Cochrane-auto: An Aligned Dataset for the Simplification of Biomedical Abstracts

Jan Bakker
University of Amsterdam
Amsterdam, The Netherlands
j.bakker@uva.nl

Jaap Kamps
University of Amsterdam
Amsterdam, The Netherlands
kamps@uva.nl

Abstract

The most reliable and up-to-date information on health questions is in the biomedical literature, but inaccessible due to the complex language full of jargon. Domain specific scientific text simplification holds the promise to make this literature accessible to a lay audience. Therefore, we create Cochrane-auto: a large corpus of pairs of aligned sentences, paragraphs, and abstracts from biomedical abstracts and lay summaries. Experiments demonstrate that a plan-guided simplification system trained on Cochrane-auto is able to outperform a strong baseline trained on unaligned abstracts and lay summaries. More generally, our freely available corpus complementing Newsela-auto and Wiki-auto facilitates text simplification research beyond the sentence-level and direct lexical and grammatical revisions.

1 Introduction

Biomedical research has the potential to directly impact people's decision-making with regards to health. However, most reliable and up-to-date sources in biomedicine contain complex language and assume a high degree of background knowledge, making them difficult to understand for the general public. Automatic text simplification approaches can be applied in an effort to make these sources more accessible. Yet, training neural models to simplify biomedical documents is a complex task which requires high quality training data.

To this end, Devaraj et al. (2021) introduced a corpus of paired (complex, simple) texts in English, derived from the Cochrane Database of Systematic Reviews. The CDSR comprises systematic reviews which are internationally recognized as the highest standard in evidence-based health care and which are accompanied by both technical abstracts and plain language summaries. Plain language summaries are written directly from the full reviews; they are not simplified versions of the abstracts.

Complex paragraph

Fifteen heterogeneous trials, involving 1022 adults with dorsally displaced and potentially or evidently unstable distal radial fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation especially in terms of patient characteristics and interventions. Methodological weaknesses among these trials included lack of allocation concealment and inadequate outcome assessment.

Simple paragraph

Fifteen trials, involving 1022 adults with potentially or evidently unstable fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation in their characteristics especially in terms of patient characteristics and the method of external fixation.

Figure 1: A complex-simple paragraph pair from Cochrane-auto.

Even so, the authors argued that portions of the lay summaries could be considered simplifications of analogous sections in the abstracts. Their corpus therefore consists of parallel technical abstracts and plain language summaries, both starting at the section describing studies and results. Nevertheless, the authors did not align the corpus at the sentence-level, and upon manual inspection, we find that roughly 29% of the simple sentences in their corpus cannot be aligned to one or more corresponding complex sentences based on its meaning. This of course limits the extent to which a large language model can benefit from training on the corpus.

In this paper, we leverage the neural alignment model proposed by Jiang et al. (2020) in order to automatically align the simple and complex sentences in the corpus. We then improve the quality of the corpus by deleting all simple sentences that are not aligned from the references. We filter out instances in which the resulting reference resembles a summarization rather than a simplification. Furthermore, we leverage the generated alignments in order to provide references not only for each complex document, but also for each sentence and

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 41–51

paragraph within the document. Hence, we present Cochrane-auto: a large, high quality dataset for the simplification of biomedical abstracts at the document-, paragraph- and sentence-level. An example is shown in Figure 1.

We validate that Cochrane-auto is a valuable resource by training simplification systems on this dataset and evaluating them against a baseline trained on the original corpus. Our results demonstrate that the plan-guided simplification system from Cripwell et al. (2023b) is indeed able to outperform the baseline after training on our dataset.

The rest of this paper continues with related work (§2), the CDSR (§3), the Cochrane corpus (§4), our new Cochrane-auto corpus (§5), our experiments (§6), and ends with the conclusion (§7) and limitations (§8).

2 Related Work

This section describes the related work on biomedical text simplification and lay summarization.

Biomedical text simplification Our approach closely follows Devaraj et al. (2021), who introduced a dataset of parallel plain language summaries and technical abstracts from the Cochrane Database of Systematic Reviews. We describe their dataset in Section 4 below. Grabar and Cardon (2018) created the CLEAR corpus, which includes 13 manually aligned Cochrane abstracts and plain language summaries in French. Ermakova et al. (2022) introduced a pilot scientific text simplification corpus of aligned sentence pairs with manual simplifications by non-experts. This pilot data set contains 147 abstracts with 648 sentences, of which 25 abstracts and 179 sentences are from the biomedical domain. Attal et al. (2023) created a set of 750 medline abstracts containing 7,643 sentences paired with expert-created sentence-level plain language adaptations. This data set is used at the TREC 2024 Plain Language Adaptation of Biomedical Abstracts (PLABA) track.[1] These earlier biomedical text simplification data sets are immensely valuable, but limited in size and restricted to sentence-level simplifications, with less freedom than observed in real-world paragraph or document level plain English summaries.

Plain English summaries Several journals, in particular in the biomedical domain, have collected plain English summaries. These plain English summaries are provided by the original authors of the paper, with varying degrees of instruction. In particular for systematic reviews very detailed instructions exist. Whiting and Davenport (2023) in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, provide detailed instructions on plain language summaries. Systematic reviews follow very strict evidence based medicine rules, such as PRISMA 2020 (Page et al., 2021). The specific guidelines for writing Cochrane Plain Language Summaries were published in 2020.[2] Our Cochrane-auto dataset contains both lay summaries from before and after the introduction of detailed template and guidelines in 2020.

Lay summarization The lay summarization task was introduced at SDProc in 2020. Chandrasekaran et al. (2020) discuss the LaySumm task of the Scholarly Document Processing Workshop at EMNLP2020.[3] LaySumm provided 572 author-generated lay summaries from a multidisciplinary collection of journals together with their corresponding full text content and abstracts. Goldsack et al. (2022) create a PLOS and e-Life corpora containing full scientific articles paired with manually created lay summaries. There was a BioLaySumm Task 1 shared task, held at the BioNLP 2023 Workshop (Demner-Fushman et al., 2023).[4] This task uses similar PLOS/e-Life corpora for a lay summarization task. Recently, Pu et al. (2024) created another SciNews corpus for plain English summarization, based on crawling scientific articles discussed in popular science news web site Science X.[5]

Prior work focused on the summarization aspects of lay summarization, whereas our paper focuses on realligning the abstracts and lay summaries, to create matching documents, paragraphs, and sentence-pairs, in a way that replicates earlier text simplification corpora.

3 The CDSR

The Cochrane Database of Systematic Reviews[6] (CDSR) comprises systematic reviews of research in health care and health policy. A *systematic review* attempts to identify, appraise and synthesize

[1] https://bionlp.nlm.nih.gov/plaba2024/

[2] https://training.cochrane.org/handbook/current/chapter-iii-s2-supplementary-material

[3] https://sdproc.org/

[4] https://biolaysumm.org/

[5] https://sciencex.com/

[6] https://www.cochranelibrary.com/cdsr/reviews

all empirical evidence that is relevant to a specific research question. Cochrane reviews are internationally recognized as the highest standard in evidence-based health care. They are written according to a comprehensive set of guidelines.[7] Each review includes a technical abstract, which is targeted at healthcare decision makers, and a plain language summary, which should be understandable for a wide range of non-expert readers.

4 The Cochrane corpus

In this section, we first give a short description of the Cochrane corpus. Second, we present a limited analysis on a selection of its contents. Third, we introduce an updated version of the corpus.

Description Devaraj et al. (2021) observed that portions of the plain language summaries in the CDSR contain roughly the same content as analogous sections in the technical abstracts. This motivated them to compile a corpus of paired (complex, simple) texts in English, comprising parallel subsets of abstracts and lay summaries from the CDSR. Each subset contains the full text from the description of studies and results onward. Abstracts adhere to a standard format, and so each complex text in the corpus covers the *Main Results* and *Authors' Conclusions* sections of the technical abstract. Plain language summaries are structured heterogeneously. Therefore, the authors made use of substring matching to determine the approximate location of the first section, paragraph or sentence (depending on the structure) describing the studies and results. They defined everything in the lay summary from that point onward as the simple text.

Analysis We randomly select ten paired (complex, simple) texts from the corpus. Next, we manually align sentences between these pairs that are equivalent or partially equivalent in meaning. As a result, we obtain 79 alignments. Of the total of 98 simple sentences in the selected texts, 68 are aligned to at least one of the 139 complex sentences. Thus, 30 out of 98 simple sentences are not aligned. While 2 of them are elaborations, the remaining 28 contain information that is present in the full review but not in the complex text. This is largely because the plain language summaries are written directly from the full review, instead of being simplified versions of the technical abstracts, and partially because the complex texts are only subsets of these abstracts. Consequently, around 29% of the sentences in the simple reference texts cannot be generated from the complex source text. This of course limits the suitability of the corpus for directly training and evaluating simplification models.

Update We run the authors' code[8] to obtain an updated version of their corpus. This corpus is based on systematic reviews that were published in the CDSR up until March 14, 2024. We apply the same preprocessing, except for the filtering of texts with more than 1,024 tokens. The resulting corpus consists of 4,468 train, 558 validation and 559 test pairs. On average, the complex and simple texts consist of 17.1 and 12.5 sentences, respectively.

5 Cochrane-auto

In this section we describe i) our alignment model, ii) our alignment procedure, iii) our alignment results, iv) the preprocessing of the resulting dataset, and v) the labeling.

5.1 Alignment model

We make use of the neural CRF alignment model proposed by Jiang et al. (2020). When provided with a (complex, simple) text pair as input, this model automatically aligns each sentence in the simple text to either one or zero corresponding sentences in the complex text. In doing so, it leverages the similar order of sentences in parallel texts and utilizes a fine-tuned BERT model to capture the semantic similarity between sentence pairs. Aligned sentences should be equivalent or partially equivalent in meaning, and multiple simple sentences may be aligned to the same complex sentence.

The authors applied their model to two simplification corpora: Newsela (Xu et al., 2015), which comprises news articles that were manually rewritten at different levels of simplification, and an updated version of the Wikipedia corpus (Zhang and Lapata, 2017), which consists of paired articles from English Wikipedia and Simple English Wikipedia. More specifically, the authors first created Newsela-manual and Wiki-manual by manually aligning 50 article groups from Newsela and 500 article pairs from Wikipedia. Then they fine-tuned BERT (Devlin et al., 2019) and trained their alignment models on train splits of these datasets. Finally, they applied their trained models to the

[7] https://training.cochrane.org/handbook

[8] https://github.com/AshOlogn/Paragraph-level-Simplification-of-Medical-Texts

	TP	FP	FN	**F1**
BERT$_{finetune}$	52	32	27	63.8
CRF Aligner	53	6	26	76.8
+ merge	56	8	23	78.3

Table 1: Performance of sentence alignment methods on 10 annotated text pairs from the Cochrane corpus.

	Cochrane-auto	Newsela-auto	Wiki-auto
Domain	Biomedical	News	General
# Doc Pairs	5,585	18,820	138,095
# Sent Pairs	35,800	813,972	685,769

Table 2: Statistics for the automatically aligned Cochrane-auto, Newsela-auto and Wiki-auto datasets.

remaining data to create the automatically aligned Newsela-auto and Wiki-auto datasets.

5.2 Alignment procedure

We create Cochrane-auto by applying the sentence alignment model that was pretrained on Wiki-manual to the updated Cochrane corpus. More precisely, we utilize the neural CRF model that we trained on Wiki-manual ourselves by running the authors' code.[9] It employs the BERT model[10] which the authors fine-tuned on the same train set to capture semantic similarity. According to Jiang et al. (2020), their fine-tuned BERT models should be able to achieve competitive performance on other monolingual parallel data, and the performance boost of adding the neural CRF model is related to the structure of the articles. Our motivation for pretraining the alignment model on Wiki-manual, and not Newsela-manual, is that the Cochrane and Wikipedia corpora both contain (complex, simple) text pairs in which the simple text is no direct simplification of the complex text.

Wiki-auto and Newsela-auto were created by first aligning paragraphs and then aligning the sentences within those paragraphs. We can also divide the texts in the updated Cochrane corpus into paragraphs based on sections and newlines. However, the sentence-level alignments between these texts generally do not reside within paragraph pairs, since these texts can be structured in a different way. We therefore apply our alignment model to the full text pairs to create Cochrane-auto.

As a result of our alignment strategy, similar sentences from different paragraphs in a simple text may be automatically aligned to the same sentence in the parallel text. For example, two simple paragraphs describing the results and conclusion may feature equivalent sentences that are both aligned to the same complex sentence; yet only one of them should be used as a reference simplification. In those cases, we leverage the fine-tuned BERT model to find the simple paragraph in which the aligned sentences have the highest similarity with the complex sentence. Then we delete all alignments between the complex sentence and the simple sentences in other paragraphs.

5.3 Alignment results

Given the paragraph alignments that were generated by Jiang et al. (2020), our trained sentence alignment model achieves an F1-score of 81.5 on the Wiki-manual test set. This is lower than the F1-score of 85.3 reported in their paper, but we do not have access to the original model weights. We also evaluate the performance of the fine-tuned BERT model alone, and find its F1-score to be 83.4. This value is obtained by computing the semantic similarity of each sentence pair within the aligned paragraphs, and aligning the pairs with a similarity higher than a threshold tuned on the dev set.

On our manually annotated subset of the Cochrane corpus, the neural CRF aligner significantly outperforms the fine-tuned BERT model. This is shown in Table 1. The higher F1-score indicates that our pretrained model is effectively able to capitalize on the structure of parallel texts in the Cochrane corpus. Since the neural CRF model normally cannot align multiple complex sentences to one simple sentence, its upper bound for the number of true positives is 68 out of 79. Nevertheless, in Section 5.5 we introduce merge operations, which can be translated into such n-to-1 alignments. Table 1 shows that adding these alignments leads to a small improvement in F1-score on our manually annotated subset.

Finally, we apply the neural CRF model to all 5,585 text pairs in the updated Cochrane corpus. This yields 39,497 automatic sentence alignments, some of which we delete as described in Section 5.2. The remaining 35,800 sentence pairs together with the corresponding document pairs

[9] https://github.com/chaojiang06/wiki-auto

[10] https://huggingface.co/chaojiang06/wiki-sentence-alignment

	Cochrane-auto	Newsela-auto	Wiki-auto
# Doc Pairs	1,085	18,319	85,123
# Para Pairs	4,171	361,964	178,982
# Sent Pairs	14,719	707,776	461,852
Avg. $\|c_i\|$	35.61	22.49	28.64
Avg. $\|s_i\|$	27.75	15.84	21.57
Avg. n	13.57	38.64	5.43
Avg. k	9.01	42.60	4.53
Avg. p	3.53	1.96	2.58

Table 3: Statistics of the datasets after preprocessing, where n is # sentences in C, and k is # sentences in S and p is # sentences per paragraph in C.

constitute Cochrane-auto. In Table 2, we compare this dataset to other automatically aligned simplification datasets. We make Cochrane-auto publicly available to foster research on the simplification of biomedical documents.

5.4 Preprocessing

For the training and evaluation of simplification systems on Cochrane-auto, we preprocess our data similarly to how Cripwell et al. (2023b) preprocessed Newsela-auto and Wiki-auto. That is, for each sentence c_i in a complex document, we use the simple sentence s_j to which it is aligned as a reference. If it is aligned to multiple s_js, we concatenate them; if it is not aligned, we use an empty string. Next, we create paragraph- and document-level references by concatenating the references for each sentence in a complex paragraph or document. Note that this may change the order of the simple sentences. Even so, we find that the resulting references are relatively coherent, as the simple sentences mostly stand on their own. Importantly, also note that simple sentences which are not aligned to any c_i are not included in any reference. Henceforth, when we refer to the simple sentences, paragraphs and documents in Cochrane-auto, we mean these references.

Let us define an instance of Cochrane-auto to be the collection of all (source, reference) pairs derived from a single text pair. We filter out instances where less than 50% of the sentences in the corresponding complex document C are aligned to any s_j. Therewith, we ensure that the remaining instances are derived from text pairs that are sufficiently similar in meaning. We also remove instances where the length of a document exceeds 1,024 tokens, or would exceed 1,024 tokens after adding the special tokens needed for the plan-guided simplification approach of Cripwell et al. (2023a). As a result, the preprocessed Cochrane-auto dataset consists of 894 train, 125 validation and 121 test instances. In Table 3, we compare the statistics of our dataset to those of the preprocessed Newsela-auto and Wiki-auto datasets, as reported by Cripwell et al. (2023b).

Copy	Rephrase	Split	Merge	Delete
8.4	45.3	4.5	6.5	35.3

Table 4: Operation class distribution for Cochrane-auto in percentages.

Figure 2 displays a short example of a complex-simple document pair from the preprocessed Cochrane-auto dataset, along with the corresponding original reference from the Cochrane corpus. In this example, the first sentence from the original reference cannot be generated based on the complex document, and as such it should not be used as a reference. Indeed, it is excluded from the new reference, because it could not be aligned to any complex sentence. Moreover, the four sentence pairs that were aligned, are correctly aligned. However, this example also shows that correctly aligned sentences may still contain information that is not present in the source sentence (*with PAD*), that the deletion and reordering of sentences may impact the discourse structure of the reference document (*None of the other*), and that it is often debatable whether the meaning of source and target sentences is similar enough to align them (the last sentence in the original reference).

5.5 Labelling

Using the same approach that Cripwell et al. (2023b) applied to Newsela- and Wiki-auto, we label each complex sentence c_i in Cochrane-auto with a simplification operation as follows:

Delete: c_i is not aligned to any s_j.

Copy: c_i is aligned to a single s_j with a Levenshtein similarity above 0.92.

Rephrase: c_i is aligned to a single s_j with a Levenshtein similarity below 0.92.

Split: c_i is aligned to multiple s_js.

Additionaly, we introduce a new simplification operation, namely ***merge***. This is motivated by the observation that one sentence in a plain lan-

Complex document

Two randomised trials with a total of 161 participants were included in this review. The studies did not report on mortality and rate of limb loss. One randomised trial with a total of 133 participants showed that there was a significant improvement in ankle brachial index (ABI) in participants who received folic acid compared with placebo (mean difference (MD) 0.07, 95% confidence interval (CI) 0.04 to 0.11, P < 0.001) and in participants who received 5-methyltetrahydrofolate (5-MTHF) versus placebo (MD 0.05, 95% CI 0.01 to 0.10, P = 0.009).A second trial with a total of 18 participants showed that there was no difference (P non-significant) in ABI in participants who received a multivitamin B supplement (mean ± SEM: 0.7 ± 01) compared with placebo (mean ± SEM: 0.8 ± 0.1). No major events were reported.

Currently, no recommendation can be made regarding the value of treatment of hyperhomocysteinaemia in peripheral arterial disease. Further, well constructed trials are urgently required.

Simple document

Two trials with 161 participants with PAD were included in this review. None of the other predefined primary outcomes (mortality and rate of limb loss) were assessed in these studies. One trial showed a significant improvement in the ankle brachial index (ABI) in participants treated daily with 400 μg folic acid or 5-methyltetrahydrofolate (5-MTHF). A second trial showed that there was no difference in ABI in participants who received a multivitamin B supplement compared with placebo.

Original reference

We looked at studies where treatments to lower homocysteine were used in people with PAD and hyperhomocysteinaemia. Two trials with 161 participants with PAD were included in this review. One trial showed a significant improvement in the ankle brachial index (ABI) in participants treated daily with 400 μg folic acid or 5-methyltetrahydrofolate (5-MTHF). A second trial showed that there was no difference in ABI in participants who received a multivitamin B supplement compared with placebo. None of the other predefined primary outcomes (mortality and rate of limb loss) were assessed in these studies. More research about the effect of homocysteine lowering therapy on the clinical progression of disease in people with PAD and hyperhomocysteinaemia is needed.

Figure 2: A complex-simple document pair from the preprocessed Cochrane-auto dataset, along with the corresponding original reference from the Cochrane corpus.

guage summary can have the same meaning as multiple complex sentences in the parallel abstract. By swapping the (complex, simple) inputs of our alignment model, we automatically align each complex sentence to one or zero simple sentences, instead of the reverse. Then we assign consecutive sentences c_i within the same paragraph the *merge* operation label if (1) they are aligned to the same simple sentence s_j, (2) one of them was already aligned to s_j and labelled *rephrase*, and (3) the other complex sentences were labelled *delete*. Because of the latter two conditions, we only add alignments from previously unaligned complex sentences to simple sentences that were already included in our references. Table 1 showed that adding these alignments can indeed lead to an improvement in alignment quality.

Table 4 shows the distribution of simplification operations for Cochrane-auto. In comparison with Newsela-auto and Wiki-auto, the classes are more imbalanced, as Cochrane-auto contains more rephrase and delete operations, and less copy and split operations. This is clearly a result of the plain language summaries being written largely independently from the technical abstracts.

6 Experiments

In this section, we train document simplification systems on Cochrane-auto and evaluate them against a baseline on the updated Cochrane corpus. We describe our planning and simplification models, our experimental setup and evaluation metrics, and our results.

6.1 Simplification models

We finetune BART (Lewis et al., 2020) to perform simplification on the documents (BART_{doc}), paragraphs ($\text{BART}_{\text{para}}$), and sentences ($\text{BART}_{\text{sent}}$) in Cochrane-auto. In doing so, we exclude sentences which are labelled *merge* from the training data for $\text{BART}_{\text{sent}}$. As a baseline, we use BART finetuned on the updated Cochrane corpus.

Furthermore, using same approach that Cripwell et al. (2023b) applied to Newsela-auto, we train a plan-guided simplification model ($\hat{O} \rightarrow \text{BART}_{\text{sent}}$) on Cochrane-auto. This is a modified version of $\text{BART}_{\text{sent}}$ that takes a control-token at the beginning of each input, representing the simplification operation (Section 5.5) that should be applied to it. Sentences which should be merged are concatenated and provided to the model together. During training, the ground-truth simplification operation labels are used as control-tokens. At inference time, the operations are predicted by a planning model.

6.2 Planning model

The task of a planning model is to predict a simplification operation for each sentence in a complex document. For example, the RoBERTa-based (Liu et al., 2019) classifier from Cripwell et al. (2023b) takes a tokenized sentence as input and outputs a

System	BARTScore ↑			BLEU ↑	FKGL ↓	SARI ↑	Length	
	P ($r \rightarrow h$)	R ($h \rightarrow r$)	F1				Tok.	Sent.
Input	-3.44	-3.01	-3.22	13.7	13.4	9.3	534.0	15.0
Reference	-0.62	-0.62	-0.62	100.0	12.6	99.6	286.8	10.8
Baseline	-3.57	-3.32	-3.44	11.4	12.6	34.1	250.5	9.1
BART_{doc}	-3.70	-3.36	-3.53	10.7	**12.3**	31.6	251.4	8.8
$\text{BART}_{\text{para}}$	-3.43	-3.25	-3.34	12.9	12.7	32.9	263.0	9.7
$\text{BART}_{\text{sent}}$	-3.26	**-3.15**	**-3.20**	**14.9**	12.5	32.0	298.8	12.1
$\hat{O} \rightarrow \text{BART}_{\text{sent}}$	**-3.21**	-3.41	-3.31	11.1	12.5	**35.5**	211.6	8.1

Table 5: **Results of document simplification systems trained on Cochrane-auto**, when evaluated on the updated Cochrane corpus. The baseline is BART trained on the updated Cochrane corpus. For BARTScore, h is the hypothesis and r is the reference.

prediction score for each operation class. We train a similar classifier to predict the label of each complex sentence in Cochrane-auto. Since our planning model must be able to predict merge operations, we also provide the subsequent sentence as input to the classifier. If the model predicts that these sentences should be merged, we label both of them with the merge operation. Otherwise, we let the classifier predict the label of the first sentence. We provide the classifier with a single sentence if that sentence appears at the end of a paragraph.

6.3 Experimental setup

We build upon the code[11] of Cripwell et al. (2023b) to train and evaluate our planning and simplification models. Moreover, we apply length-based filtering to the updated Cochrane corpus, so that it contains 3,967 train, 500 validation and 502 test pairs of $\leq$ 1,024 tokens each. We leverage this corpus to evaluate our document simplification systems and to train the baseline, while we train our other models on Cochrane-auto. After training the planning model for 10 epochs, we select the model checkpoint with the highest macro F1-score on the validation set. With regards to the simplification models, we implement early stopping based on the validation loss with a patience of 3 epochs. All other training details are the same as to those originally used by the authors of the code.

6.4 Evaluation metrics

In order to evaluate the simplifications generated by our systems, we leverage BARTScore (Yuan et al., 2021) and BLEU (Papineni et al., 2002) as analogs for meaning preservation and fluency. Furthermore, we assess readability using the Flesch-Kincaid grade level (FKGL) (Kincaid et al., 1975), and simplicity using SARI (Xu et al., 2016).

6.5 Results and Discussion

Table 5 summarizes the results of evaluating our document simplification systems on the updated Cochrane corpus. Along the dimension of readability, all of our systems obtain mean FKGL scores that are comparable to the mean reference score. However, this score is relatively high, underlining the difficulty of writing easy-to-read biomedical lay summaries. Besides, FKGL is computed based on syllable counts and sentence length, so that it does not directly capture the amount of background knowledge needed to read a text. In fact, adding statistics such as confidence intervals to a text may reduce the FKGL score, because it decreases the average amount of syllables per word. This explains why the mean readability score of the inputs is only 0.8 above that of the references.

Taking a look at the other metrics, the scores obtained by our systems appear to be relatively low. This is because, despite being written according to a comprehensive set of guidelines, there is much variety in the Cochrane references compared to the Newsela and Wikipedia references. Not only does this influence evaluation, but also does the resulting unpredictability make our trained systems relatively conservative. In addition, as discussed in Section 4, our scores are negatively impacted by the fact that parts of the references cannot be generated based on the source document. Nevertheless,

[11] https://github.com/liamcripwell/plan_simp

Simplification generated by the baseline

Two randomised trials with a total of 161 participants were included in this review. The studies did not report on mortality and rate of limb loss. One randomised trial showed that there was a significant improvement in ankle brachial index (ABI) in participants who received folic acid compared with placebo (mean difference (MD) 0.07, 95% confidence interval (CI) 1.04 to 0.11, P < 0.001). No major events were reported. Currently, no recommendation can be made regarding the value of treatment of hyperhomocysteinaemia in peripheral arterial disease. Further, well constructed trials are urgently required.

Simplification generated by $\hat{O} \rightarrow \textbf{BART}_{\textbf{sent}}$

Two randomised controlled trials with a total of 161 participants were included in this review. The studies did not report on mortality and rate of limb loss. One trial with a total of 133 participants showed that there was a significant improvement in ankle brachial index (ABI) in participants who received folic acid compared with placebo. A second trial with a total of 18 participants showed that there was no difference (P non-significant) in ABI in participants who received a multivitamin B supplement compared with placebo. No major events were reported.

Figure 3: The outputs of two document simplification systems for the complex input document in Figure 2.

we find that these evaluation scores are useful for comparing performances between systems.

To begin with, it can be seen that BART_{doc} underperforms compared to the baseline in terms of both meaning preservation and fluency, and simplicity. One reason could be that the baseline was trained on a larger dataset, as Cochrane-auto comprises only those document pairs in which at least 50% of the complex sentences were automatically aligned. Another reason is that the exclusion of unaligned sentences from the references in Cochrane-auto will to some extent have led to a loss of relevant information. This includes elaborations, sentences that were left unaligned due to alignment errors, and information that could only be aligned at the word-level rather than the sentence-level.

Furthermore, it can be observed that $\text{BART}_{\text{para}}$ and – to a larger extent – $\text{BART}_{\text{sent}}$ outperform the baseline along the dimension of fluency and meaning preservation, although they underperform along the dimension of simplicity. $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ even outperforms the baseline in terms of both SARI (simplicity) and BARTScore F1, while its BLEU score is only slightly lower than that of the baseline. These findings demonstrate that training simplification systems on Cochrane-auto, rather than the updated Cochrane corpus, can be beneficial despite all limitations mentioned above. Thus, we conclude that the creation of Cochrane-auto has indeed been a valuable contribution.

Lastly, Figure 3 displays the outputs of our baseline and $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ when they attempt to simplify the complex input document from Figure 2. It can be seen that both systems are indeed relatively conservative. Moreover, in this example, our plan-guided system is better able to determine which sentences should be kept and which ones should be deleted. Because it was trained using oracle labels, the simplification model has learned to actually delete any sentence whose predicted label is *delete*. This explains why our plan-guided simplification system generates the shortest outputs on average, especially when compared to $\text{BART}_{\text{sent}}$, which rarely deletes sentences due to its risk-avoiding nature. We conclude that having a seperate planning and simplification component has helped the system to be less conservative and thereby outperform the baseline.

7 Conclusion

In this paper, we presented Cochrane-auto: a large aligned dataset for the simplification of biomedical abstracts at the document-, paragraph- and sentence-level. Our freely available corpus complementing Newsela-auto and Wiki-auto facilitates text simplification research beyond direct lexical and grammatical revisions. Experiments demonstrated that a plan-guided simplification system trained on this corpus can outperform a strong baseline trained on unaligned abstracts and lay summaries. Future work will investigate the performance of more modern simplification systems when trained on this corpus.

8 Limitations

Our experiments are restricted to English data in the biomedical domain. There is obvious interest in looking at a more diverse set of languages, and several researchers and projects are currently working on this. This is witnessed by, for example, a recent Coling/LREC workshop devoted to this (Nunzio et al., 2024).

For those looking for very strict lexical and grammatical simplifications at the sentence-level, the plain English summaries have greater variation and incorporate the discourse structure of the entire paragraph and document. Although we filter and realign exactly as done in Wiki-auto and Newsela-auto (Jiang et al., 2020; Cripwell et al., 2023a; Bakker and Kamps, 2024), and hence have similar safeguards between aligned sentences, we observe

greater variation in Cochrane-auto. As in the other collections, our automatic alignments are imperfect, and the simple sentences that are correctly aligned may still contain information that is not present in the source sentence(s). More generally, the main limitation of our approach is that the real alignments between the complex and simple texts may not reside at the sentence-level. There are also obvious advantages to incorporating the variation and the discourse structure of the entire paragraph and document, and to further extend the scope of text simplification approaches to address all the interesting NLP challenges this presents.

As all generative models, our simplification models may suffer from creative generation (or "hallucination"), and so their outputs should not be used without manual inspection. In our text simplification setting, we can further analyse and ground the output of the model with the original source text. Hence, text simplification present an excellent setting to further study and quantify the degree of revision and additions generated by the model. This also inspired our introduction of a "merge" operator, aligning source content previously considered as delete combined with a creative insertion. As is well-known, existing evaluation measures are almost blind to detect such issues. The importance of studying and addressing these aspects is of paramount importance in future research, as they present one of the greatest challenges of generative models in NLP today.

9 Lay Summary

Many people have questions about health or medical topics. The most accurate and reliable information to answer such questions is in the biomedical literature written and used by medical experts. However, this scientific literature is very difficult to understand for non-experts. Fortunately, sometimes a special lay summary (like this one) is added to a paper to convey the main points. This is really helpful, but only few scientific articles have this, and not all the content of the articles has been "translated" for lay readers. This paper uses pairs of lay summaries and expert abstracts to create the training data for new AI models. We show that our corpus helps to build text simplification models that can automatically "translate" expert biomedical text for lay persons. This can lead to novel tools that make authoritative information from the biomedical literature directly available to non-experts.

Acknowledgments

Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Jan Bakker is funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Jan Bakker and Jaap Kamps. 2024. Beyond sentence-level text simplification: Reproducibility study of context-aware document simplification. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 27–38, Torino, Italia. ELRA and ICCL.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Dina Demner-Fushman, Sophia Ananiadou, and Kevin Cohen, editors. 2023. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Association for Computational Linguistics, Toronto, Canada.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Élise Mathurin, and Patrice Bellot. 2022. Overview of the CLEF 2022 simpletext lab: Automatic simplification of scientific texts. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 470–494. Springer.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Giorgio Maria Di Nunzio, Federica Vezzani, Liana Ermakova, Hosein Azarbonyad, and Jaap Kamps, editors. 2024. *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. 2021. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dongqi Pu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Penny Whiting and Clare Davenport. 2023. *Writing a plain language summary*, chapter 13. John Wiley & Sons, Ltd.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A Data, code, and trained models

We share all our data and the code used to create our new dataset (https://github.com/JanB100/cochrane-auto), as well as the code used to train and evaluate our simplification systems (https://github.com/JanB100/doc_simp), on GitHub. In addition, we share our pretrained planning and simplification models on HuggingFace (https://huggingface.co/janbakker) .

Cochrane-auto is freely available for research, and avoids the (almost) impossible to obtain license issues of the Newsela-auto collection. It also complements earlier direct biomedical sentence to sentence level simplification corpora with the great variation observed in human paragraph- and document-level plain English versions broadly conveying the same information.

These resources offer an easy starting point for NLP research in sentence-level, paragraph-level or document-level biomedical text simplification.

B Planning results

Copy	Rephrase	Split	Merge	Delete
3.3	51.2	0.0	0.0	45.5

Table 6: Distribution of operation classes predicted by our classifier on the updated Cochrane corpus in percentages.

Table 6 shows the distribution of operation classes predicted by our planning model on the updated Cochrane corpus. Unfortunately, the classifier never predicts *merge* and *split* operations and rarely predicts *copy* operations. This is largely a result of the infrequency of these labels in the training data.

C Cochrane-auto example

Figure 4 displays another example of a complex-simple document pair from the preprocessed Cochrane-auto dataset, along with the corresponding original reference from the Cochrane corpus.

Complex document

Fifteen heterogeneous trials, involving 1022 adults with dorsally displaced and potentially or evidently unstable distal radial fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation especially in terms of patient characteristics and interventions. Methodological weaknesses among these trials included lack of allocation concealment and inadequate outcome assessment.

External fixation maintained reduced fracture positions (redisplacement requiring secondary treatment: 7/356 versus 51/338 (data from 9 trials); relative risk 0.17, 95% confidence interval 0.09 to 0.32) and prevented late collapse and malunion compared with plaster cast immobilisation. There was insufficient evidence to confirm a superior overall functional or clinical result for the external fixation group. External fixation was associated with a high number of complications, such as pin-track infection, but many of these were minor. Probably, some complications could have been avoided using a different surgical technique for pin insertion. There was insufficient evidence to establish a difference between the two groups in serious complications such as reflex sympathetic dystropy: 25/384 versus 17/347 (data from 11 trials); relative risk 1.31, 95% confidence interval 0.74 to 2.32.

There is some evidence to support the use of external fixation for dorsally displaced fractures of the distal radius in adults. Though there is insufficient evidence to confirm a better functional outcome, external fixation reduces redisplacement, gives improved anatomical results and most of the excess surgically-related complications are minor.

Simple document

Fifteen trials, involving 1022 adults with potentially or evidently unstable fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation in their characteristics especially in terms of patient characteristics and the method of external fixation.

The complications, such a pin tract infection, associated with external fixation were many but were generally minor. Serious complications occurred in both groups.

The review concludes that there is some evidence to support the use of external fixation for these fractures. The review found that external fixation reduced fracture redisplacement that prompted further treatment and generally improved final anatomical outcome.

Original reference

Fifteen trials, involving 1022 adults with potentially or evidently unstable fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation in their characteristics especially in terms of patient characteristics and the method of external fixation. Weak methodology, such as using inadequate methods of randomisation and outcome assessment, means that the possibility of serious bias can not be excluded.

The review found that external fixation reduced fracture redisplacement that prompted further treatment and generally improved final anatomical outcome. It appears to improve function too but this needs to be confirmed. The complications, such a pin tract infection, associated with external fixation were many but were generally minor. Serious complications occurred in both groups. The review concludes that there is some evidence to support the use of external fixation for these fractures.

Figure 4: Another complex-simple document pair from the preprocessed Cochrane-auto dataset, along with the corresponding original reference from the Cochrane corpus.