Neural CRF Model for Sentence Alignment in Text Simplification

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong and Wei Xu





THE OHIO STATE UNIVERSITY
Department of Computer Science and Engineering

Georgia Tech
School of Interactive Computing
College of Computing

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

65% of the eight graders in American public schools in 2017 are not reading proficiently, and the situation is even worse for students enrolled in some urban districts.

Simplify

1) *65% of eight graders in US public schools can't read well.*
2) *The situation is worse in some urban schools.*

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

65% of the eight graders in American public schools ~~in 2017~~ are not reading proficiently, and the situation is ~~even~~ worse for students enrolled in some urban districts.

↓ **Simplify**

1) *65% of eight graders in US public schools can't read well.*
2) *The situation is worse in some urban schools.*

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

65% of the eight graders in American public schools ~~in 2017~~ are not reading proficiently, and the situation is ~~even~~ worse for students enrolled in some urban districts.

Simplify

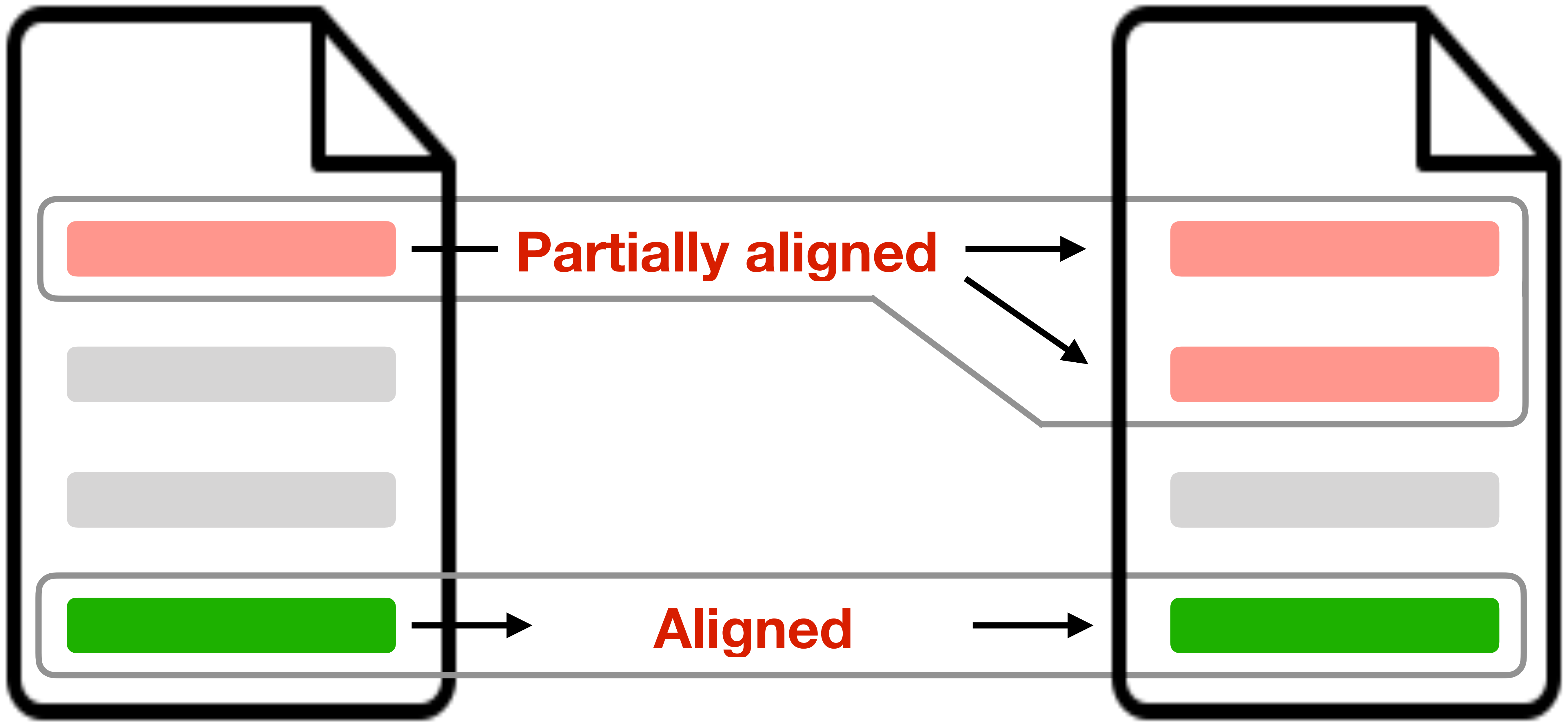
1) *65% of eight graders in US public schools can't read well.*
2) *The situation is worse in some urban schools.*

Involves a broad range of rewrite operations

(splitting, paraphrasing and deletion)

Text Simplification

- Primarily addressed by sequence-to-sequence models.
- **Training corpus** are complex-simple sentence pairs extracted by **aligning parallel articles**.



Complex article

Simple article

Wikipedia The Free Encyclopedia

newsela (Original article)

Simple English Wikipedia

newsela (Simplified article)

Weakness of Previous Work on Sentence Alignment

	Similarity metric	Alignment strategy
JaccardAlign (Xu et al., 2015)	Jaccard similarity	Greedy
MASSAlign (Paetzold et al., 2017)	TF-IDF	Dynamic programming
CATS (Štajner et al., 2018)	Lexical-similarities	Greedy

Weakness #1

Weakness #2

Weakness #1: surface-level similarity metrics, fails to capture paraphrase.

Weakness #2: native alignment strategies, do poorly on sentence splitting.

Our Solution for Sentence Alignment

- Two high-quality manually annotated sentence alignment datasets (20k / 10k sentence pairs).
- Structure prediction + BERT$_{finetune}$ → A neural CRF alignment model.

	aligned + partial vs. others*		
	Precision	Recall	F1

* Results are on the manually annotated Newsela dataset.

Our Solution for Sentence Alignment

- Two high-quality manually annotated sentence alignment datasets (20k / 10k sentence pairs).
- Structure prediction + $\text{BERT}_{finetune}$ → A neural CRF alignment model.

		aligned + partial vs. others*		
		Precision	Recall	F1
Greedy	JaccardAlign (Xu et al., 2015)	98.66	67.58	80.22
Dynamic Programming	MASSAlign (Paetzold et al., 2017)	95.49	82.27	88.39
Greedy	CATS (Štajner et al., 2018)	88.56	91.31	89.92

* Results are on the manually annotated Newsela dataset.

Our Solution for Sentence Alignment

- Two high-quality manually annotated sentence alignment datasets (20k / 10k sentence pairs).
- Structure prediction + $\text{BERT}_{finetune}$ → A neural CRF alignment model.

		aligned + partial vs. others*		
		Precision	Recall	F1
Greedy	JaccardAlign (Xu et al., 2015)	98.66	67.58	80.22
Dynamic Programming	MASSAlign (Paetzold et al., 2017)	95.49	82.27	88.39
Greedy	CATS (Štajner et al., 2018)	88.56	91.31	89.92
Threshold	$\text{BERT}_{finetune}$	94.99	89.62	92.22

* Results are on the manually annotated Newsela dataset.

Our Solution for Sentence Alignment

- Two high-quality manually annotated sentence alignment datasets (20k / 10k sentence pairs).
- Structure prediction + BERT$_{finetune}$ → A neural CRF alignment model.

		aligned + partial vs. others*		
		Precision	Recall	F1
Greedy	JaccardAlign (Xu et al., 2015)	98.66	67.58	80.22
Dynamic Programming	MASSAlign (Paetzold et al., 2017)	95.49	82.27	88.39
Greedy	CATS (Štajner et al., 2018)	88.56	91.31	89.92
Threshold	BERT$_{finetune}$	94.99	89.62	92.22
Threshold	BERT$_{finetune}$ + paragraph alignment	98.05	88.63	93.10

* Results are on the manually annotated Newsela dataset.

Our Solution for Sentence Alignment

- Two high-quality manually annotated sentence alignment datasets (20k / 10k sentence pairs).
- Structure prediction + BERT$_{finetune}$ → A neural CRF alignment model.

		aligned + partial vs. others*		
		Precision	Recall	F1
Greedy	JaccardAlign (Xu et al., 2015)	98.66	67.58	80.22
Dynamic Programming	MASSAlign (Paetzold et al., 2017)	95.49	82.27	88.39
Greedy	CATS (Štajner et al., 2018)	88.56	91.31	89.92
Threshold	BERT$_{finetune}$	94.99	89.62	92.22
Threshold	BERT$_{finetune}$ + paragraph alignment	98.05	88.63	93.10
CRF	Ours CRF aligner	97.86	91.31	95.59

+5.7

* Results are on the manually annotated Newsela dataset.

Our Contribution on Text Simplification

- Two **high-quality** text simplification datasets!
  - Newsela-Auto (666k complex-simple sentence pairs)
  - Wiki-Auto (468k complex-simple sentence pairs)
- Transformer$_{BERT}$ establishes a new **SOTA** on text simplification.

Our Work

Two manually annotated **sentence alignment** datasets
(20k / 10k sentence pairs)

Train / evaluate

Neural CRF **alignment model**
SOTA

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate

Two **text simplification** datasets
Newsela-Auto and Wiki-Auto
(666k / 468k sentence pairs)

Train / evaluate

Seq2Seq generation models
for **text simplification**
SOTA

Sentence Alignment

Text Simplification

Our Work


Two manually annotated
sentence alignment datasets
(20k / 10k sentence pairs)
Train / evaluate
Neural CRF **alignment model**
SOTA
Apply the trained alignment model to the entire
Newsela and Wikipedia corpora to generate
Two **text simplification** datasets
Newsela-Auto and Wiki-Auto
(666k / 468k sentence pairs)
Train / evaluate
Seq2Seq generation models
for **text simplification**
SOTA
Sentence Alignment
Text Simplification

newsela Corpus (Xu et al. 2015)

- Newsela is an U.S. education company based in New York.
- **1932 news articles** rewritten by professional editors for school children.
- Each article is simplified into 4 different readability levels.

- But, only document-aligned.

We manually align sentences for article pairs at adjacent reading levels in 50 article groups (20,343 sentence pairs).

Annotating Sentence Alignment in newsela

Step 1: Align paragraph using CATS* tool kit and manually correct errors.

Step 2: Crowdsource alignment labels for sentence pairs on Figure-Eight

- Classify sentence pairs into *aligned* / *partially aligned* / *not aligned*
- Inter-annotator agreement: 0.807 (Cohen Kappa)

Step 3: Verify the crowdsourcing labels by $\times 4$

We also manually align sentences for Wikipedia, please check our paper!

* CATS: A Tool for Customised Alignment of Text Simplification Corpora, Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, Simone Paolo Ponzetto, LREC 2018.

Crowdsourcing Annotation Interface

Sentence A

Professors from Bard teach the classes.

Sentence B

Professors from nearby Bard College teach the classes

What's the relationship between Sentence A and Sentence B ?

○ **A and B are equivalent**

- A and B are equivalent (convey the same meaning, though one sentence can be much shorter or simpler than the other sentence)

○ **A , B are partially overlapped**

- A and B are partially overlap (share information in common, while some important information differs/missing).

○ **A and B are mismatched**

- The two sentences are completely dissimilar in meaning.

Comments (Optional)

If you have any comment about this HIT, please type it here

Our Work

Two manually annotated **sentence alignment** datasets
(20k / 10k sentence pairs)

Train / evaluate

Neural CRF **alignment model**

SOTA

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate

Two **text simplification** datasets
Newsela-Auto and Wiki-Auto
(666k / 468k sentence pairs)

Train / evaluate

Seq2Seq generation models
for **text simplification**

SOTA

Sentence Alignment

Text Simplification

Neural CRF Alignment Model

Step 1: Paragraph alignment algorithm

- Based on sentence similarity and vicinity information.
- Significantly improve alignment accuracy (+3 points in precision)

Step 2: Sentence alignment model

Algorithm 1: Pairwise Paragraph Similarity

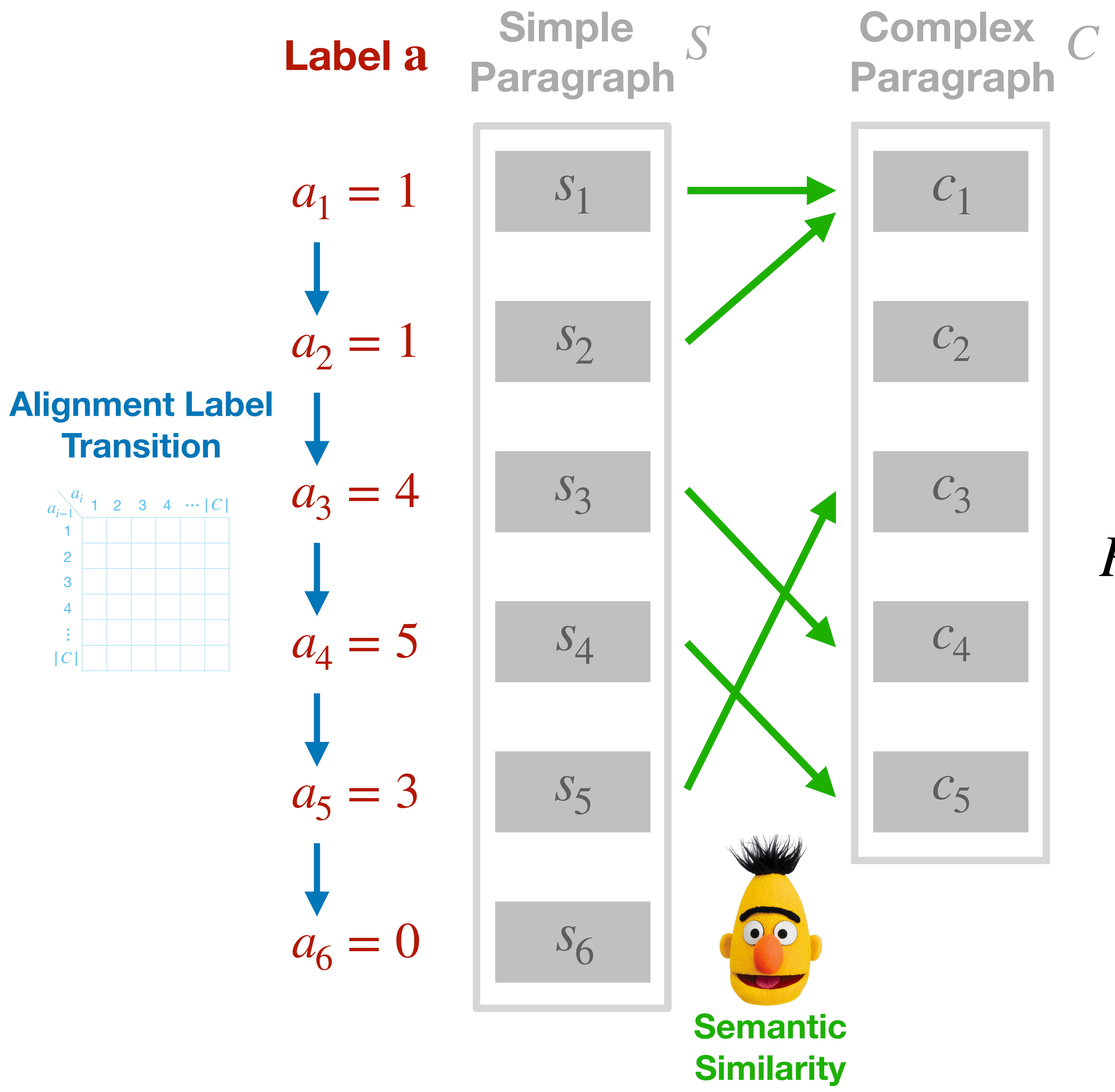
Initialize: $simP \in \mathbb{R}^{2\times k\times l}$ to $0^{2\times k\times l}$
for $i \leftarrow 1$ **to** k **do**
  for $j \leftarrow 1$ **to** l **do**
    $simP[1,i,j] = \underset{s_p \in S_i}{\text{avg}} \left(\max_{c_q \in C_j} simSent(s_p, c_q) \right)$
    $simP[2,i,j] = \max_{s_p \in S_i, c_q \in C_j} simSent(s_p, c_q)$
  end
end
return $simP$

Algorithm 2: Paragraph Alignment Algorithm

Input : $simP \in \mathbb{R}^{2\times k\times l}$
Initialize: $alignP \in \mathbb{I}^{k\times l}$ to $0^{k\times l}$
for $i \leftarrow 1$ **to** k **do**
  $j_{max} = \underset{j}{\text{argmax}}\, simP[1,i,j]$
  if $simP[1,i,j_{max}] > \tau_1$ *and* $d(i,j_{max}) < \tau_2$ **then**
    $alignP[i,j_{max}] = 1$
  end
  for $j \leftarrow 1$ **to** l **do**
    if $simP[2,i,j] > \tau_3$ **then**
      $alignP[i,j] = 1$
    end
    if $j > 1$ & $simP[2,i,j] > \tau_4$ & $simP[2,i,j-1] > \tau_4$ & $d(i,j) < \tau_5$ & $d(i,j-1) < \tau_5$ **then**
      $alignP[i,j] = 1$
      $alignP[i,j-1] = 1$
    end
  end
end
return $alignP$

Screenshots of paragraph alignment algorithm

Neural CRF Sentence Alignment Model



Semantic Similarity

Alignment Label Transition

$$\Psi(\mathbf{a}, S, C) = \sum_{i=0}^{|S|} sim(s_i, c_{a_i}) + T(a_i, a_{i-1})$$

$$P(\mathbf{a} \mid S, C) = \frac{\exp(\Psi(\mathbf{a}, S, C))}{\sum_{\mathbf{a} \in \mathbf{A}} \exp(\Psi(\mathbf{a}, S, C))}$$

Linear-chain CRF

all possible alignments (dynamic programming)

Evaluation on Sentence Alignment*

- 50 manually annotated article groups (0.5 million sentence pairs) in Newsela.
- 35 train / 5 dev / 10 test, evaluate on article pairs at adjacent readability level.

		aligned + partial vs. others		
		Precision	Recall	F1
Greedy	JaccardAlign (Xu et al., 2015)	98.66	67.58	80.22
Dynamic Programming	MASSAlign (Paetzold et al., 2017)	95.49	82.27	88.39
Greedy	CATS (Štajner et al., 2018)	88.56	91.31	89.92
Threshold	$\text{BERT}_{\textit{finetune}}$	94.99	89.62	92.22
Threshold	$\text{BERT}_{\textit{finetune}}$ + paragraph alignment	98.05	88.63	93.10
CRF	Ours CRF aligner	97.86	91.31	95.59

+5.7

* See our paper for full evaluation on two classification tasks and two new datasets.

Our Work

Two manually annotated **sentence alignment** datasets (20k / 10k sentence pairs)

Train / evaluate

Neural CRF **alignment model**

SOTA

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate

Two **text simplification** datasets Newsela-Auto and Wiki-Auto (666k / 468k sentence pairs)

Train / evaluate

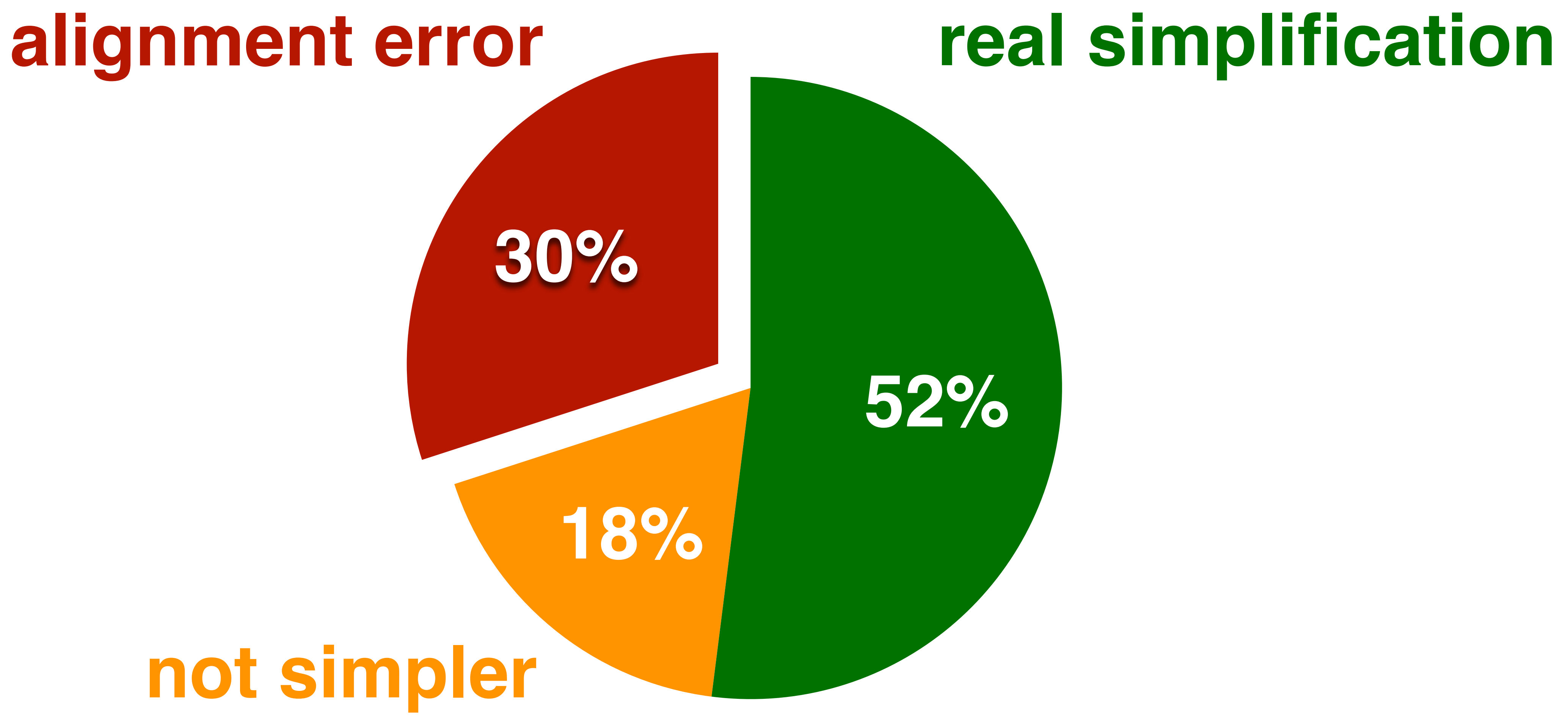
Seq2Seq generation models for **text simplification**

SOTA

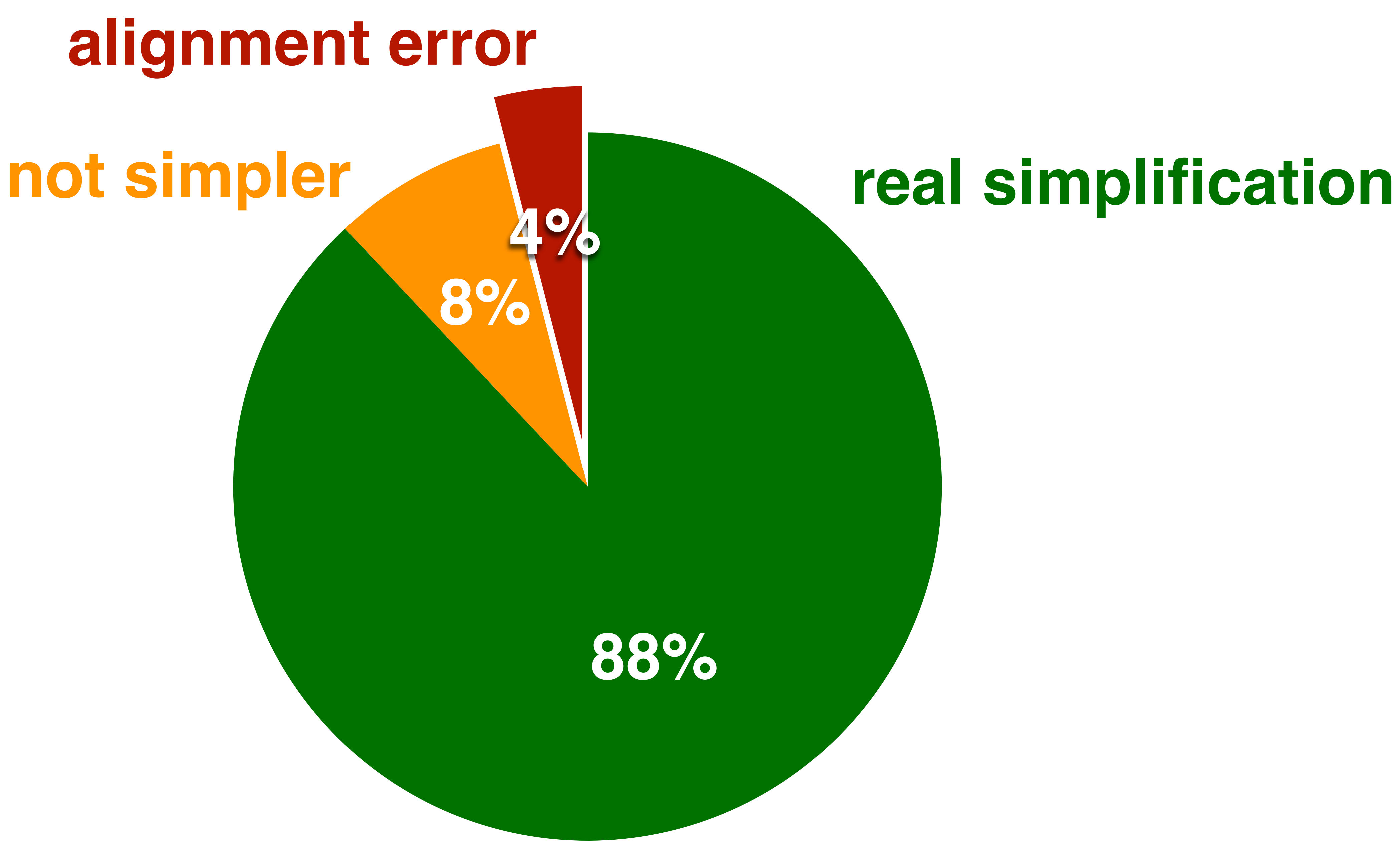
Sentence Alignment

Text Simplification

New Corpora Contain Way Fewer Errors*



Wiki-Large
(Zhang and Lapata, 2017)

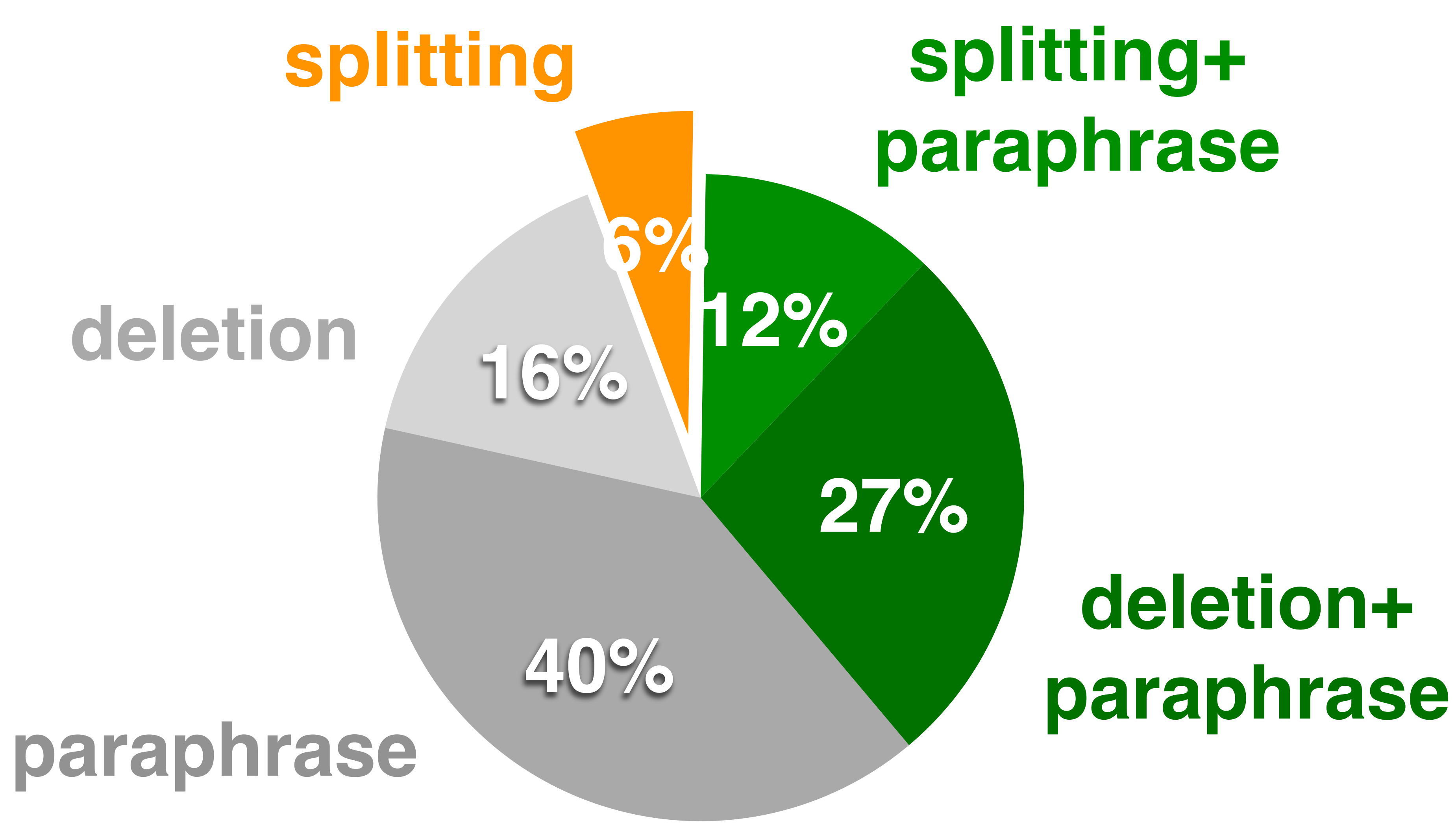


Wiki-Auto (this work)
1.6 times larger

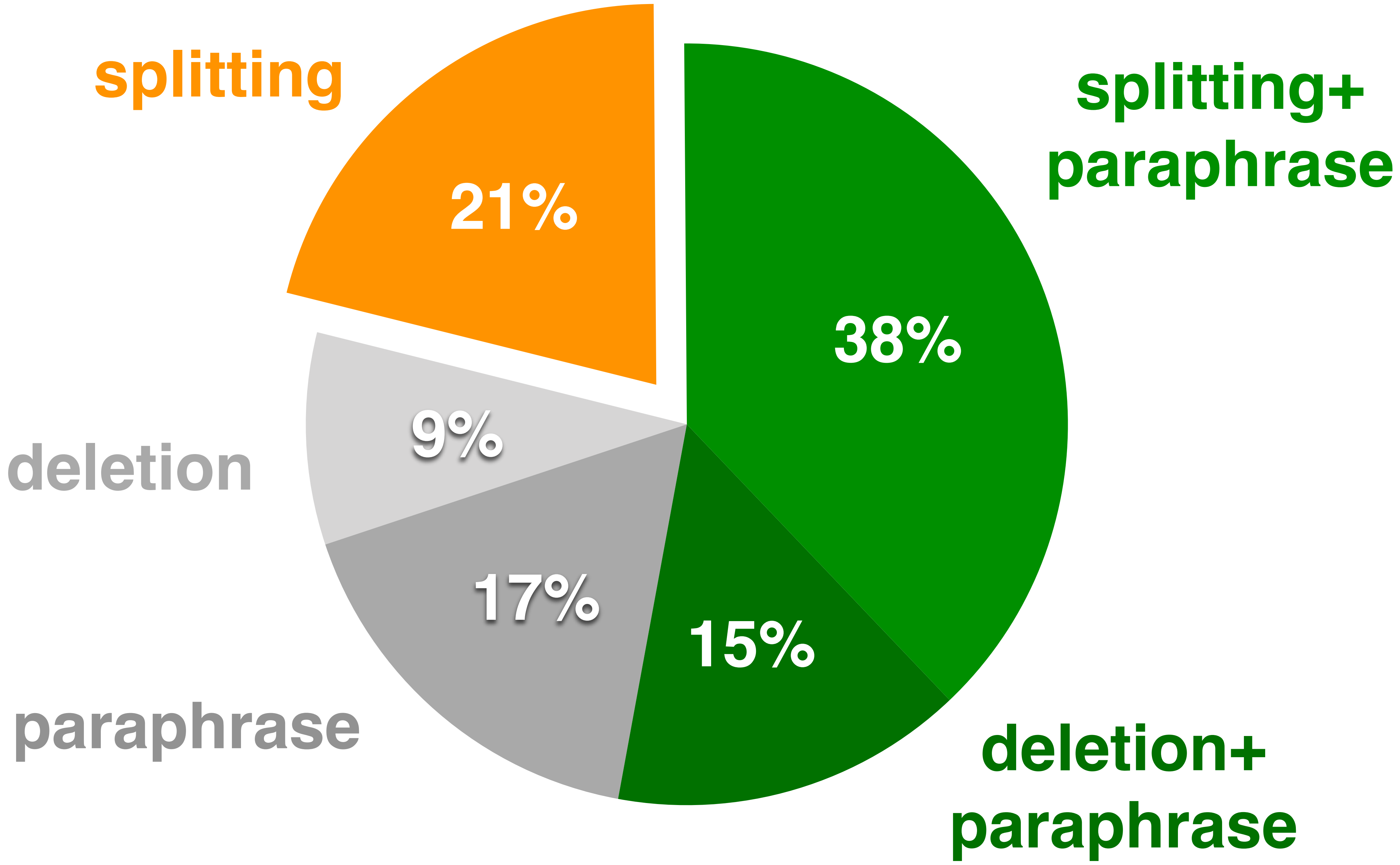
Wiki-Auto has 75% less defective pairs (alignment error + not simpler).

* Based on manual inspection on 100 random sampled sentences from each dataset.

New Corpora Contain More High-quality Simplification*



Newsela
(Xu et al., 2015)



Newsela-Auto (this work)
4.7 times larger

Newsela-Auto has much more splitting and complex re-writes.

* Based on manual inspection on 100 random sampled sentences from each dataset.

Our Work

Two manually annotated **sentence alignment** datasets
(20k / 10k sentence pairs)

Train / evaluate

Neural CRF **alignment model**
SOTA

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate

Two **text simplification** datasets
Newsela-Auto and Wiki-Auto
(666k / 468k sentence pairs)

Train / evaluate

Seq2Seq generation models
for **text simplification**
SOTA

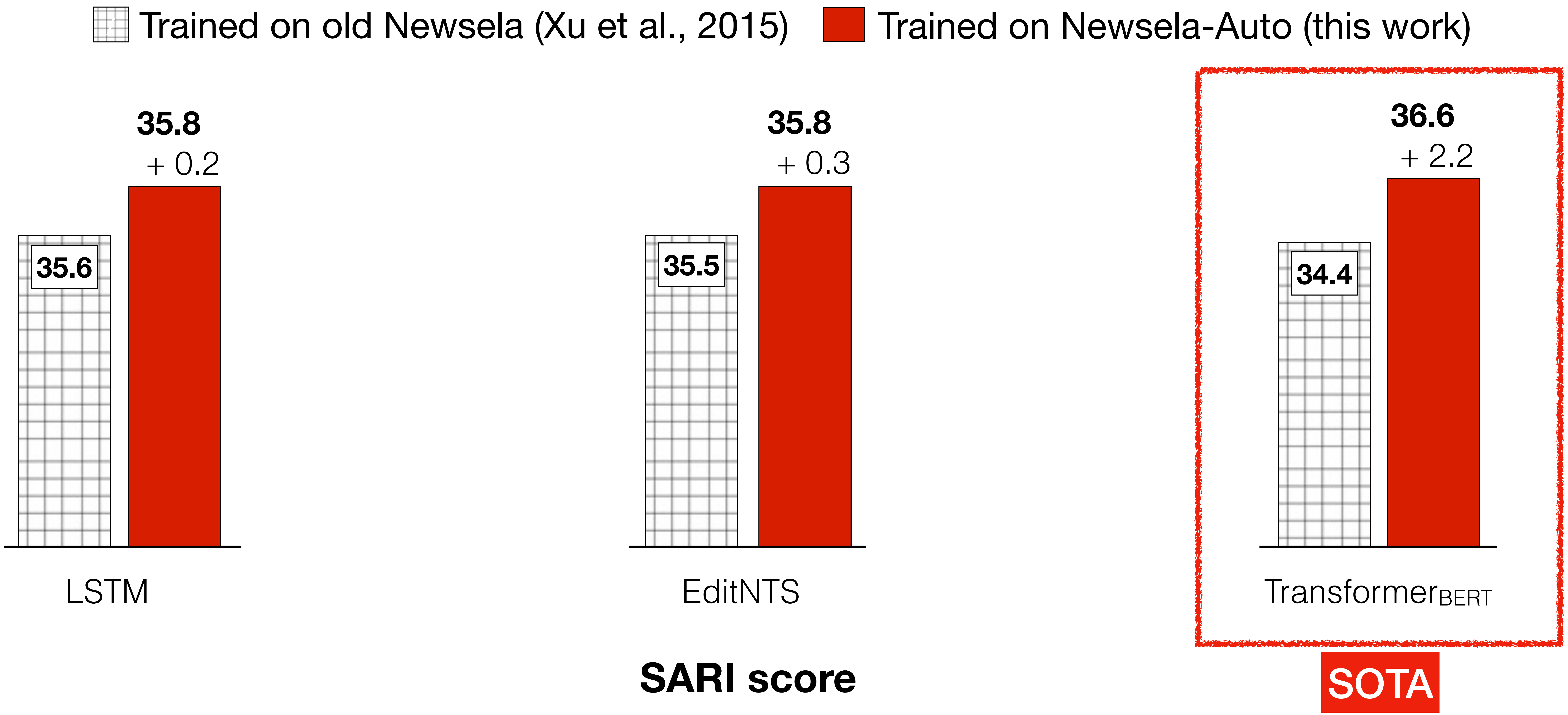
Sentence Alignment

Text Simplification

Experiments on Text Simplification

- Transformer$_{BERT}$ (Rothe et al., 2020)
- Baseline models
    - LSTM
    - EditNTS (Dong et al., 2019)
    - Rerank (Kriz et al., 2019)
- Datasets
    - This work: Newsela-Auto and Wiki-Auto
    - Old: Newsela (Xu et al., 2015) and Wiki-Large (Zhang and Lapata, 2017)

Automatic Evaluation on Text Simplification*

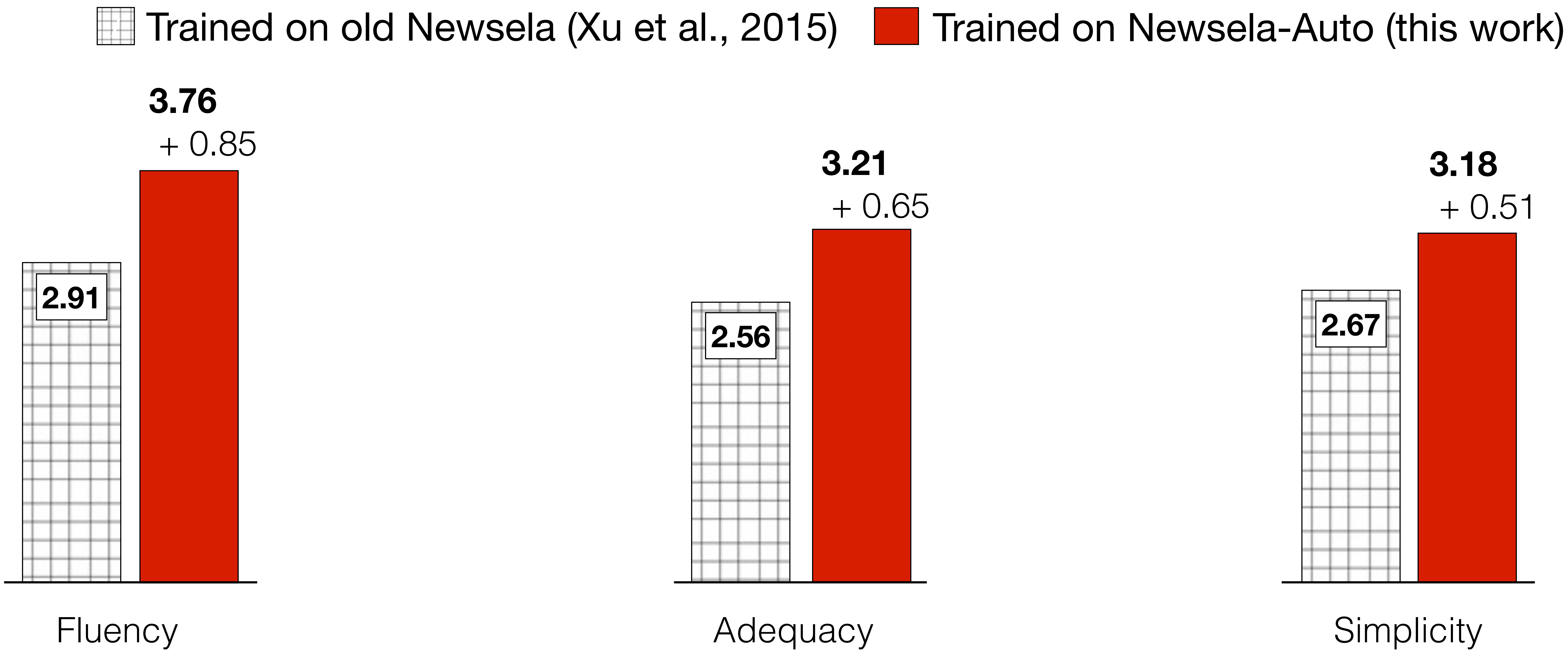


SARI score

Main evaluation metric for text simplification (Xu et al., 2016)

* Evaluate on the Newsela-Auto (this work) test set.

Human Evaluation on Text Simplification*

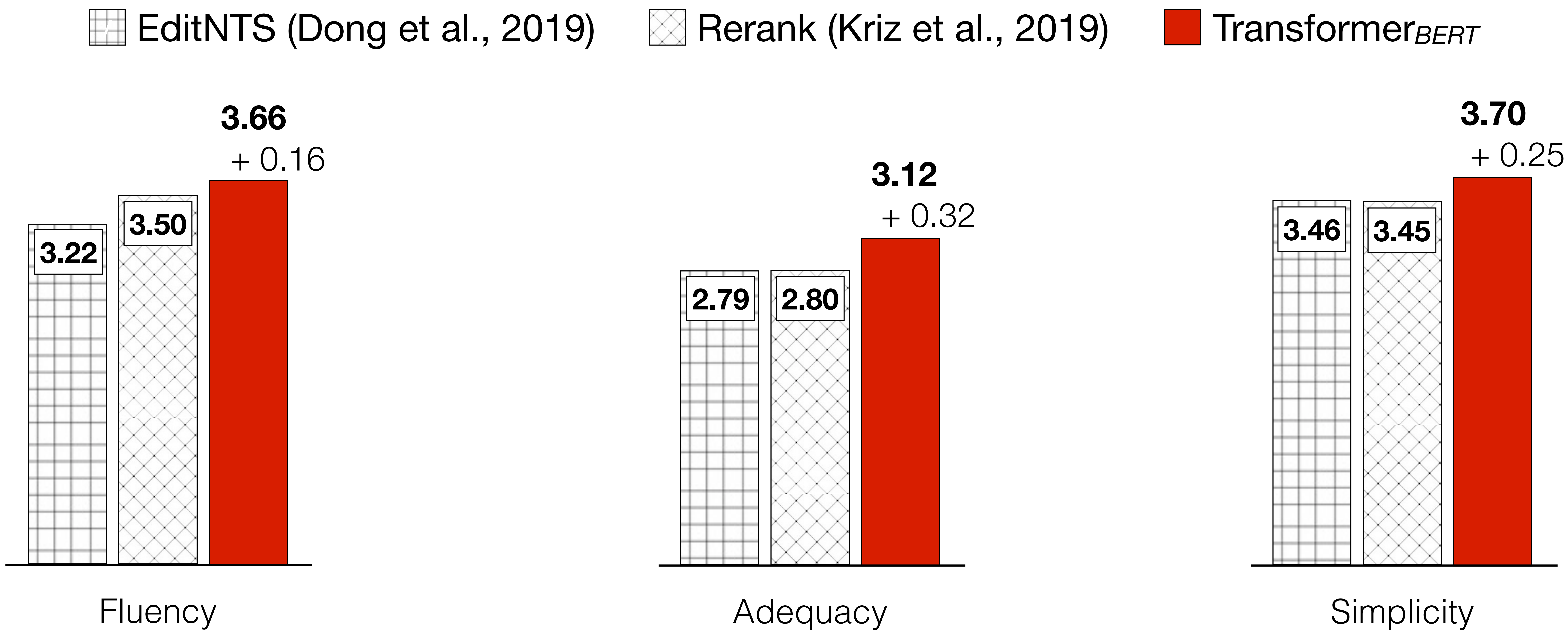


Transformer$_{BERT}$ model

(In 5-point Likert scale)

* Evaluate on the Newsela-Auto (this work) test set.

Human Evaluation on Text Simplification*



Transformer$_{BERT}$ trained on Newsela-Auto dataset is new SOTA in human evaluation.

See our paper for auto and human evaluation on the Wiki-Auto dataset.

* Evaluate on the Old Newsela (Xu et al., 2015) test set.

Takeaways

- Two **high-quality** text simplification datasets!
    - Newsela-Auto (666k complex-simple sentence pairs)
    - Wiki-Auto (468k complex-simple sentence pairs)
- PyTorch code for our **text simplification models** is also available!
- Check the code/data at https://github.com/chaojiang06/wiki-auto
- Contact: Chao Jiang (jiang.1530@osu.edu)

Backup Slides

Crowdsourcing Annotation Interface

Sentence A

Professors from Bard teach the classes.

Sentence B

Professors from nearby Bard College teach the classes

What's the relationship between Sentence A and Sentence B ?

○ **A and B are equivalent**

- A and B are equivalent (convey the same meaning, though one sentence can be much shorter or simpler than the other sentence)

○ **A , B are partially overlapped**

- A and B are partially overlap (share information in common, while some important information differs/missing).

○ **A and B are mismatched**

- The two sentences are completely dissimilar in meaning.

Comments (Optional)

If you have any comment about this HIT, please type it here