Abstract

Most existing work on text simplification is limited to sentence-level inputs, with attempts to iteratively apply these approaches to document-level simplification failing to coherently preserve the discourse structure of the document. We hypothesise that by providing a high-level view of the target document, a simplification plan might help to guide generation. Building upon previous work on controlled, sentence-level simplification, we view a plan as a sequence of labels, each describing one of four sentence-level simplification operations (copy, rephrase, split, or delete). We propose a planning model that labels each sentence in the input document while considering both its context (a window of surrounding sentences) and its internal structure (a token-level representation). Experiments on two simplification benchmarks (Newsela-auto and Wiki-auto) show that our model outperforms strong baselines both on the planning task and when used to guide document-level simplification models.

1 Introduction

Text simplification aims to transform a given text into a simpler version of itself that preserves the core meaning such that it can be better understood by a wider audience (Gooding, 2022). Simplification has also been shown to be a useful preprocessing step for downstream NLP tasks such as relation extraction (Miwa et al., 2010; Niklaus et al., 2016) and machine translation (Chandrasekar et al., 1996; Mishra et al., 2014; Li and Nenkova, 2015; Štajner and Popovic, 2016).

Previous research has mostly considered the simplification of isolated sentences. Much work has focused on training a statistical or a neural model on pairs of complex and simplified sentences assuming that such models will learn to perform simplification operations (e.g. sentence splitting, lexical simplification or syntactic rephrasing) implicitly from the inductive bias present in the training data (Zhang and Lapata, 2017; Nisioi et al., 2017; Jiang et al., 2020). However, because the training data is obtained using distant supervision techniques and is often imbalanced in terms of simplification operations (many of which occur infrequently (Jiang et al., 2020)), system outputs have been found to be overly conservative, often making no changes or being limited to the paraphrasing of short word sequences (Alva-Manchego et al., 2017; Maddela et al., 2021). In addition, these systems provide limited capacity for controllability and are unable to express alternative variants of the simplified text (Alva-Manchego et al., 2017; Cripwell et al., 2021).

In response, controllable simplification systems have been proposed which either constrain attributes of the output (length, amount of paraphrasing, lexical and syntactic complexity) (Martin et al., 2020) or explicitly specify which simplification operation to perform (Alva-Manchego et al., 2017; Dong et al., 2019; Malmi et al., 2019; Scarton et al., 2020; Maddela et al., 2021; Cripwell et al., 2022).

To guide the simplification of full documents, we combine the power of data-driven neural generative models with strategies from controllable simplification. Our hypothesis is that document-level simplification can be facilitated by a plan specifying how each complex input sentence should be transformed to yield a simplified version of that document - should it be copied, deleted, split or rewritten?

We make the following contributions: We present a model for predicting document simplification plans which leverages both the context of sentences and their internal structure (the words they consist of). We create the data necessary to train this model by labelling complex sentences in simplification corpora with the simplification operation that relates it to the corresponding simplified sentence. We compare our planning model
with several alternative neural architectures and we briefly examine the impact of planning on document simplification.

Experiments on two simplification benchmarks (Newsela-auto and Wiki-auto) show that our model outperforms strong baselines both on the planning task and when used to guide document-level simplification models. \(^1\)

2 Related Work

Document-Level Simplification. There is limited existing work on document-level text simplification. Early attempts largely applied sentence-level techniques iteratively over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b). However, this is generally viewed as insufficient for certain operations and maintaining the discourse coherence of the document (Siddharthan, 2003; Alva-Manchego et al., 2019b).

There are several works that address subproblems of simplification that only consider a limited set of operations, like paraphrasing and sentence re-ordering (Lin et al., 2021), insertion (Srikanth and Li, 2021) or deletion (Zhong et al., 2020; Zhang et al., 2022). Others fully address simplification but only extend inputs to the level of paragraphs without clearly differentiating the problem from the sentence-level (Laban et al., 2021; Devaraj et al., 2021).

Recently, Sun et al. (2020) proposed a sentence-level model (SUC) that uses an encoding of surrounding sentences as context information to influence the simplification. They use two extra encoders to build a representation of the two preceding and two following sentences, which are attended over in their encoder-decoder generative model. However, when applied to the document-level task, their system was unable to outperform any baseline systems (Sun et al., 2021).

Operation Prediction. Revision-based simplification models learn to predict edit operations to apply at the token-level rather than generating the entire simplification from scratch (Alva-Manchego et al., 2017; Dong et al., 2019; Kumar et al., 2020; Omelianchuk et al., 2021; Dehghan et al., 2022). This has the benefit of providing more control and interpretability over generative approaches, often at the cost of the ability to perform major structural changes. It also allows some systems to leverage non-autoregressive generation strategies, resulting in faster inference times (Malmi et al., 2019; Omelianchuk et al., 2021).

Some works have attempted to predict rewrite operations at the sentence-level. Applying a binary classifier to predict whether simplification should be performed has been found to improve SARI results, reducing conservatism and spurious transformations (Scarton et al., 2020; Garbacea et al., 2021). Others have proposed multi-class systems to predict sentence-level operations that are then used to condition a generative model (Scarton and Specia, 2018; Scarton et al., 2020; Cripwell et al., 2022). These show some capacity for general improvement over end-to-end systems, while also significantly improving performance for specific operations (e.g. splitting in the case of Cripwell et al. (2022)).

At the document-level, there has been limited interest to date. However, there are recent works specifically looking at predicting sentence deletions (Zhong et al., 2020; Zhang et al., 2022). Both of these use features of the discourse structure from surrounding sentences to identify likely deletion candidates.

We bring all of these methods together by proposing a system that uses both sentence and document-level information to predict a multi-class, sentence-level operation plan over an entire document.

3 Problem Formulation

Let \( C \) denote an English language document. The aim of document-level simplification is to produce a text \( S \) that simplifies the input document \( C \).

As a plan can provide a high-level view of a document, we hypothesize that a document-level simplification model that is based on a plan specifying a simplification operation for each input sentence should fare better than a simplification model that directly simplifies an entire document.

We therefore decompose simplification into a two-stage generation process:

\[
p(S \mid C) = p(S \mid C, P)p(P \mid C)
\]

where input document \( C = c_1 \ldots c_n \) is a sequence of complex sentences, \( S = s_1 \ldots s_k \) is a sequence of simplified sentences and \( P = o_1 \ldots o_n \) is a sequence of sentence-level simplification operations for \( C \).

\(^1\)Pretrained models, code, and data are available at https://github.com/liamcripwell/plan_simp.
We consider three simplification operations proposed in previous work on sentence simplification (copy, rephrase, and split) to which we add delete, an operation that is needed to account for the fact that, contrary to sentence simplification, document-level simplification can require for a sentence present in the input document to be excluded from the resulting simplified document.

Given the input document $C$, the first-stage model aims to predict the sequence of simplification operations $P$ that should be applied to each individual sentence in that document. The second-stage model generates the output simplified document $S$ conditioned on the input document $C$ and its accompanying simplification plan $P$.

In this work, we focus on the planning stage, comparing different architectures and demonstrating the impact of planning on three possible document-level simplification models. We leave the exploration of alternative, more complex architectures for the simplification stage to future work.

In Table 1, we provide statistics of each dataset after preprocessing, where $n$ is # sentences in $C$ and $k$ is # sentences in $S$.

<table>
<thead>
<tr>
<th></th>
<th>Wiki-auto</th>
<th>Newsela-auto</th>
</tr>
</thead>
<tbody>
<tr>
<td># Doc Pairs</td>
<td>85,123</td>
<td>18,319</td>
</tr>
<tr>
<td># Sent Pairs</td>
<td>461,852</td>
<td>707,776</td>
</tr>
<tr>
<td>Avg. $</td>
<td>C</td>
<td>$</td>
</tr>
<tr>
<td>Avg. $</td>
<td>S</td>
<td>$</td>
</tr>
<tr>
<td>Avg. $</td>
<td>c</td>
<td>$</td>
</tr>
<tr>
<td>Avg. $</td>
<td>s</td>
<td>$</td>
</tr>
<tr>
<td>Avg. $n$</td>
<td>5.43</td>
<td>38.64</td>
</tr>
<tr>
<td>Avg. $k$</td>
<td>4.53</td>
<td>42.60</td>
</tr>
</tbody>
</table>

Table 1: Statistics of each dataset after preprocessing, where $n$ is # sentences in $C$ and $k$ is # sentences in $S$.

### 4 Data

In this section we introduce the datasets used, explain how annotation is performed for each complex sentence and describe other preprocessing steps.

**Dataset.** For all experiments, we utilise Wiki-auto and Newsela-auto (Jiang et al., 2020), two datasets of English documents paired with their simplification. These datasets were derived from WikiLarge (Zhang and Lapata, 2017) and Newsela (Xu et al., 2015) by aligning the input document with the output simplified document at both the sentence and the paragraph level.

WikiLarge gathers three simplification datasets which were automatically-collated from English Wikipedia and Wikipedia simple (Zhu et al., 2010; Woodsend and Lapata, 2011b; Kauchak, 2013). Newsela consists of news articles, each manually rewritten at five different levels of simplification, corresponding to discrete reading levels (0-4) of increasing simplicity. Aligned pairs are created by pairing every article version with each other version corresponding to a higher reading level. Because of this, there can be up to four aligned document pairs that contain the same document as either the input or the output.

The types of operations present in different reading level pairings differ significantly, with adjacent level transitions being extremely conservative (no instances of deletion throughout entire dataset). To mitigate any issues arising from this, all models we train with Newsela-auto receive a control-token at the start of the input which specifies the target reading level.

We do not use the D-Wikipedia dataset from Sun et al. (2021) as it does not contain sentence/paragraph alignments and is poorly formatted. In particular, all text is lower-cased and pretokenized in a way that makes it difficult to accurately parse sentences. There are also regular formatting issues at points where references exist in the source article.

**Annotating Complex Sentences.** Using the pairs $(c_i, s_j)$ of complex and simplified sentences available in Wiki-auto and Newsela-auto, we heuristically assign a silver simplification opera-
tion label to each complex sentence \( c_i \) in these two datasets as follows:

**Delete:** \( c_i \) is not aligned to any \( s_j \).

**Copy:** \( c_i \) is aligned to a single \( s_j \) with a Levenshtein similarity above 0.92.

**Rephrase:** \( c_i \) is aligned to a single \( s_j \) with a Levenshtein similarity below 0.92.

**Split:** \( c_i \) is aligned to multiple \( s_j \)s.

**Preprocessing.** Wiki-auto contains many document pairs with wildly different sizes. We therefore clip all complex documents after the last aligned paragraph. Many simple articles resemble a summarization, rather than a simplification of the complex article (lots of deletion, often consisting of about one sentence from each paragraph in \( C \)). Because of this, we also remove documents where more than 50% of aligned sentences are labelled as *delete*. Finally, we remove all articles that exceed 1024 tokens (so that we can fit them into a BART baseline generative model).

For Newsela-auto, article pairs are much more even in length as they are manually created to be gradual, direct simplifications of each other. We perform the same length-based filtering to exclude documents that will not fit into a baseline generative model.

**Train/Dev/Test Split.** For both datasets we use a train/validation/test split of 92.5/2.5/5. This is applied at the document-level so that sentences from the same document will not exist across different sets. For Newsela, this means that all reading level versions of a single article will exist within the same set.

Table 1 and Figure 1 give some statistics and a graphical description of the two datasets after pre-processing.

5 Planning

We present our model and four alternative models we explored for comparison. Training details are given in Appendix A.

5.1 Model (Contextual Classifier)

Given some input document \( C = c_1 \ldots c_n \) consisting of \( n \) complex sentences \( c_i \), the task of the planner is to predict a sequence \( \hat{P} = \hat{o}_1 \ldots \hat{o}_n \) of \( n \) simplification operations with \( \hat{o}_i \in \{\text{copy, rephrase, split, delete}\} \).

One challenge with this is that the operations have different, sometimes conflicting requirements.

By construction, splitting is mostly context independent as it is mainly determined by the input sentence’s internal structure: a sentence will be split only if it has the appropriate syntactic (e.g., The man who sleeps snores \( \rightarrow \) The man sleeps. He snores.) or discourse (e.g., John went shopping after he left work \( \rightarrow \) John left work. Afterwards he went shopping.) structure. For sentence splitting, context (the other sentences in the input document) has little impact.

In contrast, deletion and to a lesser extent, copy and rephrase are mostly context dependent. Intuitively, a sentence can only be omitted in the simplified text in cases where it is either redundant with, or of minor semantic import relative to, other sentences in the document. That is, while for splitting, internal sentence structure is the key factor, for deletion, it is the semantics of the input sentence and how it relates to that of the other sentences which matters most.

We model these different requirements by using a token level encoder for the target document sentence \( c_i \) (the input sentence to be labelled with a simplification operation) and a sentence level representation of the context where each \( c_p \in c_1 \ldots c_{i-1}, c_{i+1} \ldots c_n \) is represented by a sentence level embedding using SBERT. In this way both the internal structural information needed to capture splitting operations and the contextual information required by the other operations are provided. Specifically, we propose a model for planning that combines a classifier with cross-attention over the (dynamic or static) context and two types of positional embeddings. Figure 2 illustrates our model.
architecture.

**Classifier with Cross-attention over the Context.** We build upon a RoBERTa classifier architecture to enable conditioning upon the surrounding sentences in the document. We do this by inserting an additional cross-attention layer between the self-attention and the feed-forward layer of each transformer block, allowing the model to attend to a latent representation of the surrounding sentences, \(Z_i\).

**Context Representation.** To obtain \(Z_i\), we take a fixed window of radius \(r\), extract the \(r\) sentences on either side of the target sentence to be simplified and concatenate the representation of each of these sentences. Each context sentence is encoded with the pretrained Sentence-BERT (SBERT) model\(^2\) (Reimers and Gurevych, 2019) and combined with custom learnt positional embeddings.\(^3\)

To better simulate autoregressive inference, we consider a strategy where the left context consists of previously simplified sentences, rather than complex ones. We refer to this as *dynamic context*. At training time, we use the ground truth simplifications

\[
\text{Context}_{i,r} = \text{Concat}(s_{j-r..j-1}, c_{i..i+r}) \tag{1}
\]

where \(j \in \{1, \ldots, |S|\}\) is the index of the first sentence aligned to \(c_i\) in the simple document \(S\).

During inference, the simplifications generated at preceding timesteps \(\hat{s}_{j-r..j-1}\) are used.

**Positional Embeddings.** We use custom positional embeddings to encode both information about document, and relative context-window positions. These are each handled by a dedicated embedding layer and added to the representations of the corresponding context sentence.

Document positional embedding indices are simply the document quintile (1-5) that a given sentence falls into. We use quintiles as this will ensure that all indices are encountered within the input document. The context positional embedding indices are the relative distance of a given sentence from the input sentence \(c_i\), adjusted to be within \(\mathbb{N}_0\) :

\[
\text{ContextPosIdxs} = \{p - i + r \mid p \in \{i - r, \ldots, i + r\}\}
\]

**Initialisation.** Given that the cross-attention layers must be trained from scratch, the start of training can see a lot of instability in the model, potentially making it more difficult to model context-independent features of the input sentence. To account for this, we initialise the RoBERTa layers with weights from a context-independent classifier.

### 5.2 Alternative Models

We compare our model with four alternative models. The different inputs/outputs of the models are illustrated in Figure 3.

**Classifier.** We fine-tune pretrained RoBERTa-base (Liu et al., 2019), which has 12 hidden layers and a hidden size of 768, and
Wiki-auto
Model C R S D Micro Macro
EncDec full 26.9 42.2 36.0 51.8 43.2 40.8
Tagger + Dec 29.3 54.5 30.0 51.8 47.7 44.4
Tagger 38.6 54.2 31.7 58.5 30.6 45.8
Classifier 42.1 52.9 42.6 49.0 48.4 46.7
Dyn. Context 44.8 57.9 42.4 54.8 52.8 50.0
+ docpos 43.7 55.4 43.6 56.7 52.3 49.9

Newsela-auto
Model C R S D Micro Macro
EncDec full 26.1 10.8 11.7 9.0 12.2 11.5
Tagger + Dec 72.2 73.9 75.9 79.7 75.0 75.4
Tagger 71.4 72.7 74.1 78.4 73.7 74.1
Classifier 77.0 75.6 78.0 78.5 77.4 77.8
Dyn. Context 79.3 77.3 82.8 81.4 79.7 80.2
+ docpos 80.0 78.1 83.6 82.0 80.3 80.8

Table 2: Planning Accuracy. Dyn. Context is the contextual classifier described in Section 5.1 with \( r = 13 \), dynamic context and weights initialised using the classifier weights (C: Copy, R: Rephrase, S: Split, D: Delete).

We consider a model that frames the problem as a sequence tagging task over the full document, predicting the entirety of \( \hat{P} \) at once. Each \( c_i \) is encoded using the same SBERT model as the contextual classifier, with the input document \( C \) therefore being represented as a sequence of sentence embeddings. In contrast to the classifier, the tagger makes predictions based both on the input sentence to be classified and on the context. However, because the input representation at each index is for an entire sentence we lose some resolution with respect to token-level content. The approach is thus less adapted for splitting.

**Tagger + Dec.** We also consider an autoregressive variant of the tagger that better models the dependencies between predicted tags. Here, we include a 1-layer decoder and condition each prediction both on the input document and on the previously predicted operation tags for the earlier sentences. This approach is somewhat similar to Dong et al. (2019); Malmi et al. (2019), except we abstract to the document-level and do not require explicit realisation, as this will be handled downstream by the simplification model.

**EncDec full.** Finally, we experiment with an encoder-decoder variant that conditions on a token-level representation of the input, thereby combining a global view and a token-level representation of the input document. We use sentence separator tokens to delimit each sentence in the input document.

5.3 Evaluation Metrics

To evaluate the performance of the various planners we use F1-score, considering each individual prediction at the sentence-level. We report the F1 for each operation class as well as both the micro and macro averages. The micro F1 weights all examples equally, whereas the macro re-weights examples such that each class is represented equally in the final score. Given the class imbalances in the data, we regard macro F1 as our primary metric.

5.4 Results

Table 2 summarizes the results.

Compared to the various baselines, our model consistently shows best results on both datasets. The improvement over the context-free classifier is slightly less on Newsela-auto however. We conjecture that the much larger dataset and additional guidance provided by the reading levels allows the classifier to achieve rather high accuracy without document-level context. We also note that the context-free classifier is markedly outperformed by other models with respect to delete, which confirms the intuition that context modeling particularly matters for this operation.

Of the four baselines, EncDec full performs worst presumably because the very long input (the whole context is modelled at the token level) challenges the attention mechanism which tends to become blurry as the length of the input increases. This is particularly apparent on the longer Newsela documents.

The tagger models, which both use sentence-level encodings of the complex document, perform
worse than the classifier. This highlights the importance of having a token-level modeling of the input sentences.

We observe a strong difference in terms of absolute scores between the two datasets. This is likely a result of Wiki-auto being an inferior simplification corpus (discussed in Section 4).

Next, we examine the impact of our modeling choices using ablation (Table 3) and focusing on the higher-quality, Newsela-auto dataset. Our best model is one with dynamic left-context, a context radius of 13, document position embeddings and weight initialisation. We see (Sub-table a) that each of these components help improve performance (document position appears less important with a larger context window). Sub-tables b-d show that using a dynamic rather than a static context increases results by up to +6.7 Macro F1, while increasing the context radius from 9 to 13 sentences mostly improves performance when dynamic context is used. Using document positional embeddings also generally improves results (Sub-table d).

### 6 Simplification

To assess whether document plans can help improve simplification models, we experiment with two simple document-level simplification models and compare their performance with and without a preceding planning step.

#### 6.1 Simplification Models

All models use the BART model (Lewis et al., 2020) fine-tuned on aligned text pairs.  

We consider two variants for document-level simplification: (i) **Doc-BART**, which is finetuned on full document pairs; and (ii) **Sent-BART** which is finetuned on sentence pairs and iteratively applied to each input sentence at test time.

We compare these to various plan-guided (PG) systems whereby one of our planners predicts an $\hat{o}_i$ for each $c_i$ and is given as a control-token to a sentence-level BART simplification model. In the case of the dynamic planner, $\hat{o}_i$ is predicted based on the sequence of previously simplified sentences $\hat{s}_{i-r} \ldots \hat{s}_{i-1}$.

Training details are given in Appendix B.

#### 6.2 Evaluation

To measure meaning preservation and fluency, we use BARTScore (Yuan et al., 2021), a state-of-the-

---

Table 3: Ablations on Newsela-auto TestSet.

<table>
<thead>
<tr>
<th>Model</th>
<th>Copy</th>
<th>Rephrase</th>
<th>Split</th>
<th>Delete</th>
<th>Micro</th>
<th>Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Ablation on Best Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyn, $r = 13$, +init, +docpos</td>
<td>80.0</td>
<td>78.1</td>
<td>83.6</td>
<td>82.0</td>
<td>80.3</td>
<td>80.8</td>
</tr>
<tr>
<td>-docpos</td>
<td>79.3</td>
<td>77.3</td>
<td>82.8</td>
<td>81.4</td>
<td>79.7</td>
<td>80.2</td>
</tr>
<tr>
<td>-init</td>
<td>74.9</td>
<td>72.1</td>
<td>77.8</td>
<td>75.2</td>
<td>74.6</td>
<td>75.0</td>
</tr>
<tr>
<td>-init, -docpos</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td>(b) Dynamic vs. Static Context</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stat, $r = 9$</td>
<td>71.3</td>
<td>69.5</td>
<td>75.4</td>
<td>73.3</td>
<td>72.0</td>
<td>72.4</td>
</tr>
<tr>
<td>Stat, $r = 13$</td>
<td>72.2</td>
<td>65.3</td>
<td>69.9</td>
<td>68.3</td>
<td>68.5</td>
<td>68.9</td>
</tr>
<tr>
<td>Dyn, $r = 9$</td>
<td>73.1</td>
<td>70.1</td>
<td>75.5</td>
<td>75.9</td>
<td>73.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td>(c) With vs without Initialisation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyn, $r = 9$</td>
<td>73.1</td>
<td>70.1</td>
<td>75.5</td>
<td>75.9</td>
<td>73.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Dyn, $r = 9$ +init</td>
<td>79.3</td>
<td>78.0</td>
<td>82.7</td>
<td>79.8</td>
<td>79.7</td>
<td>80.0</td>
</tr>
<tr>
<td>Dyn, $r = 13$</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$ +init</td>
<td>79.3</td>
<td>77.3</td>
<td>82.8</td>
<td>81.4</td>
<td>79.7</td>
<td>80.2</td>
</tr>
<tr>
<td>(d) Window Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stat, $r = 9$</td>
<td>71.3</td>
<td>69.5</td>
<td>75.4</td>
<td>73.3</td>
<td>72.0</td>
<td>72.4</td>
</tr>
<tr>
<td>Stat, $r = 13$</td>
<td>72.2</td>
<td>65.3</td>
<td>69.9</td>
<td>68.3</td>
<td>68.5</td>
<td>68.9</td>
</tr>
<tr>
<td>Dyn, $r = 9$</td>
<td>73.1</td>
<td>70.1</td>
<td>75.5</td>
<td>75.9</td>
<td>73.1</td>
<td>73.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$</td>
<td>75.6</td>
<td>72.0</td>
<td>77.7</td>
<td>77.1</td>
<td>75.1</td>
<td>75.6</td>
</tr>
<tr>
<td>Dyn, $r = 9$ +docpos</td>
<td>73.8</td>
<td>72.9</td>
<td>77.2</td>
<td>75.8</td>
<td>74.6</td>
<td>74.9</td>
</tr>
<tr>
<td>Dyn, $r = 13$ +docpos</td>
<td>74.9</td>
<td>72.1</td>
<td>77.8</td>
<td>75.2</td>
<td>74.6</td>
<td>75.0</td>
</tr>
<tr>
<td>Dyn, $r = 9$ +init +docpos</td>
<td>79.4</td>
<td>78.0</td>
<td>83.1</td>
<td>82.0</td>
<td>80.1</td>
<td>80.6</td>
</tr>
<tr>
<td>Dyn, $r = 13$ +init +docpos</td>
<td>80.0</td>
<td>78.1</td>
<td>83.6</td>
<td>82.0</td>
<td>80.3</td>
<td>80.8</td>
</tr>
</tbody>
</table>

---

4We use the pretrained facebook/bart-base model from https://huggingface.co/facebook/bart-base.
Table 4: Results of document simplification systems on Newsela-auto. For BARTScore, $s$ is the source, $h$ is the hypothesis, and $r$ is the reference.

art summarization metric that has proved effective on many other text generation tasks. We also compute SMART (Amplayo et al., 2022), a new metric that considers sentences as the primary unit of comparison. It was shown to be highly effective for document summarization and does not use any neural model, making it very fast to compute (we use the SMARTL+CHRF version). We cannot use other model-based metrics, such as BERTScore or QuestEval, as these do not support texts longer than 512 tokens.

To assess simplicity, we use the Flesch-Kincaid grade level (FKGL), a document-level metric used to measure text readability, which has been found to have the highest correlation with simplicity measures of human-written simplifications (Scialom et al., 2021). We also report the popular SARI (Xu et al., 2016). The EASSE python library (Alvamanchego et al., 2019a) is used for calculation of FKGL and SARI. We include results for other popular metrics in Appendix D.

At test time we generate sequences using beam search with a beam size of 5 and a maximum length of 1024 tokens. We enforce a minimum length for Doc-BART, which is tuned on the validation set.

We do not conduct a human evaluation as we intend the focus of this work to be on the planning component and include simplification results only to confirm its efficacy. We leave a more in-depth investigation of the interaction between planning and document-level simplification to future work.

6.3 Results

Results can be seen in Table 4.

PG$_{Dyn}$ achieves the highest results of all systems. Using the silver operation labels (PG$_{Oracle}$) leads to a substantial further increase in performance across every metric, highlighting the impact of planning and pointing to the possibility of further improvements to be made.

Using either PG$_{Dyn}$ or PG$_{Clf}$ yields generally better results than Sent-BART. Both systems achieve better FKGL and SARI, suggesting greater output simplicity. Sent-BART achieves much higher source-oriented BARTScore (faithfulness) than even the references, suggesting some conservativity in its transformations.

PG$_{Clf}$ achieves slightly higher recall BARTScore than PG$_{Dyn}$, while also generating the longest outputs, both in terms of tokens and sentences. This suggests it is less effective at identifying sentences for deletion, confirming our hypothesis that context is key for deletion. We can see here that the rank order of SMART matches that of BARTScore, suggesting it is similarly suited for simplification.

Both PG$_{Tag}$ and PG$_{Tag+Dec}$ perform quite badly relative to the other PG systems and Sent-BART. However, Doc-BART is by far the worst performing system, presumably a result of it failing to properly handle the long document lengths.

7 Conclusion

In this paper we present an approach to document simplification that decomposes the task into a two-stage process of planning and generation. We propose a planning system that is able to take document context and structure into account to produce a coherent high-level simplification plan. By using this plan to guide a sentence-level simplification model, we are able to outperform end-to-end systems in terms of both meaning preservation and simplicity.
We leave for future work the development of dedicated simplification models that can leverage a document-level plan while also considering contextual information directly during generation.

8 Acknowledgements

We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the French National Research Agency (Gar- dent; award ANR-20-CHIA-0003, XNLP "Multilingual, Multi-Source Text Generation").

Experiments presented in this paper were carried out using the Grid 5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

9 Limitations

All models we propose are only trained with English datasets and therefore do not extend to other languages. There are also additional high-level simplification operations for which our planning framework does not offer support, such as sentence reordering, insertion, and fusion.

Furthermore, the NewseLA dataset which we use in our experiments requires a license to use, meaning that researchers cannot fully reproduce our work without first obtaining said license from NewseLA Inc. Due to this constraint and the low-quality alignments observed within the Wiki- auto dataset, we strongly encourage any work towards producing new open-access datasets for the document-level simplification task.

References


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


Sian Gooding. 2022. On the ethical considerations of text simplification. In Ninth Workshop on Speech


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/).


A Training Details for the Planning Models

Each model was trained with a learning rate of $1 \times 10^{-5}$, a batch size of 32 and a dropout rate of 0.1. We ran experiments on a computing grid with $2 \times$ Nvidia A40 GPUs (45GB memory).

For the contextual classifier, we test with $r$ values of 9 and 13, subject to findings in Appendix C. All layers in common with the standard RoBERTa architecture are initialised with the RoBERTa-base pretrained weights. All added positional embedding layers are also initialised with the pretrained weights from the RoBERTa-base positional embedding layer. All other layers are randomly initialised.

B Training Details for the Simplification Models

For all generative models, we used a learning rate of $3 \times 10^{-5}$, a batch size of 16, and performed dropout with a rate of 0.1 and early stopping. The network has 6 layers in each of the encoder and decoder, with a hidden size of 768. All models were trained on a computing grid using $2 \times$ Nvidia A40 GPUs (45GB memory) in under 24 hours.

C Context Window Size

To determine the optimal context window size for the contextual planner we ran a series of experiments with varying values of the radius, $r$. We used 100,000 random examples from the Newsela-auto (non-adjacent reading levels) training set and trained a model with each of the configurations for 5 epochs. Results can be seen in Figure 4.

The deletion operation is most affected by the inclusion of context, with performance rapidly rising as $r$ grows to 13. The rephrase operation appears to slowly degrade in performance as $r$ increases, while the other two operations show no obvious pattern. We also observe that $r = 9$ produces the highest macro F1.

D Extra Evaluation Results

For clarity, we provide scores for a wider range of simplification evaluation metrics that were not included in the main body of the paper in Table 5. These mostly include popular metrics used for sentence simplification that we do not believe adapt as well to the document-level setting, do not provide further insight into system differences, or have not received much support in the literature. Specifically, we include BLEU (Papineni et al., 2002), and full operation scores for both SARI and D-SARI (Sun et al., 2021). For D-SARI, we apply the document-level penalties on top of the base EASSE implementation of SARI.

We can see that the main SARI differences between the context-free planner and Sent-BART is that Sent-BART achieves higher keep, while the planner achieves higher add. This suggests that Sent-BART is likely more conservative in edits. Further, as the planner does not have access to contextual content, it is likely failing to consistently copy/delete the correct parts of the text.

E Example Planner Predictions

Figure 5 shows example snippets of planner model outputs. We have selected representative extracts that highlight the strengths and weaknesses of the main models. We do not include outputs from Tagger as they are virtually identical to Tagger+Dec in most cases and therefore do not provide further insight.

F Example Simplification

Figure 6 shows system output examples for the simplification models. We only show texts from Wiki-auto as they are easier to showcase due to their shorter length, as well as their being licensing restrictions for Newsela content.
<table>
<thead>
<tr>
<th>System</th>
<th>BLEU ↑</th>
<th>D-SARI ↑</th>
<th>add</th>
<th>keep</th>
<th>delete</th>
<th>SARI ↑</th>
<th>add</th>
<th>keep</th>
<th>delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>46.2</td>
<td>8.76</td>
<td>0.0</td>
<td>26.29</td>
<td>0.0</td>
<td>20.52</td>
<td>0.0</td>
<td>61.56</td>
<td>0.0</td>
</tr>
<tr>
<td>Doc-BART</td>
<td>31.13</td>
<td>30.60</td>
<td>16.54</td>
<td>25.01</td>
<td>50.24</td>
<td>47.07</td>
<td>20.41</td>
<td>55.40</td>
<td>65.40</td>
</tr>
<tr>
<td>Sent-BART</td>
<td>70.74</td>
<td>66.27</td>
<td>53.89</td>
<td>71.95</td>
<td>72.95</td>
<td>73.02</td>
<td>55.91</td>
<td>83.66</td>
<td>79.48</td>
</tr>
<tr>
<td>PG_{Tag}</td>
<td>48.08</td>
<td>42.96</td>
<td>31.70</td>
<td>44.01</td>
<td>53.17</td>
<td>56.13</td>
<td>35.61</td>
<td>65.61</td>
<td>67.18</td>
</tr>
<tr>
<td>PG_{Tag+Dec}</td>
<td>48.12</td>
<td>43.31</td>
<td>31.57</td>
<td>44.68</td>
<td>53.69</td>
<td>56.06</td>
<td>35.54</td>
<td>65.54</td>
<td>67.11</td>
</tr>
<tr>
<td>PG_{Clf}</td>
<td>70.84</td>
<td>62.97</td>
<td>56.31</td>
<td>65.15</td>
<td>67.47</td>
<td>73.83</td>
<td>57.62</td>
<td>83.56</td>
<td>80.32</td>
</tr>
<tr>
<td>PG_{Dyn}</td>
<td>72.41</td>
<td>67.42</td>
<td>56.83</td>
<td>71.82</td>
<td>73.61</td>
<td>75.00</td>
<td>58.88</td>
<td>84.75</td>
<td>81.36</td>
</tr>
<tr>
<td>PG_{Oracle}</td>
<td>78.97</td>
<td>77.02</td>
<td>63.44</td>
<td>83.92</td>
<td>83.70</td>
<td>80.74</td>
<td>65.22</td>
<td>89.94</td>
<td>87.05</td>
</tr>
</tbody>
</table>

Table 5: Extra results for document simplification experiments on Newsela-auto.

![Figure 5: Example planning results for various models. Subfigures show representative snippets from Newsela-auto test-set documents. The silver labels are shown above in yellow, and system outputs are shown on the rows below with correct predictions in green and incorrect predictions in red. Clf_{Dyn} is our best performing model, the contextual classifier with dynamic context. Figure 5a shows a case where there are lots of context-agnostic operations (rephrase, split) resulting in poor performance from Tagger+Dec. Figure 5b shows a varied snippet where Clf_{Dyn} appears to be the best at identifying both rephrase and split, as well as delete. Figures 5c and 5d show that Tagger+Dec is capable of performing well in situations demanding a lot of context-dependent operations (copy, delete).](image-url)
Silvano "Nano" Campeggi (January 23, 1923 – August 29, 2018) was an Italian artist who designed and produced the artwork for the posters of many classic Hollywood films. His iconic images are associated with the golden era of Hollywood and Campeggi is now generally regarded as the most important graphic artist and poster designer in the history of American cinema. In the following decades, Campeggi designed and produced the poster and advertising graphics for over 3000 films, working not only under contract with the MGM studios, but also with Warner Brothers, Paramount, Universal, Columbia Pictures, United Artists, RKO, Twentieth-Century Fox and several other movie studios. Sixty-four of the films he illustrated won Oscars, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany’s", and "Gigi". Campeggi died on 29 August 2018, at the age of 95.

Table 6: Simplification outputs for a specific document pair example. Although Newsela-auto is the focus of our simplification experiments, we can only include example documents from Wiki-auto due to licensing constraints.