	Abstract A		Abstract B	
Perspectives	**Target**	**Generated**	**Target**	**Generated**
Grammaticality	4.25	4.50	3.50	4.00
Meaning Preservation	3.75	4.75	3.50	4.50
Understandability	3.75	3.50	2.75	2.50
Correctness of Key Information	3.50	4.50	4.00	4.00

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 1. 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 so-

luta corporis saepe accusamus earum explicabo sint,