Document-level Text Simplification with Coherence Evaluation

Laura Vásquez-Rodríguez[1,2,*], **Matthew Shardlow**[4],
Piotr Przybyła[5,6], **Sophia Ananiadou**[2,3]
[1]Idiap Research Institute, Martigny, Switzerland
[2]National Centre for Text Mining, The University of Manchester, Manchester, UK
[3]Artificial Intelligence Research Center (AIRC), Tokyo, Japan
[4]Department of Computing and Mathematics,
Manchester Metropolitan University, Manchester, UK
[5]Universitat Pompeu Fabra, Barcelona, Spain
[6]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
laura.vasquez@idiap.ch, m.shardlow@mmu.ac.uk
piotr.przybyla@upf.edu, sophia.ananiadou@manchester.ac.uk

Abstract

We present a coherence-aware evaluation of document-level Text Simplification (TS), an approach that has not been considered in TS so far. We improve current TS sentence-based models to support a multi-sentence setting and the implementation of a state-of-the-art neural coherence model for simplification quality assessment. We enhanced English sentence simplification neural models for document-level simplification using 136,113 paragraph-level samples from both the general and medical domains to generate multiple sentences. Additionally, we use document-level simplification, readability and coherence metrics for evaluation. Our contributions include the introduction of coherence assessment into simplification evaluation with the automatic evaluation of 34,052 simplifications, a fine-tuned state-of-the-art model for document-level simplification, a coherence-based analysis of our results and a human evaluation of 300 samples that demonstrates the challenges encountered when moving towards document-level simplification.

1 Introduction

Text Simplification (TS) is the process of transforming text into a simpler variant that is easier to understand for wider audiences (Rello et al., 2013a; Collantes et al., 2015; Xu et al., 2015; Paetzold and Specia, 2016; Scarton et al., 2018; Cao et al., 2020). Simplifications can vary depending on the audiences' needs and expertise. For example, people with disabilities, such as dyslexia, have a better understanding of content with shorter word lengths (Rello et al., 2013b), however, this aspect is not necessarily relevant for non-native speakers (Paetzold and Specia, 2016).

In the past decade, most of the research efforts in automatic TS have focused on simplification at the sentence level, without considering the impact of TS at a document level. When multiple sentences in text are simplified, the overall quality of the text is affected (Siddharthan, 2003). Incorrect simplifications impact the overall text meaning and create disruptions to its structure (e.g., a sentence split without using adequate sentence connectors). Sentence simplification usually does not take into account the wider context to which sentences belong. Nevertheless, most of the practical applications of TS are motivated by the target audience needing to understand complete documents rather than isolated sentences. In general, sentences are evaluated within the scope of the sentence that is being simplified without considering possible disruptions that can happen between the nearby context (e.g., sentences left unconnected or unrelated sentences).

The generation and evaluation of document-level simplification[1] have been explored to a limited extent (Section 2). Meanwhile, the discourse features, such as cohesion, coherence and anaphora, have been widely considered in related fields (Maruf et al., 2021). We choose coherence to measure the relatedness of sentences in document-level TS, because of the availability of annotated data.[2]

Coherence is a logical and structured relationship between co-located sentences. This relationship can be at a local level between nearby sentences. This is called local coherence. In contrast,

[*] Work done as a PhD student at the University of Manchester, United Kingdom.

[1] We refer to "document-level simplification" to multiple sentences or paragraphs, given the nature of existing TS datasets beyond sentence level. We report average numbers per document in Table 1.

[2] In the future, the support of additional evaluation metrics could be needed. These would address possible issues that arise in a document-level scenario such as the overdeletion and reordering of sentences, which could also affect the coherent aspects of the text.

this relation could be observed at a broader level, such as sentences in each section of a scientific paper, where they belong to a common topic. We refer to this scenario as global coherence (Jurafsky and Martin, 2021). When a sentence is incoherent, the logical relationship between the events is disrupted, such as in Example 1 (Li and Jurafsky, 2017). This example is readable, simple, and grammatically correct, but there is no logical sequence of events or discourse elements.

Hui went to a restaurant
She ordered a pizza (1)
She read a menu and sat down

In this paper, we contribute to the transition from sentence-level to document-level TS (DocS), carrying out experiments at a paragraph level to understand the possible challenges of this setting. To achieve this, we enhance a state-of-the-art sentence simplification model to perform DocS with paragraph-level data from the general and medical domains. In addition, we evaluate our system outputs using DocS metrics, such as coherence, readability, and simplicity, to validate the suitability of simplifications when multiple sentences are present. We summarise our main contributions as follows:

1. The evaluation of local coherence at the document level using state-of-the-art neural models. This task has not been explored before in the field of TS.

2. A state-of-the-art model for simplification generation at the document level, fine-tuned with paraphrasing data.

3. A manual analysis of the results and a human evaluation of simplifications that highlights the challenges and limitations faced when performing TS at the document level, including the evaluation of coherence.

2 Related Work

In the past, TS at the document level has scarcely been explored despite the known need for simplification methods and evaluation metrics beyond sentence-level (Alva-Manchego et al., 2019). Nevertheless, recently, new directions have been explored leveraging existing methods and resources from sentence-level domain (Siddharthan, 2003; Alva-Manchego et al., 2019; Sun et al., 2021; Cripwell et al., 2023b; Sun et al., 2023; Cripwell et al., 2023a; Joseph et al., 2023).

Similarly for document-level corpora, there have been limited efforts to alleviate the lack of parallel texts at a document level (Xu et al., 2015; Vajjala and Lučić, 2018). Recently, datasets for the general (Sun et al., 2020; Laban et al., 2023) and medical (Devaraj et al., 2021; Joseph et al., 2023) domains have been proposed, aligning existing corpora such as Wikipedia and Cochrane reviews.[3] These resources include complex and simpler variants of a text, which are leveraged for TS. The creation of new parallel corpora for document-level simplification is also increasing beyond English (Rios et al., 2021; Hauser et al., 2022; Trienes et al., 2022; Aumiller and Gertz, 2022), which enhances opportunities for cross-lingual settings.

In relation to the evaluation metrics, sentence-level research has typically relied on the following automatic metrics: SARI (Xu et al., 2016) for simplicity, Flesch–Kincaid Grade Level (FKGL) (Kincaid et al., 1975) for readability and BLEU (Papineni et al., 2002) for grammaticality. However, BLEU, typically used in TS and summarisation, has been discouraged due to its poor performance in simplification operations, such as sentence splitting, and its negative correlation with simplicity (Sulem et al., 2018). Similarly, there are also limitations considered for FKGL (Tanprasert and Kauchak, 2021). However, we still use this metric in our work to compare with previous work. At a document level, Sun et al. (2021) proposed D-SARI evaluation metric that considers additional document-level penalties for system outputs (e.g., simplifications that outnumber the gold standard in sentence count).

2.1 Coherence as a Metric for Evaluation

Document-level evaluation is used for several NLP applications (e.g., machine translation (Maruf et al., 2021), summarisation (Fabbri et al., 2021) and simplification (Devaraj et al., 2022)), covering a wide range of discourse phenomena, such as anaphora, cohesion and coherence. In particular, coherence has been considered for applications including summarisation and essay rating, where the relationship (e.g., common entities and topics) between sentences is relevant.

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

The evaluation of coherence has typically been analysed using methods, such as entity-grids (Barzilay and Lapata, 2008; Joty et al., 2018), graphs (Mesgar and Strube, 2015, 2016) and Rhetorical Structure Theory (Šnajder et al., 2019; Guz et al., 2020). Unfortunately, manual assessment of coherence is challenging and laborious. Therefore, artificially augmented data have been used, where an ordered paragraph is considered coherent, but its randomly reordered counterpart is assumed not to be (Mohiuddin et al., 2021). To improve this practice, Lai and Tetreault (2018) proposed the Grammarly Corpus of Discourse Coherence (GCDC), which is manually annotated by experts and non-experts (i.e., MTurk workers).

Overall, TS at a document level has been barely explored, mainly because of the low corpora availability and challenges in evaluation. We introduce coherence as an automatic metric for the first time in TS, using existing state-of-the-art coherence models trained on professionally-created corpora. Furthermore, beyond the limitations of the existing evaluation resources for DocS and the difficulty it represents for evaluators to assess coherence, we share a detailed analysis of challenges encountered when using coherence as an evaluation metric.

3 Methods

We describe the adaptation of sentence-level TS methods into a document-level setting. We trained a sentence-level state-of-the-art TS model using paragraphs (Section 3.1) for discourse generation (i.e., longer, well-structured and logically simplified texts). There is no limitation on the simplifications that can occur at the document level, which means that we can expect modifications at a lexical, syntactic or semantic level, inferred from the training data. After texts are generated, we evaluate our simplifications (Section 3.2) through document-level metrics for simplicity, readability and coherence. We demonstrate our selected methods in Figure 1.

3.1 Model

Our proposed coherence-aware TS approach extends sentence simplification models for document-level simplification. We generate simplification of multiple sentences by retraining the sentence simplification model on longer passages (i.e., paragraph-level or document-level data). We select the Multilingual Unsupervised Sentence Simplification by Mining Paraphrases (MUSS) model (Martin et al., 2022), a multilingual model designed for sentence simplification. Although MUSS was designed to output individual sentences, its underlying architecture is the language generation model BART (Lewis et al., 2020). BART is capable of generating longer outputs if trained for a specific task (e.g., summarisation (Goldsack et al., 2022)) by changing its constraints, such as the number of tokens in the output.

3.2 Evaluation

One of the main challenges of document-level TS is evaluation. When simplification of multiple sentences of text is performed, the continuity of the discourse can be disrupted, affecting the semantic narrative of the text. Since there is no single evaluation metric to capture all possible variations caused by simplification, we relied on different metrics to approximate the performance of our model.

We measured readability using FKGL (Kincaid et al., 1975), simplicity using D-SARI metric (Sun et al., 2021) and coherence using a neural approach (Section 3.2.1). We clarify that despite the well-known criticism of simplification evaluation metrics, we used D-SARI and FKGL as a baseline for comparison with previous work. Also, we discarded SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002) as evaluation baselines since they only deal with sentence-level TS. We expect that our initial efforts towards evaluation at a document level contribute to the development of TS.

3.2.1 Coherence

Since the aforementioned metrics (i.e., FKGL and D-SARI) do not measure any semantic component of the discourse structure, we selected coherence as a complementary evaluation metric. For the evaluation of coherence, we have selected a neural model trained on data annotated by experts as proposed by Lai and Tetreault (2018). We measured the coherence of the original text, the predicted simplification, and the gold-standard simplification to understand how coherence is affected during simplification. For this task, we selected the Paragraph Sequence (ParSeq) model (Lai and Tetreault, 2018). Its architecture consists of 3 stacked LSTMs. Each layer consists of sequences of word embeddings that represent sentences (layer 1), paragraphs (layer 2) and documents (layer 3).

The document represented in the last layer will be scored with a coherence label. This model con-

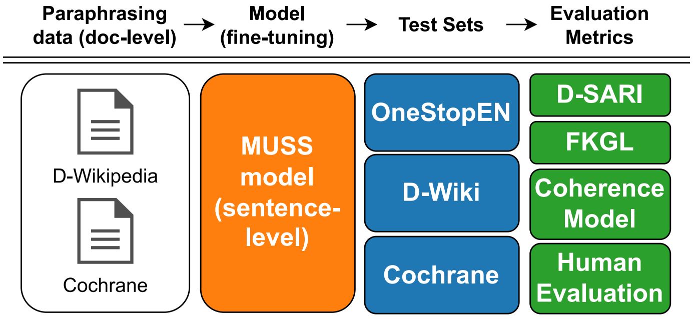


Figure 1: MUSS model fine-tuning with coherence evaluation.

siders the natural division of paragraphs (i.e. paragraph breaks) as an element to consider for the evaluation of coherence. The model was trained on the GCDC dataset (Lai and Tetreault, 2018), individually for each data source (i.e., Clinton, Enron, Yahoo and Yelp).

We adapted the model provided by the authors to our setting in order to evaluate our simplified outputs. We did not perform major modifications to the model; our changes were focused on the data processing stage to align with the expected format in the original model. The goal of this coherence model was to evaluate our predictions by assigning values to determine the quality of simplification in terms of coherence using the following scale: -1 (low coherence), 0 (medium coherence), and 1 (high coherence). Once the coherence model was trained, we scored the system outputs generated from our proposed baselines (Section 4.2).

The main limitation of these models is that although their training procedure is straightforward, the accuracy of the model is not high (Table 2). However, assessing coherence is challenging, even for humans (Lai and Tetreault, 2018). We still find coherence assessment valuable, especially when these models can help to discriminate between extremes (i.e., high or low coherence).

4 Experiments

We adapted three sentence simplification models (Section 4.2) for DocS by training them with paragraph-level data (Section 4.1). We evaluated our models with metrics that consider discourse factors to understand the impact on simplification when multiple sentences are involved (Section 4.3).

4.1 Datasets

We trained our sentence simplification models using corpora from the general and medical domains. These simplification models were trained using the training and validation sets. For the general domain, we used D-Wikipedia[4] dataset and for the medical domain, we used Cochrane[5], a paragraph-level dataset built by aligning the relevant sections of the Cochrane Database of Systematic Reviews (CDSR) abstracts and their plain language summaries. We show statistics of the selected datasets in Table 1.

With respect to the availability of datasets, document-level simplification resources are scarce. To alleviate this, we use “plain language”[6] as an alternative to the simple language. For Cochrane dataset, the plain language summary is a simpler version of the original, however, it may not be simple enough for many audiences. Tailoring simplification to a specific audience will require an additional step of personalisation, which is beyond this work.

Given our selected datasets and our training stage, we evaluated our model with the test sets available for D-Wikipedia and Cochrane. We also tested the models using the OneStopEnglish Corpus[7] to understand how well the model can generalise to external data. This dataset is divided into levels of complexity: Advanced, Intermediate and Elementary. We selected the samples from the Elementary level and Advanced level articles where the difference in complexity is more considerable.

For the evaluation of coherence, we used the released code and the dataset by Lai and Tetreault (2018) to train the proposed models since no trained models were made available. Lai and Tetreault (2018) introduced several coherence mod-

[4] https://github.com/RLSNLP/Document-level-text-simplification

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

[6] As defined by Cochrane in their guide: https://training.cochrane.org/guidance-writing-cochrane-plain-language-summary.pdf

[7] https://github.com/nishkalavallabhi/OneStopEnglishCorpus

Dataset	Subset	Samples	Sentences	Sent/Doc	Words (W)	W/Sent
Cochrane	train	3 568	51 280	14.37	1 478 770	28.84
	valid	411	5 788	14.08	168 365	29.09
	test	480	6 984	14.55	197 480	28.28
D-Wikipedia	train	132 546	652 644	4.92	18 776 870	28.77
	valid	3 000	14 764	4.92	425 317	28.81
	test	8 000	40 062	5.01	1 155 679	28.85
OneStopEN	all	2 623	7 115	2.71	182 224	25.61

Table 1: Datasets statistics. We report the total number of documents, sentences and words.

Dataset	Train Samples	Test Samples	Accuracy
Clinton	1000	200	42.00%
Enron	1000	200	48.50%
Yahoo	1000	200	52.00%
Yelp	1000	200	48.00%
All	4000	800	40.50%

Table 2: Coherence datasets statistics and classification task accuracy for the ParSeq Model

els, trained on four datasets: Yahoo[8], Clinton[9], Enron[10] and Yelp[11]. We selected the Yahoo dataset of the GCDC corpus[12], which consists of 369 texts for training and 76 texts for testing. This corpus was created using the Yahoo Questions and Answers dataset, which is freely available upon request for research purposes. We also performed experiments using all datasets combined. However, we did not get any improvement with respect to the Yahoo dataset, which initially performed well on the original paper benchmarks. For the replication of this experiment, we trained the ParSeq model with the original train split and tested it on the held-out dataset. For the "All" category, we created a combined dataset with all the available train and test splits. We present the results in Table 2 with the coherence evaluation for each dataset. These coherence models were trained to classify texts in low, medium and high coherence. For our experiments, we classified the outputs of the simplification models and report the normalised scores for simplification assessment as described in Section 4.3.

4.2 Models

We adapted the MUSS model (Martin et al., 2022) to generate document-level simplifications by removing sentence-level constraints (i.e., token limits for the output). We also updated the existing sentence-level evaluation of the original model to document-level, using the document-level evaluation metric D-SARI and the test sets from D-Wikipedia and Cochrane instead of SARI metric and ASSET (Alva-Manchego et al., 2020).[13]

The original MUSS model was fine-tuned in multiple datasets and languages. Among these available models, we selected the *Mined* model as a baseline, which was trained using mined paraphrases from the CCNet (Wenzek et al., 2020), a subset of an open source snapshot of the WWW. The model was trained with multiple sequences (i.e., groups of sentences with less than 300 characters) and it was designed to perform at a sentence level, which will make it a useful reference to compare to the document-level counterparts. This model is openly available,[14] avoiding the need to replicate the training stage. We decided not to use the *Mined+WikiLarge* model, since it was trained on a sentence-level dataset Wikilarge (Zhang and Lapata, 2017), which diverges from our objective of document-level TS.

On the basis of these resources, we tested the following combinations:

- **Mined+D-Wikipedia**: *Mined* model fine-tuned with D-Wikipedia train and validation sets.

- **Mined+Cochrane**: *Mined* model fine-tuned with Cochrane train and validation sets.

[8] L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part):https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[9] https://foia.state.gov/Search/Results.aspx?collection=Clinton_Email

[10] https://www.cs.cmu.edu/~./enron/

[11] https://www.yelp.com/dataset

[12] https://github.com/aylai/GCDC-corpus

[13] We make our code available in Github: https://github.com/lmvasque/ts-coherence

[14] https://github.com/facebookresearch/muss

- **Mined+D-Wikipedia+Cochrane**: *Mined+D-Wikipedia* model fine-tuned with Cochrane train and validation sets.

4.2.1 Training Details

We performed our training using 1 NVIDIA HGX A100 SXM4 80GB GPU and the same hyperparameters as in the original work by Martin et al. (2022). Training for the Cochrane and D-Wikipedia datasets took 1.3 days and 4 hours respectively. We used this hardware for convenience and due to time constraints, but these jobs can be replicated using a GPU with 32 GB of RAM.

4.3 Evaluation

We evaluated our models using readability, simplicity and coherence evaluation metrics. To calculate FKGL scores, we used the textstat[15] Python package. For D-SARI, we adapted the available code to score our simplification outputs, since the original code evaluates a single text and its gold standard at the same time. In addition, we also analysed the lengths of our predictions and references to further understand the impact on the D-SARI evaluation metric.

Finally, we evaluated coherence before and after simplification. The original GCDC corpus coherence ratings were given using the values of 1 (high), 2 (medium), or 3 (low). We used normalised coherence scores as follows: 1 for high coherence, 0 for medium coherence and -1 for low coherence. We used this scale to make it easier to understand by humans, as it seemed more natural for us. This however does not affect any of the computational aspects of the work. The coherence scores were calculated individually for each sample, and we report the average value for all the samples in the test set as shown in Table 3.

5 Results

We evaluated the results for simplification quality (D-SARI) and readability (FKGL) in Table 4 and Table 5, respectively. We also included D-SARI underlying scores related to three simplification operations: keep (D_{keep}), delete (D_{del}) and add (D_{add}). Since D-SARI is a relatively recent simplification metric, we performed a detailed analysis of the impact of the difference in lengths between predictions and references in the calculation of this metric. We aim to understand the underlying penalties from D-SARI, demonstrating how document-level TS models are likely to generate an output of different lengths, affecting the reliability of this metric. As a reference, we include our analysis in Appendix A.

In terms of readability, the FGKL metric (lower is better) in the model *Mined+D-Wikipedia+Cochrane*, showed the worst performance when evaluated using the Cochrane test set, with a score of 12.69. This result is mainly because the Cochrane articles are in the medical domain, where the vocabulary tends to be more complex and the sentences are longer. We also calculated the FKGL score of the gold reference corresponding to this simplification and we achieved a similar value of 12.43 for the Cochrane test set. The *Mined+D-Wikipedia* model showed the best readability results.

We selected the OneStopEnglish dataset as an external dataset for model generalisation. As shown in Figure 2f in the Appendix, almost all predictions are shorter than the gold standard simplification. Therefore, all values for D_{add} are low, due to the penalty of LP_1. In terms of readability, all models evaluated with OneStopEnglish test set showed better performance compared to Cochrane test set.

We evaluated our selected measure of coherence in our models' predictions, including comparisons between the inputs (complex), predictions (simple) and gold-standard simplifications. As shown in Table 3, our coherence predictions in OneStopEnglish, D-Wikipedia and Cochrane texts are affected by simplification. For OneStopEnglish and Cochrane, the coherence scores in all our predictions were lower than the complex text. In the D-Wikipedia test set, coherence values were lower only for the complex text for the model *Mined+Cochrane*. Since this test set was automatically aligned from Wikipedia, it may already have coherence limitations resulting from its original text.

Also, we note that for professionally written samples (OneStopEnglish) the coherence is significantly high, especially for the complex texts, which are more elaborated. Cochrane and OneStopEnglish gold-standard have a higher coherence, which may be related to the fact that these are written by professional authors, rather than created by crowdsourcing community as the D-Wikipedia dataset. Overall, gold-standard values

[15] https://pypi.org/project/textstat/

Model	Simple (prediction)	Simple (gold-reference)	Complex	Test Set
Mined	**0.167**	0.056	**0.222**	OneStopEnglish
Mined+D-Wikipedia	0.019			
Mined+Cochrane	0.0			
Mined+D-Wikipedia+Cochrane	-0.037			
Mined	0.041	-0.005	0.031	D-Wikipedia
Mined+D-Wikipedia	**0.098**			
Mined+Cochrane	0.028			
Mined+D-Wikipedia+Cochrane	0.041			
Mined	**0.047**	**0.061**	0.055	Cochrane
Mined+D-Wikipedia	-0.013			
Mined+Cochrane	-0.07			
Mined+D-Wikipedia+Cochrane	-0.053			

Table 3: Document level TS Models coherence evaluation. We evaluated each text with the following value of coherence: 1 (high), 0 (medium) and -1 (low). We report the average value for each model and test set.

Model	Test	D-SARI↑	D_{keep} ↑	D_{del} ↑	D_{add} ↑
Mined	OneStopEnglish	25.46	14.77	61.46	0.16
Mined+D-Wikipedia		24.67	11.71	61.85	0.46
Mined+Cochrane		**26.05**	**14.88**	**62.74**	**0.53**
Mined+D-Wikipedia+Cochrane		23.14	7.21	61.91	0.32
Mined	D-Wikipedia	26.67	19.56	**59.78**	0.68
Mined+D-Wikipedia		**32.51**	**27.43**	59.31	**10.77**
Mined+Cochrane		22.87	18.1	49.22	1.30
Mined+D-Wikipedia+Cochrane		22.39	12.74	53.27	1.17
Mined	Cochrane	**33.16**	17.09	**82.07**	0.33
Mined+D-Wikipedia		30.53	13.12	77.75	0.71
Mined+Cochrane		32.98	**18.06**	78.55	**2.32**
Mined+D-Wikipedia+Cochrane		32.14	16.20	78.55	1.68

Table 4: Document-level evaluation using D-SARI (complex, simple and reference).

are also higher than most of our predictions, except for D-Wikipedia, which again, is likely to have noisy alignments (e.g., no simplification, incorrect complex-simple pairs), affecting its coherence.

5.1 Manual Analysis of Coherence

To analyse complex, reference, and simplified sentences, we automatically scored 34,052 simplifications from all baselines using our selected coherence model. We summarised the scores of the evaluated texts in Table 3. Then, we performed a manual review of ∼50 samples with the goal of evidencing possible coherence issues. We selected texts that were negatively affected by the simplification process. A total of 12,585 samples were ranked as "Low" coherence. Additionally, we verified that their complex counterparts had a "Medium" or "High" coherence to ensure that it was not originally incoherent. This analysis was manually performed by the first author of this paper.

Our analysis confirmed the difficulty of distinguishing outputs between high coherence and medium coherence, as explained by Lai and Tetreault (2018). In some cases, the models may also assign low scores to complex sentences and references. This may be due to the fact that most of these texts are automatically aligned (except for OneStopEnglish) and also, because of the fair accuracy of the coherence model as shown in Table 2. Additionally, to support our findings, we analysed a set of low-coherence samples to highlight the potential issues related to coherence that can occur after simplification. We compared a set of complex sentences with their simple counterparts generated by the proposed simplification systems. We report below our analysis of the selected samples, including a summary of the coherence issues found, as shown in Table 6.

1. **Unconnected content:** content that differs from the original topic of the complex text. In Example 1 there is a 'review' or 'evaluation' which has no connection with the biography of the Nepalese actor. Also, in the first sentence in Example 3 it is not clear whether males earn more than women (when the original and remaining text state otherwise). These pitfalls are also referred to in TS research as factuality

Model	Test	$FKGL_c \downarrow$	$FKGL_s \downarrow$	$FKGL_r \downarrow$
Mined	OneStopEnglish	10.71	10.84	7.89
Mined+D-Wikipedia			10.51	
Mined+Cochrane			**9.84**	
Mined+D-Wikipedia+Cochrane			9.91	
Mined	D-Wikipedia	**10.14**	9.60	**7.10**
Mined+D-Wikipedia			**7.95**	
Mined+Cochrane			9.81	
Mined+D-Wikipedia+Cochrane			9.53	
Mined	Cochrane	10.40	12.24	12.43
Mined+D-Wikipedia			**11.37**	
Mined+Cochrane			12.00	
Mined+D-Wikipedia+Cochrane			12.69	

Table 5: Document-level evaluation using FKGL (complex, simple and reference).

Example #	Issue	Model	Test Set
1	unconnected ideas, words or phrase repetition	Mined+D-Wiki+Cochrane	D-Wikipedia
2	change in sentence order	Mined+D-Wiki	
3	unconnected ideas, words or phrase repetition	Mined+D-Wiki+Cochrane	OneStopEnglish
4	words or phrase repetition	Mined+D-Wiki+Cochrane	
5	unconnected ideas, lack of connectives, non-logical entities	Mined+D-Wiki	Cochrane
6	lack of connectives, words or phrase repetition	Mined+D-Wiki+Cochrane	

Table 6: Summary of coherence issues present in the manual analysis. We report the most representative issues found in Table 8 and 9, including information about the trained model and the test set used for evaluation.

evaluation (Devaraj et al., 2022) before and after simplification.

2. **Words or phrase repetition:** words or phrases can also show nonsense repetitions, such as "film film film" or "performed and performed" in Example 1 or 'in-human-induced climate' in Example 4. Similar situation for Example 6.

3. **Lack of connectives:** although sentences can have a related topic (i.e., topically coherent), they lack adequate connections between sentences. In Example 5 most of the sentences are introduced by "this is done", or sentences starting with "this". There is no fluent narrative in this text.

4. **Non-logical entities:** subjects or entities could be completely disconnected from the context, such as the word "motorage" in a clinical study (Example 5), lacking lexical coherence.

5. **Change in sentence order:** sentences in a text can keep their same content, but changing their original order and extracting them from the original context leads to less coherent ideas, such as in Example 2.

6 Human Evaluation

Due to the limitations of the coherence neural models, we further evaluate their performance against human criteria to better understand the existing gap with respect to automatic metrics. We performed a human evaluation of 300 samples of automatically simplified text, divided into 5 sets of 20 paragraphs; each set was annotated by 3 evaluators. For the evaluation, we recruited 15 annotators working within the NLP domain (staff and PhD students from the University of Manchester and Manchester Metropolitan University). We selected the *Mined+Cochrane* model evaluated in the Cochrane dataset and the *Mined+D-Wikipedia* model evaluated in the D-Wikipedia. We had a total of 50 unique texts for each model. We selected these models to measure coherence in both domains (medical and general) in their best setting (within their own test sets).[16]

As a result of our human evaluation, we noticed that texts from the general domain were perceived as more coherent than those from the medical domain. While some of our annotators had experience with texts from the medical domain, these are still significantly technical and seem incoherent for some of them. However, most of the texts from both domains were rated as high coherence. We also correlated the automatic scores for each of

[16] We explain in detail the proposed task in Appendix B

the evaluated texts. The correlation between automatic metrics and human evaluation was 0.029 and -0.085 for the general and medical domains.

The correlation between the coherence estimation of human annotators and the trained model is clearly weak.[17] The main reason is that the model has to operate outside its original domain: it was trained on documents written by human authors but was evaluated on the machine-generated text of simplifications. Designing architectures and training strategies for coherence assessment models that operate with good performance on substantially different data is a direction for future research.

7 Discussion

We trained the sentence-level MUSS model using paragraph-level data, evaluated with TS metrics and coherence. In our results, we observed the generation of longer sentences, in comparison to the original model. In addition, we saw an improvement in readability for the *Mined+D-Wikipedia* model using the D-Wikipedia test set compared to the other baselines. The *Mined+Cochrane* had the lowest performance, most likely since it belongs to the medical domain.

The reliable evaluation of TS remains a challenge. We noticed that the use of D-SARI evaluation is significantly affected by the penalties from differences in the number of words and sentences. This leaves other aspects of simplification unattended, mainly at a discourse level such as the generation of coherent, topically-related simplifications. When simplification is performed at the document level, there are more opportunities for elaboration (Srikanth and Li, 2021), but also, for shortening the content when it is explained in simpler words. Due to this, it is unlikely to find a strong correlation (i.e., equality in length) between the size of the predictions and the gold standard. This is one of the main weaknesses of traditional TS metrics (e.g., FKGL, D-SARI), which rely mostly on length aspect. Our analysis was done to demonstrate this limitation further and as a motivation for discourse-level evaluation metrics for TS.

The evaluation of coherence has shown new directions that could be explored to address this need. When simplification is performed beyond the sentence level, it disrupts the flow of ideas in the text and leaves sentences in paragraphs unconnected. As shown in Table 3, there is a decrease in coherence for both professionally and non-professionally written corpora for most of the models, which means simplification cannot be done without considering this aspect. In general, our samples were classified from medium to low coherence. Thereby, there is an opportunity to improve the coherence models to have more notable gaps and a more fine-grained analysis between the proposed categories. These coherence models could have alternative neural architectures, including larger annotated datasets by professional annotators.

As we mentioned earlier, coherence itself is a challenging factor to assess. This applies not only to automated evaluation methods but also to humans, especially to non-trained experts (Lai and Tetreault, 2018) when classifying average samples (e.g., medium level of coherence). However, there is value in classifying simplifications as an additional aspect to consider for document-level TS. By performing a comparison between our inputs, predictions and the gold standard we obtain a valuable notion of coherence in model evaluation. The evaluation of coherence is a first step, among the possible discourse elements that must be assessed during simplification, such as better readability (Martinc et al., 2021) and factuality (Devaraj et al., 2022).

8 Conclusion

In this study, we demonstrated that with the models and resources available, implementing discourse-aware simplification models becomes possible. We implemented a document-level model by extending a state-of-the-art sentence TS model and included different evaluations from a document-level perspective. The evaluation of DocS based on coherence is necessary, but it remains a challenge due to the subjective nature of this task. Nevertheless, the assessment of coherence represents a viable tool for detecting those simplifications that are unclear, inconsistent or lack a consistent narrative.

In the future, we expect to explore additional directions towards discourse elements such as cohesion and anaphora to support coherence evaluation for TS. We will also consider the implementation of alternative coherence models to improve coherence assessment and its generalisation for other domains within TS. Finally, we will consider baselines in which documents are simplified sentence-by-sentence to compare against our DocS systems, which consider context in the generation step.

[17] We include our annotator agreement analysis and their feedback in Appendix C.

9 Lay Summary

Text Simplification (TS) is a research area that makes text more understandable for wider audiences. A complex text can be transformed into a more simple variety, based on the needs of specific populations. These audiences include people with disabilities, non-native speakers or people with minimum expertise in areas such as healthcare, law and news. Text is simplified by changing difficult words, writing sentences in a more simple structure (e.g., shorter, avoiding passive voice) and explaining technical terms.

In recent years, simplification research has only been limited to the transformation of sentences. However, we could also make documents more accessible to the general public, such as the simplification of scientific papers, legal contracts and news, rather than just individual sentences. This is a challenging step, as there is limited annotated data by people trained to simplify documents. Also, the evaluation requires a lot of time and effort, and the automatic evaluation metrics are not reliable.

In this work, we proposed the SimDoc simplification system. This model combines different aspects of language such as simplicity, readability and coherence to achieve the simplification of documents. The aspect of coherence expresses the logical relationships between sentences from the same topic (e.g., a story or a news article). We contribute with our research by including a professionally annotated dataset adapted to different levels of readability. We also include a benchmark that evaluates large language models incrementally, starting with no data to larger sets of simplification examples. These large language models have been trained to automatically generate text, but they do not know how to simplify text until we show similar examples. Finally, we carry out a detailed analysis of the system outputs showing the limitations and future work of our solution.

The simplification of text considering simplicity, readability and coherence is encouraging, which motivates the research community to continue towards the direction of document-level simplification. Eventually, this will make knowledge more accessible and universal to wider communities. However, the simplification of documents is a challenging area of research. The evaluation of coherence can be improved using more professionally annotated data and from multiple domains. Although our method is tested in the news domain, it would not necessarily perform well in the medical domain. Also, the evaluation of coherence beyond the existing classification (low, medium and high) could be more granular, opening an opportunity to expand the benefit of this research to more audiences.

Acknowledgments

We would like to thank Nhung T.H. Nguyen for her valuable discussions and comments. Also, we thank Joel Tetreault for sharing the GCDC dataset and the related code. Laura Vásquez-Rodríguez's work was funded by the Kilburn Scholarship from the University of Manchester. Piotr Przybyła's work was supported by the Polish National Agency for Academic Exchange through a Polish Returns grant number PPN/PPO/2018/1/00006.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Miguel Collantes, Maureen Hipe, Juan Sorilla, Laurenz Tolentino, and Briane Paul Samson. 2015. Simpatico: A text simplification system for senate and house bills. In *Proceedings of the 11th National Natural Language Processing Research Symposium*, pages 26–32, Manila, Philippines. National University.

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.

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.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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*, page 10589–10604.

Grigorii Guz, Peyman Bateni, Darius Muglich, and Giuseppe Carenini. 2020. Neural RST-based evaluation of discourse coherence. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 664–671, Suzhou, China. Association for Computational Linguistics.

Renate Hauser, Jannis Vamvas, Sarah Ebling, and Martin Volk. 2022. A multilingual simplified language news corpus. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 25–30, Marseille, France. European Language Resources Association.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J. Ramanathan, Wei Xu, Byron C. Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts.

Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 558–568, Melbourne, Australia. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2021. Discourse Coherence. *Draft of December 29, 2021*, pages 1–25.

J. Peter Kincaid, Robert P. Fishburne, R 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. In *Institute for Simulation and Training*, pages 1–49.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

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.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Louis Martin, Angela Fan, de la Clergerie, Antoine Bordes, and Benoit Sagot. 2022. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54(2).

Mohsen Mesgar and Michael Strube. 2015. Graph-based coherence modeling for assessing readability. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 309–318, Denver, Colorado. Association for Computational Linguistics.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Tasnim Mohiuddin, Prathyusha Jwalapuram, Xiang Lin, and Shafiq Joty. 2021. Rethinking coherence modeling: Synthetic vs. downstream tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3528–3539, Online. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Understanding the lexical simplification needs of non-native speakers of English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 1–10, New York, USA. Association for Computing Machinery.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction – INTERACT 2013*, pages 203–219, Berlin, Heidelberg. Springer Berlin Heidelberg.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. SimPA: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Advaith Siddharthan. 2003. Preserving discourse structure when simplifying text. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, pages 103–110, Budapest, Hungary. Association for Computational Linguistics.

Jan Šnajder, Tamara Sladoljev-Agejev, and Svjetlana Kolić Vehovec. 2019. Analysing rhetorical structure as a key feature of summary coherence. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 46–51, Florence, Italy. Association for Computational Linguistics.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Renliang Sun, Zhe Lin, and Xiaojun Wan. 2020. On the helpfulness of document context to sentence simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1411–1423, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355, Toronto, Canada. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

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 Analysis of D-SARI penalties

In the Cochrane test set (Figure 2d), we noticed an increase in the length of the majority predictions (mostly between 100 and 200 words), compared the observations of the D-Wikipedia dataset (Figure 2e), in which the length range is more variable (from 0 to 200 words). In the D-Wikipedia test set (Figure 2e), there is a large group of predictions that are longer than the reference, but the majority are shown on the right of the red line, which means that the predictions are still shorter than the reference. This behaviour is even more evident in the OneStopEnglish (Figure 2f) test set, which has longer input articles, therefore, no predictions are longer than the reference. These patterns are also consistently repeated in the sentence-based analysis as well in Figures 2b, 2a and 2c.

With respect to Table 4 results, for D-SARI (where 0 is the lowest and 100 the highest value), we found that the *Mined* model has the highest score of 33.16 using the Cochrane (C) test set. As shown in Figure 2d, the *Mined* model predictions are even shorter than the gold standard simplifications, relative to other models. This leads to significant penalties (LP_1) in D_{add}, with a score of 0.33. Nonetheless, D_{keep} and D_{del} scores, 17.09 and 82.07, respectively, are less affected for the inverse case, where the gold standard is longer than the predictions (LP_2). Additionally, we can see that most of the datasets show a low score for D_{add}, for having smaller predictions than the reference. In the case of the OneStopEnglish corpus, the D-SARI scores are lower for all the models. This dataset has a larger difference between the simple and complex versions and the content is completely new to the models.

Regarding the sentence count, there is no clear correlation between the number of sentences in the gold standard and the predictions (i.e., they do not have the same number of sentences), directly affecting D_{keep} with SLP penalty in the difference in sentence numbers. The difference in sentence count affects the Cochrane test set for the *Mined+Cochrane* model (18.06) than the D-Wikipedia test set for *Mined+D-Wikipedia* (27.43) in D_{keep} scores. The Cochrane dataset is created from the alignment of an extended abstract (with multiple sections, e.g., background, objectives, results), whereas their plain language summaries may consist of a few paragraphs or a less structured format (Devaraj et al., 2021). Since its content may differ significantly, more penalties (SLP) are present in D_{keep} due to the high variability in the number of sentences.

Model	Test Set	Human	Auto	Corr.
Mined+ D-Wiki	D-Wikipedia	0.613	-0.060	0.029
Mined+ Cochrane	Cochrane	0.147	-0.100	-0.085

Table 7: Human Evaluation for general and medical domain, including automatic scores from the neural coherence models. The coherence score values range from 1 (high) to -1 (low). *Corr* stands for Correlation.

B Human Evaluation: Task Definition

The proposed task consisted in classifying texts into two categories. Unlike the automatic evaluation of coherence, we performed the evaluation using 2 categories (low, high) rather than 3 categories (low, medium, high). Previous research (Lai and Tetreault, 2018) has demonstrated the difficulty of modelling an intermediate class in human evaluation, leading to the inaccurate classification of texts, especially for those annotators that are not professionally trained. We requested our annotators to evaluate the coherence of 20 texts each in a spreadsheet. Similarly to Lai and Tetreault (2018), we also provided a definition for coherence to the annotators.

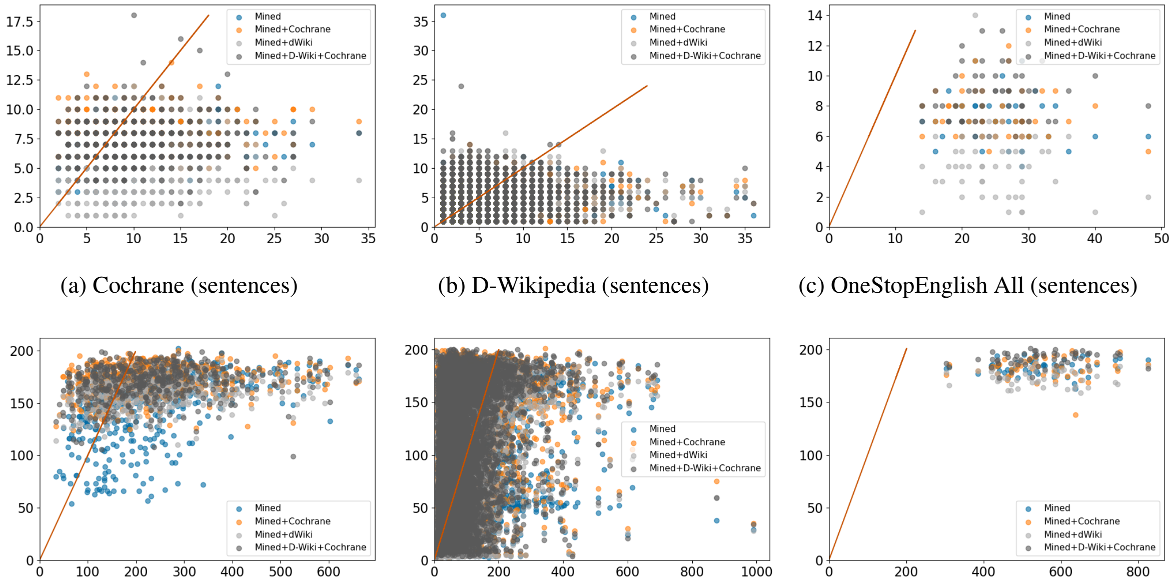
The annotators could ask questions anytime and provide feedback once the evaluation was completed, if any. We present our results in Table 7. While the evaluation was done using categorical values (high, low), we normalised our evaluation as with the automatic evaluation (1 for high and -1 for low coherence). We report the average values of each model.

C Annotator Agreement

We calculated the Fleiss' kappa values to measure the agreement between the annotators, using the pyirr[18] Python package. For the general domain, we had an agreement of 0.402 (*Mined+D-Wikipedia*), while in the medical domain, it scored 0.019 (*Mined+Cochrane*).

For *Mined+D-Wiki* texts, the agreement was fair, while Cochrane showed slight agreement between annotators. As mentioned in Section 6, the varied experience of the annotators in the different domains may have affected the final agreement on

[18] https://pypi.org/project/pyirr/



(a) Cochrane (sentences)

(b) D-Wikipedia (sentences)

(c) OneStopEnglish All (sentences)

(d) Cochrane (words)

(e) D-Wikipedia (words)

(f) OneStopEnglish All (words)

Figure 2: D-SARI metric is affected by this length of sentences and words in their predictions and gold-standard. In this Figure, we analysed the difference in the count of words and sentences between predictions (*y-axis*) vs gold-standard simplifications (*x-axis*). The red line marks the limit where observations have the same number of words or sentences. Predictions that are smaller than their gold-standard are shown to the right of the line.

the evaluations. A more segregated and detailed definition of the task (i.e., with examples) could also have helped on better annotators' accuracy.

While our human evaluation may have some limitations, we have learned lessons for the improvement of future evaluation. Overall, the annotators' feedback can be summarised as follows:

1. Provide concrete examples of high and low coherence texts with the coherence definition.

2. Include additional post-processing steps, which could help the annotators to focus on coherence only and not other aspects of language (e.g., grammar).

3. Give a more strict definition for coherence since different people can consider different coherence levels as satisfactory.

4. Keep texts short, since longer texts could be more difficult to evaluate.

We understand that coherence is challenging to evaluate. When using the current coherence model, we often see cases in which the differences in coherence scores are not significant between each other in our model outputs. Hence, with a minimal difference, untrained readers could be confused on defining coherence in a subjective way. Therefore, the evaluation and quality of the simplifications should be supported with a human evaluation.[19]

D Error Analysis

We performed a manual inspection of the evaluated samples to further understand the limitations of the automatic evaluation in comparison with the human evaluation. To that end, we split the samples into 2 groups: where automatic metrics agreed with humans and also, the cases in which they differ.

Since our systems are sentence-based pre-trained models, some of the outputs were quite short (1-2 sentences). These samples, such as Example 2, were mostly categorised as incoherent by the automatic metrics, although they formed coherent sentences (which was also confirmed with the human evaluation).

john caspar wild (1804-1846) was a swiss-american artist . he was a lithographer and painter . (2)

However, when collections of multiple short sentences were gathered together, there was an agree-

[19] We have included an error analysis of the evaluated samples in the Appendix D

Table 8: Simplification samples from evaluations, examples 1 to 3

Example 1: Complex vs system (Mined+DWiki+Cochrane) - D-Wikipedia test set
Complex: hari bansha acharya ( ) is a nepalese actor , comedian , singer and writer . he is known for his method acting . he is one half of the comedy duo maha jodi along with fellow comedian madan krishna shrestha . he is known for his performance as arjun in the 1997 patriotic drama film “ balidaan “ . he also wrote and performed in the film “ shatru gatey “ .
System: the purpose of this review was to evaluate the role of hari bansha acharya in the 1997 film film film balidaan and performed and performed in the film shatruy . he is known for his performance as arjun in the 1997 review film film balidaan was performed and performed in the film shatru gatey .
Example 2: Complex vs system (Mined+DWiki) - D-Wikipedia test set
Complex: tapestry is a form of textile art , traditionally woven by hand on a loom . tapestry is weft-faced weaving , in which all the warp threads are hidden in the completed work , unlike cloth weaving where both the warp and the weft threads may be visible . in tapestry weaving , weft yarns are typically discontinuous ; the artisan interlaces each coloured weft back and forth in its own small pattern area . it is a plain weft-faced weave having weft threads of different colors worked over portions of the warp to form the design.most weavers use a natural warp thread , such as wool , linen or cotton . the weft threads are usually wool or cotton but may include silk , gold , silver , or other alternatives .
System: it is a plain weft-faced weave having weft threads of different colors worked over portions of the warp to form the design . unlike cloth weaving where both the warp and the weft threads may be visible in the completed work , in tapestry weaving the weft threads are typically discontinuous .
Example 3: Complex vs system (Mined+DWiki+Cochrane) - OneStopCorpus test set
Complex: male bosses are being paid bonuses double the size of those given to female colleagues in identical jobs – a disparity that means men enjoy salary top-ups of £141,500 more than women over the course of a working lifetime . the figures , released by the chartered management institute (cmi) , reveal that men in uk management roles earned average bonuses of £6,442 in 2012 compared with £3,029 for women . in the most senior roles , female directors received bonuses of £36,270 over the past 12 months , compared with £63,700 awarded to male directors .
System: this review found that men in uk management roles earned average less of £3,029 for women over the course of a working lifetime . in the most senior roles , female directors received greater rates of £36,270 over the past 12 months , compared with £63,700 given with male directors . however , men enjoy salary top-ups of £141 , more than women over the course of a working lifetime than women over the course of a working lifetime .

ment between the system and humans, such as Example 3. This case was mostly present in general domain test sets.

the 1917 south american championship of nations was the second south american championship of nations tournament . it was held in montevideo , uruguay from september 30 to october 14 , 1917 . uruguay won the title after winning 1-0 over argentina . ángel romano was the top scorer of the tournament with 4 goals . uruguay was the defending champions . (3)

Finally, grammar and punctuation can also affect the coherence outcome. Texts such as Example 4, have an automatic rating of -1 (low), while the human evaluation had a full agreement of 1 (high).

lionel blair (born henry lionel ogus ; 12 december 1928) is a british actor , choreographer , tap dancer and television presenter . he is best known for presenting “ name that tune “ in the united kingdom . (4)

For the medical domain, there was a significant disagreement between annotations. Texts can vary in complexity and length and the experience of the annotators as well. We did not find any specific pattern for coherence in the Cochrane test set, except for those already mentioned.

Table 9: Simplification samples from evaluations, examples 4 to 6)

Example 4: Complex vs system (Mined+DWiki+Cochrane) - OneStopCorpus test
Complex: low-income countries will remain on the front line of human-induced climate change over the next century , experiencing gradual sea-level rises , stronger cyclones , warmer days and nights , more unpredictable rainfall , and larger and longer heatwaves , according to the most thorough assessment of the issue yet . the last major united nations (un) assessment , in 2007 , predicted runaway temperature rises of 6°c or more by the end of the century . that is now thought unlikely by scientists , but average land and sea temperatures are expected to continue rising throughout this century , possibly reaching 4°c above present levels – enough to devastate crops and make life in many cities unbearably hot .
System: the aim of this review was to assess the evidence from randomised controlled trials to determine the effects of human-induced climate change in human-induced climate change over the next century , experiencing gradual sea-level rises , stronger cyclones , warmer days and nights , more unpredictable rainfall , and larger and longer heat groups according to the most thorough assessment of the intervention yet . the evidence is current to august 2013. the review concludes that there is now thought unlikely by scientists , but average land and sea temperatures are expected to continue through rising throughout this century , possibly reaching 4°c above present levels – enough to allow life in many cities involving hot .
Example 5: Complex vs system (Mined+DWiki) - Cochrane test set
Complex: we included two trials (116 women) comparing planned home versus hospital management for pprom . overall , the number of included women in each trial was too small to allow adequate assessment of pre-specified outcomes . investigators used strict inclusion criteria and in both studies relatively few of the women presenting with pprom were eligible for inclusion . women were monitored for 48 to 72 hours before randomisation . perinatal mortality was reported in one trial and there was insufficient evidence to determine whether it differed between the two groups (risk ratio (rr) 1.93 , 95 % confidence interval (ci) 0.19 to 20.05) . there was no evidence of differences between groups for serious neonatal morbidity , chorioamnionitis , gestational age at delivery , birthweight and admission to neonatal intensive care . there was no information on serious maternal morbidity or mortality . there was some evidence that women managed in hospital were more likely to be delivered by caesarean section (rr (random-effects) 0.28 , 95 % ci 0.07 to 1.15) . however , results should be interpreted cautiously as there is moderate heterogeneity for this outcome (i^2 = 35 %) . mothers randomised to care at home spent approximately 10 fewer days as inpatients (mean difference -9.60 , 95 % ci -14.59 to -4.61) and were more satisfied with their care . furthermore , home care was associated with reduced costs . the review included two relatively small studies that did not have sufficient statistical power to detect meaningful differences between groups . future large and adequately powered randomised controlled trials are required to measure differences between groups for relevant pre-specified outcomes . special attention should be given to the assessment of maternal satisfaction with care and cost analysis as they will have social and economic implications in both developed and developing countries .
System: motorage is a term used to describe a condition that affects two trials . it is a type of problem that affects two trials . this is done by putting a lot of conditions in a certain area . this is done to make a lot of changes in a certain area . this is done by looking at a certain number of places where they have to do not have to have a problem . this means that they have to do not have to be able to do with a problem . this is because it is not possible to have a problem that they have to do not have to have a problem . this is because it is not possible to have a lot of conditions that are not caused by people who have to do not have to have a problem . this is because it is not possible that some people have to do not have to know what they have to do . this is because people who have to do not have to know what they have to do .
Example 6: Complex vs system (Mined+Cochrane) - Cochrane test
Complex: six studies (including 142 participants) were eligible for inclusion . two compared three-times-a-week prophylactic administration with on-demand treatment in children with hemophilia . pooled results from these two studies showed a rate ratio of 0.30 (95 % confidence interval ; 0.12 to 0.76) for all bleedings and 0.22 (95 % confidence interval 0.08 to 0.63) for joint bleedings favouring prophylaxis . results on the number of patients with preserved joints after three to seven years of follow-up were not pooled due to significant heterogeneity . three of the remaining four studies evaluated hemophilia a ; one showed a statistically significant decrease in frequency of joint bleeds with prophylaxis compared to placebo , with a rate difference of -10.73 (95 % confidence interval -16.55 to -4.91) bleeds per year . two studies compared two prophylaxis regimens , failing to demonstrate an advantage of one regimen over the other in terms of bleeding frequency . the fourth study evaluated hemophilia b and showed fewer joint bleeds with weekly (15 iu/kg) versus bi-weekly (7.5 iu/kg) prophylaxis , rate difference -3.30 (95 % confidence interval -5.50 to -1.10) bleeds per year . non-significant increases in both inhibitor and infectious complications were observed in patients on prophylaxis , which occurred more often when using long-term venous access . there is strong evidence from randomised controlled trials and observational trials that prophylaxis preserves joint function in children with hemophilia as compared to on-demand treatment . there is insufficient evidence from randomised controlled trials to confirm the observational evidence that prophylaxis decreases bleeding and related complications in patients with existing joint damage . well-designed randomised controlled trials and prospective observational controlled studies are needed to establish the best prophylactic regimen and to assess the effectiveness of prophylactic clotting factor concentrates in adult patients .
System: this review included six studies (including 142 participants) comparing three-times-a-week prophylactic administration with on-demand treatment in children with hemophilia . the review includes six studies with a total of 142 participants . two studies evaluated the effects of prophylactic treatment on the frequency of joint bleeds with on-demand treatment in children with hemophilia . two studies compared two prophylaxis regimens , failing to demonstrate an advantage of one regimen over the other in terms of bleeding frequency . the fourth study evaluated hemophilia b and showed fewer joint bleeds with weekly . in the number of patients with preserved joints after three to seven years of follow-up were not pooled due to significant heterogeneity . three of the remaining four studies evaluated hemophilia a ; one showed a statistically significant decrease in frequency of joint bleeds with prophylaxis compared to placebo.