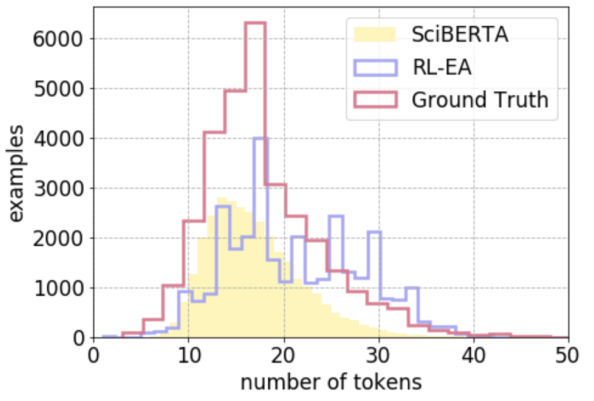

Figure 6: Distribution of model summary length (test set).

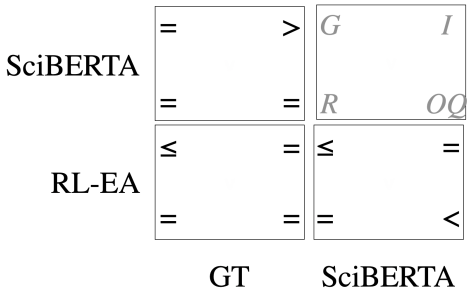

Figure 7: Statistical significance test on the values in Tab. 8. The four positions correspond to **G**rammaticality, **I**nformativeness, **R**elevance and **O**verall **Q**uality respectively, as shown in upper left box. "=" means no statistical difference, ">" means the row performs significantly better than the column at the significance level $\alpha = 0.05$, whereas "⩽" indicates the same at $\alpha = 0.01$.

We use the two-tailed Wilcoxon signed-rank test to compute the statistical significance between the systems (Fig. 7). Though SciBERTA performs the best in all the evaluation dimensions as well as the overall average, it is significantly better than the ground truth only in terms of the informativeness and outperforms RL-EA regarding grammaticality and overall quality. Except for grammaticality, there is no statistically-significant difference detected between RL-EA and the ground truth.

As an indication of the degree to which participants agreed in their judgements, Fig. 8 shows the mean scores given by each judge to each of the three conditions.[8] We can see that the judgements by A_2 pattern in a different way to those of participants A_1 and A_3, which are more similar to each other. This might be due to the different backgrounds of the participants, as the former is a materials scientist by training, whereas the latter two are chemists.

Anecdotally, the participants independently re-

[8] In a Latin Square design, agreement can only be measured within experimental groups, not across them. However, as the experimental group size is 1 in our 3×3 setting, this is of no use to us here.

Dimension	Prompt	Rating range $0 < - - - - - > 5$	
Grammaticality	Are the individual sentences of the summary well-written and grammatical?	Disagree	Agree
Informativeness	To which degree does the summary contain false or misleading information?	A lot	Zero
Relevance	Does the summary capture the main points of the abstract?	Disagree	Agree
Overall quality	If I were a journal editor, I would accept this summary for enticing readers to the website.	Disagree	Agree

Table 7: Human evaluation of the question-answering rating system.

dimension	SciBERTA	RL-EA	GT
Grammaticality	**4.52**	3.29	4.24
Informativeness	**4.81**	4.62	4.38
Relevance	**3.55**	3.52	3.50
Overall quality	**3.48**	2.71	3.33
Average	**4.09**	3.53	3.86

Table 8: Mean scores of human evaluation on RSC-SUM.

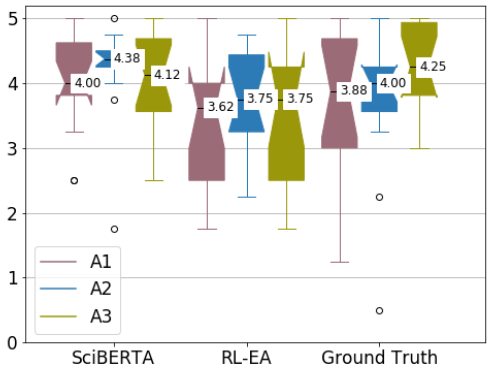

Figure 8: Participants' individual mean scores, by conditions.

ported that the fact that the summaries were uncased would be a barrier to use of automatic summarisation in chemistry, but that it didn't cause them trouble in the small number of examples they were looking at. For example, "no" (stop word), "No" (nobelium), and "NO" (nitrogen oxide), all have vastly different meanings.

One explicitly stated reason for the lower scores of human summaries is that they often contain grammatical errors. In their feedback, the participants noted that they were generous when grading the grammaticality as the summaries are mostly right regardless of the case. Nevertheless, when looking closely at a few ground truth summaries, we noticed that they seem to have been marked down for grammatical mistakes. The overarching reason is probably that writing the summaries was an after-acceptance task for the authors, which may not have had their full attention. For example in the Appendix A.2 Tab. 10 we can see that the ground truth summary consists of the motivation sentence, while every method accurately picked out the contribution sentence. Of these, the one by SciBERTA is the most grammatically pleasing. As the input source document is the abstract of a paper, authors have a tendency to start with motivational statements, e.g. "we synthesised something and then studied its properties, applications etc". One participant remarked that they tended to mark down the summaries that only focused on this aspect of the abstract. Conversely, some ground truth summaries draw on the full paper and the domain knowledge of the authors, and so could contain information beyond the abstract. While we do not have specific feedback on this point, this could also lead to an undesirable score.

7 Discussion

In this paper we introduced a corpus that consists of titles and abstracts of papers in the scientific domain as input text, and author-written table-of-contents summaries. The summaries are meant to be distinct from the titles and abstracts, but at least one fifth of the training data contains significant extractive elements. Overall the corpus compares favourably to others in its abstractive qualities. It contains scientific terminology in the chemistry domain, and some of the terminology in the summaries does not occur in the available input text. Consequently, as part of a larger effort of enriching chemical NLP, in the future we will also release about 40K full text open access articles that have corresponding TOC summaries.

We tested three state-of-the-art deep summarisation methods and found that a transformer-based method that uses pretrained scientific BERT embeddings produces the best overall results in both quantitative evaluation and a qualitative study with three domain expert participants. It is surprisingly resistant to the strong extractive component in the training data, and produces novel content despite the short input text. A qualitative study showed that automatic summaries yield acceptable results and are in some aspects significantly better than author-written summaries. However, in order to be truly useful, the summarisation methods and associated embeddings need to be adapted to deal with the cased text, which may lead to difficulties with an expanded vocabulary. In conclusion, this study paves the way for future exploration of summarisation and semantics in the chemistry domain.

8 Acknowledgements

The authors would like to thank Aileen Day and Peter Corbett who participated in our evaluation and the reviewers for their thoughtful comments.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Florian Boudin, Juan-Manuel Torres-Moreno, and Patricia Velázquez-Morales. 2008. An efficient statistical approach for automatic organic chemistry summarization. In *International Conference on Natural Language Processing*, pages 89–99. Springer.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Peter Corbett and John Boyle. 2018. Chemlistem: chemical named entity recognition using recurrent neural networks. *Journal of cheminformatics*, 10(1):59.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159.

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

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Lezan Hawizy, David M Jessop, Nico Adams, and Peter Murray-Rust. 2011. Chemicaltagger: A tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3(1):17.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

N Moratanch and S Chitrakala. 2017. A survey on extractive text summarization. In *2017 international conference on computer, communication and signal processing (ICCCSP)*, pages 1–6. IEEE.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. On the difficulty of training recurrent neural networks.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Joseph J Pollock and Antonio Zamora. 1975. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4):226–232.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. 2020. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688 – 702.e13.

Bingjun Sun, Prasenjit Mitra, C Lee Giles, and Karl T Mueller. 2011. Identifying, indexing, and ranking chemical formulae and chemical names in digital documents. *ACM Transactions on Information Systems (TOIS)*, 29(2):1–38.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Leonid Zaslavsky, Daniel Lowe, Chih-Hsuan Wei, Zhiyong Lu, and Evan Bolton. 2017. Improving chemical names matching for verification, rating, and validation of pubchem compound records. In *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, volume 253. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA.

Hakime Öztürk, Arzucan Özgür, Philippe Schwaller, Teodoro Laino, and Elif Ozkirimli. 2020. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today*, 25(4):689–705.

A Appendices

A.1 SciBERTA hyperparameter setting

Table 9 provides the specific hyperparameter setting for the SciBERTA system.

hyperparameter	SciBERTA
training_steps	100,000
warmup steps	10,000
max_pos	512
batch size	8
grad_accum_cnt	5
dropout	0.2
learning rate	0.002
learning rate$_{dec}$	0.01

Table 9: Hyperparameter settings in SciBERTA.

A.2 Summary examples

The real examples generated by the three systems are shown in Tables 11–14. Compared to training on full RSCSUM, training on RSCSUM-T80 could lead to mistakes using PGN-E and RL-EA models.

Candidate$_{PGN-E}$**:** hydrogen peroxide (h2o2) plays a significant role in regulating the redox balance in the living body . this work illustrated a high sensitivity hydrophilic photoacoustic probe for ratiometric imaging of h2o2 in living mice .
Candidate$_{RL-EA}$**:** a high sensitivity hydrophilic photoacoustic probe for ratiometric imaging of hydrogen peroxide in vitro and vivo .
Candidate$_{SciBERTA}$**:** a hydrophilic photoacoustic probe was developed for ratiometric imaging of h2o2 in vitro and vivo .
Ground truth: Hydrogen peroxide (H2O2) plays a significant role in regulating the redox balance in the living body .
Abstract: Enhancing hydrophilicity of photoacoustic probes for effective ratiometric imaging of hydrogen peroxide . Hydrogen peroxide (H2O2) plays a significant role in regulating the redox balance in the living body . Compared with traditional imaging techniques , ratiometric photoacoustic imaging is no doubt a superior choice for H2O2 visualization . However , the difficult design of ratiometric probes with only one activatable agent that exhibits two changeable absorption peaks under H2O2 activation remains a big challenge . In this work , we developed a near - infrared absorbing probe , which responded to H2O2 selectively and permitted the ratiometric photoacoustic imaging of H2O2 in living mice . The probe was constructed from an Aza-BODIPY backbone attached with a benzeneboronic acid pinacol ester moiety though a quaternization reaction . Oligo(ethylene glycol) (OEG) was introduced into hydrophobic Aza-BODIPY to enhance the water - solubility of the probe . The OEG-Aza-BODIPY-BAPE probe exhibited sensitive and specific ratiometric PA signals towards H2O2 . In vivo experiments also showed that the OEG-Aza-BODIPY-BAPE probe can be used for ratiometric PA imaging . Overall , our work illustrated a high sensitivity hydrophilic photoacoustic probe for ratiometric imaging of hydrogen peroxide in vitro and vivo .

Table 10: An example "C8AY01644E" produced by PGN-E, RL-EA and SciBERTA on full RSCSUM. Overlaps are highlighted by different colours.

Candidate$_{full}$**:** β-amyloid (aβ) plays a central role in alzheimer's disease (ad), but the specific molecular mechanism and associated structures remain unknown.
Candidate$_{T80}$**:** β-amyloid (aβ) plays a central role in alzheimer's disease (ad), but the specific molecular mechanism and associated structures remain unknown in contrast to structured conformations and associated structures.
Ground truth: Direct correlation of Alzheimer patient data to a spectrum of NMR structures and chemical properties of beta amyloid (Aβ) variants allows identification of conformation - dependent disease properties .
Abstract: Pathogenic properties of Alzheimer's β-amyloid identified from structure – property patient-phenotype correlations. β-Amyloid (Aβ) plays a central role in Alzheimer's disease (AD), but the specific molecular mechanism and associated structures remain unknown. We compiled patient data for carriers.....We conclude that disordered monomers are likely to be pathogenically important in contrast to structured conformations.....

Table 11: An example "C4DT03122A" produced by PGN-E training on full RSCSUM and RSCSUM-T80 respectively. The underlined phrase is a repetitive expression.

Candidate$_{full}$**:** the addition of inbr3 to the oxidative sonogashira cross - coupling reaction of 2 - ethynylaniline with (e) - trimethyl (3,3,3 - trifluoroprop -1-enyl) silane led the subsequent cyclization of these 1,3 - enynes under palladium catalysis provides access to the corresponding indoles bearing a 3,3,3 - trifluoroprop -1-enyl group at their 2 - position .
Candidate$_{T80}$**:** the addition of inbr3 to the oxidative sonogashira cross - coupling reaction of 2 - ethynylaniline with (e) - trimethyl (3,3,3 - trifluoroprop -1-enyl) silane led the oxidative sonogashira coupling of 2 - ethynylanilineselectronic with a affords a dramatic enhancing effect of inbr3 .
Ground truth: A dramatic enhancing effect of InBr3 was observed towards the oxidative Sonogashira cross - coupling reaction of 2-ethynylaniline with (E)-trimethyl(3,3,3-trifluoroprop-1-enyl)silane .
Abstract: The addition of InBr3 to the oxidative Sonogashira cross - coupling reaction of 2-ethynylaniline with (E)-trimethyl(3,3,3-trifluoroprop-1-enyl)silane led to a dramatic increase in the reactivity to afford the corresponding 1,3-enynes bearing a trifluoromethyl group on their terminal sp2 carbon . The subsequent cyclization of these 1,3-enynes under palladium catalysis provides access to the corresponding indoles bearing a 3,3,3-trifluoroprop-1-enyl group at their 2- position .

Table 12: An example "C5OB02558C" produced by RL-EA on full RSCSUM and RSCSUM-T80 respectively. The strikethrough denotes a wrong expression.

Candidate$_{full}$**:** a novel visible light promoted carbodifluoroalkylation of allylic alcohols is disclosed. a series of difluoro 1,5 - dicarbonyl compounds were obtained through a tandem radical addition and 1,2-aryl migration process.
Candidate$_{T80}$**:** a novel visible light promoted carbodifluoroalkylation of allylic alcohols was developed via a tandem radical addition and 1,2-aryl migration process , which proceeds via a radical intermediate.
Ground truth: A novel visible light promoted carbodifluoroalkylation of allylic alcohols is disclosed.
Abstract: Visible light promoted carbodifluoroalkylation of allylic alcohols via concomitant 1,2-aryl migration. A novel visible light promoted carbodifluoroalkylation of allylic alcohols is disclosed. A series of difluoro 1,5-dicarbonyl compounds were obtained through a tandem radical addition and 1,2-aryl migration process. Mechanistic analysis indicated that the 1,2-aryl rearrangement proceeded via a radical intermediate.

Table 13: An example "C5CC01189B" produced by SciBERTA on full RSCSUM and RSCSUM-T80 respectively.