Simplifying Documents and Evaluating Simplification Models

Claire Gardent

Joint work with Liam Cripwell

CNRS / LORIA

Example

Complex Input Document

Owls are birds from the order of Strigiformes, comprising over 200 species of mostly solitary and nocturnal birds of prey typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Simplified Output Document

Owls are birds. There are over 200 species and are all animals of prey. Most of them are solitary and nocturnal. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Avg nb of sentences in Input Document: 39

Simplification Operations

Owls are birds from the order of Strigiformes, comprising over 200 species of mostly solitary and nocturnal birds of prey typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Owls are birds. There are over 200 species ...

Sentence Splitting

Document Simplification

Owls are birds from the order of Strigiformes, comprising over 200 species of ***mostly solitary and nocturnal birds of prey*** typified by an upright stance, binocular vision, binaural hearing, and sharp talons. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Owls are birds. There are over 200 species ***and are all animals of prey. Most of them are solitary and nocturnal*** . Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Rephrasing

Document Simplification

Owls are birds from the order of Strigiformes, comprising over 200 species of mostly solitary and nocturnal birds of prey typified by an upright stance, binocular vision, binaural hearing, and sharp talons. ***Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish*** .

Owls are birds. There are over 200 species and are all animals of prey. Most of them are solitary and nocturnal. ***Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits)***.

Rephrasing

Document Simplification

Owls are birds from the order of Strigiformes, comprising over 200 species of mostly solitary and nocturnal birds of prey ***typified by an upright stance, binocular vision, binaural hearing, and sharp talons*** . Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.

Owls are birds. There are over 200 species and are all animals of prey. Most of them are solitary and nocturnal. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Deletion

Why Simplify ?

- To aid reader comprehension (Mason, 1978; Williams et al., 2003; Kajiwara et al., 2013)
    - Adult vs children
    - Native vs non Native
    - Reading disability
    - Expert vs non-Expert
- Useful preprocessing step for downstream NLP tasks such as
    - relation extraction (Miwa et al., 2010; Niklaus et al., 2016)
    - machine translation (Chandrasekar et al., 1996; Mishra et al., 2014; Li and Nenkova, 2015; Mishra et al., 2014; Štajner and Popovic, 2016).

Outline

Document Simplification

- Guiding Sentence Simplification using Controls

  $$\textit{Complex Sentence} \rightarrow \hat{o}$$
  $$\hat{o}\textit{, Complex Sentence} \rightarrow \textit{Simple Sentence}$$

- Modeling Context

Controlled Simplification

CONTROL, Complex Sentence $\rightarrow$ *Simple Sentence*

Controlled Simplification

CONTROL, Complex Sentence $\rightarrow$ *Simple Sentence*

split, "*John saw a man who walks*" $\rightarrow$ "John saw a man. The man walks"

Controlled Simplification

CONTROL, Complex Sentence $\rightarrow$ *Simple Sentence*

CONTROL : ***split, copy, rephrase, delete***

Document Simplification

A ***planning*** approach

> ***PLAN*** - A sequence of simplification operations for the input document

Document Simplification

A ***planning*** approach

> ***PLAN*** - A sequence of simplification operations for the input document

A ***context-based, structural*** approach

> Simplification operations are predicted based on a sentence ***context*** and ***structure***
>
> ***CONTEXT*** - The text surrounding a sentence
>
> ***STRUCTURE*** - The words making up a sentence

Outline

Document Simplification

- Guiding Sentence Simplification using Controls

  $$\textit{Complex Sentence} \rightarrow \hat{o}$$
  $$\hat{o}, \textit{ Complex Sentence} \rightarrow \textit{Simple Sentence}$$

- Modeling Context

Evaluating Simplification Models

- A new, reference-less metric for simplicity
- The trade-off between simplification and meaning preservation

Controlled Sentence Simplification

Cripwell et al. NAACL Findings 2022

End-To-End Neural Simplification

- Encoder-Decoder Models
- Trained on parallel Corpora of (C,S) pairs
  - Wiki: English Wikipedia / Simple English Wikipedia
  - Newsela
- Implicitly learn simplification operations from the training data

End-To-End Simplification

Shortcomings

- Noisy training data

End-To-End Simplification

Shortcomings

- Noisy training data
- Some simplification operations are rare
  (Jiang et al., 2020)

End-To-End Simplification

Shortcomings

- Noisy training data
- Some simplification operations are rare
  (Jiang et al., 2020)
- Overly conservative models
  (Alva-Manchego et al., 2017; Maddela et al., 2021)

Controlled Sentence Simplification

A ***classifier*** which, given a sentence, predicts a simplification operation

- copy (no simplification needed)
- rephrase
- split

A ***pipeline simplification model*** which generates a simplification based on a predicted simplification operation

$$\textit{Complex Sentence} \rightarrow \textit{CONTROL}$$

$$\textit{CONTROL, Complex Sentence} \rightarrow \textit{Simple Sentence}$$

Controlled Sentence Simplification

A ***classifier*** which, given a sentence, predicts a simplification operation

- copy (no simplification needed)
- rephrase
- split based on syntax
- split based on discourse structure

Results

- Outperforms end-to-end baselines and previous controllable systems.
- Performs splits much more often than existing systems, and knows when to perform minimal edits.

Document Simplification

Cripwell et al. EACL 2023

Previous work

Sentence-level simplification iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)

> ***Low discourse coherence***
> (Siddharthan, 2003; Alva-Manchego et al., 2019b).

Previous work

Sentence-level simplification iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)

> ***Low discourse coherence***
> (Siddharthan, 2003; Alva-Manchego et al., 2019b).

Sub-problems of simplification

- paraphrasing and sentence re-ordering (Lin et al., 2021)
- insertion (Srikanth and Li, 2021) or
- deletion (Zhong et al., 2020; Zhang et al., 2022).

> ***Only consider a limited set of operations***

Previous work

Sentence-level simplification iteratively applied over a document (Woodsend and Lapata, 2011a; Alva-Manchego et al., 2019b)

> ***Low discourse coherence***
> (Siddharthan, 2003; Alva-Manchego et al., 2019b).

Sub-problems of simplification

- paraphrasing and sentence re-ordering (Lin et al., 2021)
- insertion (Srikanth and Li, 2021) or
- deletion (Zhong et al., 2020; Zhang et al., 2022).

> ***Only consider a limited set of operations***

A sentence-level model that uses context information to influence document simplification (Sun et al. 2020)

> ***Underperform the baseline*** (Sun et al. 2021)

Plan-Guided Document Simplification

Plan-Guided (PG) pipeline

First PLAN a sequence of simplification operations
Input D $\Rightarrow$ Simplification Plan

$c1, \ldots, c_n \Rightarrow \hat{o}, \ldots, \hat{o}_n$

then SIMPLIFY
Input S + Simplification Operation $\Rightarrow$ Simplified S

$c_i, \hat{o}_i \Rightarrow s_i$

Planning Simplification Operations

$$c_1, \ldots, c_n \Rightarrow \hat{o}, \ldots, \hat{o}_n$$

Planning Simplifications

$$c_1, \ldots, c_n \Rightarrow \hat{o}, \ldots, \hat{o}_n$$

Given some input document $C = c_1, \ldots, c_n$ the task of the planner is to ***predict a sequence of n simplification operations*** $\hat{P} = \hat{o}, \ldots, \hat{o}_n$ with $\hat{o}_i \in$ {copy, rephrase, split, **delete**}

Challenges

Simplification Operations have different requirements

Splitting

- mainly depends on the ***input sentence's internal structure***

  *The man **who** sleeps snores → The man sleeps. He snores.*

  *John went shopping **after** he left work → John left work. Afterwards he went shopping.*

Challenges

Simplification Operations have different requirements

Splitting

- mainly depends on the ***input sentence's internal structure***

  *The man **who** sleeps snores → The man sleeps. He snores.*

  *John went shopping **after** he left work → John left work. Afterwards he went shopping.*

Deletion, copy and rephrase

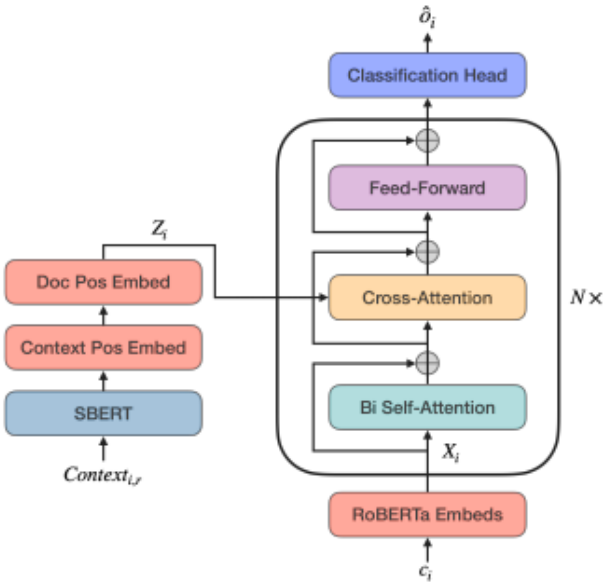
- are mostly ***context dependent*** .

  A sentence can only be omitted if it is either **redundant** with, or of **minor semantic import** relative to, other sentences in the document

Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier



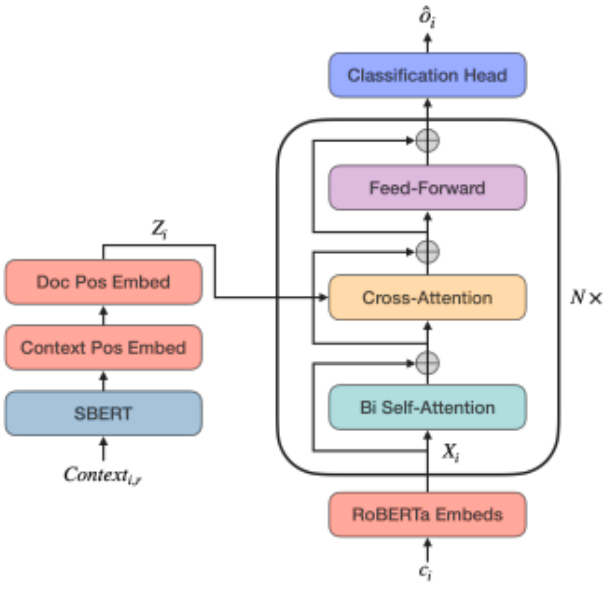
Planning Model

RoBERTa classifier with cross-attention over the context

- layers initialised with weights from a context-independent classifier

Internal structure

- ***Token level*** encoder for c_i



Planning Model

RoBERTa classifier with cross-attention over the context

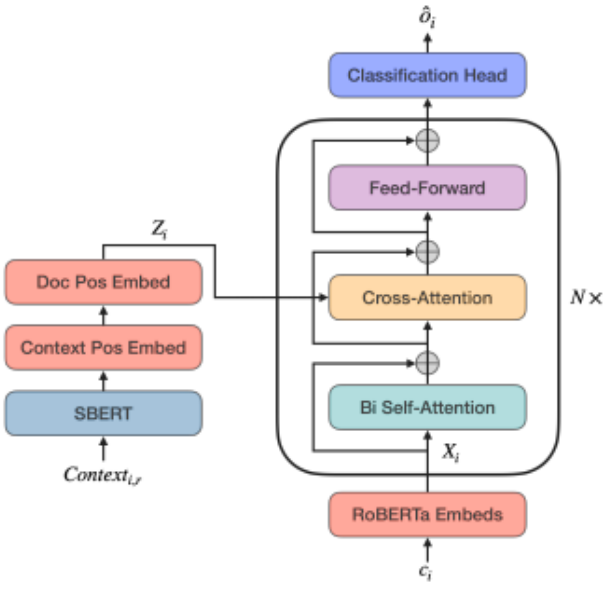
- layers initialised with weights from a context-independent classifier

Internal structure

- ***Token level*** encoder for c_i

Context

- fixed window of Sentence level embedding (SBERT) for ***surrounding sentences***



Planning Model

RoBERTa classifier with cross-attention over the context

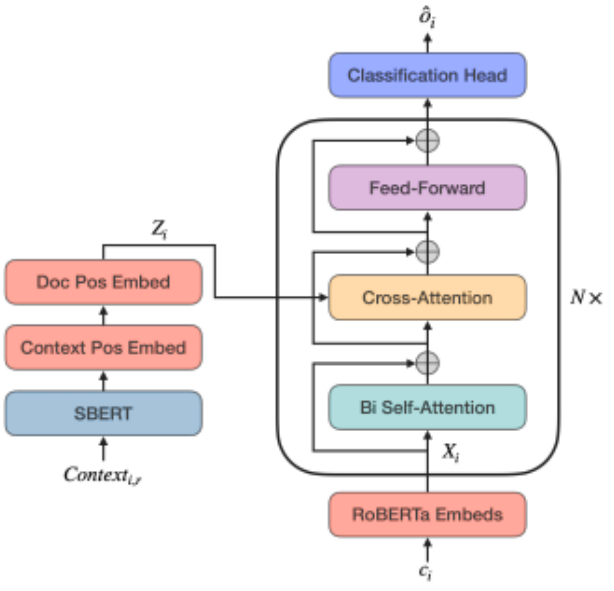
- layers initialised with weights from a context-independent classifier

Internal structure

- ***Token level*** encoder for c_i

Context

- fixed window of Sentence level embedding (SBERT) for ***surrounding sentences***
- The left context is ***dynamically*** updated with previously simplified sentences



Planning Model

RoBERTa classifier with cross-attention over the context

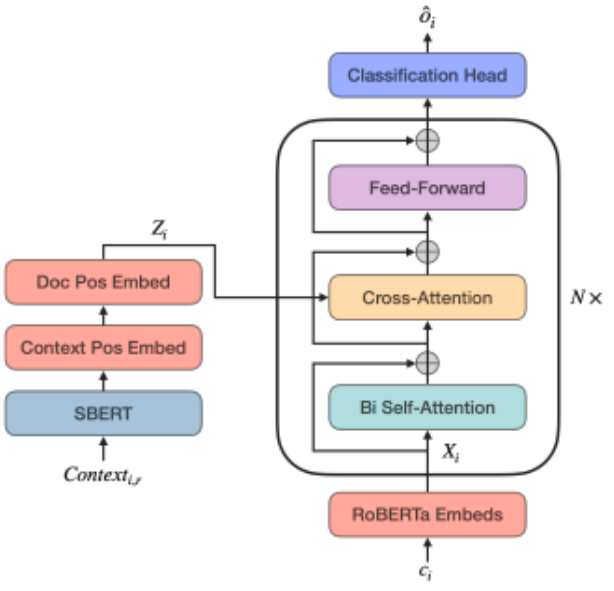
- layers initialised with weights from a context-independent classifier

Internal structure

- ***Token level*** encoder for c_i

Context

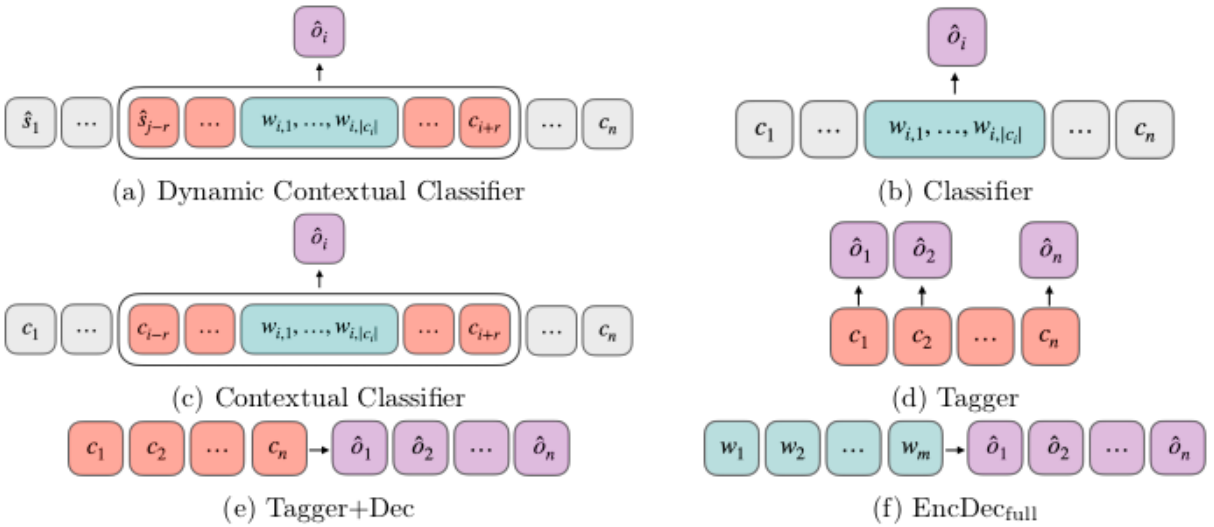
- fixed window of Sentence level embedding (SBERT) for ***surrounding sentences***
- The left context is ***dynamically*** updated with previously simplified sentences



Context positional embedding: relative distance of a given sentence from the input sentence c_i

Document positional embedding: the document quintile (1-5) that a given sentence falls into

Alternative Models

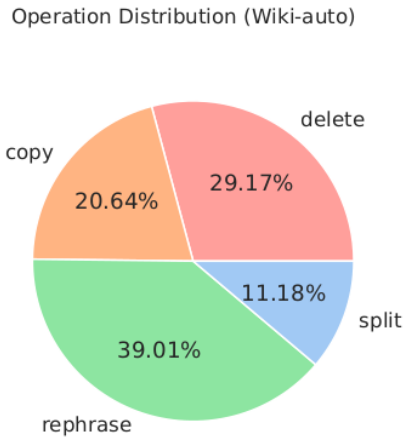


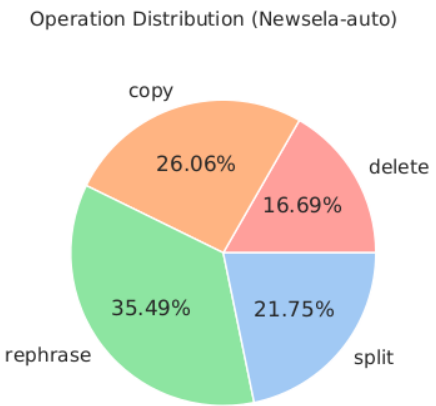
Dynamic Contextual Classifier: our model
Contextual Classifier: Static left context
Classifier: no context
Tagger: Sequence tagging on SBERT representations (no internal structure)

Tagger-Decoder: Each prediction is conditioned on the input document and on the previously predicted operation tags. SBERT encodings.
EncDec$_{full}$: Same as Tagger-Decoder but with token encodings

Data

(C, S) pairs with C a complex document and S its simplification (sentences are aligned)





	Wiki-auto	Newsela-auto
# Doc Pairs	85,123	18,319
# Sent Pairs	461,852	707,776
Avg. $\|C\|$	155.51	868.98
Avg. $\|S\|$	97.72	674.94
Avg. $\|c_i\|$	28.64	22.49
Avg. $\|s_i\|$	21.57	15.84
Avg. n	5.43	38.64
Avg. k	4.53	42.60

- n: the number of sentences in C
- k: the number of sentences in S

Labeling the data

$(C, S) \rightarrow (C, S, o)$

Delete

- c_i is not aligned to any s_j .
  The complex sentence c_i is not aligned to any sentence s_j in the simplified version.

Labeling the data

$(C, S) \rightarrow (C, S, o)$

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.
  The complex sentence c_i is aligned to a similar sentence s_j in the simplified version

Labeling the data

$(C, S) \rightarrow (C, S, o)$

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.
  The complex sentence c_i is aligned to a sentence s_j in the simplified version but differs from it.

Labeling the data

$(C, S) \rightarrow (C, S, o)$

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
  The complex sentence c_i is aligned to several sentences in the simplified version.

Planning Accuracy Results

	Wiki-auto						Newsela-auto					
Model	C	R	S	D	Micro	**Macro**	C	R	S	D	Micro	**Macro**
EncDec$_{full}$	26.9	42.2	36.0	51.8	43.2	40.8	26.1	10.8	11.7	9.0	12.2	11.5
EncDec	29.3	54.5	30.0	51.8	47.7	41.4	72.2	73.9	75.9	79.7	75.0	75.4
Tagger	38.6	54.2	31.7	**58.5**	50.6	45.8	71.4	72.7	74.1	78.4	73.7	74.1
Classifier	42.1	52.9	42.6	49.0	48.4	46.7	77.0	75.6	80.0	78.5	77.4	77.8
Dyn. Context	**44.8**	**57.9**	42.4	54.8	**52.8**	**50.0**	79.3	77.3	82.8	81.4	79.7	80.2
+ docpos	43.7	55.4	**43.6**	56.7	52.3	49.9	**80.0**	**78.1**	**83.6**	**82.0**	**80.3**	**80.8**

- Our model consistently shows best results on both datasets.

Planning Accuracy Results

	Wiki-auto						Newsela-auto					
Model	C	R	S	D	Micro	**Macro**	C	R	S	D	Micro	**Macro**
EncDec$_{full}$	26.9	42.2	36.0	51.8	43.2	40.8	26.1	10.8	11.7	9.0	12.2	11.5
EncDec	29.3	54.5	30.0	51.8	47.7	41.4	72.2	73.9	75.9	79.7	75.0	75.4
Tagger	38.6	54.2	31.7	**58.5**	50.6	45.8	71.4	72.7	74.1	78.4	73.7	74.1
Classifier	42.1	52.9	42.6	49.0	48.4	46.7	77.0	75.6	80.0	78.5	77.4	77.8
Dyn. Context	**44.8**	**57.9**	42.4	54.8	**52.8**	**50.0**	79.3	77.3	82.8	81.4	79.7	80.2
+ docpos	43.7	55.4	**43.6**	56.7	52.3	49.9	**80.0**	**78.1**	**83.6**	**82.0**	**80.3**	**80.8**

- Our model consistently shows best results on both datasets.
- The ***context-free classifier under-performs for deletions***, which confirms the intuition that context modeling particularly matters for this operation.

Planning Accuracy Results

	Wiki-auto						Newsela-auto					
Model	C	R	S	D	Micro	**Macro**	C	R	S	D	Micro	**Macro**
EncDec$_{full}$	26.9	42.2	36.0	51.8	43.2	40.8	26.1	10.8	11.7	9.0	12.2	11.5
EncDec	29.3	54.5	30.0	51.8	47.7	41.4	72.2	73.9	75.9	79.7	75.0	75.4
Tagger	38.6	54.2	31.7	**58.5**	50.6	45.8	71.4	72.7	74.1	78.4	73.7	74.1
Classifier	42.1	52.9	42.6	49.0	48.4	46.7	77.0	75.6	80.0	78.5	77.4	77.8
Dyn. Context	**44.8**	**57.9**	42.4	54.8	**52.8**	**50.0**	79.3	77.3	82.8	81.4	79.7	80.2
+ docpos	43.7	55.4	**43.6**	56.7	52.3	49.9	**80.0**	**78.1**	**83.6**	**82.0**	**80.3**	**80.8**

- Our model consistently shows best results on both datasets.
- The ***context-free classifier under-performs for deletions***, which confirms the intuition that context modeling particularly matters for this operation.
- ***EncDec full performs worst*** presumably because the very long input (the whole context is modelled at the token level) challenges the attention mechanism

Planning Accuracy Results

	Wiki-auto						Newsela-auto					
Model	C	R	S	D	Micro	**Macro**	C	R	S	D	Micro	**Macro**
EncDec$_{full}$	26.9	42.2	36.0	51.8	43.2	40.8	26.1	10.8	11.7	9.0	12.2	11.5
EncDec	29.3	54.5	30.0	51.8	47.7	41.4	72.2	73.9	75.9	79.7	75.0	75.4
Tagger	38.6	54.2	31.7	**58.5**	50.6	45.8	71.4	72.7	74.1	78.4	73.7	74.1
Classifier	42.1	52.9	42.6	49.0	48.4	46.7	77.0	75.6	80.0	78.5	77.4	77.8
Dyn. Context	**44.8**	**57.9**	42.4	54.8	**52.8**	**50.0**	79.3	77.3	82.8	81.4	79.7	80.2
+ docpos	43.7	55.4	**43.6**	56.7	52.3	49.9	**80.0**	**78.1**	**83.6**	**82.0**	**80.3**	**80.8**

- Our model consistently shows best results on both datasets.
- The ***context-free classifier under-performs for deletions***, which confirms the intuition that context modeling particularly matters for this operation.
- ***EncDec full performs worst*** presumably because the very long input (the whole context is modelled at the token level) challenges the attention mechanism
- The encoder-decoder and the tagger, which both use a ***sentence level encoding of the complex sentence*** to be classified perform worse than the classifier - this highlights the importance of having a ***token-level modeling of the input sentence*** .

Ablations

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

Plan-Guided Document Simplification

$$c_i, \hat{o}_i \Rightarrow s_i$$

Plan Guided Document Simplification

Predict simplification operations

$$c1, \ldots, c_n \Rightarrow \hat{o}, \ldots, \hat{o}_n$$

Simplify each input sentences using controls

$$c_i, \hat{o}_i \Rightarrow s_i$$

Document Simplification Models

Fine-tuned on ***sentence pairs*** and iteratively applied to each input sentence

- Plan-Guided (PG): ***pipeline***

$$c_i, \hat{o}_i \Rightarrow s_i$$

Document Simplification Models

Fine-tuned on ***sentence pairs*** and iteratively applied to each input sentence

- Plan-Guided (PG): ***pipeline***

  $$c_i, \hat{o}_i \Rightarrow s_i$$

- Sent-BART: ***end-to-end***

  $$c_i \Rightarrow s_i$$

Document Simplification Models

Fine-tuned on ***sentence pairs*** and iteratively applied to each input sentence

- Plan-Guided (PG): ***pipeline***

  $$c_i, \hat{o}_i \Rightarrow s_i$$

- Sent-BART: ***end-to-end***

  $$c_i \Rightarrow s_i$$

Fine-tuned on ***full document pairs***

- Doc-BART

  $$DOC \Rightarrow SIMPLIFIED$$

Document Simplification Models

Fine-tuned on ***sentence pairs*** and iteratively applied to each input sentence

- Plan-Guided (PG): ***pipeline***

  $$c_i, \hat{o}_i \Rightarrow s_i$$

- Sent-BART: ***end-to-end***

  $$c_i \Rightarrow s_i$$

Fine-tuned on ***full document pairs***

- Doc-BART

  $$DOC \Rightarrow SIMPLIFIED$$

Evaluation Metrics

Summarization metrics

- BARTScore (Yuan et al., 2021)
- SMART (Amplayo et al., 2022)

SARI (Xu et al., 2016)

- Most popular simplification metric.
- Computes n-gram edits between input, output, and references.

FKGL (Kincaid et al., 1975)

- Readibility metrics
- Uses surface-level statistics like syllable counts and sentence length.

Results

System	BARTScore ↑				SMART ↑			FKGL ↓	SARI ↑	Length	
	Faith. ($s \rightarrow h$)	P ($r \rightarrow h$)	R ($h \rightarrow r$)	F1	P	R	F1			Tokens	Sents
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
Doc-BART	-2.48	-2.68	-2.76	-2.72	61.9	43.9	50.6	10.01	47.07	600.8	20.7
Sent-BART	**-1.86**	-1.63	-1.56	-1.60	78.9	80.1	79.3	5.03	73.02	666.4	42.6
PG_{Tag}	-1.95	-2.22	-2.18	-2.20	5.07	62.0	62.6	61.6	56.13	657.4	41.8
PG_{EncDec}	-1.94	-2.22	-2.18	-2.20	62.2	62.5	61.6	5.09	56.06	654.2	41.4
PG_{Clf}	-1.91	-1.68	**-1.53**	-1.60	77.8	**81.2**	79.3	**4.95**	73.83	688.8	44.5
PG_{Dyn}	-1.91	**-1.60**	-1.54	**-1.57**	**80.2**	81.0	**80.5**	4.98	**75.00**	667.2	42.6
PG_{Oracle}	-1.93	**-1.39**	**-1.40**	**-1.40**	**85.5**	**85.0**	**85.3**	**4.91**	**80.74**	655.6	42.1

- ***Pipeline*** (PG Dyn) achieves the highest results of all systems.

Results

System	BARTScore ↑				SMART ↑			FKGL ↓	SARI ↑	Length	
	Faith. ($s \rightarrow h$)	P ($r \rightarrow h$)	R ($h \rightarrow r$)	F1	P	R	F1			Tokens	Sents
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
Doc-BART	-2.48	-2.68	-2.76	-2.72	61.9	43.9	50.6	10.01	47.07	600.8	20.7
Sent-BART	**-1.86**	-1.63	-1.56	-1.60	78.9	80.1	79.3	5.03	73.02	666.4	42.6
PG_{Tag}	-1.95	-2.22	-2.18	-2.20	5.07	62.0	62.6	61.6	56.13	657.4	41.8
PG_{EncDec}	-1.94	-2.22	-2.18	-2.20	62.2	62.5	61.6	5.09	56.06	654.2	41.4
PG_{Clf}	-1.91	-1.68	**-1.53**	-1.60	77.8	**81.2**	79.3	**4.95**	73.83	688.8	44.5
PG_{Dyn}	-1.91	**-1.60**	-1.54	**-1.57**	**80.2**	81.0	**80.5**	4.98	**75.00**	667.2	42.6
PG_{Oracle}	-1.93	**-1.39**	**-1.40**	**-1.40**	**85.5**	**85.0**	**85.3**	**4.91**	**80.74**	655.6	42.1

- ***Pipeline*** (PG Dyn) achieves the highest results of all systems.
- ***Improving planning*** (PG Oracle) would substantially increase performance (PG Oracle)

Results

System	BARTScore ↑				SMART ↑			FKGL ↓	SARI ↑	Length	
	Faith. $(s \rightarrow h)$	P $(r \rightarrow h)$	R $(h \rightarrow r)$	F1	P	R	F1			Tokens	Sents
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
Doc-BART	-2.48	-2.68	-2.76	-2.72	61.9	43.9	50.6	10.01	47.07	600.8	20.7
Sent-BART	**-1.86**	-1.63	-1.56	-1.60	78.9	80.1	79.3	5.03	73.02	666.4	42.6
PG_{Tag}	-1.95	-2.22	-2.18	-2.20	5.07	62.0	62.6	61.6	56.13	657.4	41.8
PG_{EncDec}	-1.94	-2.22	-2.18	-2.20	62.2	62.5	61.6	5.09	56.06	654.2	41.4
PG_{Clf}	-1.91	-1.68	**-1.53**	-1.60	77.8	**81.2**	79.3	**4.95**	73.83	688.8	44.5
PG_{Dyn}	-1.91	**-1.60**	-1.54	**-1.57**	**80.2**	81.0	**80.5**	4.98	**75.00**	667.2	42.6
PG_{Oracle}	-1.93	**-1.39**	**-1.40**	**-1.40**	**85.5**	**85.0**	**85.3**	**4.91**	**80.74**	655.6	42.1

- ***Pipeline*** (PG Dyn) achieves the highest results of all systems.
- ***Improving planning*** (PG Oracle) would substantially increase performance (PG Oracle)
- E2E ***simplification of full document*** (Doc-BART) yields poor results

Example output

Complex	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.
Simple	Silvano "Nano" Campeggi (January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. Campeggi was known for his poster designs for "Casablanca", "Singin' in the Rain", and "Breakfast at Tiffany's". Campeggi died on August 29, 2018 in Florence from respiratory failure at the age of 95.
Doc-BART	Silvano "Nano" Campeggi (January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies, including "Casablanca", "Ben-Hur", "Singin' in the Rain", "An American in Paris", "West Side Story", "Exodus", "Breakfast at Tiffany's", "Gigi", and "".
Sent-BART	Silvano "Nano" Campeggi (23 January 1923 – 29 August 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. Campeggi is generally regarded as the most important graphic artist and poster designer in the history of American cinema. Campeggi designed and produced the poster and advertising graphics for over 3000 movies, 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 movies 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.
PG_{Dyn}	Silvano "Nano" Campeggi (January 23, 1923 – August 29, 2018) was an Italian artist. He designed and produced the artwork for the posters of many classic Hollywood movies. Sixty-four of the movies 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.

Context-Aware Document Simplification

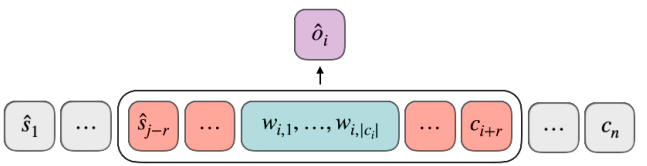
Cripwell et al. Findings of ACL 2023

Context-Aware Simplification

PG (plan-guided) pipeline

First PLAN,
Input D $\Rightarrow$ Simplification Plan
$c1, \ldots, c_n \Rightarrow \hat{o}, \ldots, \hat{o}_n$

PLANNING is Context-Aware ...

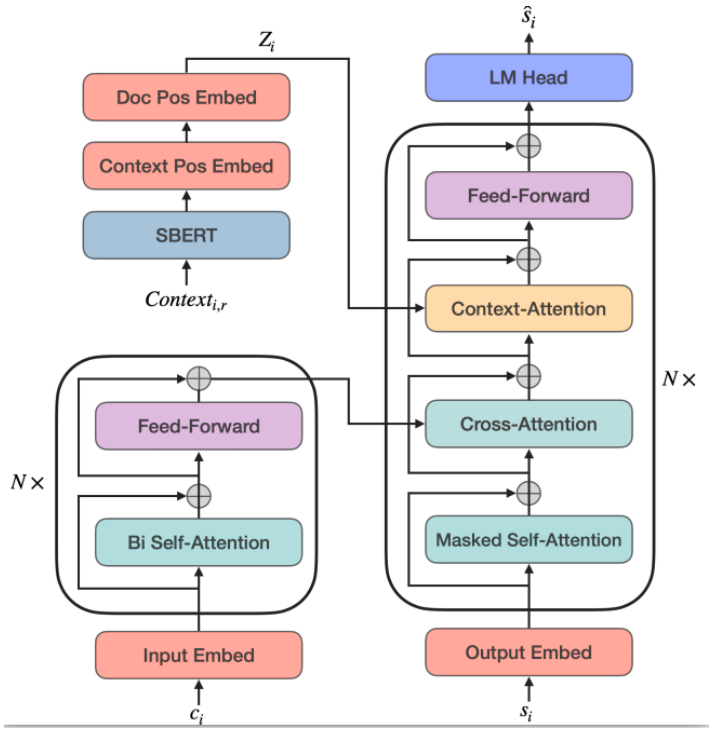


then SIMPLIFY
Input S + Simplification Operation $\Rightarrow$ Simplified S
$c_i, \hat{o}_i \Rightarrow s_i$

... but SIMPLIFICATION is not

Context-Aware BART (ConBART)

- Modification of the BART architecture
- Generation is ***conditioned on both an input sentence*** c_i ***and a representation of the document context*** Z_i of that sentence
- Same ***context modeling*** as for planner (SBERT encoding of the neighbouring sentences)



Contexts and Models

Text-Only Models (BART, LED)

- Input = sentence, paragraph or document
- Models: BART (sentence, paragraph) and LongFormer(document, paragraph)

Contexts and Models

Text-Only Models (BART, LED)

- Input = sentence, paragraph or document
- Models: BART (sentence, paragraph) and LongFormer(document, paragraph)

Contextual Model (ConBART)

- Input: sentence + context window of n sentences (SBART embeddings)
- Model: context-aware modification of BART

Contexts and Models

Text-Only Models (BART, LED)

- BART: input = sentence or paragraph
- LongFormer(LED): inpu = document or paragraph

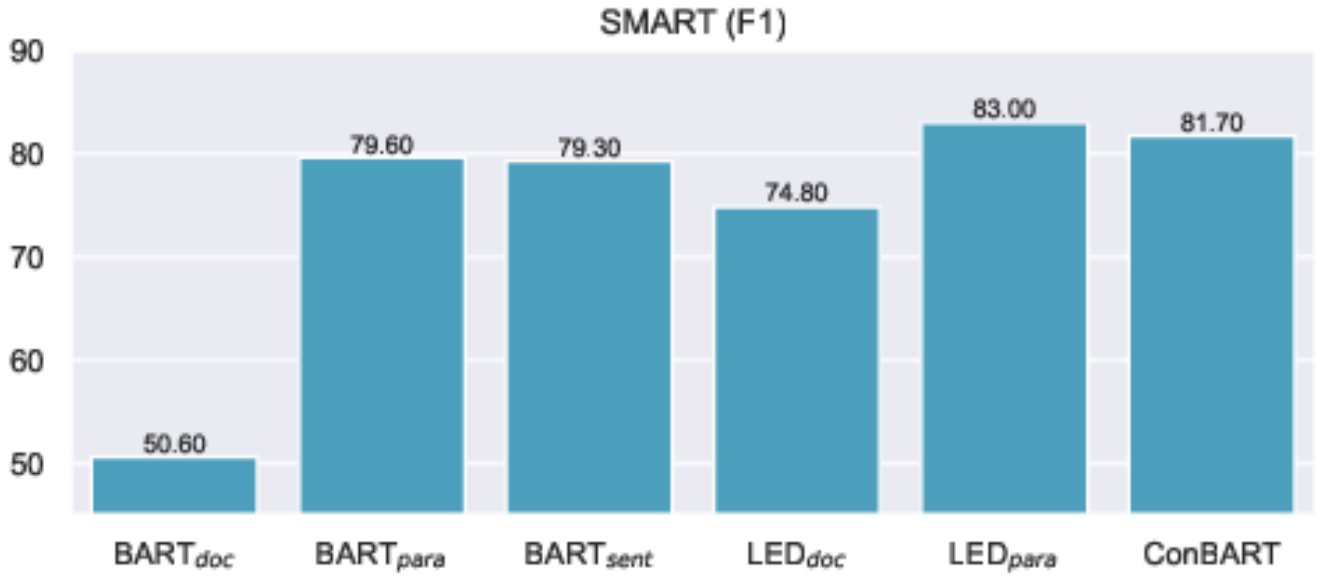
Contextual Model (ConBART)

- Input: sentence + context window of n sentences (SBART embeddings)
- Model: context-aware modification of BART

Plan-Guided Pipelines ($\hat{O} \rightarrow M$)

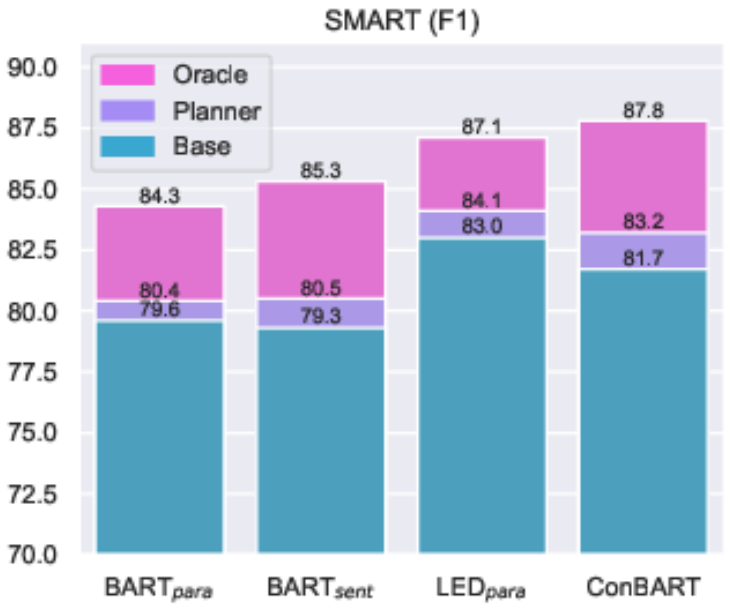
- $\hat{O}$, a predicted simplification plan
- M, a simplification model (BART, LED, ConBART)

Which context helps most ?



- The best two models use a medium size context (ConBART window, LED$_{para}$ paragraphs)
- Full Document context does not work well (BART$_{d}$, LED$_{d}$)
- For longer contexts, LongFormers are necessary (BART$_{X}$ vs. LED$_{X}$)

Does planning help ?



- Planning systematically improves performance
- Planning needs improving: the model simplifying based on the oracle plan has much higher performance

Human Evaluation

- All systems achieve high fluency.
- $\hat{O} \rightarrow LED_{para}$ achieves the highest adequacy.
- $\hat{O} \rightarrow LED_{para}$ and $\hat{O} \rightarrow ConBART$ achieve highest simplicity.

Generalising to OOD Data

- Plan-guidance helps on unseen domains.
- Paragraph-level models are likely less adaptable to unseen domains as paragraph length varies across corpora and models tend to be biased towards paragraph length of training data.

System	Flu	Ad	Simp	**Mean**
LED_{para}	**0.932**	0.632	0.664	0.743
$\hat{O} \rightarrow \text{LED}_{\text{para}}$	0.890	0.684	0.760	0.778
$\hat{O} \rightarrow$ ConBART	0.890	**0.760**	**0.764**	**0.805**

Example Output

Input

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.

Output

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.

Example Output

Input

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.

Output

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.

Example Output

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*.

Example Output

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.

Evaluation

Cripwell et al. EMNLP 2023

Evaluation

- Most popular evaluation metrics require ***multiple high-quality references***
  - something not readily available for simplification
  - makes it difficult to evaluate on unseen domains.

Evaluation

- Most popular evaluation metrics require ***multiple high-quality references***
    - something not readily available for simplification
    - makes it difficult to evaluate on unseen domains.
- Many metrics evaluate simplification quality by ***combining multiple criteria (fluency, adequacy, simplicity)***
    - high scores could be spurious indications of simplicity (Scialom et al.2021)

Evaluation

We propose a new learned evaluation metric (SLE) which focuses on simplicity, outperforming almost all existing metrics in terms of correlation with human judgements.

- Most popular evaluation metrics require multiple high-quality references -- something not readily available for simplification -- which makes it difficult to test performance on unseen domains.
- Furthermore, most existing metrics conflate simplicity with correlated attributes such as fluency or meaning preservation.

Metric	Simplification	Semantic	Ref-less
BLEU	✗	✗	✗
BERTScore	✗	✓	✗
QUESTEVAL	✗	✓	✓
SARI	✓	✗	✗
FKGL	✓	✗	✓
LENS	✓	✓	✗
SLE	✓	✓	✓

SLE - Simplicity Level Estimate

- A ***learned*** metric, trained to estimate the simplicity of a sentence.
- ***Reference less***
- Highly correlated with human judgements of simplicity (Competitive with the best performing ***reference-based metric***)
- Can be used as
    - an absolute measure of simplicity
    - to measure error with respect to a target simplicity level.
    - a ***relative measure of simplicity gain compared to the input***

SLE Model

- Regression Model (high score = high simplicity)

  Sentence $\rightarrow$ *Simplicity Score*

- Trained on Newsela
  1,130 ***documents*** labelled with five discrete reading levels (0-4)

- Trained to model ***sentence level*** simplicity level

Document vs Sentence Level Simplicity

Not all sentences in a document have the same simplicity level

There is some overlap in terms of simplicity level across adjacent levels

$\rightarrow$ Merely training to minimize error with respect to these labels would likely result in ***mode collapse within levels (peaky, low-entropy distribution)*** and strong ***overfitting*** to the Newsela corpus.

$\rightarrow$ To allow the model to better differentiate between sentences from the same reading level, we apply

- Label softening
- Document-level Optimisation

Label Softening

We interpolate regression labels throughout adjacent class regions according to their Flesch-Kincaid grade level (FKGL), a readability metric often used in education as a means to judge the suitability of books for students (high values = high complexity).

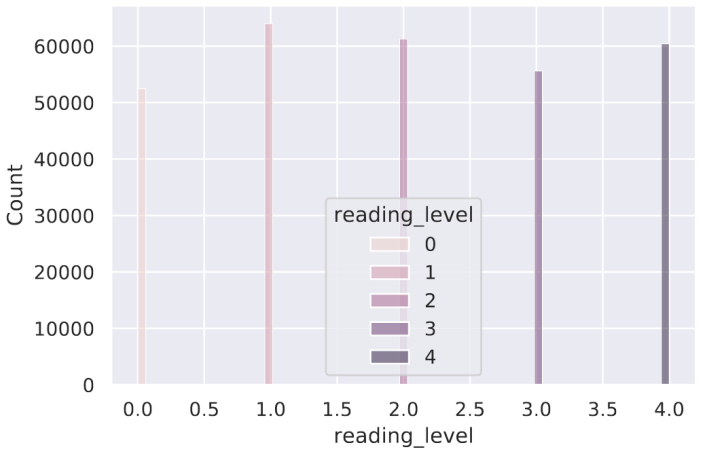
Revised FKGL score (Negative FKGL rescaled)
$f_{L,i}$: the FKGL score of sentence x_i
f_L: the set of negative FKGL scores for sentences belonging to some reading level

$$f'_{L,i} = 2 \cdot \frac{f_{L,i} - \min f_L}{\max f_L - \min f_L}$$

Revised Simplicity score
$\bar{f'_L}$ is the mean of f'_L
$l_{L,i}$ is the reading level for the ith sentence of L

$$l'_{L,i} = f'_{L,i} - \bar{f'_L} + l_{L,i}$$

Before and After Label Softening



(a) Original

(b) Softened

Document-level Optimisation

Motivation

The labels of individual sentences are likely noisy, but approach the document label on average.

Method

- We keep sentences from each given document together
- We perform early stopping with respect to the ***document-level MAE*** (Mean Absolute Error)

Evaluation

SLE scores

- Mean Absolute Error with respect to the original quantized reading levels
- Document-level error when averaging all sentence estimates from a given document (Doc-MAE) - to verify whether SLEs approximate true document-level simplicity labels in aggregate.
- F1 score (Classification) after rounding estimate.

Correlation with human judgments of Simplicity

- Using ΔSLE
- Measures simplicity gain wrt the input

$$\Delta\mathrm{SLE}(\hat{y}) = \mathrm{SLE}(\hat{y}) - \mathrm{SLE}(x)$$

SLE Scores

Model	MAE ↓	Doc-MAE ↓	F1 ↑
$\text{SLE}_{\mathbb{Z}}$ (quantized)	0.825	0.544	0.401
SLE (softened)	0.924	0.448	0.402

- As expected, soft labels worsens MAE with respect to the original reading levels
- Document-level MAE is improved, suggesting that quantized labels lead to more extreme false negatives
- When treated as a classification task both systems show similar performance (F1).

  *SLE is better able to approximate document-level simplicity ratings on average, with little to no drawback at the sentence level

Correlation with Human Judgements

Two datasets of human simplicity ratings

Simplicity-DA

- 600 system outputs, each with 15 ratings and 22 references
- Simplicity hard to assess on non fluent, non adequate output (High correlation between simplicity and fluency/adequacy scores (Pearsons r Fluency: 0.771, adequacy: 0.758))
- We only keep Simplicity-DA outputs with high human fluency and meaning preservation (ratings at least 0.3 std. devs above the mean).

Human-Likert

- 100 human-written sentence simplifications, each with $\sim$60 simplicity ratings and 10 references.

Correlation with Human Judgements

Metric	Human-Likert	Simplicity-DA✓
LENS	**0.531****	**0.429****
SARI	0.395**	0.109
BERTScore	0.389**	0.142
BLEU	0.333**	0.084
ΔSLE	**0.516****	**0.381****
$\Delta\text{SLE}_{\mathbb{Z}}$	0.479**	0.328**
FKGL	0.354**	0.260*
QUESTEVAL	0.134	0.090

ΔSLE outperforms all existing metrics except Lens, but is ***reference less*** and uses a ***smaller network architecture*** than Lens and BERTScore.

On Simplicity-DA\cmark, metrics follow a similar rank order except for certain metrics dropping substantially (SARI, BERTScore, BLEU).

Evaluation (Document Simplification)

Open Challenges for Simplification Evaluation

Trade-off *Meaning Preservation / Conservativity / Simplicity*

- How can we measure all aspects ?

Open Challenges for Simplification Evaluation

Trade-off *Meaning Preservation / Conservativity / Simplicity*

- How can we measure all aspects ?

Out-of Domain Evaluation

- How well does a simplification model generalise to unseen text type ?

Open Challenges for Simplification Evaluation

Trade-off *Meaning Preservation / Simplicity*

- How can we measure each dimension ?

Out-of Domain Evaluation

- How well does a simplification model generalise to unseen text type?

  $\rightarrow$ Reference-less metrics

Test Sets

In domain: Newsela News articles

- 1,130 ***documents*** manually rewritten at five discrete reading levels (0-4)

OOD: Wikipedia

- 1K documents
- at least 10 sentences and 3 paragraphs.
- 19 of the most common semantic types, grouped into 5 broad categories

Models

Trained on Newsela

Text-Only Models

- LED_{para} - paragraph-level Longformer

Plan-Guided Models

- $\hat{O} \rightarrow LED_{\text{para}}$

  Longformer model conditioned on simplification plan

- $\hat{O} \rightarrow$ ConBART

  BART conditioned on simplification plan and document context

ChatGPT (No API), Zero-shot

Evaluating Meaning Preservation

Used for summarisation

- skewed for precision
- we use their recall version

SummaC

- an NLI entailment-based metric
- compute an NLI entailment matrix between each of the M input sentences and N output sentences.
- Score for each output sentence computed by Convolution
- Sentence scores are then averaged.

QAFactEval

- a QA-based metric)
- Questions and correct answers are first generated from the summary
- Answers are predicted from the input document.
- Score = average of these answer overlap scores

Entity Matching between input and output

Evaluating Conservativity

High score for meaning preservation can be obtained by overly conservative models.

Simplifications are slightly shorter than their inputs and often contain more sentences (splitting)

- Average lengths of outputs (no. of tokens and sentences)
- BLEU with respect to the input

Evaluating Simplicity

FKGL

SLE

$$\mathrm{SLE}(Y) = \frac{1}{|Y|} \sum_{i=1}^{|Y|} \mathrm{SLE}(y_i)$$

- Mean of sentences' scores
- Y, document Y
- y_i, the ith sentence of document Y

ϵSLE

$$\epsilon\mathrm{SLE} = \frac{1}{N} \sum_{i=1}^{N} \left| \mathrm{SLE}(\hat{Y}_i) - l_i \right|$$

- Mean absolute error (MAE) between the predicted and target document reading levels.
- Estimates of how much the document simplicity level divergers from the target reading level
- l_i, a target simplicity level

Brief Summary of In-Domain Results

End-to-End, Text Only Models (LED_{para})

- Meaning preserving
- Conservative (high BLEU, long output)
- Low simplicity scores

Plan-Guided models

- Meaning Preservation results not too far from LED_{para}
- Simplify: Length and BLEU close to reference
- Still Conservative; higher faithfulness scores than the references

Brief Summary of Out-Of-Domain Results

Tests Newsela trained Models on Wiki data (no reference)

End-to-End, Text Only Models (LED_{para})

- produces very short texts (different from In-Domain Results), overfit to Newsela text length ?

Plan-Guided models

- have good simplicity and meaning preservation scores

ChatGPT

- generates very short texts

Brief Summary of Human Evaluation

- 250 paragraphs from the test set that contain between 3-6 sentences.
- (complex paragraph, generated simplification)
- Binary rating on Fluency, Meaning Preservation, Simplification
- Final score = proportion of positive ratings

Text Only Models

- underperforms on meaning preservation and simplicity

Plan-Guided Models

- are better overall

Conclusion and Perspectives

Conclusion and Future Work

Planning

- seem to help improve document level simplification and generalising to new domains

Simplification metrics

- there is a need for a metric which correctly captures the tradeoff between meaning preservation and simplification

Types of Simplification

- Here (Newsela): simplification in terms of school level
- What about: expert/layman, disadvantaged users ?

LLMs

- How well do they simplify ?
- Can prompting helps diversifying simplification (generate simplifications for diverse users)?

Questions ?