| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | Flesch-Kincaid | Gunning | Coleman-Liau |
|---|---|---|---|---|---|---|
| Oracle extractive | **53.56**$_{\pm 0.58}$ | **25.54** $_{\pm 0.78}$ | **49.56**$_{\pm 0.65}$ | 14.85 | 13.45 | 16.13 |
| BERT extractive | 26.60$_{\pm 0.51}$ | 11.11$_{\pm 0.41}$ | 24.59$_{\pm 0.47}$ | **13.44** | **13.26** | **14.40** |
| Pointer generator | 38.33$_{\pm 0.61}$ | 14.11$_{\pm 0.46}$ | 35.81$_{\pm 0.60}$ | 16.36 | 15.86 | 15.90 |
| BART | 52.53$_{\pm 0.51}$ | 21.83$_{\pm 0.52}$ | 49.75$_{\pm 0.52}$ | 13.59 | 14.16 | 14.45 |
| BART+CNN/DM | 52.46$_{\pm 0.48}$ | 21.84$_{\pm 0.50}$ | 49.70$_{\pm 0.50}$ | 13.73 | 14.33 | 14.60 |
| BART+PubMed | 52.66$_{\pm 0.48}$ | 21.73$_{\pm 0.48}$ | 49.97$_{\pm 0.51}$ | **13.30** | **13.80** | **14.28** |
| BART+CNN/DM+PubMed | **53.02**$_{\pm 0.48}$ | **22.06**$_{\pm 0.49}$ | **50.24**$_{\pm 0.49}$ | 13.60 | 14.11 | 14.41 |

Table 3: Test set performance evaluated by ROUGE and readability score. BART model pretrained on CNN/DM and PubMed is the best-performing model based on ROUGE, while BART model pretrained on PubMed is the best one based on readability score (Best model performance is in bold). $x_{\pm}$ indicates 95% interval: $[x-, x+]$

**Intermediate pre-training** To compensate for the limited training data, we added intermediate pre-training steps for the BART model before finetuning. We first experimented with adding labeled data for summarization task in other domains. We adopted the CNN/DM dataset (**?**), which contains about 287K document-summary pairs, and BART is among the best-performing systems for this task. Secondly, we tried to pre-train BART with an unlabeled biomedical corpus to expose the model to medical domain-specific language. We used the PMC articles dataset [8] which contains 300K PubMed abstracts. Following the BART paper, we corrupted these documents using several transformations, including text substitution and sentence shuffling. BART was then trained on the corrupted abstracts to reconstruct the original PubMed abstracts. Lastly, we combined these two strategies to train BART on CNN/DM and PubMed sequentially before finetuning it on our dataset.

**Training details** All experiments were run using a single NVIDIA Tesla V-100 GPU. All models were developed using `PyTorch`. We used `neural-summ-cnndm-pytorch`[9] to implement the pointer-generator model. The batch size was set to 4. Other hyper-parameters were set to default values. We built the BERT extractive model using code released by the authors.[10] The learning rate was set to $2 \times 10^{-3}$ and the batch size 140. Other hyper-parameters were set to default values. We used the Fairseq [11] BART implementation. All BART models were trained using the Adam optimizer. The learning rate was set to $3 \times 10^{-5}$, and learning decay was applied. The minimum length of the generated summaries was set to 100, and the maximum length was set to 700.

## Results

**Automated evaluation** ROUGE and readability results on the CDSR test set are shown in Table **??**. We compare the seven methods described above: Oracle extractive, BERT extractive, pointer-generator, BART, BART pre-trained on CNN/DM, BART pre-trained on Pubmed abstracts, and BART pre-trained on both CNN/DM and PubMed abstracts.

The oracle extractive method, as an upper bound for the extractive approach, produces the best ROUGE-1 and ROUGE-2 scores. However, it obtains approximately the same level of readability as the source text in our test set (Table **??**) , which indicates that selecting the reference sentences will only result in a summary that is as difficult to read as the original abstract. In contrast, the BERT-based extractive model achieves better readability scores while performing worst in terms of ROUGE scores. This demonstrates that, in practice, training the model to extract the correct content from the original abstract might be difficult, even though the model learns to extract shorter and easier sentences.

Among the 5 abstractive models, the pointer-generator model performs significantly worse in both ROUGE and readability, emphasizing the importance of pre-training for our task. BART-based models achieve surprisingly good performance in terms of both summarization and readability, suggesting contemporary NLP models have the potential to perform the task, and to help the general public access professional medical information. Additionally, BART pre-trained on CNN/DM and PubMed abstracts achieves the best performance in ROUGE, and BART pre-trained only on PubMed abstracts obtained the lowest readability. This demonstrates the usefulness of either adding task-relevant labeled data or domain-specific unlabeled data. However, our strategies for adding such data are quite straightforward, and we lacked resources to do hyperparameter search for the relatively expensive pre-training procedure. Therefore, we only see marginal improvement compared with the BART model. We will aggregate more relevant data, and develop better pre-training strategies to improve the performance in future work.

**Human evaluation** Table **??** shows the human evaluation results. Intriguingly, human evaluators rated the model-generated summaries with comparable or even higher scores for all the four aspects, and for both abstract A and B. The average Kendalls coefficient (**?**) for the two biomedical abstracts among all evaluators' inter-rater aggrement is 0.62. Kendalls coefficient ranges from -1 to 1, indicating low to high association. Considering the subjectivity of the rating task, this number indicated high human agreement for the tasks. While larger scale study is required, this work

[8] https://www.kaggle.com/cvltmao/pmc-articles

[9] https://github.com/lipiji/neural-summ-cnndm-pytorch/

[10] https://github.com/nlpyang/presumm

[11] https://github.com/pytorch/fairseq

Table 4: Human evaluation scores of the expert-generated summaries (*Target*) and the model-generated summaries (*Generated*) for two abstracts from the test set. Generated abstracts from BART+CNN/DM+PubMed model have better scores in grammaticality, meaning preservation, and correctness of key information.

provides preliminary evidence that automatically-generated plain language summaries are readable and interpretable to non-expert human readers.

### Qualitative analysis

We present the output of our best two models in the last two columns of Table **??**. This provides evidence that the best-performing models can address some transformations, and generate grammatical and meaningful outputs. Specifically, out of the five listed phenomena, we observed that model-generated summaries could achieve three transformation types to some extent, including removing unnecessary details, jargon explanation and sentence structure simplification. Some capabilities the model demonstrated are encouraging for future research. For example, it learned to explain the term RCT from similar examples in the training data.

On the downside, the models are still struggle with some difficult transformations, such as relevant background explanation. This ability is harder to learn, and our dataset might not contain the required background knowledge. Therefore, external knowledge might be also useful. Furthermore, we also see risks in using the current abstractive models to generate reliable information for the public. For example, in the example of sentence structure simplification, *BART+PubMed* changed the meaning of the original sentence: the source sentence claims an association between the pattern of blood flow with poor prognosis, while the generated sentence focuses on the Doppler ultrasonography. *BART+CNN/DM+PubMed* performs better in this case.

## Discussion

Automated lay language summarization of biomedical scientific reviews requires both summarization and the acquisition of domain knowledge. Previously, available datasets were constructed at sentence level. However, sentence-level simplification or transformation does not require the complex strategies used by experts when rendering biomedical literature understandable to a lay audience. Therefore, we consider the document-level dataset as an important outcome of our work, which can be useful for future research on this topic. Abstractive models are more practical than extractive ones, since extractive summaries are written in the same professional language as their source documents. The best performing model is BART pre-trained on both CNN/DM and PubMed abstracts, which preserves key information (based on ROUGE) while dropping the reading requirements a year or two (based on readability scores).

Human evaluation is necessary for our task. There is a considerable gap between the automatic evaluation merics and human judgement. Despite being widely used to evaluate summarization systems, ROUGE is not practical for our task because it can neither capture the required transformation phenomena nor assess difficulty in understanding. Similarly, lower readability scores do not imply understandability. Readability scores consider only the surface forms, without considering the complexity introduced by medical abbreviations and domain-specific concepts. Human evaluation is the most robust method to evaluate the performance. However, aside from the small number of participants, the survey questions need a formal validity. Further studies are required to find that BART-derived summaries were more appealing to human raters on several fronts hold when more abstracts and human raters are involved.

## Conclusion

We propose a novel plain language summarization task at the document level and construct a dataset to support training and evaluation. The dataset is of high quality, and the task is challenging due to typical transformation phenomena in this domain. We tried both extractive and abstractive summarization models, and obtained best performance with a BART model pre-trained further on CNN/DM and PubMed, as evaluated by automated metrics. Human evaluation suggests the automatically generated summaries may be at least as acceptable as their professionally authored counterparts.

Unde commodi consequatur eaque nisi laboriosam, amet eius illo aspernatur dolorem quam officiis ex, unde deserunt quisquam maiores molestias quas provident ratione culpa quidem enim?Reprehenderit earum consequuntur asperiores tempora, dolor quo nobis consequuntur commodi recusandae incidunt maiores.Voluptatum officiis vero maiores assumenda earum quidem quae at, repellat enim et ipsa ad, itaque voluptate dolorum adipisci, saepe ducimus soluta eaque quasi atque fuga, at tenetur fuga minus adipisci facilis repellendus doloremque placeat?Dolor vero nobis quaerat doloremque exercitationem, nam animi error inventore dolores iusto optio velit eveniet, sunt corporis repudiandae quam nihil ea quis dolores tenetur expedita saepe, veniam repellendus ullam necessitatibus fugit reiciendis alias eaque ipsa iste, unde perspiciatis excepturi.Facilis magni amet placeat neque ratione sunt inventore iure quasi vero porro, voluptas explicabo eius, eum repellat provident deserunt soluta corporis saepe accusamus earum explicabo sint,