Context-Aware Document Simplification

Liam Cripwell, Joël Legrand, Claire Gardent

▶ **To cite this version:**

Liam Cripwell, Joël Legrand, Claire Gardent. Context-Aware Document Simplification. Findings of the Association for Computational Linguistics: ACL 2023, ACL, Jul 2023, Toronto, Canada. pp.13190-13206, 10.18653/v1/2023.findings-acl.834 . hal-04369783

HAL Id: hal-04369783
https://hal.science/hal-04369783v1

Submitted on 2 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Context-Aware Document Simplification

Liam Cripwell
Université de Lorraine
CNRS/LORIA
liam.cripwell@loria.fr

Joël Legrand
Université de Lorraine
Centrale Supélec
CNRS/LORIA
joel.legrand@inria.fr

Claire Gardent
CNRS/LORIA
Université de Lorraine
claire.gardent@loria.fr

Abstract

To date, most work on text simplification has focused on sentence-level inputs. Early attempts at document simplification merely applied these approaches iteratively over the sentences of a document. However, this fails to coherently preserve the discourse structure, leading to suboptimal output quality. Recently, strategies from controllable simplification have been leveraged to achieve state-of-the-art results on document simplification by first generating a document-level plan (a sequence of sentence-level simplification operations) and using this plan to guide sentence-level simplification downstream. However, this is still limited in that the simplification model has no direct access to the local inter-sentence document context, likely having a negative impact on surface realisation. We explore various systems that use document context within the simplification process itself, either by iterating over larger text units or by extending the system architecture to attend over a high-level representation of document context. In doing so, we achieve state-of-the-art performance on the document simplification task, even when not relying on plan-guidance. Further, we investigate the performance and efficiency tradeoffs of system variants and make suggestions of when each should be preferred.

1 Introduction

Text simplification transforms a given text into a simpler version of itself that can be understood by a wider audience, while preserving the same core meaning (Gooding, 2022). It has also proven useful as a preprocessing step for downstream NLP tasks such as machine translation (Chandrasekar et al., 1996; Mishra et al., 2014; Li and Nenkova, 2015; Štajner and Popovic, 2016) and relation extraction (Miwa et al., 2010; Niklaus et al., 2016).

Most previous work has focused on sentence-level simplification by training neural models on complex/simple sentence pairs under the assumption that they will learn to perform required operations (e.g. sentence splitting, lexical substitution or syntactic rephrasing) implicitly from the training data (Zhang and Lapata, 2017; Nisioi et al., 2017; Jiang et al., 2020). However, the imbalanced representation of simplification operations throughout popular datasets, and the overly-conservative models arising from their use, have led to attempts at controllable simplification to achieve more variation and diversity in output texts (Alva-Manchego et al., 2017; Cripwell et al., 2021; Maddela et al., 2021).

Recently, strategies from controllable simplification have been leveraged to achieve state-of-the-art results on the document simplification task (Cripwell et al., 2023). Specifically, by using a planning model capable of considering the sentences surrounding a complex sentence, a sentence-level simplification model can be guided such that the structure of the resulting document remains more coherent. Despite this success, the sentence simplification model still has no direct access to document context which we believe limits the extent to which it can accurately produce simplified sentences that are consistent with the larger document.

As such, we propose various systems that allow access to some representation of surrounding content within the simplification module, while still allowing for the possibility of plan-guidance. We show that in doing so, we are able to achieve state-of-the-art document simplification performance on the Newsela dataset, even without relying on a generated plan. Further, we investigate the performance and efficiency tradeoffs of various system variants. [1]

Our key contributions are (i) a detailed investigation of how document context, input text and simplification plans impact document-level simplifica-

[1] Pretrained models, code, and data are available at https://github.com/liamcripwell/plan_simp.

Findings of the Association for Computational Linguistics: ACL 2023, pages 13190–13206
July 9-14, 2023 ©2023 Association for Computational Linguistics

tion and (ii) several state of the art models for document simplification. We show in particular that document level simplification is improved by combining a representation of the local context surrounding complex sentences with a simplification plan indicating how complex sentences should be simplified (whether they should be deleted, rephrased, split or copied).

2 Related Work

Context in Controlled Text Generation The use of external context within controlled text generation pipelines has seen recent success in areas outside of simplification. Li et al. (2021) control review generation by using document and sentence-level plans in the form of knowledge graph subgraphs. Smith et al. (2020) control the style of generated dialog responses by conditioning on a desired style token appended to other contextual utterances. Hazarika et al. (2022) modulate the amount of attention paid to different parts of a dialog context and show that using contextual encoding of question phrases can guide a model to more often generate responses in the form of questions. Slobodkin et al. (2022) consider summarisation where salient spans are first identified before being used to control the generation, while Narayan et al. (2023) first generate a summarisation plan consisting of question-answer pairs.

Simplification Planning Certain controllable sentence simplification works have approached simplification as a planning problem whereby an operation plan is first generated before being realised downstream to form the simplified text. The first of these are revision-based models that predict a sequence of token-level operations (delete, substitute, etc.), allowing for more control and interpretability (Alva-Manchego et al., 2017; Dong et al., 2019; Kumar et al., 2020; Omelianchuk et al., 2021; Dehghan et al., 2022). Others have taken a sentence-level approach by predicting a high-level operation (sentence split, rephrase, etc.) and using this to condition more typical neural systems (Scarton and Specia, 2018; Scarton et al., 2020; Garbacea et al., 2021; Cripwell et al., 2022).

Recently, the latter approach was leveraged for document simplification where it obtained state-of-the-art performance (Cripwell et al., 2023). Here, a sequence of sentence-level operations is predicted for an entire document and then used to iteratively condition a sentence-level simplification model. The system considers both local (token representation of the sentence) and global document context (sequence of sentence-level encodings) when predicting an operation for a given sentence.

Document-Level Simplification. Initial attempts at document simplification simply applied sentence simplification methods iteratively over documents (Woodsend and Lapata, 2011; Alva-Manchego et al., 2019b; Sun et al., 2021). However, it was noted this alone is insufficient for performing certain operations, often leading to poor discourse coherence in the output (Siddharthan, 2003; Alva-Manchego et al., 2019b).

Various sub-problems of document simplification have been approached in isolation, such as sentence deletion (Zhong et al., 2020; Zhang et al., 2022), insertion (Srikanth and Li, 2021), and reordering (Lin et al., 2021). Sun et al. (2021) took a holistic approach by iteratively applying a sentence-level model, but with additional encoders to embed the two preceding and following sentences, which are used as additional input during generation. However, this was unable to outperform baselines.

Recently, Cripwell et al. (2023) achieved state-of-the-art performance by producing a document simplification framework capable of performing all of the most common operations. Specifically, they use both high-level document context and sentence-level features to generate a plan specifying which operations to be performed on each sentence in a given document, which is then used to condition a sentence simplification model.

3 Problem Formulation

The goal of text simplification is to generate a text S that simplifies an input text C. In the document-level case, $C = c_1 \dots c_n$ is a sequence of complex sentences and $S = s_1 \dots s_m$ is a sequence of simple sentences. Cripwell et al. (2023) further decompose this task into a two-stage process wherein a generated plan conditions the simplification:

$$P(S \mid C) = P(S \mid C, O)P(O \mid C)$$

where $O = o_1 \dots o_n$ is a simplification plan, i.e. a sequence of sentence-level simplification operations for C (*copy*, *rephrase*, *split*, or *delete*). The motivation here is that the plan provides a high-level description of how to transform C into S,

which can in turn be used to guide the iterative generation of the simplified document across sentences.

Although the use of such plans has shown improved results, little attention has been given to how the generation stage itself can be modified to improve document-level simplification. In this work, we investigate whether further changes can be made to simplification models in order to make better use of high-level plans, or alternatively, whether it is possible to forego the planning stage entirely by incorporating high-level document context into the generative model directly.

Terminology and Notations. We use the following terminology and notational conventions:

- $C = c_1 \dots c_n$ is a complex document of n sentences;
- p_i is the ith paragraph from the complex document C;
- $S = s_1 \dots s_m$ is a ground-truth simplified version of C, containing m sentences;
- $\hat{S} = \hat{s}_1 \dots \hat{s}_{m'}$ is a predicted simplification of C, generated by a simplification model;
- o is a simplification operation with value *copy*, *rephrase*, *split*, or *delete*;
- $\hat{O} = \hat{o}_1 \dots \hat{o}_n$ is a predicted simplification plan stipulating specific sentence-level operations that should be applied to each $c_i \in C$ so as to arrive at some $\hat{S}$;
- Z_i is a high-level representation of the document context for c_i. It is a sequence of vector encodings for a fixed window of sentences surrounding c_i within C.

4 Data

For all experiments, we use Newsela-auto (Jiang et al., 2020) which is currently the highest-quality document-level simplification dataset available. It consists of 1,130 English news articles from the original Newsela (Xu et al., 2015) dataset which are each manually rewritten at five different levels of simplification, corresponding to discrete reading levels (0-4) of increasing simplicity. It also includes both sentence and paragraph alignments for each document pair. Like previous works, for all our models we prepend a control-token to the input specifying the target document reading level.

We use the same filtered version of Newsela-auto used in Cripwell et al. (2023), along with the same train/validation/test splits to allow for model comparison. This also includes plan labels, consisting of an operation (*copy*, *rephrase*, *split*, or *delete*) assigned to each sentence pair. Statistics of this data can be seen in Table 1.

	Train	Validation	Test
# Document Pairs	16,946	457	916
# Paragraph Pairs	335,018	9,061	17,885
# Sentence Pairs	654,796	17,688	35,292

Table 1: **Statistics of the filtered Newsela-auto dataset from Cripwell et al. (2023).** There is a train/validation/test split of 92.5%/2.5%/5%, assigned at the document-level (i.e. sentences and paragraphs from the same document will be contained within the same set). All reading-level variations of a specific article are also contained within the same set.

5 Models

We distinguish three model categories: (i) models whose sole input is text and which simplify a document either by iterating over its sentences/paragraphs or by handling the entire document as a single input; (ii) models that take both a complex sentence and some representation of its document context as input and simplify a document by iterating over its sentences; and (iii) models that are guided by a plan via control-tokens denoting sentence-level simplification operations prepended to the input sequence. These are illustrated in Table 2 and presented in more detail in the following subsections. Additional training details are outlined in Appendix A.

5.1 Text-Only Models

The most basic group of models we test are those that simply take a text sequence as input. We use baseline models trained to take entire documents or individual sentences. We also experiment with using paragraph inputs, the results of which we believe should scale better to the document-level than isolated sentences. Because paragraphs contain a wider token-level representation of local context this might provide enough information to maintain coherency in the discourse structure of the final document.

BART. We finetune BART (Lewis et al., 2020) to perform simplification at the document (**BART$_{\text{doc}}$**),

System	Input						
	Text			Context	Plan		
	Document	*Paragraph*	*Sentence*		*Document*	*Paragraph*	*Sentence*
BART	C	p_i	c_i	-	-	-	-
LED	C	p_i	-	-	-	-	-
ConBART	-	-	c_i	Z_i	-		
PG_{Dyn} (2023)	-	-	c_i	-	-	-	$\hat{o}_i$
$\hat{O} \rightarrow$ BART	C	p_i	c_i	-	$\hat{O}$	$\hat{o}_{j..j+\lvert p_i \rvert}$	$\hat{o}_i$
$\hat{O} \rightarrow$ LED	C	p_i	-	-	$\hat{O}$	$\hat{o}_{j..j+\lvert p_i \rvert}$	-
$\hat{O} \rightarrow$ ConBART	-	-	c_i	Z_i	-	-	$\hat{o}_i$

Table 2: **Different system types** and the specific forms of text, context, and plan inputs they consume. C is a complex document, c_i is the ith sentence of C, and p_i is the ith paragraph of C. $\hat{O}$ is a predicted document simplification plan, $\hat{o}_i$ is the individual operation predicted for the ith sentence, and $\hat{o}_{j..j+|p_i|}$ is the plan extract for a specific paragraph p_i, where j is the index of the first sentence in p_i.

sentence (**$BART_{sent}$**), and paragraph (**$BART_{para}$**) levels. [2] Both $BART_{sent}$ and $BART_{para}$ are applied iteratively over a document and outputs are concatenated to form the final simplification result.

Longformer. Encoder-decoder models like BART often produce worse outputs and become much slower the longer the input documents are. Longformer (Beltagy et al., 2020) is one proposal that aims to overcome these limitations by using a modified self-attention mechanism that scales linearly with sequence length. We finetune a Longformer encoder-decoder to perform the simplification on documents (**LED_{doc}**) and paragraphs (**LED_{para}**). [3]

5.2 Context-Aware Model (ConBART)

We propose a context-aware modification of the BART architecture (**ConBART**) that is able to condition its generation on both an input sentence c_i and a high-level representation of its document context Z_i (a sequence of vectors representing surrounding sentences in the document). This is done via extra cross-attention layers in each decoder attention block that specifically focus on Z_i. The ConBART architecture is illustrated in Figure 1.

We produce Z_i by employing the same context representation strategy used for planning in Cripwell et al. (2023). Specifically, the document context is obtained by taking a fixed window of sentences surrounding the target c_i, encoding them with Sentence-BERT (SBERT, (Reimers and Gurevych, 2019)), and applying custom positional embeddings to represent location within the document.

By generating the plan autoregressively, it is also possible to use previously simplified sentences within the left context of the current complex sentence, a method we refer to as *dynamic context*. In this case, the window of sentences represented within Z_i is defined:

$$Context_{i,r} = Concat(\hat{s}_{i-r..i-1}, c_{i..i+r}) \quad (1)$$

where r is the context window radius and $\hat{s}_i$ is the simplification output for the ith sentence c_i. We use the same recommended setting of $r = 13$.

The intuition behind the ConBART architecture is that the contextual information should allow for the simplification model to implicitly learn useful features of the discourse structure of the document in a similar way to the planner in Cripwell et al. (2023).

5.3 Plan-Guided Systems

Existing System. We compare with the state-of-the-art system proposed by Cripwell et al. (2023), **PG_{Dyn}**, which consists of a standard sentence-level BART model that is guided by a planner, which predicts the simplification operation to be applied each input sentence given a left and right context window of sentence representations, Z_i. The planner uses dynamic document context, allowing it to auto-regressively update the left context part of Z_i during planning as each sentence is simplified (see Equation 1).

[2] All models are initialised with the pretrained *facebook/bart-base* model from https://huggingface.co/facebook/bart-base.

[3] All models are initialised with the pretrained *allenai/led-base-16384* model from https://huggingface.co/allenai/led-base-16384.

System	BARTScore ↑			SMART ↑			FKGL ↓	SARI ↑	Length	
	P ($r \rightarrow h$)	R ($h \rightarrow r$)	F1	P	R	F1			Tok.	Sent.
Input	-2.47	-1.99	-2.23	63.2	62.7	62.8	8.44	20.5	866.9	38.6
Reference	-0.93	-0.93	-0.93	100	100	100	4.93	99.9	671.5	42.6
BART_{doc}	-2.68	-2.76	-2.72	61.9	43.9	50.6	10.01	47.1	600.8	20.7
$\text{BART}_{\text{sent}}$	-1.63	-1.56	-1.60	78.9	80.1	79.3	5.03	73.0	666.4	42.6
$\text{BART}_{\text{para}}$	-1.85	-1.49*	-1.67	77.2	82.8*	79.6	5.28	73.7	752.8	45.6*
LED_{doc}	-1.68	-1.73	-1.70	75.3	74.9	74.8	4.87	68.7	643.7	41.5
LED_{para}	-1.61	-1.40*	-1.50*	81.1	**85.5***	83.0*	5.15	76.9*	712.9	44.9*
ConBART	-1.59	-1.50	-1.54*	81.2	82.5*	81.7	5.01	75.8	669.8	42.8
PG_{Dyn} (2023)	-1.60	-1.54	-1.57	80.2	81.0	80.5	4.98	75.0	667.2	42.6
$\hat{O} \rightarrow$ ConBART	-1.52*	-1.45*	-1.48*	82.8*	84.0*	83.2*	4.96	78.3*	671.6	43.0
$\hat{O} \rightarrow \text{BART}_{\text{para}}$	-1.75	-1.47*	-1.61	79.4	81.9	80.4	5.11	74.9	715.3	42.7
$\hat{O} \rightarrow \text{LED}_{\text{para}}$	**-1.50***	**-1.42***	**-1.46***	**83.7***	84.9*	**84.1***	5.09	**78.5***	683.1	42.8
$\text{PG}_{\text{Oracle}}$ (2023)	-1.39*	-1.40*	-1.40*	85.5*	85.0*	85.3*	4.91	80.7*	655.6	42.1
$O \rightarrow$ ConBART	**-1.32***	**-1.32***	**-1.32***	**88.0***	**87.7***	**87.8***	4.92	**83.8***	659.6	42.3
$O \rightarrow \text{BART}_{\text{para}}$	-1.60	-1.36*	-1.48*	83.6*	85.3*	84.3*	5.07	79.7*	706.2	42.3
$O \rightarrow \text{LED}_{\text{para}}$	-1.36*	-1.33*	-1.35*	87.0*	87.3*	87.1*	5.03	82.3*	673.6	42.4

Table 3: **Results of document simplification systems on Newsela-auto.** For BARTScore, h is the hypothesis and r is the reference. Scores significantly higher than PG_{Dyn} are denoted with $*$ ($p < 0.005$). Significance was determined with Student's t-tests.

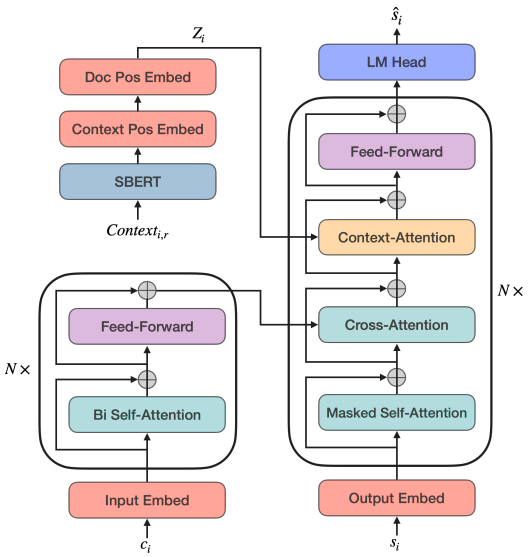


Figure 1: **ConBART model architecture.** The added context attention layer is shown in yellow, which allows for cross-attention over high-level document content, Z_i.

Pipelines. We construct pipeline systems that consist of each of our proposed models, guided by a document plan generated by the planner from Cripwell et al. (2023) (the same as is used by PG_{Dyn}). For this, we use modified versions of each simplification model that are trained to take an operation control-token at the beginning of each text input.

We refer to each of these pipeline systems as $\hat{O} \rightarrow h$, where h is the simplification model. We also report results where the ground-truth/oracle plans are used to condition models ($O \rightarrow h$).

Note that because the planner updates its document context autoregressively at the sentence-level, this does not interface perfectly with paragraph-level simplification models. As such, for pipelines using a paragraph-level simplification model ($\hat{O} \rightarrow \text{BART}_{\text{para}}$, $\hat{O} \rightarrow \text{LED}_{\text{para}}$), we only update the planner's context after each paragraph has been processed. Thus, for those paragraph level models, the left context of a complex sentence c_i is only simplified up to the first sentence of the paragraph containing c_i, i.e.

$$Context_{i,r} = Concat(\hat{s}_{i-r..j-1}, c_{j..i+r}) \quad (2)$$

where j is the index of the first sentence within the same paragraph as c_i, assuming $j > i - r$.

We also experimented with multi-task systems that are trained to perform both planning and sim-

plification within a single model, therefore not requiring a pipeline setup. However, this ultimately proved unsuccessful (further details in Appendix B).

6 Evaluation

6.1 Automatic Evaluation

Text simplification is often evaluated on the basis of 3 criteria: adequacy (or meaning preservation), fluency, and simplicity. For automatic evaluation, we use BARTScore (Yuan et al., 2021) and SMART (Amplayo et al., 2022) as analogs for both adequacy and fluency. Both are reference-based metrics that have previously been used for document simplification as well as other text generation tasks.

For assessing simplicity, we use both the Flesch-Kincaid grade level (FKGL) and SARI (Xu et al., 2016). FKGL is a document-level metric of text readability that has the highest correlation with human judgements (Scialom et al., 2021), while SARI is a simplification metric that has become a staple in the sentence-level simplification literature. We use EASSE (Alva-Manchego et al., 2019a) to calculate both of these.

At test time we generate sequences using beam search with a beam size of 5 and a maximum length of 1024 tokens.

6.2 Human Evaluation

Historically, automatic evaluation of long-form text generation has been very difficult to perform (Howcroft et al., 2020; Thomson and Reiter, 2020). As such, we conduct a human-evaluation of proposed systems to more accurately gauge performance.

As full documents are very long and difficult to compare, we conduct evaluations at the paragraph-level. For each comparison, a complex paragraph is shown next to an extract from a generated simplification corresponding to that paragraph. Evaluators are then asked to judge whether the generated text (i) is fluent (**fluency**); (ii) preserves the core meaning of the input (**adequacy**); and (iii) is simpler to read/understand (**simplicity**).

Using the test set, we randomly sample 33 complex paragraphs from each non-adjacent reading-level transition pairing, for a total of 198 paragraphs. We take the references and outputs from 4 high performing systems (PG_{Dyn}, LED_{para}, $\hat{O} \rightarrow \text{LED}_{\text{para}}$, $\hat{O} \rightarrow$ ConBART) for each (990 outputs in total) and have an annotator rate them on each of the 3 criteria. Because we use a large pool of annotators we impose a binary answering scheme (yes/no) in order to avoid the inter-annotator subjectivity that is inherent when using a Likert scale. The proportion of positive results is used as the final score for a given system.

Further details of the human evaluation are given in Appendix C.

7 Results and Discussion

Results are shown in Table 3. We also report results for other commonly used metrics in Appendix D.

Context Awareness Matters. Considering all metrics, we find that text-only models that take as input either a sentence ($\text{BART}_{\text{sent}}$) or a whole document ($\text{BART}_{\text{doc}}$, LED_{doc}) underperform models whose input is more local to the input sentence, either because they work at the paragraph level (LED_{para}) or because they take both the complex sentence and its local document context as input (ConBART). In other words, models that have access to a *local document context* (LED_{para}, ConBART) perform best overall.

LED vs BART. LED models ($\text{LED}_{\text{doc/para}}$) outperform their standard counterpart ($\text{BART}_{\text{doc/para}}$) showing that modified self-attention is not only more efficient but also more precise than standard self-attention in the case of long input.

The Utility of Planning. Plan guided models (4th horizontal block in Table 3) outperform their standard couterpart on all metrics, showing that a predicted plan has a positive impact on simplification. This is further supported by the fact that models guided by an oracle plan (5th block) provide even greater performance.

Comparison with the State-of-the-Art (PG_{Dyn}). $O \rightarrow$ ConBART is similar to PG_{Dyn} in that, in both cases, a document is simplified by iterating over its sentences and prediction is guided by the local context of the sentence to be simplified. A key difference is that in PG_{Dyn}, this context is exclusively used to predict a simplification operation, while in $O \rightarrow$ ConBART it is additionally used to condition the generation of the simplified sentences. We find that adding this extra control results in significantly better scores compared to the state-of-the-art PG_{Dyn} model. This illustrates that document context has utility for both planning (predicting the cor-

System	Fluency			Adequacy			Simplicity			Mean
	Minor	Major	All	Minor	Major	All	Minor	Major	All	
Reference	90.9*	**96.0**	93.4	80.8	70.7*	75.8	83.8*	82.8*	83.3	84.2
PG_{Dyn} (2023)	91.9*	94.9	93.4	**83.8**	73.7	78.8	88.9	85.9	87.4	86.5
LED_{para}	**98.0**	92.9	**95.5**	81.8	80.8	81.3	**92.9**	85.9	**89.4**	**88.7**
$\hat{O} \rightarrow LED_{para}$	90.9*	**96.0**	93.4	80.8	**82.8**	**81.8**	83.8*	90.9	87.4	87.5
$\hat{O} \rightarrow$ ConBART	89.9**	**96.0**	92.9	81.8	79.8	80.8	86.9	**91.9**	**89.4**	87.7

Table 4: **Human evaluation results for selected simplification systems.** The *minor* group includes those examples with a reading-level transition of 2 levels (e.g. 0-2, 1-3, etc.), whereas the *major* class includes those of 3-4 levels. Each of these groups make up half of the entire set. Ratings significantly different from the highest score in each column are denoted with $*$ ($p < 0.05$) and $**$ ($p < 0.01$). Significance was determined with two proportion Z-tests.

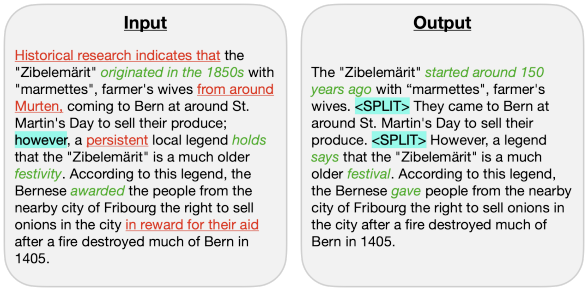
rect simplification operation) and realisation (simplifying a given sentence). While $\hat{O} \rightarrow LED_{para}$ achieves the best overall results of any system, it is slightly outperformed by $O \rightarrow$ ConBART when oracle plans are used, suggesting that an improved planner would provide better simplifications when used by $\hat{O} \rightarrow$ ConBART over $\hat{O} \rightarrow LED_{para}$.

Human Evaluation Results from the human evaluation are shown in Table 4. To better identify where each model excels, we report separate scores for test paragraph pairs with minor (reading-level transition of 2) and major (>2) degrees of simplification, as well as total average scores.

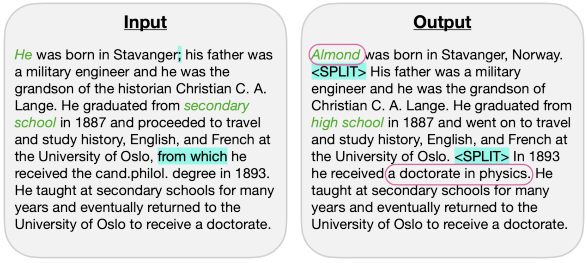
On fluency, all of the systems achieve very high ratings, which is unsurprising given the recognised ability of large language models (LLMs) to produce highly fluent texts. For adequacy, $\hat{O} \rightarrow LED_{para}$ achieves the highest overral score, closely followed by LED_{para} and $\hat{O} \rightarrow$ ConBART. In terms of simplicity, LED_{para} and $\hat{O} \rightarrow$ ConBART equally achieve the highest score. Across all criteria, LED_{para} achieves the highest average ratings, although very few scores are significantly better than other systems.

When considering performance differences between the minor and major simplification groups, we observe some clear trends. Systems that are not guided by a high-level document plan or do not have access to some contextual information during generation (PG_{Dyn} and LED_{para}) perform notably worse on examples requiring major simplification than they do on minor cases. Conversely, the models with both of these features appear to either perform equally as well or even excel on major cases. This suggests potential conservativity in the simplifications performed by PG_{Dyn} and LED_{para}.

Another interesting observation is the relatively



(a) a good result



(b) a poor result

Figure 2: **Example WikiLarge simplification output extracts from $\hat{O} \rightarrow LED_{para}$** (a target reading-level of 3 was used in each case). Note that these are small extracts from larger documents shown in Appendix E. Deletions are underlined and in red; rephrasings are *italicised and in green*; splitting points are highlighted in cyan; and factual errors are circled.

low ratings given to the references compared to the system outputs. In particular, they receive a much lower adequacy score than any other system on major cases. This could perhaps be a result of the systems generating outputs that bear more of a resemblance to the inputs than those written by humans (see faithfulness BARTScores in Appendix D). For instance, human editors might have been able to confidently delete more content, or refer to some of the information in different paragraphs which the evaluators were not privy to. Despite this, the references still receive fluency ratings competitive with the other systems.

Example Simplifications Figure 2 shows some example simplification outputs from $\hat{O} \rightarrow \text{LED}_{\text{para}}$. These are paragraph-level extracts from larger document outputs, which are provided in Appendix E. Due to licensing constraints imposed by Newsela, we use out-of-domain documents from the Wiki-Large dataset (Zhang and Lapata, 2017) in these examples.

8 Model Efficiency

There are various others factors to consider when comparing systems, beyond their raw performance. For instance, the size of the model(s) and how much time/resources are required for each to perform inference are important practical considerations that must be made when selecting a model for real-world use. As such, we compare each system based on the time taken to simplify the test set and their total parameter counts. Table 5 shows these results.

In our case, any system that uses a plan requires a second model, approximately doubling the number of parameters that must be loaded. These pipeline setups also naturally add to overall inference time. Further, both plan pipelines and ConBART make use of dynamic context, which imposes an autoregressive bottleneck on the simplification of individual documents.

Because of the linearly scaling attention mechanism, Longformer-based models are the fastest of proposed systems. Because of this and its overall high performance, we recommend LED_{para} in situations where time or computing resources are at all limited. Alternatively, $\hat{O} \rightarrow$ ConBART offers a good compromise that provides the high performance of a plan-guided system while mitigating further increases to inference time. This is because it uses the same autoregressive protocol as the planner and can therefore share the generated context representations.

All inference processes were run on a single Nvidia A40 GPU, using a batch size of 16, 32 CPU workers for data loading, and a beam size of 5 for generation. Appendix F provides details on the specific algorithm used to handle dynamic context generation for appropriate models.

	Inference Time ↓	**# Params ↓**
BART_{doc}	182.6	140
$\text{BART}_{\text{sent}}$	54.0	140
$\text{BART}_{\text{para}}$	68.9	140
LED_{doc}	49.1	162
LED_{para}	45.9	162
ConBART	74.7	156
PG_{Dyn} (2023)	76.6	154+140
$\hat{O} \rightarrow \text{BART}_{\text{para}}$	119.1	154+140
$\hat{O} \rightarrow \text{LED}_{\text{para}}$	103.3	154+162
$\hat{O} \rightarrow$ ConBART	82.7	154+156

Table 5: **Model efficiency statistics.** All times are in milliseconds and model parameters are in millions. Inference times are calculated on the test set and normalised by the total number of sentences (i.e. # ms per sentence).

9 Conclusion

We develop a range of document simplification models that are able to use different combinations of text, context, and simplification plans as input, with several models outperforming the previous state-of-the-art both on automatic metrics and according to human judgements. Our results show that a high-level representation of the document can be useful for low-level surface realisation as well as global planning. Further, simplification models with access to local document context, either by working at the paragraph level or handling an additional input representation, lead to better meaning preservation than those that operate on individual sentences. We conclude by evaluating the model efficiency of each system and making recommendations for their selection under different circumstances.

10 Limitations

Newsela Dataset One limitation to this study is our use of the Newsela dataset. Because this requires a license to access, researchers cannot fully reproduce our work without first obtaining permission from Newsela Inc. Unfortunately there is currently no other large dataset offering high quality

aligned documents for simplification under an open source license. The only other datasets so far used for document-level simplification are based on WikiLarge, which has very poor and inconsistent alignments at the document-level (Xu et al., 2015; Sun et al., 2021; Cripwell et al., 2023).

Paragraph-Level Human Evaluation In order to reduce complexity, our human evaluation was performed on paragraphs rather than full documents. As a result, there is a potential limit to the accuracy of human judgements when certain discourse phenomena are present. For example, important information may be excluded from a specific output paragraph (therefore prompting a low adequacy rating), but this could actually be present in a different part of the true simplified document.

Monolinguality This study focused entirely on simplification for English-language documents. Reproducing the proposed systems for use on other languages would require dedicated datasets of similar scale, along with sentence/paragraph alignments and operation labels (which likely do not currently exist). Further, the nature of simplification in other languages may differ quite a lot from English with respect to the types of operations that are performed, potentially reducing the suitability of the proposed framework.

Generalised Target Audience We approach this study with our definition of "simplification" being based on that of a generalised audience, following the standard set out by the assigned reading-levels of the Newsela dataset. Existing works often outline the intent for their systems to be used to simultaneously assist a wide array of different target users, such as those with cognitive impairments, non-native speakers, and children (Maddela et al., 2021; Garbacea et al., 2021; Sun et al., 2021). However, they rarely go into any detail about which simplification strategies work for each of these different groups or perform human evaluation with annotators from the same target demographics (Gooding, 2022). As such, we acknowledge that using our systems for a specific demographic might prove insufficient to enable their consumption of media without first making further revisions to support their precise needs.

References

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019a. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

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

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2022. Smart: Sentences as basic units for text evaluation.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2021. Discourse-based sentence splitting. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 261–273, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2022. Controllable sentence simplification via operation classification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2091–2103, Seattle, United States. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. 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.

Mohammad Dehghan, Dhruv Kumar, and Lukasz Golab. 2022. GRS: Combining generation and revision in unsupervised sentence simplification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 949–960, Dublin, Ireland. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097, Online. Association for Computational Linguistics.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Devamanyu Hazarika, Mahdi Namazifar, and Dilek Hakkani-Tür. 2022. Zero-shot controlled generation with encoder-decoder transformers.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. 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.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. 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.

Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021. Knowledge-based review generation by coherence enhanced text planning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 183–192, New York, NY, USA. Association for Computing Machinery.

Junyi Jessy Li and Ani Nenkova. 2015. Detecting content-heavy sentences: A cross-language case study. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1271–1281, Lisbon, Portugal. Association for Computational Linguistics.

Zhe Lin, Yitao Cai, and Xiaojun Wan. 2021. Towards document-level paraphrase generation with sentence rewriting and reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1033–1044, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Kshitij Mishra, Ankush Soni, Rahul Sharma, and Dipti Sharma. 2014. Exploring the effects of sentence simplification on Hindi to English machine translation system. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 21–29, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 788–796, Beijing, China. Coling 2010 Organizing Committee.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint.

Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 170–174, Osaka, Japan. The COLING 2016 Organizing Committee.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91,

Vancouver, Canada. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

C. Scarton, P. Madhyastha, and L. Specia. 2020. Deciding when, how and for whom to simplify. © 2020 The Author(s) and IOS Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial Licence (http://creativecommons.org/licenses/by-nc/4.0/).

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification.

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*, Budapest, Hungary. Association for Computational Linguistics.

Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst, and Ido Dagan. 2022. Controlled text reduction.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2020. Controlling style in generated dialogue.

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.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

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.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

K. Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *AAAI*.

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.

Bohan Zhang, Prafulla Kumar Choubey, and Ruihong Huang. 2022. Predicting sentence deletions for text simplification using a functional discourse structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Dublin, Ireland. Association for Computational Linguistics.

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.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9709–9716.

A Training Details

For all simplification models, we used a learning rate of $2e^{-5}$, a batch size of 16, and a 0.1 dropout rate. All models were trained on a computing grid using 2× Nvidia A40 GPUs (45GB memory) until convergence or a maximum of 48 hours.

For ConBART and planning pipelines we use the same settings as Cripwell et al. (2023) for construction of the high-level document context. Specifically, this includes a fixed context window radius of size 13 and use of a dynamic context mechanism.

B Multi-Task Systems

We also experimented with models that are explicitly trained to perform both the planning and simplification tasks using the same network. As high-level plans appear to improve the performance of simplification models, we hypothesise that learning both tasks in tandem could benefit overall performance. The motivation for this approach is to potentially produce a model that is capable of yielding similar or better simplification performance to the pipeline systems but with a more efficient single-model setup.

Specifically, these models were trained to generate the simplified text prefixed by a predicted plan in the form of operation-specific tokens. This was tested with both ConBART (**ConBART$_{\textbf{prefix}}$**) and a document-level Longformer (**LED$_{\textbf{prefix}}$**). In the case of the Longformer we also test a variant that generates the plan tokens as sentence separators (**LED$_{\textbf{sep}}$**). Results are shown in Table 6.

Unfortunately, from our experiments none of these seemed to result in performance exceeding those of simplification-only models. Improvement could perhaps be reached given the correct tuning of hyperparameters and loss weightings, however we did not have the time or resources to pursue this further in this study.

C Human Evaluation Details

The Newsela-auto paragraph alignments were used to identify valid references for each test paragraph. In order to align correct extracts from generated system outputs we took different steps depending on the system. For paragraph-level models (those using LED$_{\text{para}}$), we simply use the full simplification output for each source paragraph. For sentence-level models (ConBART, PG$_{\text{Dyn}}$), we first used the alignments to identify which paragraph the source sentence belongs to, then concatenated their simplification results.

Human judgements were crowdsourced on the MTurk platform. We sourced workers from English speaking countries (AU, CA, GB, IE, NZ, US) and paid them \$0.2 USD for each individual evaluation. We ran an initial test ourselves and timed how many evaluations could be completed within an hour. According to this, subjects should earn approximately \$18 USD per hour (which is above the minimum wage in all of these countries). The form and instructions presented to human evaluators is shown in Figure 3.

D Additional Evaluation Results

In Table 7 we provide additional results for popular automatic evaluation metrics that were not included in the main text. Specifically, we include BLEU (Papineni et al., 2002), and full operation-specific scores for SARI. In general, the results are similar to those in Table 3, with $\hat{O} \rightarrow$ LED$_{\text{para}}$ and $\hat{O} \rightarrow$ ConBART achieving the best results.

Faithfulness BARTScore is included for clarity rather than being a direct estimation of output quality. It shows how semantically similar system outputs are to their inputs, roughly equating to a measurement of conservativity.

E Example Simplifications

Figure 4 shows several example simplifications by the $\hat{O} \rightarrow$ LED$_{\text{para}}$ system on full documents. Due to licensing constraints imposed by Newsela, we use out-of-domain documents from the WikiLarge dataset here. As these are Wikipedia articles they are quite different in tone than the Newsela articles as well as being much shorter in length. Regardless, we still believe this provides clarity on the types of editing performed by the model.

F Dynamic Context Algorithm

Algorithm 1 shows the process used to handle dynamic context generation for appropriate models. As each document needs to be simplified autoregressively at the sentence level, we construct batches of sentences with the same index from different documents in order to speed up processing. Note that this could potentially be further optimised (e.g. via parallelism) and merely serves as a reasonable baseline algorithm.

System	BARTScore				SMART ↑			FKGL ↓	SARI ↑	Length	
	Faith. ($s \rightarrow h$)	P ↑ ($r \rightarrow h$)	R ↑ ($h \rightarrow r$)	F1 ↑	P	R	F1			Tok.	Sent.
Input	-0.93	-2.47	-1.99	-2.23	63.2	62.7	62.8	8.44	20.52	866.9	38.6
Reference	-1.99	-0.93	-0.93	-0.93	100	100	100	4.93	99.99	671.5	42.6
$\text{LED}_{\text{prefix}}$	-1.83	-1.72	-2.00	-1.86	73.5	67.6	69.8	4.97	63.14	604.0	38.0
LED_{sep}	-1.82	-1.80	-1.88	-1.84	72.6	70.5	71.1	5.06	62.64	640.4	40.2
$\text{ConBART}_{\text{prefix}}$	-1.96	-1.62	-1.60	-1.61	80.4	79.6	79.9	4.90	74.31	643.6	41.5

Table 6: Results of multi-task systems on the Newsela-auto test set.

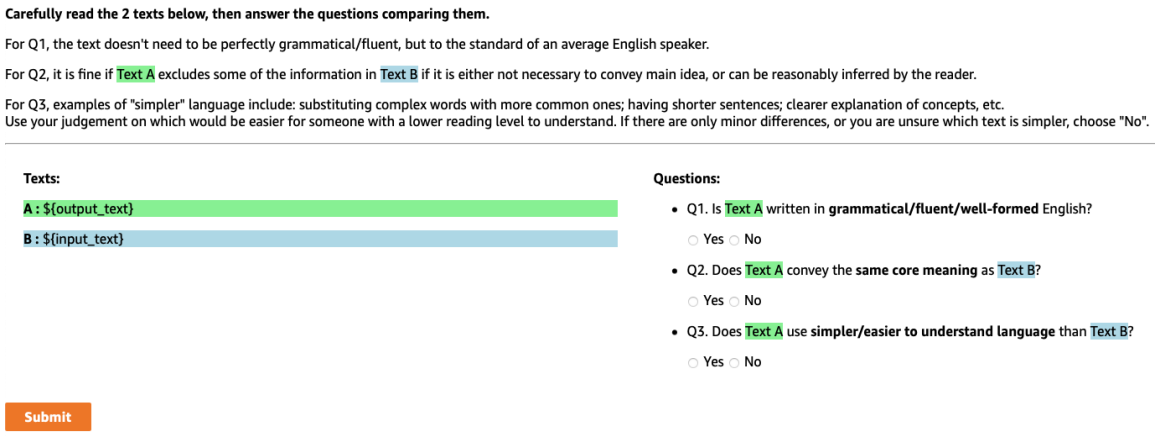


Figure 3: Submission form used in human evaluation.

System	BARTScore	BLEU ↑	SARI ↑	add	keep	delete
	Faith. ($s \rightarrow h$)					
Input	-0.93	46.2	20.5	0.0	61.6	0.0
Reference	-1.99	100	100	100	100	100
BART_{doc}	-2.48	31.1	47.1	20.4	55.4	65.4
$\text{BART}_{\text{sent}}$	-1.86	70.7	73.0	55.9	83.7	79.5
$\text{BART}_{\text{para}}$	-2.11	68.6	73.7	57.8	82.6	80.8
LED_{doc}	-1.90	63.7	68.7	52.2	78.2	75.7
LED_{para}	-1.86	74.5*	76.9*	64.3*	85.0	81.5
ConBART	-1.89	73.7	75.8	61.4	84.9	81.2
PG_{Dyn} (2023)	-1.91	72.4	75.0	58.9	84.8	81.4
$\hat{O} \rightarrow$ ConBART	-1.92	76.0*	78.3*	64.6*	86.8*	**83.4**
$\hat{O} \rightarrow \text{BART}_{\text{para}}$	-2.05	71.3	74.9	58.5	84.7	81.4
$\hat{O} \rightarrow \text{LED}_{\text{para}}$	-1.87	**76.8***	**78.5***	**65.1***	**87.3***	83.0
$\text{PG}_{\text{Oracle}}$ (2023)	-1.93	78.9*	80.7*	65.2*	89.9*	87.1*
$O \rightarrow$ ConBART	-1.93	**82.6***	**83.8***	**70.8***	**91.7***	**88.7***
$O \rightarrow \text{BART}_{\text{para}}$	-2.09	76.1*	79.7*	64.1*	88.7*	86.3*
$O \rightarrow \text{LED}_{\text{para}}$	-1.90	81.4*	82.3*	69.6*	90.6*	86.7*

Table 7: Extra automatic evaluation results on Newsela-auto. For BARTScore, s is the source text and h is the hypothesis. Scores significantly higher than PG_{Dyn} are denoted with $*$ ($p < 0.005$). Significance was determined with Student's t-tests.

Input

The Zibelemärit is an annual market with aspects of a fair in the old town of Bern, Switzerland. It takes place the fourth Monday in November.

Historical research indicates that the "Zibelemärit" *originated in the 1850s* with "marmettes", farmer's wives from around Murten, coming to Bern at around St. Martin's Day to sell their produce; however, a persistent local legend *holds* that the "Zibelemärit" is a much older *festivity*. According to this legend, the Bernese *awarded* the people from the nearby city of Fribourg the right to sell onions in the city in reward for their aid after a fire destroyed much of Bern in 1405.

As the name indicates, it is mainly onions that are sold on the "Zibelemärit". Bernese farmers, who are proud of their decorative onion tresses and onion wreaths, also sell other onion products on the market, including Zwiebelkuchen (onion pie), onion soup and onion sausages. Decorative chains of sugar onions are also popular with children.

The "Zibelemärit" opens very early in the day, at around 03:00 to 04:00. Later in the morning, the narrow alleys are usually packed tight with people, which is what the Bernese call the "Gstungg". A general confetti battle in which mostly children *participate ensues* at four o'clock in the afternoon, *officially ending the market*.

Output

The Zibelemärit is an annual market in the old town of Bern, Switzerland. It takes place the fourth Monday in November.

The "Zibelemärit" *started around 150 years ago* with "marmettes", farmer's wives. <SPLIT> They came to Bern at around St. Martin's Day to sell their produce. <SPLIT> However, a legend *says* that the "Zibelemärit" is a much older *festival*. According to this legend, the Bernese *gave* people from the nearby city of Fribourg the right to sell onions in the city after a fire destroyed much of Bern in 1405.

In this country, it is mainly onions that are sold on the "Zibelemärit." Bernese farmers also sell other products, including Zwiebelkuchen (onion pie), onion soup and onion sausages. Decorative chains of sugar onions are also popular with children.

The "Zibelemärit" opens very early in the day, at around 03:00 to 04:00. Later in the morning, the narrow alleys are usually packed tight with people, which is what the Bernese call the "Gstungg." A general confetti battle in which mostly children *fight breaks out* at four o'clock in the afternoon *to end the market*.

Input

Glenn Edward Greenwald (born March 6, 1967) is an American journalist and author.

He is best known for a series of reports published from June 2013 by "The Guardian" newspaper detailing the United States and British global surveillance programs, and based on *classified documents disclosed* by Edward Snowden. Greenwald and the team he worked with won both a George Polk Award and a Pulitzer Prize for those reports.

He has written several best-selling books, including "No Place to Hide". Before the Snowden file *disclosures*, Greenwald was *considered one of the most influential* opinion columnists in the United States. After working as a constitutional attorney for ten years, he began *blogging* on national security issues before becoming a "Salon" *contributor* in 2007 and *then* for "The Guardian" in 2012. He now writes for (and has co-edited) "The Intercept", which he founded in 2013 with Laura Poitras and Jeremy Scahill.

Greenwald's work on the Snowden story was featured in the documentary "Citizenfour", which won *the 2014 Academy Award for Best Documentary Feature*. Greenwald *appeared on-stage with director Laura Poitras and Snowden's girlfriend, Lindsay Mills, when the Oscar was given*. In the 2016 Oliver Stone feature film "Snowden", Greenwald was played by actor Zachary Quinto.

Output

Glenn Greenwald is an American journalist and author.

He is best known for a series of reports published from June 2013 by the Guardian newspaper. <SPLIT> They are based on *documents leaked* by Edward Snowden.

He has written several best-selling books, including "No Place to Hide." Before the Snowden file *leaks*, Greenwald *was one of the most respected* opinion columnists in the United States. He began writing about national security issues before becoming a "Salon" *writer* in 2007 and *a writer* for "The Guardian" in 2012. He now writes for The Guardian.

Greenwald's work on the Snowden story was featured in the documentary "Citizenfour". <SPLIT> The movie won *an Academy Award*. Greenwald *worked with director Laura Poitras and Snowden's girlfriend, Lindsay Mills, to make the documentary*. The 2016 Oliver Stone feature, "Snowden," was played by Zachary Quinto.

Input

Christian Lous Lange (17 September 1869 – 11 December 1938) was a Norwegian historian, teacher, and political scientist. He was one of the world's *foremost exponents* of the theory *and practice of internationalism*.

He was born in Stavanger; his father was a military engineer and he was the grandson of the historian Christian C. A. Lange. He graduated from *secondary school* in 1887 and proceeded to travel and study history, English, and French at the University of Oslo, from which he received the cand.philol. degree in 1893. He taught at secondary schools for many years and eventually returned to the University of Oslo to receive a doctorate.

His intellectual support of internationalism was demonstrated by the first of a three volume historical treatise ("Histoire de l'internationalisme"). The work, published in 1919, was said to contribute to the ideological preparation for the League of Nations. For this and all his other activities, Lange *shared* the Nobel Peace Prize with Hjalmar Branting in 1921.

Lange died at the age of 69, one day after the 17th anniversary of his receipt of the Nobel Prize. He was the father of Carl Viggo Manthey Lange, Halvard Lange and Christian August Manthey Lange.

Output

Christian Lous Lange (17 September 1869 – 11 December 1938) was a Norwegian historian, teacher, and political scientist. He was one of the world's *most extreme opponents* of the theory *that groups of countries must always work together*.

Almond was born in Stavanger, Norway. <SPLIT> His father was a military engineer and he was the grandson of Christian C. A. Lange. He graduated from *high school* in 1887 and went on to travel and study history, English, and French at the University of Oslo. <SPLIT> In 1893 he received a doctorate in physics. He taught at secondary schools for many years and eventually returned to the University of Oslo to receive a doctorate.

Lange *won* the Nobel Peace Prize with Hjalmar Branting in 1921.

Lange died at the age of 69. He was the father of Carl Viggo Manthey Lange, Halvard Lange and Christian August Manthey Lange.

Figure 4: Example simplification outputs for $\hat{O} \rightarrow \text{LED}_{\text{para}}$, illustrating both strong and poor performances (a target reading-level of 3 was used for all examples). Input documents are taken from WikiLarge due to licensing constraints around sharing Newsela content. Deletions are underlined and in red; rephrasings are *italicised and in green*; splitting points are highlighted in cyan; and factual errors are circled.

Algorithm 1 Generation strategy for systems using a dynamic context mechanism. Inference is performed autoregressively in batches containing 1 sentence per document. At the end of each time step, the simplified sentences are encoded for use within the context of the next step. This naturally extends to the paragraph-level case by replacing sentences with paragraphs.

procedure DYNAMICGENERATION(*test_set*)
  $g \leftarrow$ *load_planner*()
  $h \leftarrow$ *load_simplifier*()
  max_idx $\leftarrow \max_{C \in \textit{test\_set}} |C|$
  for $i \leftarrow 1$ to *max_idx* **do**
    sents $\leftarrow \{c_i \mid C \in \textit{test\_set}\}$ ▷ ith sentence from each document
    context $\leftarrow$ *load_context*(*sents*)
    if pipeline system **then**
      plans $\leftarrow g(\textit{sents}, \textit{context})$
      sents $\leftarrow$ *plans* + *sents* ▷ Prepend plans to texts
    end if
    preds $\leftarrow h(\textit{sents,context})$
    context $\leftarrow$ *update_context*(*preds*)
  end for
end procedure

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
10

☑ A2. Did you discuss any potential risks of your work?
10, "generalised target audience"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
10

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
7, 10

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The dataset, Newsela, is a well established dataset that has already been used numerous times within the literature.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

5, 8

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
8, Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.2, Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
7, 8, Appendix C

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

6.2, Appendix B

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix C

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We do not collect any personal data from the MTurk workers, only binary answers to concrete questions pertaining to presented text extracts.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Again, no personal data was taken from the human evaluators and therefore there were no obvious ethical risks associated with the evaluation tasks.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C