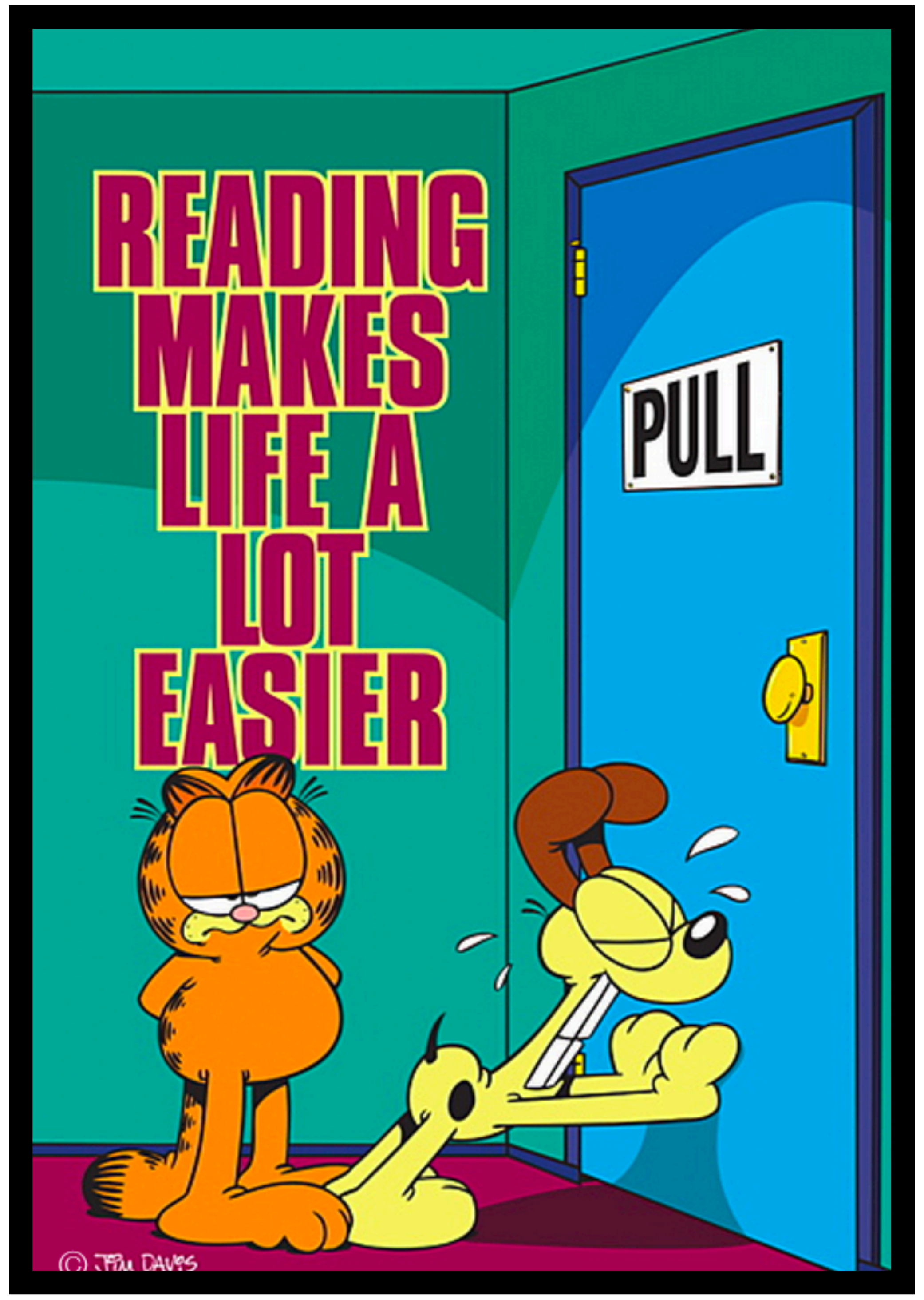


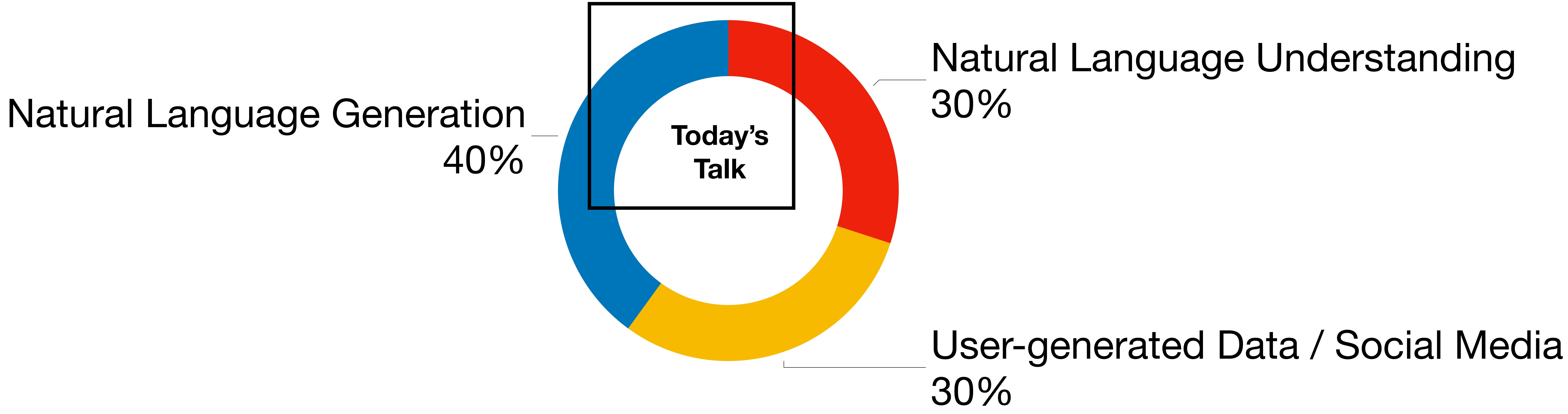

(Image Source: Garfield)

# Automatic Text Simplification

Wei Xu
School of Interactive Computing
Georgia Institute of Technology
@cocoweixu @cocoxu

Georgia Tech

# My Research


Natural Language Generation
40%
Today's
Talk
Natural Language Understanding
30%
User-generated Data / Social Media
30%

# Text Simplification

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

**Science**

## Preserved on ancient teeth, a fossilized microbial world

By Deborah Netburn, Los Angeles Times
Published: 03/05/2014 Word Count: 682

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive. And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Throughout most of the history of archaeology, researchers have considered calcified plaque disposable -- often removing it from skeletons in the process of cleaning them.

# Text Simplification

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

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive. And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Throughout most of the history of archaeology, researchers have considered calcified plaque disposable -- often removing it from skeletons in the process of cleaning them.

# Text Simplification

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

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive.

And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Throughout most of the history of archaeology, researchers have considered calcified plaque disposable -- often removing it from skeletons in the process of cleaning them.

# Text Simplification

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

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive.

**split**

The buildup of plaque can trap the bacteria that live in our mouths.

It turns them into tiny fossils.

And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

**paraphrase**

Even after death, these micro-fossils don’t break down.

Throughout most of the history of archaeology, researchers have considered calcified plaque disposable -- often removing it from skeletons in the process of cleaning them.

**deletion**

# Text Simplification

It involves a complex combination of rewrite operations, e.g., deletion, paraphrasing, reordering.

The layers of ~~calcified~~ plaque entomb the bacteria that also ~~live~~ in our mouths -- turning them into small fossils ~~even when we are alive.~~

split

The buildup of plaque can trap the bacteria that live in our mouths.

It turns them into tiny fossils.

# Text Simplification

It involves a complex combination of rewrite operations, e.g., deletion, paraphrasing, reordering.

The layers of ~~calcified~~ plaque entomb the bacteria that also ~~live~~ in our mouths -- turning them into small fossils ~~even when we are alive.~~

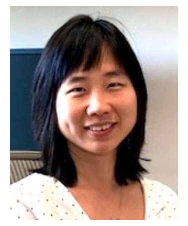
split

The buildup of plaque can trap the bacteria that live in our mouths.

It turns them into tiny fossils.

- **Learn more about how professional editors do this task:**
  - "Discourse Level Factors for Sentence Deletion in Text Simplification" (AAAI 2020)



  - "Text Simplification for Language Learners: A Corpus Analysis" (Petersen & Ostendorf, 2007)
- **Other related work:**
  - "Annotation and Classification of Sentence-level Revision Improvement" (Afrin & Litman, 2018)

# Human Text Simplification

Professional editors rewrite news articles into 4 different readability levels for grade 3-12 students.



Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” in TACL (2015)

# Why Text Simplification?

It can help a lot of people!

- Children (Leonardo et al., 2018)
- Second language learners (Housel et al., 2020)
- Deaf and hard-of-hearing students (Alonzo et al., 2020)
- People with dyslexia (Rello at al. 2013)
- People with autism spectrum disorder (González-Navarro et al., 2014)

using Newsela

our collaborators at RIT

- and many others …

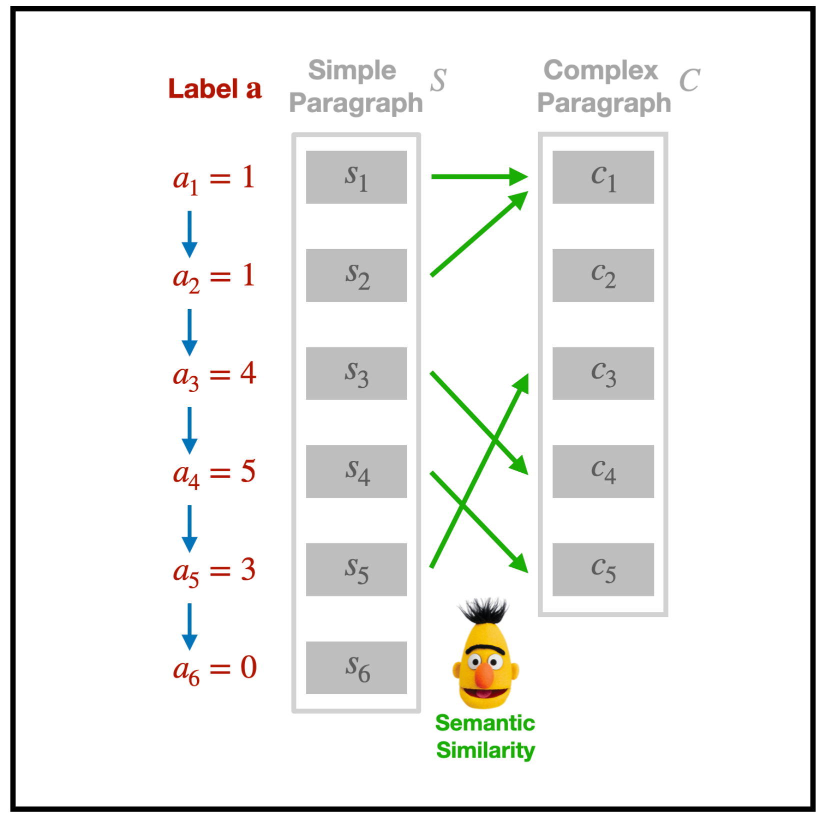
# Today’s Talk

**How to create machine learning models to mimic the professional editing process?**

### SARI Metric & Turk Corpus

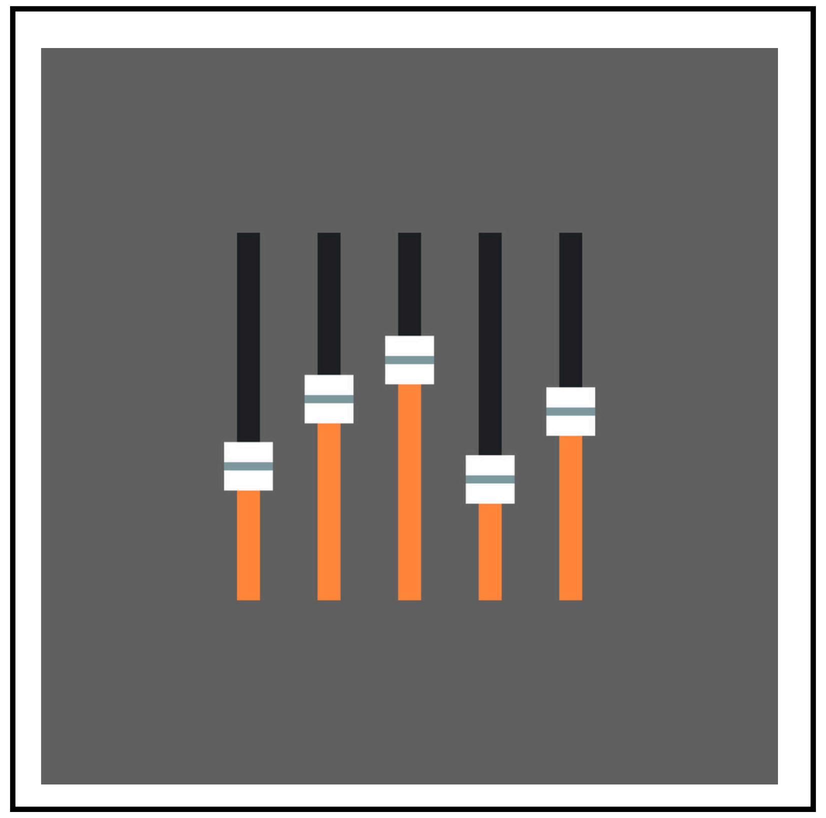
(TACL 2015 & 2016)

### Neural CRF Sentence Aligner



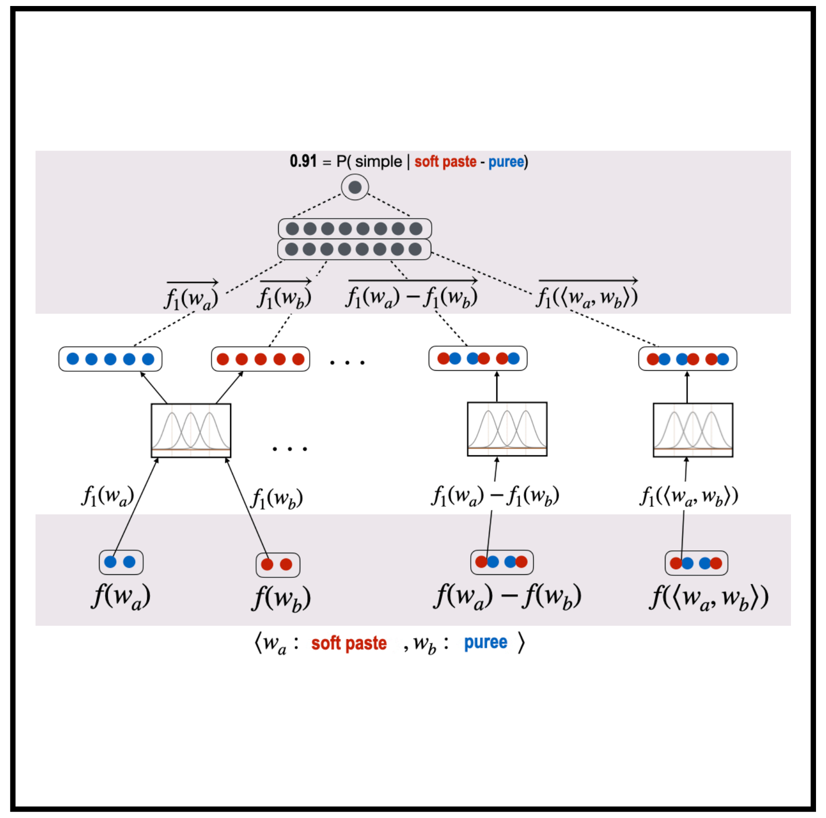

(ACL 2020)

### Controllable Text Generation



(new work)

### Neural Readability Ranking



(EMNLP 2018)

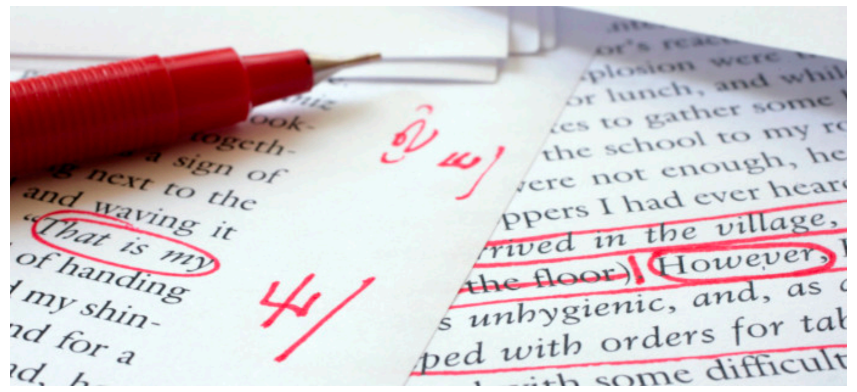
# Part 0 — Preliminary

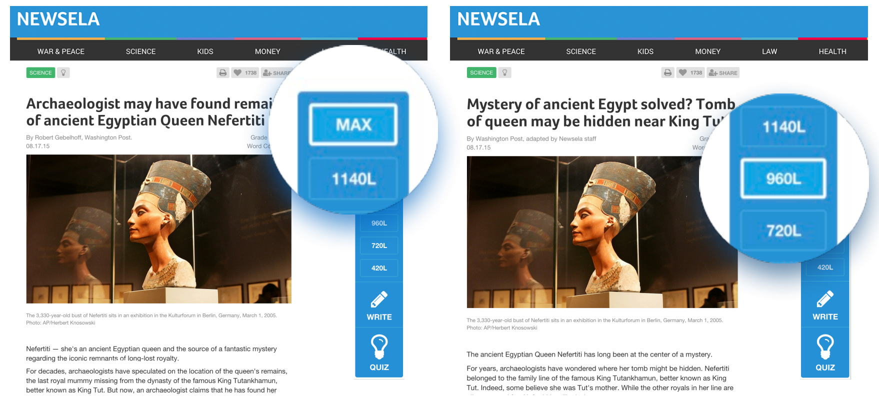
**Problems in Current Text Simplification Research: New Data Can Help**
Xu et al. (TACL 2015)

**Optimizing Statistical Machine Translation for Simplification**
Xu et al. (TACL 2016)

# Automatic Text Simplification

It is a great benchmark for natural language generation (NLG) models.







**complicated rewriting**

(covers other text-to-text tasks: compression, style transfer, summarization, etc. )

**limited training data**

(compares to machine translation)

**for social good!**

(helps children, people w/ disability, legal & medical documents, etc.)

# Automatic Text Simplification

A brief history …

**rule-based methods**

1997 Chandrasekar & Srinivas
1999 Dras (PhD thesis)
2000 Carroll, Minnen, Pearce, Canning, Devlin
2002 Canning (PhD thesis)
2004 Siddharthan (PhD thesis)

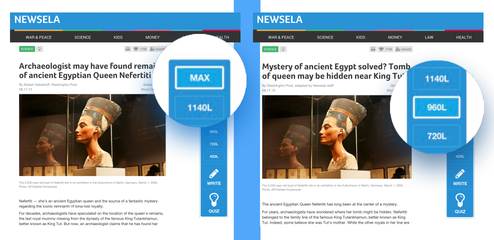
Simple English Wikipedia

**2010 Zhu, Bernhard, Gurevych**

**statistical machine translation**

2011 Woodsend & Lapata
2011 Coster & Kauchak
2012 Wubben, van den Bosch, Krahmer
2014 Narayan & Gardent
2014 Siddharthan (Survey)
2014 Angrosh, Nomoto, Siddharthan
2014 Narayan (PhD thesis)

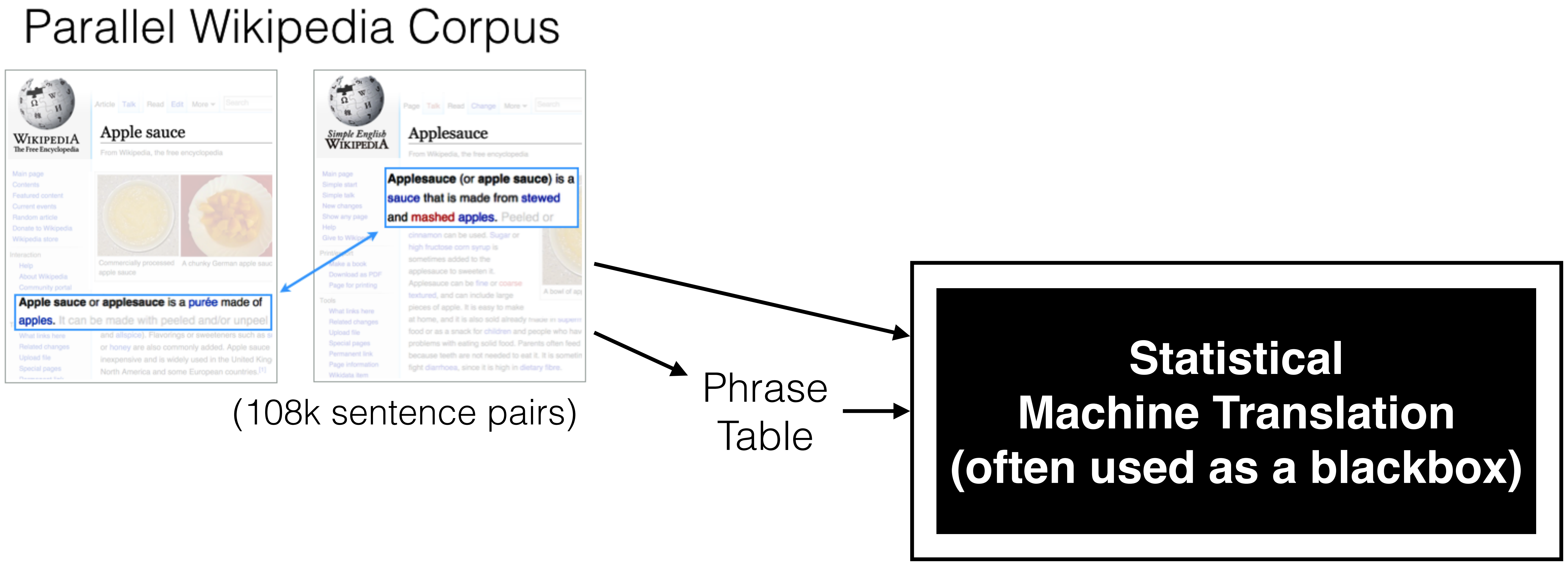
newsela®



**2015 Xu, Callison-Burch, Napoles**
"Problems in Current Text Simplification Research: New Data Can Help" (TACL 2015)

**2016 Xu, Napoles, Pavlick, Chen, Callison-Burch**
"Optimizing Statistical Machine Translation for Simplification" (TACL 2016)

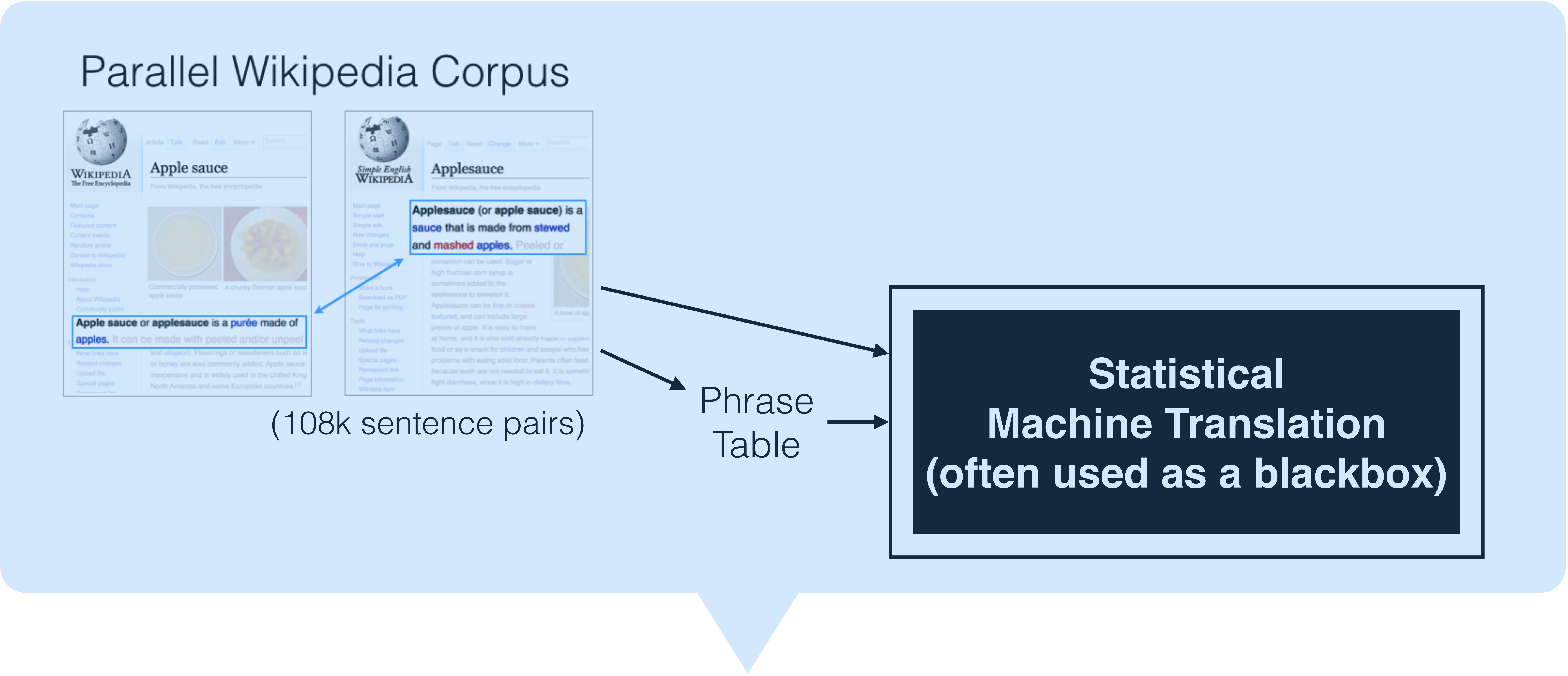
# Text Simplification — previous work




**This is the standard approach since 2010.**

Wei Xu, Chris Callison-Burch, Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help" in TACL (2015)

# Text Simplification — previous work




**I showed that this setup is suboptimal, and how to fix it!**

Wei Xu, Chris Callison-Burch, Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help" in TACL (2015)

# SARI Metric & Turk Corpus

It made possible to train/test state-of-the-art statistical & neural generation models.

**Large-scale Paraphrases**
(lexical, phrasal, syntactic)

**Tuning Data**
(crowdsourced multi-references)

amazon mechanical turk™
Artificial Artificial Intelligence

**Learning Objective**

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \bar{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \bar{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(R \cap \bar{I})}$$

**Pairwise Ranking Optimization**

$$\begin{aligned} g(i,j) > g(i,j') &\Leftrightarrow h_{\mathbf{w}}(i,j) > h_{\mathbf{w}}(i,j') \\ &\Leftrightarrow h_{\mathbf{w}}(i,j) - h_{\mathbf{w}}(i,j') > 0 \\ &\Leftrightarrow \mathbf{w} \cdot \mathbf{x}(i,j) - \mathbf{w} \cdot \mathbf{x}(i,j') > 0 \\ &\Leftrightarrow \mathbf{w} \cdot (\mathbf{x}(i,j) - \mathbf{x}(i,j')) > 0 \end{aligned}$$

**Feature Functions**
(readability, language modeling, etc.)

Wei Xu, Courtney Napoles, Ellie Pavlick, Chris Callison-Burch. “Optimizing Statistical Machine Translation for Simplification” in TACL (2016) tensorflow / tensor2tensor

Used by 340 | Watch 444 | Star 9.1k | Fork 2.4k

Code | Issues 481 | Pull requests 3 | Security | Insights

Branch: master tensor2tensor / tensor2tensor / utils / sari_hook.py / Jump to

Find file | Copy path

lukaszkaiser [V2] Make checkpointing work, support multiple targets. f9f63c3 on Feb 7

2 contributors

253 lines (210 sloc) 9.68 KB

Raw | Blame | History

You're using jump to definition to discover and navigate code. Opt out Beta

Learn more or give us feedback

```
# coding=utf-8
# Copyright 2019 The Tensor2Tensor Authors.
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

"""SARI score for evaluating paraphrasing and other text generation models.

The score is introduced in the following paper:

   Optimizing Statistical Machine Translation for Text Simplification
   Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen and Chris Callison-Burch
   In Transactions of the Association for Computational Linguistics (TACL) 2015
```

**SARI is added to TensorFlow by Google AI group in Feb 2019.**

Google AI TensorFlow

# Automatic Text Simplification

Now, primarily addressed by sequence-to-sequence neural network models.

**Input sentece:**

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship

→

**seq2seq models**
**(RNN, Transformer)**

→

**Generated Output:**

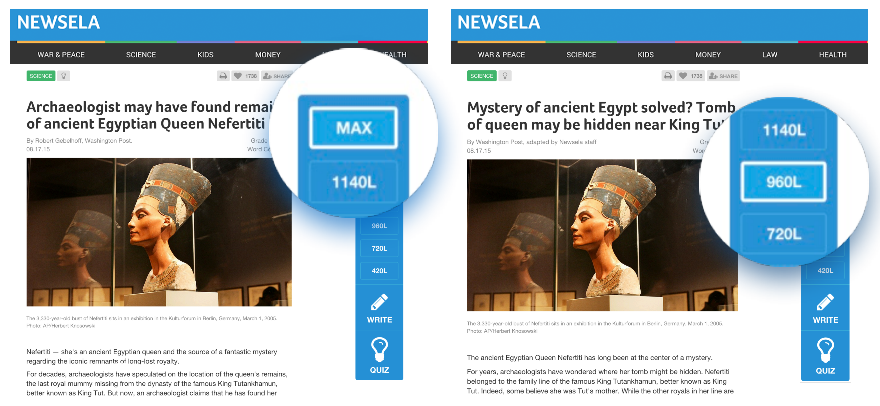
Scientists have found documents in Portugal.

They have also found out who owned the ship.

- **Some example works:**

  - LSTM model (Nisioi et al. 2017)
  - Transformer model (Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, Parmanto Bambang, 2018)

# Part 1 — High-quality Training Data

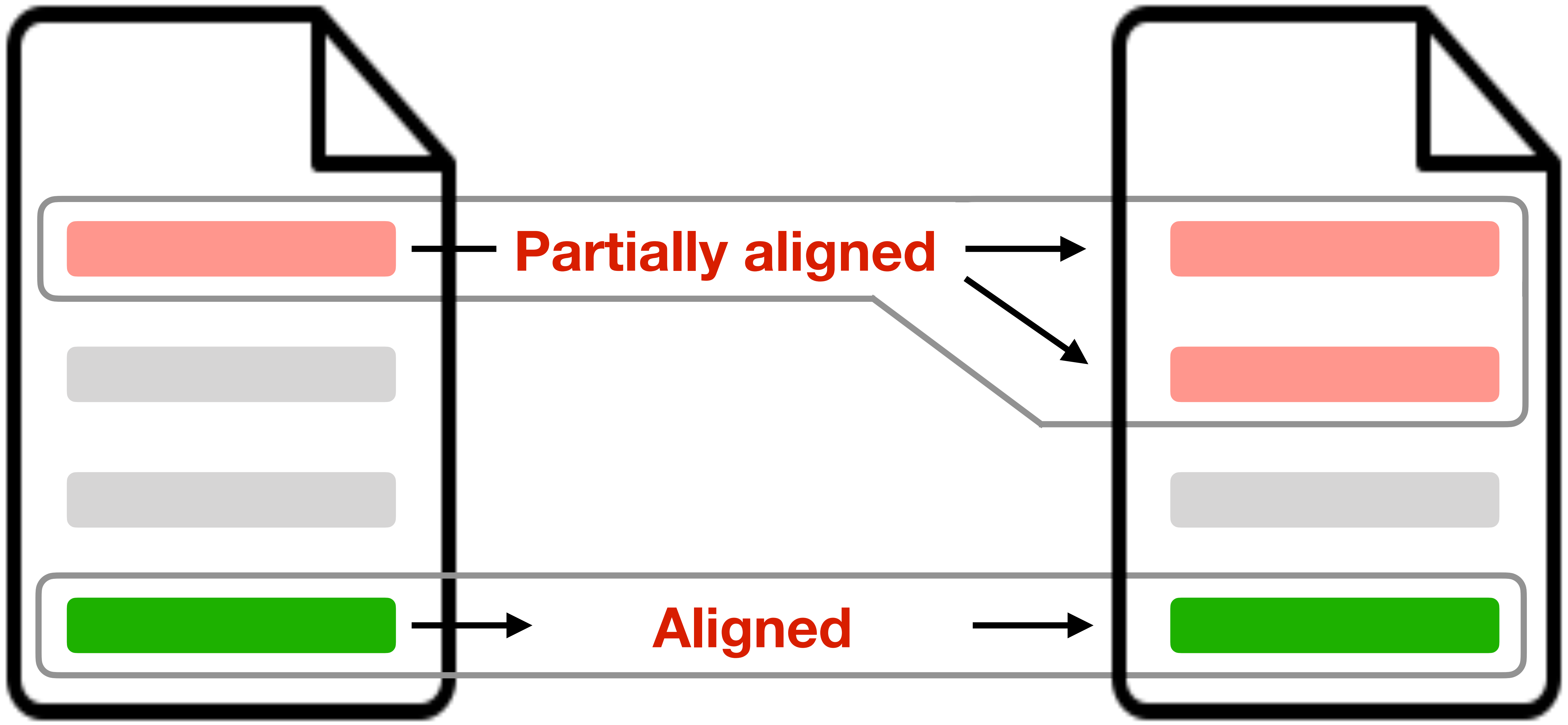



Simple English
WIKIPEDIA

**Neural CRF Model for Sentence Alignment in Text Simplification**
Chao Jiang et al. (ACL 2020)

# Automatic Text Simplification

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




**Complex article**

Wikipedia The Free Encyclopedia

newsela (Original article)

**Simple article**

Simple English Wikipedia

newsela (Simplified article)

# Prior Work for 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 + 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 + 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 |

* 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 |
| Threshold | $\text{BERT}_{finetune}$ + 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 | Our CRF aligner | 97.86 | 91.31 | 95.59 |

+5.7

* Results are on the manually annotated Newsela 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

# 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

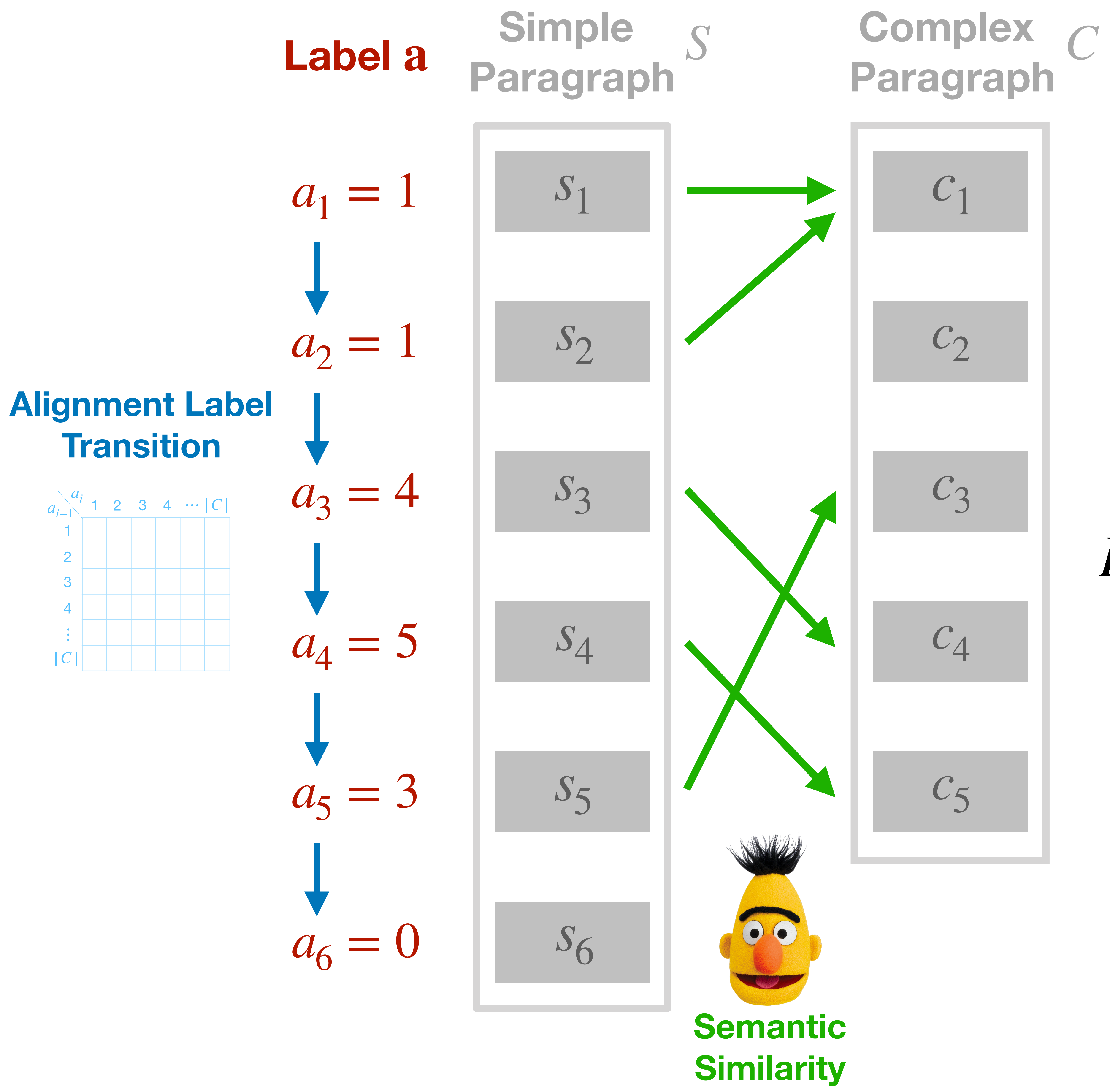
**Initialize:** $simP \in \mathbb{R}^{2\times k\times l}$ to $0^{2\times k\times l}$
**for** $i \leftarrow 1$ **to** $k$ **do**
  **for** $j \leftarrow 1$ **to** $l$ **do**
    $simP[1,i,j] = \underset{s_p \in S_i}{\text{avg}} \left( \underset{c_q \in C_j}{\max} \; simSent(s_p, c_q) \right)$
    $simP[2,i,j]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




Semantic Similarity

Alignment Label Transition

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

**Linear-chain CRF**

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

all possible alignments (dynamic programming)

# Evaluation on Sentence Alignment*

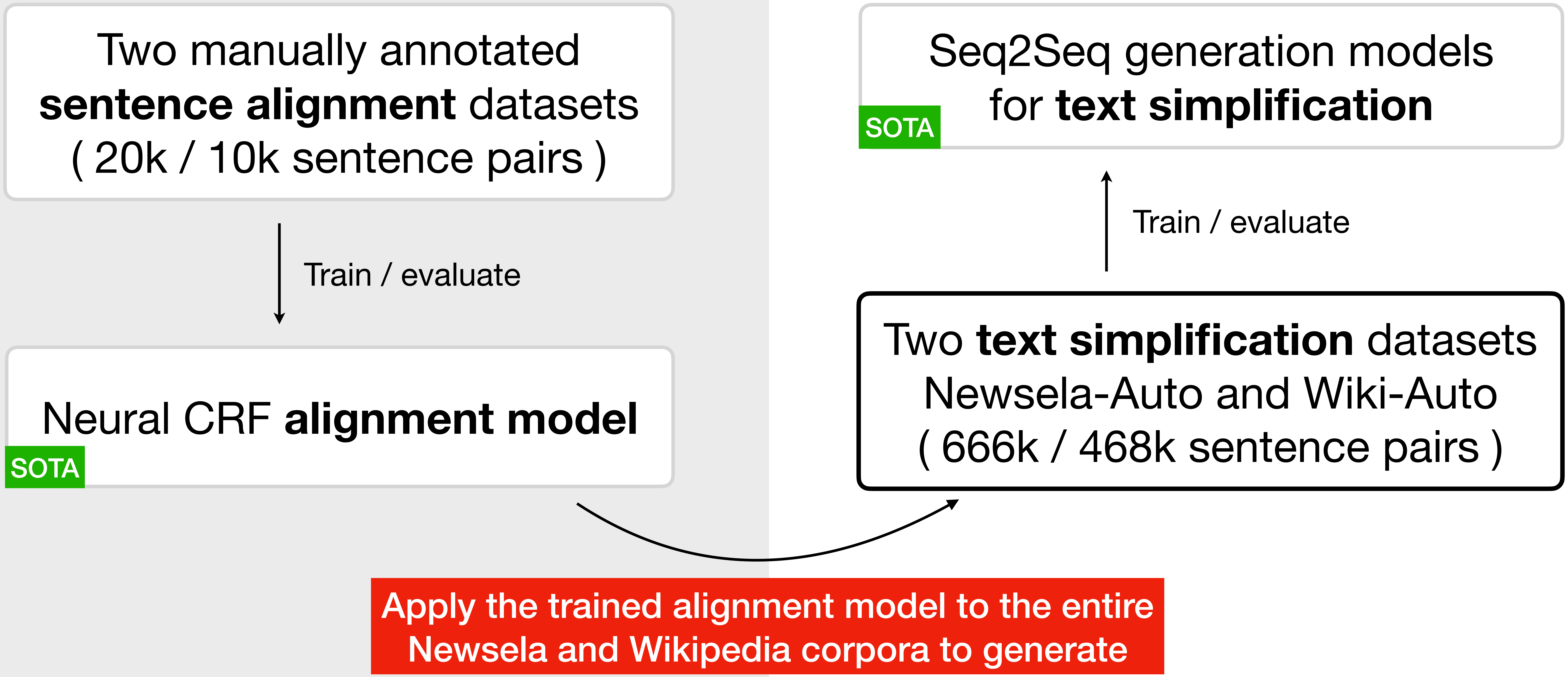
- 50 manually annotated article groups (0.5 million sentence pairs) in Newsela.
- 35 train / 5 dev / 10 test, evaluate on article pairs at adjacent readability level.

| | | 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 | $\text{BERT}_{finetune}$ | 94.99 | 89.62 | 92.22 |
| Threshold | $\text{BERT}_{finetune}$ + paragraph alignment | 98.05 | 88.63 | 93.10 |
| CRF | Our CRF aligner | 97.86 | 91.31 | 95.59 |

+5.7

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

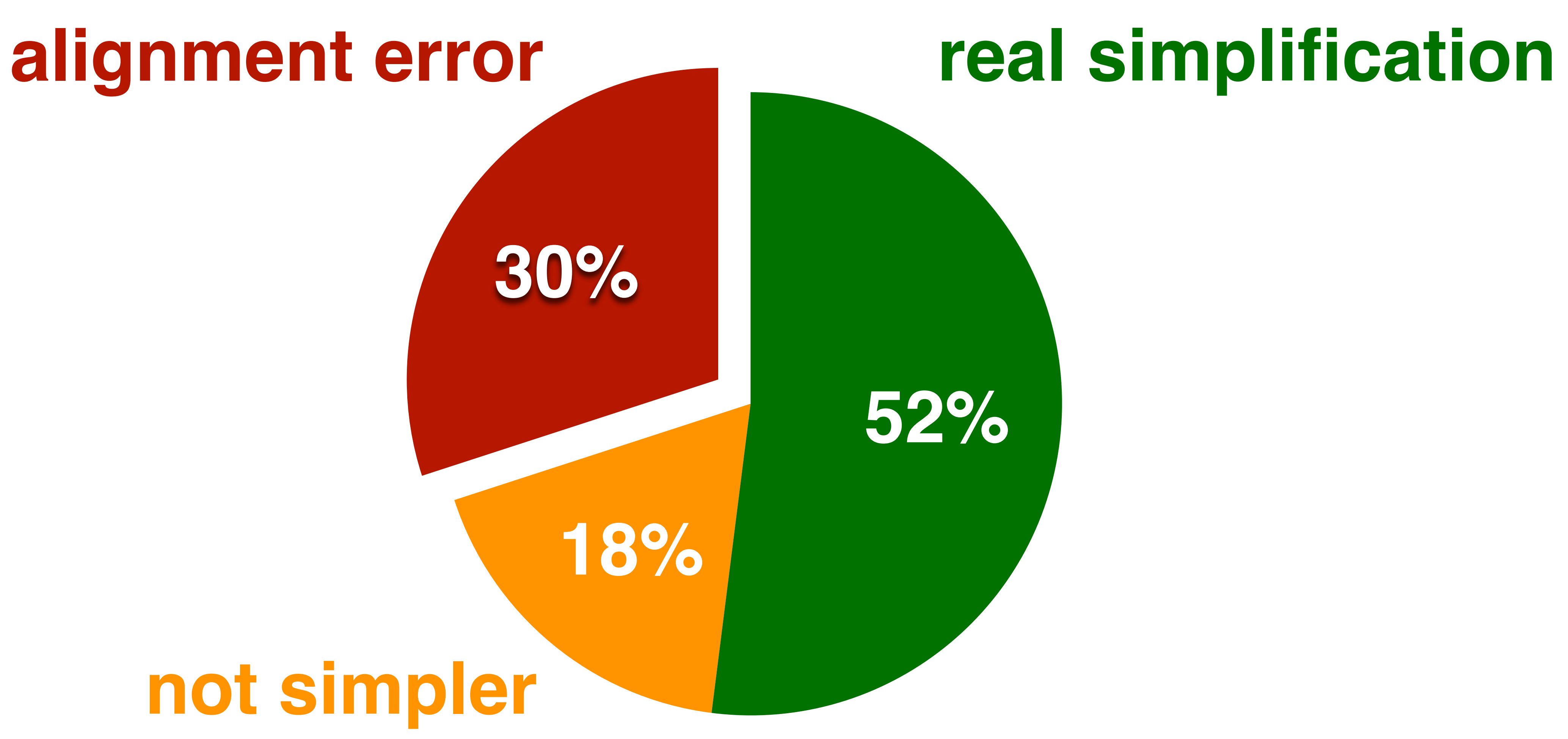
# Our Work



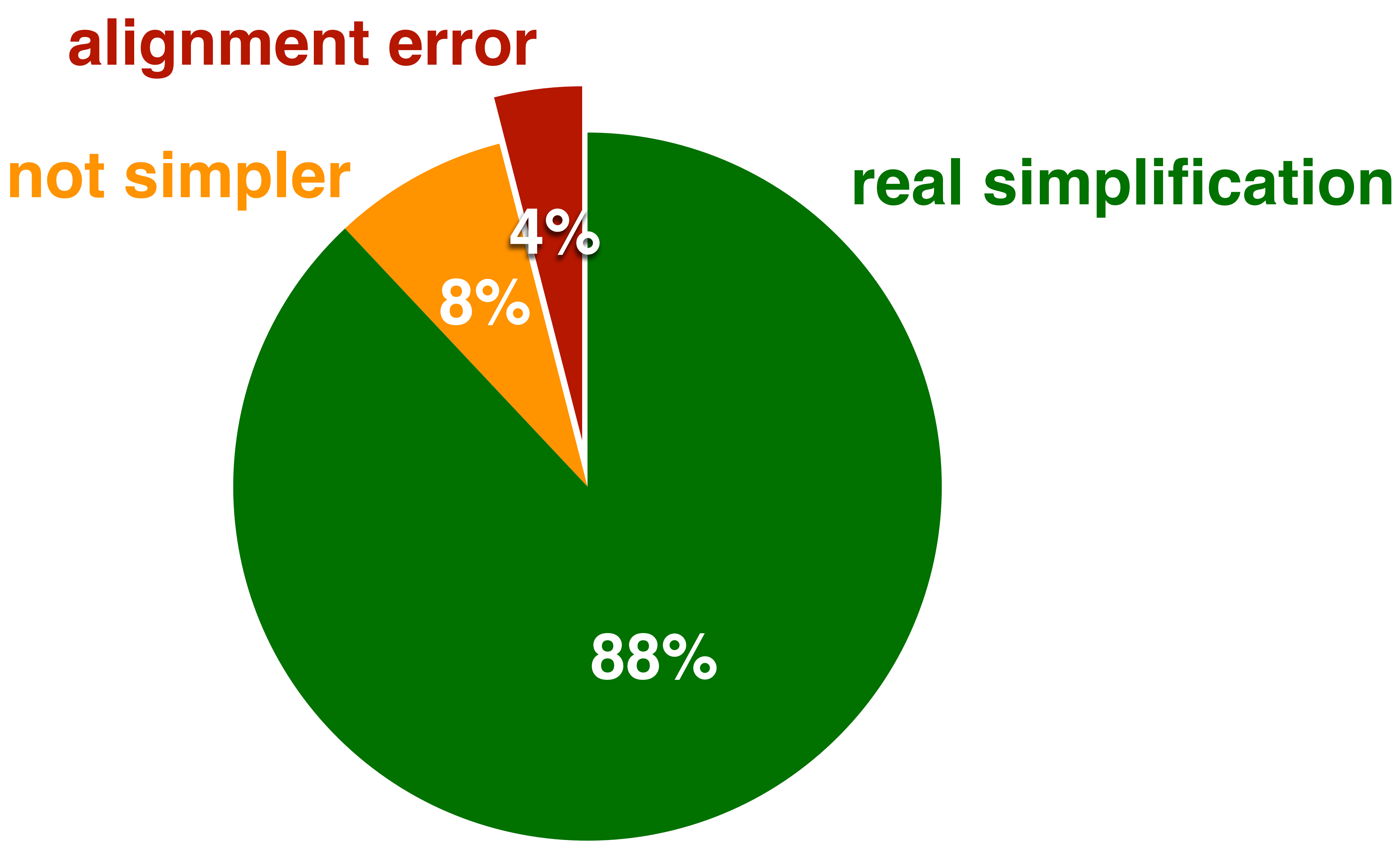

Sentence Alignment

Text Simplification

# New Corpora Contain Way Fewer Errors*




**Wiki-Large**
(Zhang and Lapata, 2017)

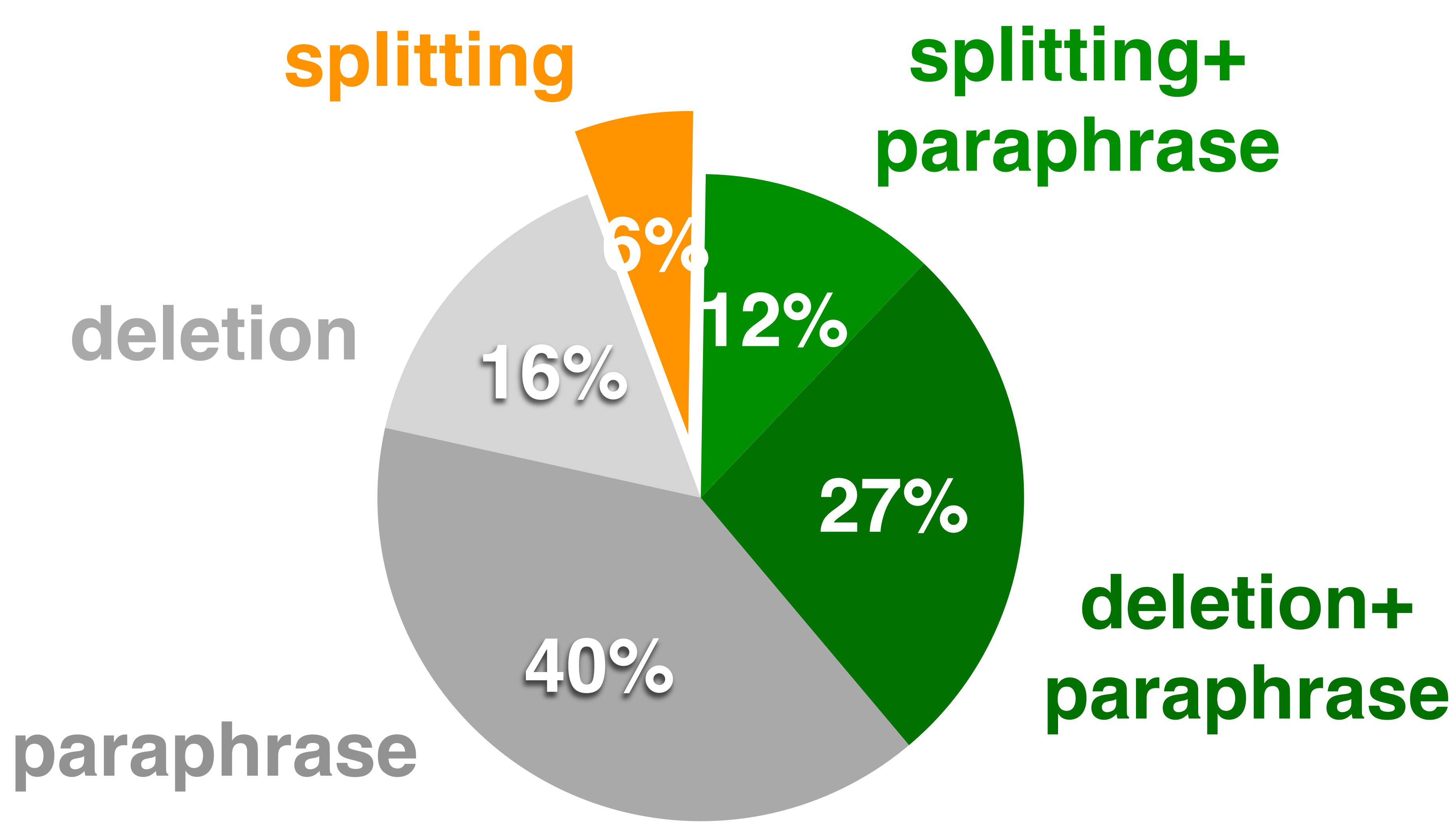



**Wiki-Auto (our work)**
1.6 times larger

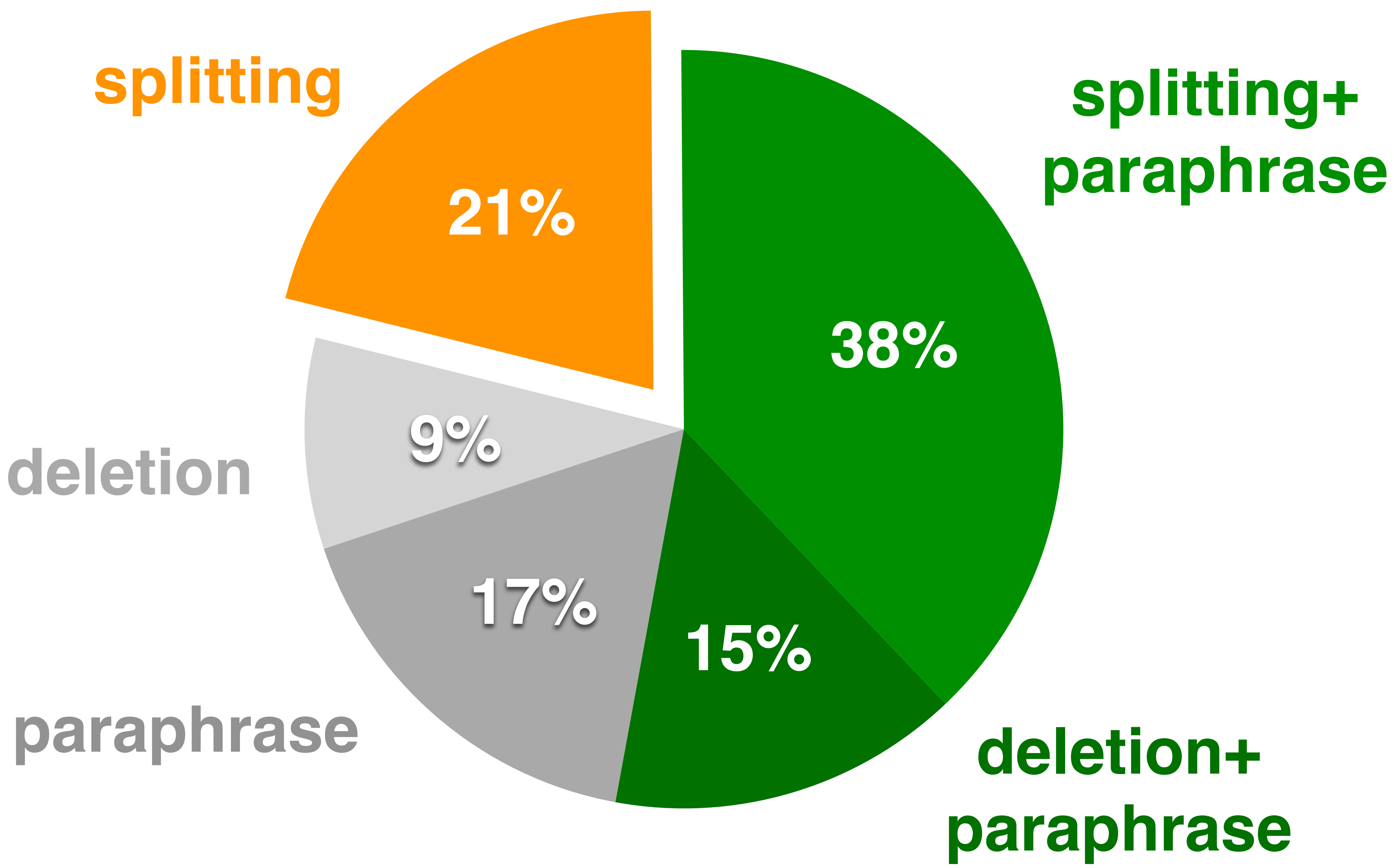
Wiki-Auto has 75% less defective pairs (alignment error + not simpler).

* 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

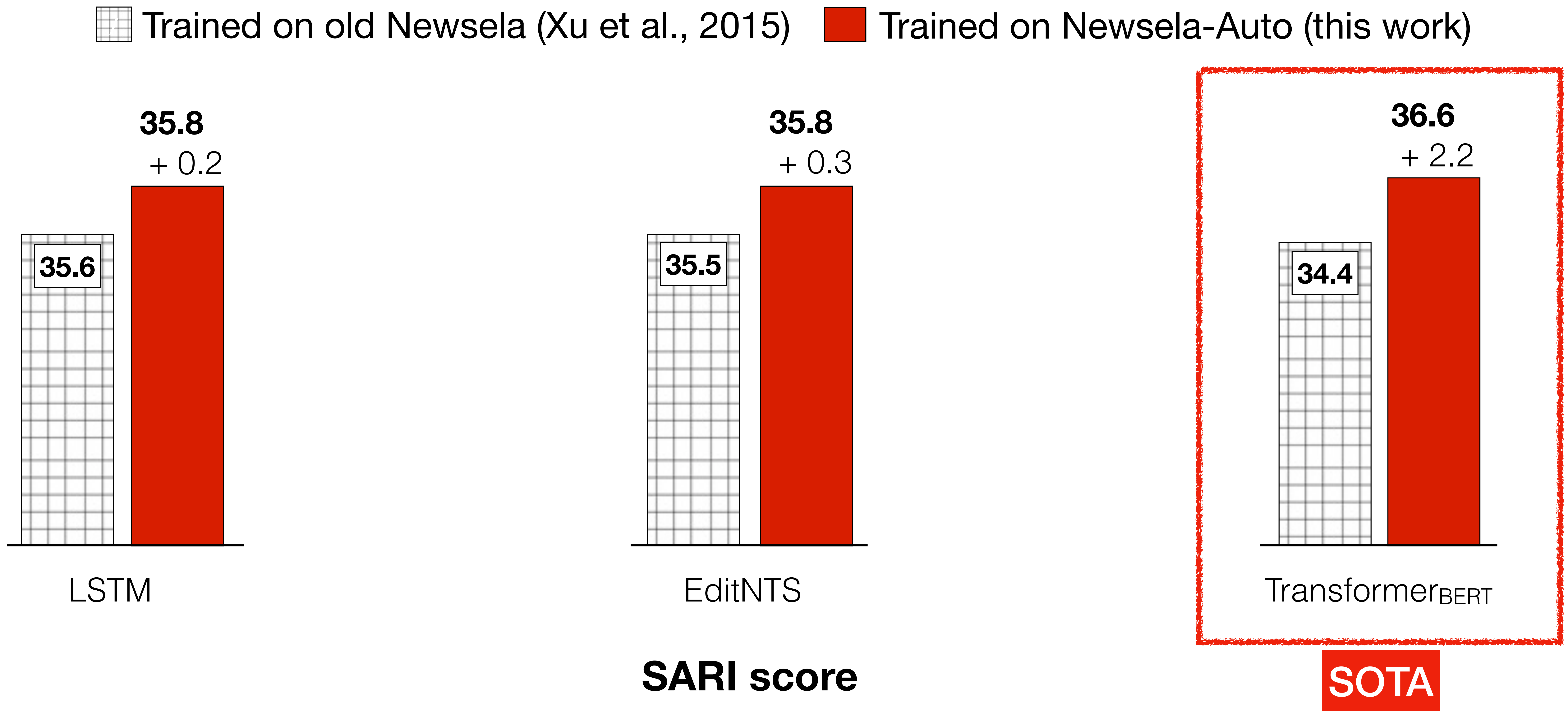
Sentence Alignment

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
  - Our work: Newsela-Auto and Wiki-Auto
  - Old: Newsela (Xu et al., 2015) and Wiki-Large (Zhang and Lapata, 2017)

# Automatic Evaluation on Text Simplification*

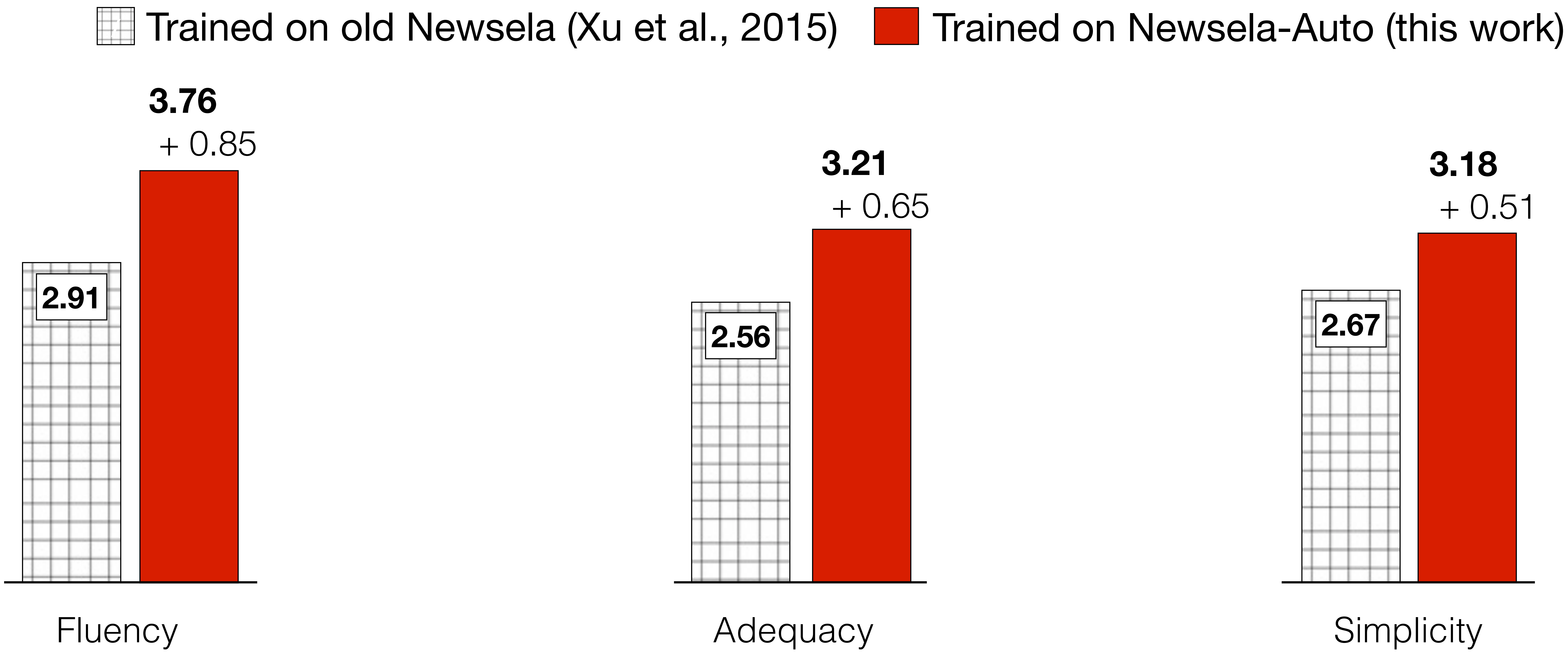



**SARI score**

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

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

# Human Evaluation on Text Simplification*

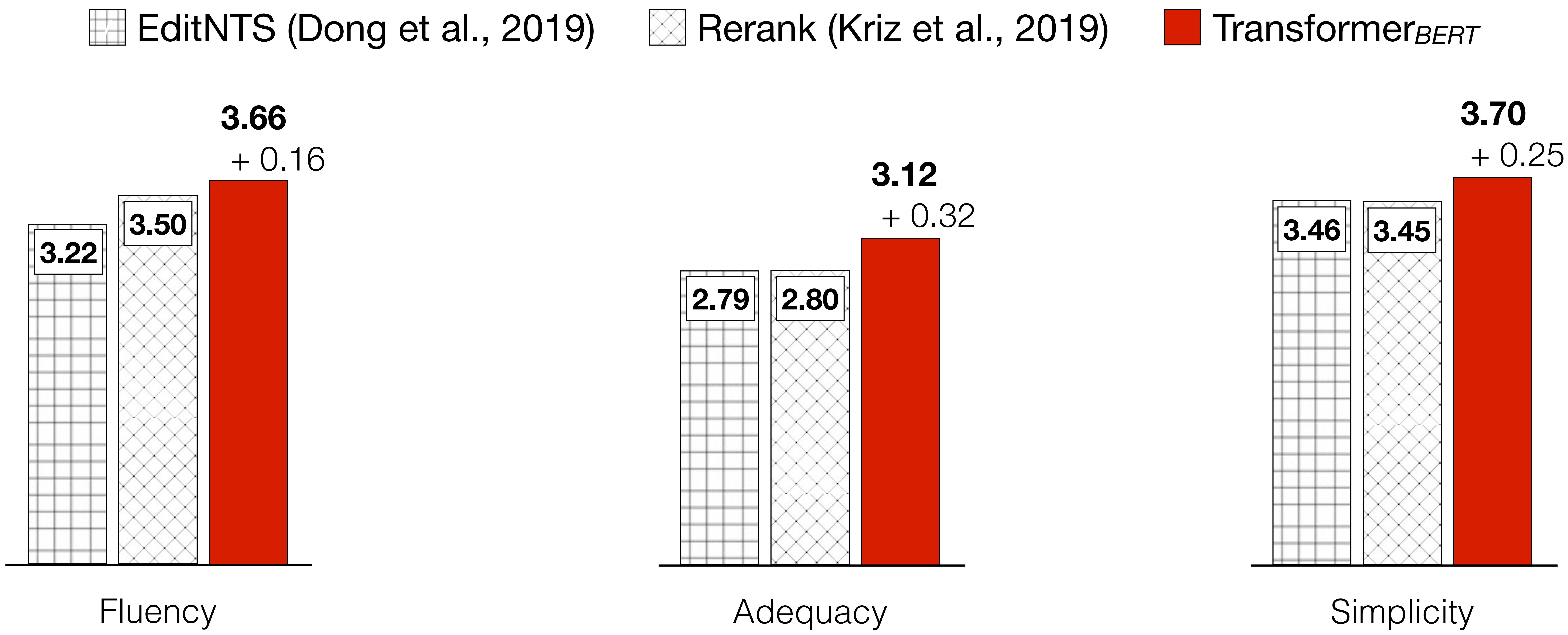



**Transformer$_{BERT}$ model**

(In 5-point Likert scale)

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

# Human Evaluation on Text Simplification*



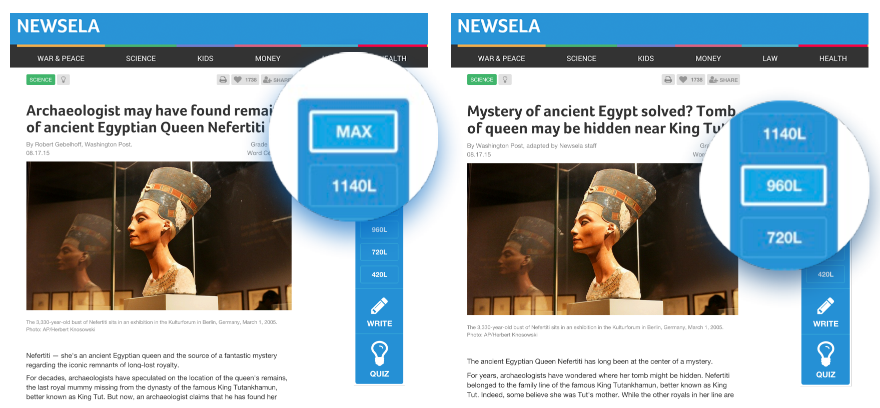

**Transformer$_{BERT}$ trained on Newsela-Auto dataset is new SOTA in human evaluation.**

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

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

# Open Source

Code and data are available at - https://github.com/chaojiang06/wiki-auto




*Simple English*
WIKIPEDIA

**Neural CRF Model for Sentence Alignment in Text Simplification**

Chao Jiang et al. (ACL 2020)

- **Other related work:**
  - "Clue: Cross-modal Coherence Modeling for Caption Generation" (Malihe Alikhani et al., 2020)

# Part 2 — Better Controllable Generation Model

**Controllable Text Simplification with Explicit Paraphrasing**

Mounica Maddela et al. (new work)

# Automatic Text Simplification

Now, primarily addressed by sequence-to-sequence neural network models.

**Input sentece:**

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship

→

**seq2seq models**
**(RNN, Transformer)**

→

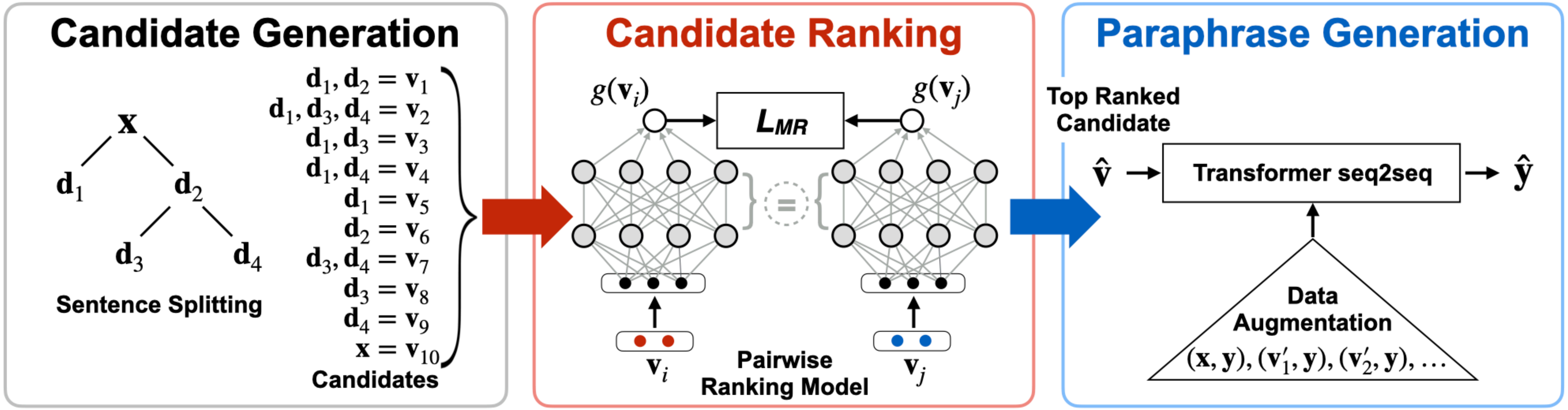
**Generated Output:**

Scientists have found documents in Portugal.

They have also found out who owned the ship.

# Controllable Text Generation

We study how to incorporate linguistic rules with neural network models.




**INPUT** $\mathbf{x}$ : *The exhibition, which opened Oct. 8 and runs through Jan. 3, features 27 self-portraits.* **REFERENCE** $\mathbf{y}$ : *The show started Oct. 8. It ends Jan. 3.*

$\mathbf{d}_1$ : *The exhibition features 27 self-portraits.* $\mathbf{d}_2$ : *The exhibition opened Oct. 8 and runs through Jan. 3.*

$\mathbf{d}_3$ : *The exhibition opened Oct. 8.* $\mathbf{d}_4$ : *The exhibition runs through Jan. 3.* $\hat{\mathbf{v}} = \mathbf{v}_7$ : *The exhibition opened Oct. 8. The exhibition runs through Jan. 3.*

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (new work)

# Controllable Text Generation

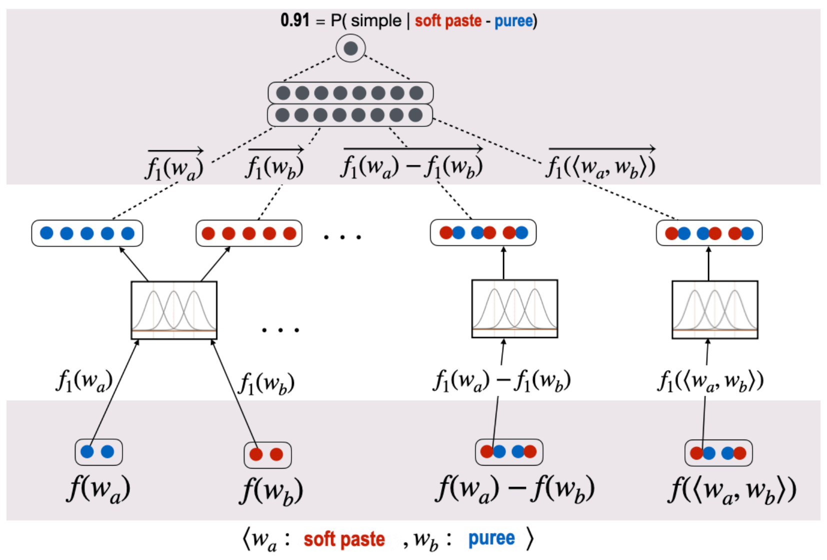
We can control the degree of sentence splitting, deletion, and paraphrasing.

| | **System Outputs** |
|---|---|
| Complex | *Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship.* |
| Simple | *Since 2010, experts have been figuring out who owned the ship.* |
| Hybrid-NG | *since 2010, the project **scientists** have uncovered documents in portugal that have **about** who **owns** the ship.* |
| LSTM | *since 2010, **scientists** have uncovered documents in portugal that have revealed who owned the ship.* |
| Transformer$_{bert}$ | ***they discovered** that the ship **had been important**.* |
| EditNTS | *since 2010, project researchers have uncovered documents in portugal**.** **have** revealed who owned the ship.* |
| Our Model ($cp$ = 0.6) | ***scientists** have **found** a **secret deal**. **they have discovered** who owned the ship.* |
| Our Model ($cp$ = 0.7) | ***scientists** have **found** documents in portugal**. they have also found out** who owned the ship.* |
| Our Model ($cp$ = 0.8) | ***scientists** have **found** documents in portugal**. they have discovered** who owned the ship.* |
| Complex | *Experts say China's air pollution exacts a tremendous toll on human health.* |
| Simple | *China's air pollution is very unhealthy.* |
| Hybrid-NG | *experts say the **government's** air pollution exacts a toll on human health.* |
| LSTM | *experts say china's air pollution exacts a tremendous toll on human health.* |
| Transformer$_{bert}$ | *experts say china's pollution has a tremendous **effect** on human health.* |
| EditNTS | *experts say china's air pollution **can cause** human health.* |
| Our Model ($cp$ = 0.6) | *experts say china's air pollution **is a big problem** for human health.* |
| Our Model ($cp$ = 0.7) | *experts say china 's air pollution **can cause a lot of damage** on human health.* |
| Our Model ($cp$ = 0.8) | *experts say china 's air pollution **is** a **huge** toll on human health.* |

Table 5: Examples of system outputs. **Red** marks the errors; **blue** marks good paraphrases. $cp$ is a soft constraint that denotes number of words that can be copied from the input.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (new work)

# Part 3 — A Lightweight Method




**A Word-Complexity Lexicon and a Neural Readability Ranking Model for Lexical Simplification**

Maddela Mounica et al. (EMNLP 2018)

# Lexical Simplification

*Applesauce is a purée made of apples.*

**Pairwise neural ranking model to better measure readability.**

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Lexical Simplification

*Applesauce is a* ***purée*** *made of apples.*

↓

*Applesauce is a* ***liquidized sauce*** *. It is made of apples.*
***soft paste***
***thick liquid***

**Complex Word Identification - Substitution Generation - Substitution Ranking**

**Pairwise neural ranking model to better measure readability.**

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Lexical Simplification

*Applesauce is a* ***purée*** *made of apples.*

*Applesauce is a* ***soft paste*** / ***thick liquid*** / ***liquidized sauce*** *. It is made of apples.*

complexity

**Complex Word Identification - Substitution Generation - Substitution Ranking**

- **Other related work:**
  - "Phrasal Substitution of Idiomatic Expressions" (Liu & Hwa, 2016)

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model

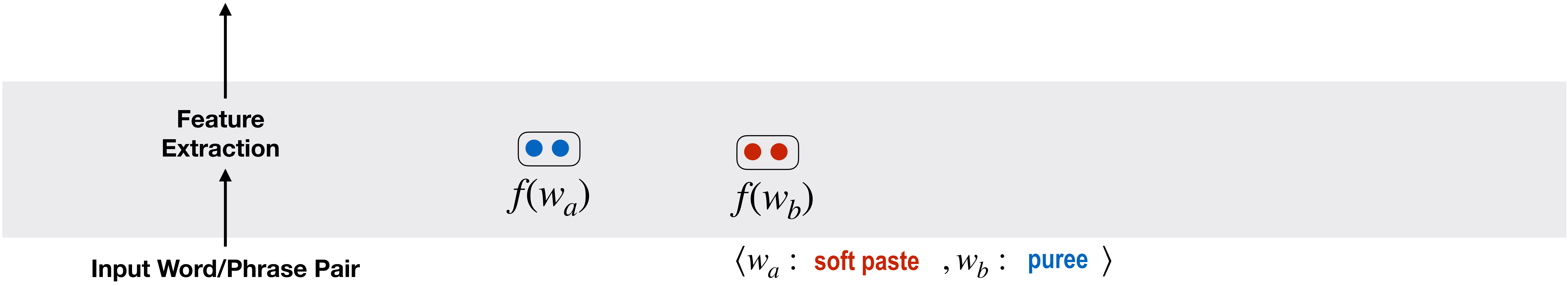
to better measure the readability of words and phrases.

**Input Word/Phrase Pair** $\langle w_a :$ soft paste $, w_b :$ puree $\rangle$

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model

to better measure the readability of words and phrases.



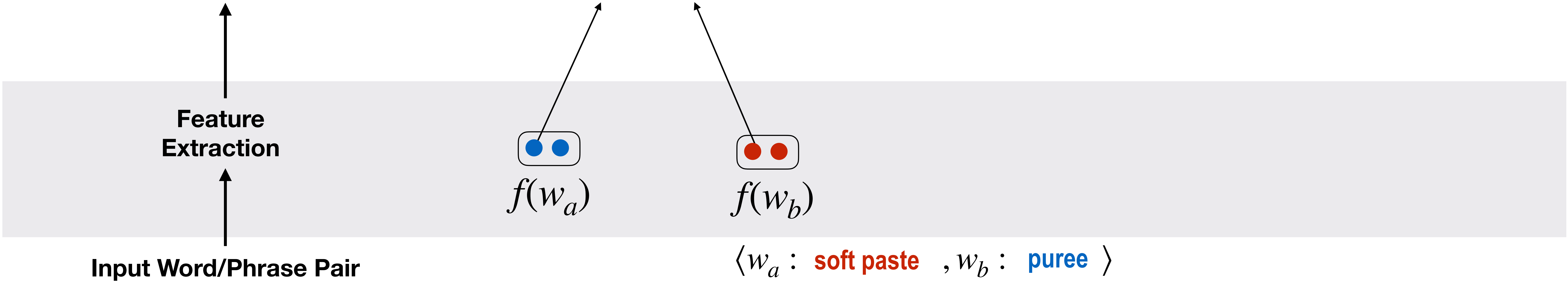

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model

to better measure the readability of words and phrases.

**Word-Complexity Lexicon Score**
**+ 0/1 binary indicator**

word length
word frequency
number of syllables
ngram language model probabilities




Mounica Maddela, Wei Xu. “A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification” in EMNLP (2018)

# The Word-Complexity Lexicon

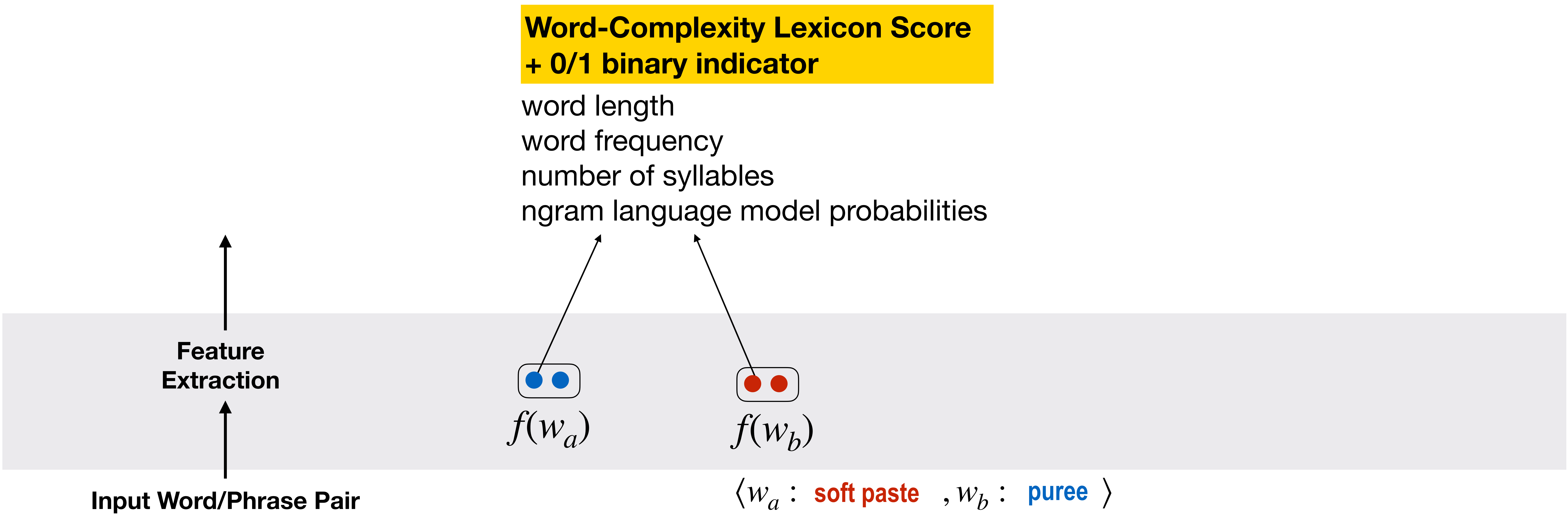
- 15,000 most frequent English words from Google 1T ngram corpus
- Rated on a 6-point Likert scale
    - 11 annotators (non-native speakers)
    - 5 ~ 7 ratings for each word
    - 2.5 hours to rate 1000 words

| Word | Score | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|---|
| *muscles* | 1.6 | 2 | 1 | 2 | 2 | 1 |
| *pattern* | 2.4 | 2 | 3 | 1 | 1 | 3 |
| *educational* | 3.2 | 3 | 3 | 3 | 3 | 4 |
| *cortex* | 4.2 | 4 | 4 | 4 | 4 | 5 |
| *assay* | 5.8 | 6 | 6 | 6 | 5 | 6 |

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model

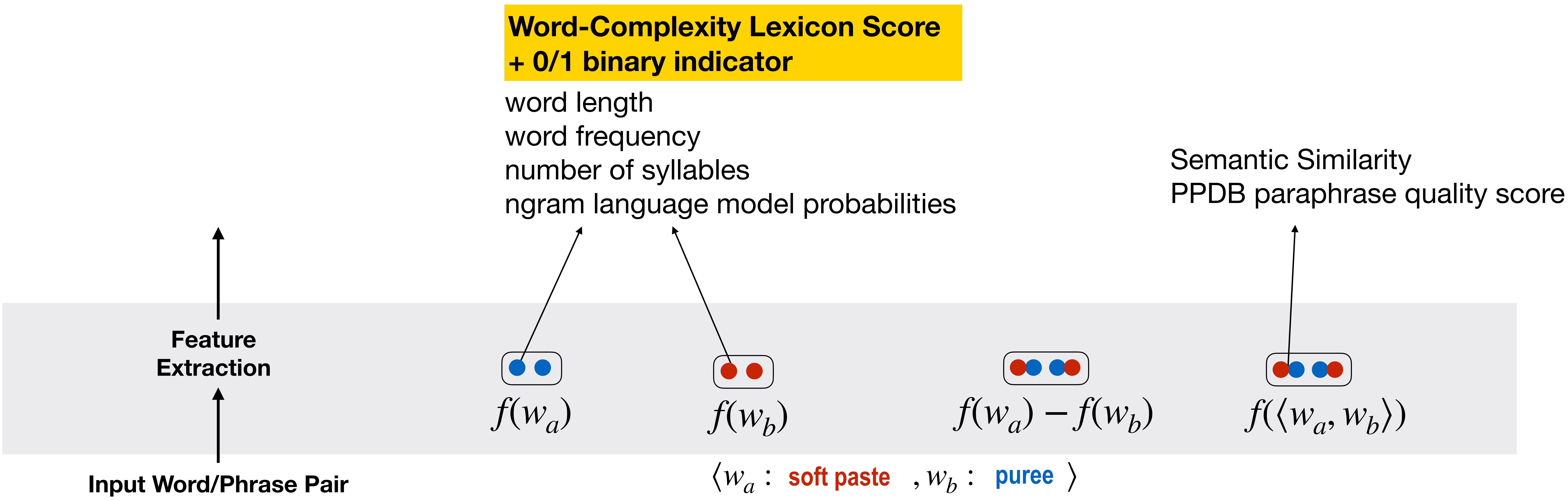
to better measure the readability of words and phrases.




Mounica Maddela, Wei Xu. “A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification” in EMNLP (2018)

# Neural Readability Ranking Model

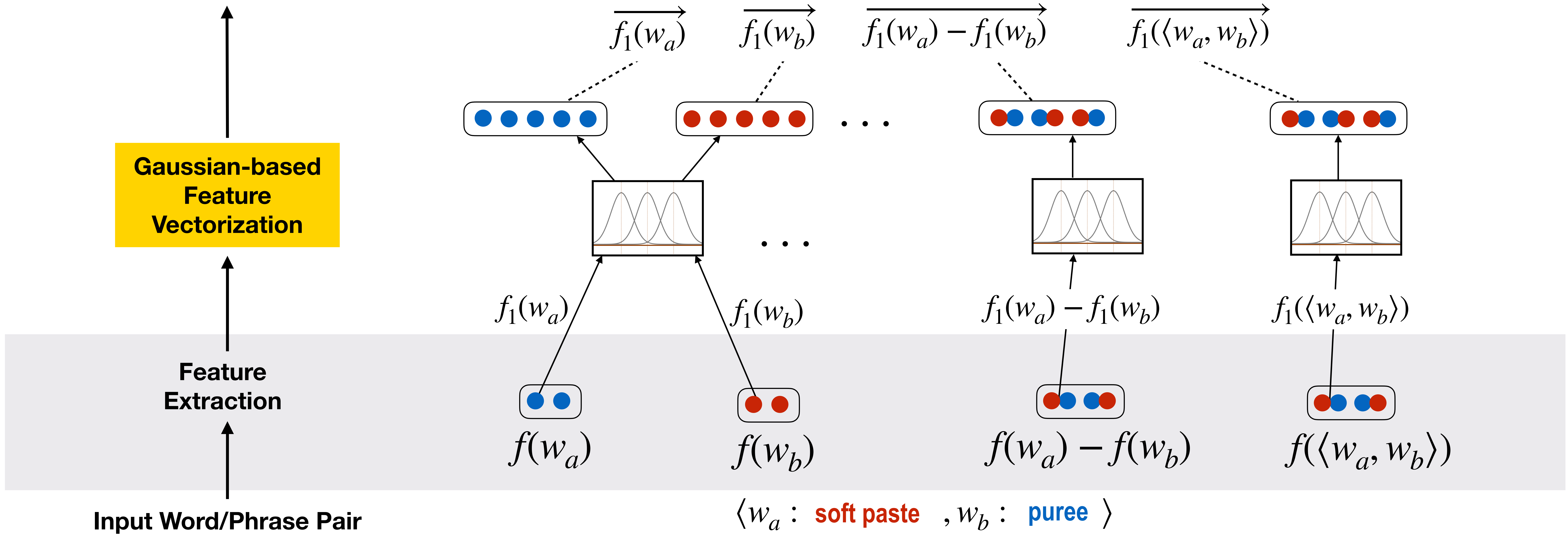
to better measure the readability of words and phrases.




Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model

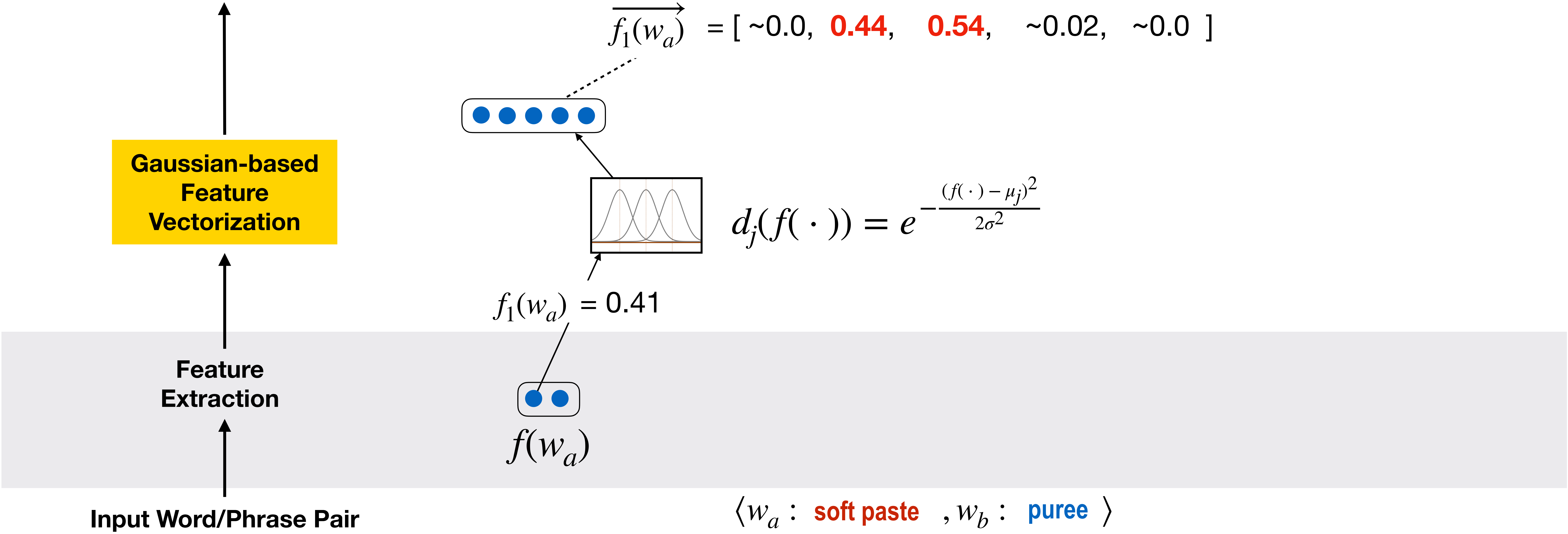
to better measure the readability of words and phrases.




Mounica Maddela, Wei Xu. “A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification” in EMNLP (2018)

# Neural Readability Ranking Model

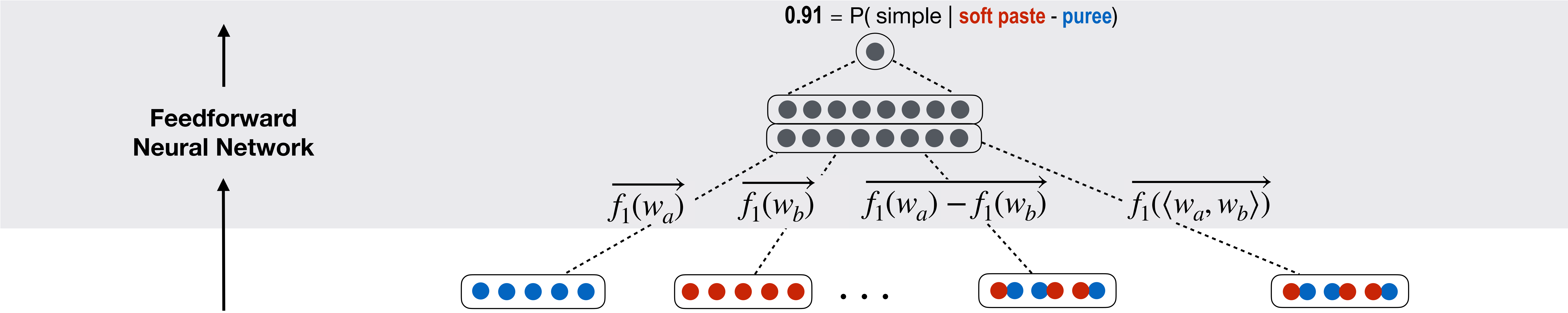
to better measure the readability of words and phrases.




Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model

to better measure the readability of words and phrases.




**P > 0** $\Rightarrow$ $w_a$ is simpler than $w_b$

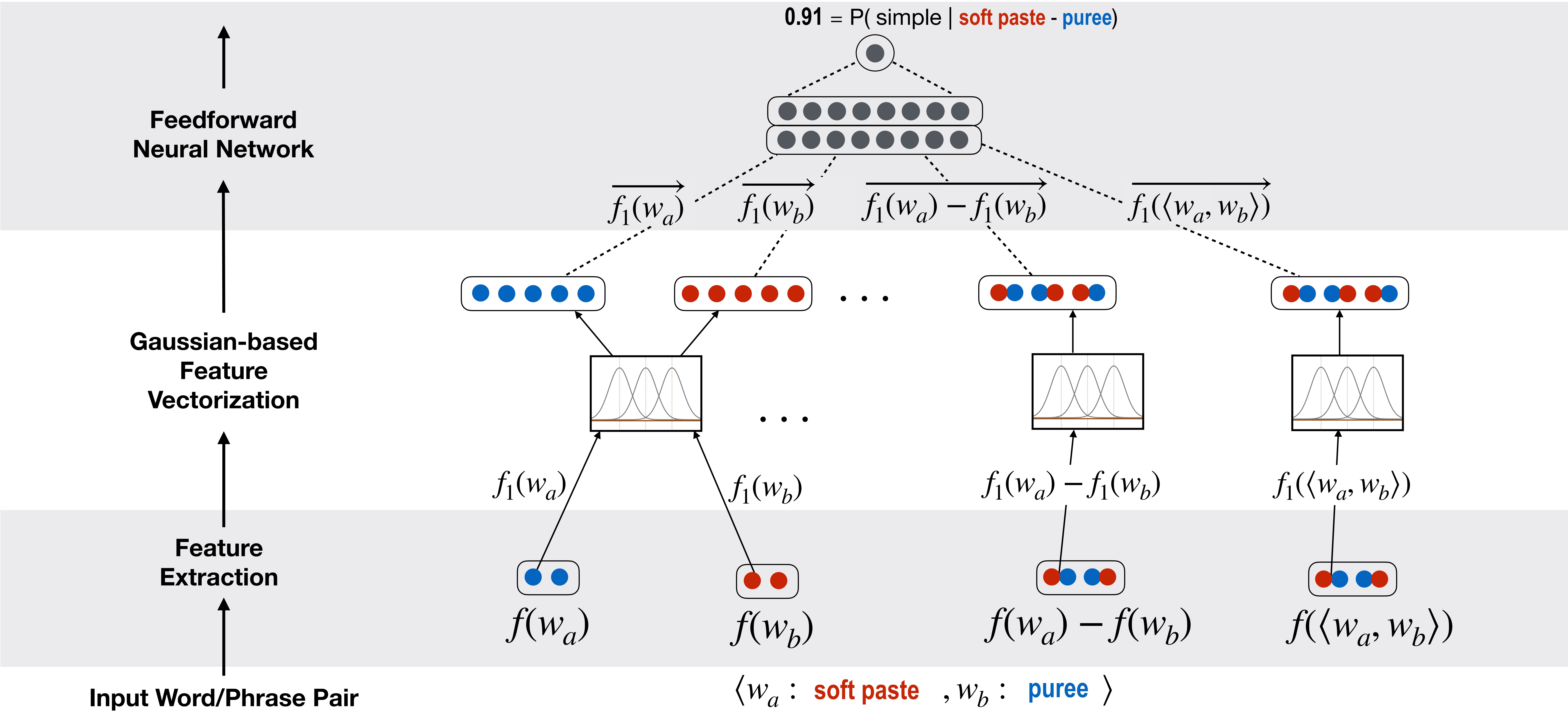
**P < 0** $\Rightarrow$ $w_a$ is more complex than $w_b$

**| P |** indicates complexity difference

Mounica Maddela, Wei Xu. “A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification” in EMNLP (2018)

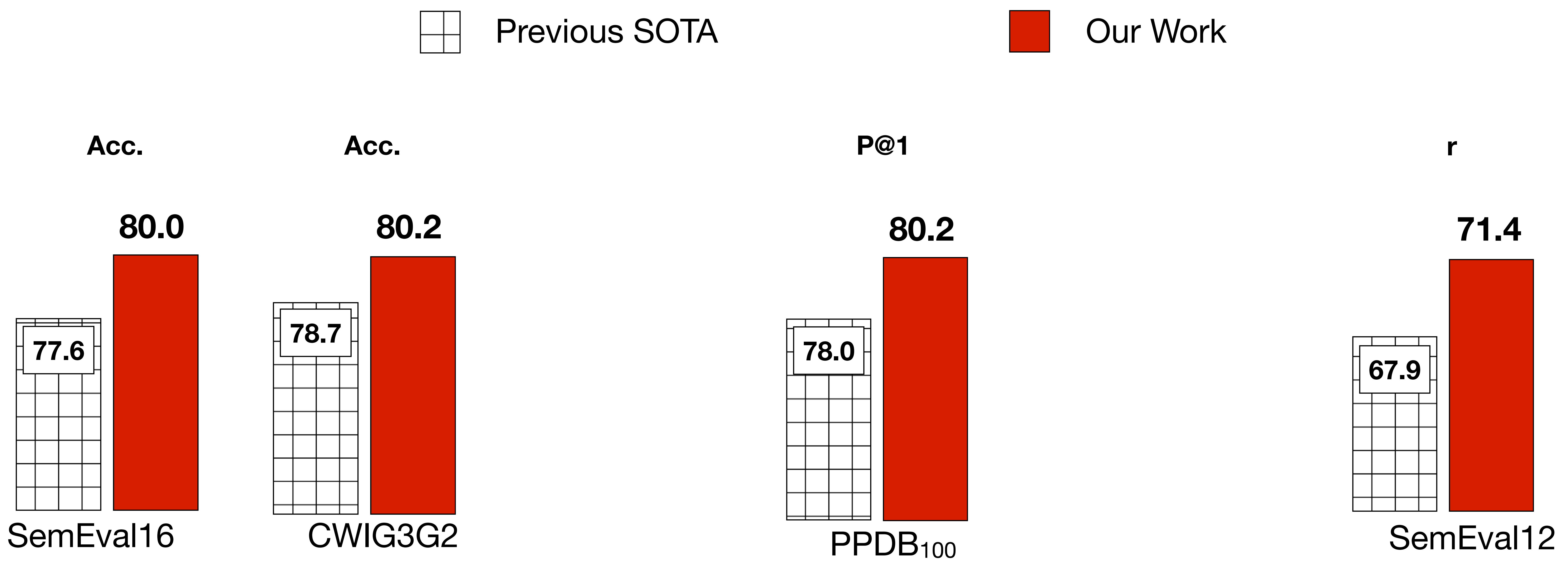
# Neural Readability Ranking Model

to better measure the readability of words and phrases.




Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Neural Readability Ranking Model




**Complex Word Identification - Substitution Generation - Substitution Ranking**

**Improved the state-of-the-art significantly for all lexical simplification tasks.**

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# SimplePPDB++

A database of 14.1 million paraphrase rules with improved complexity ranking scores.

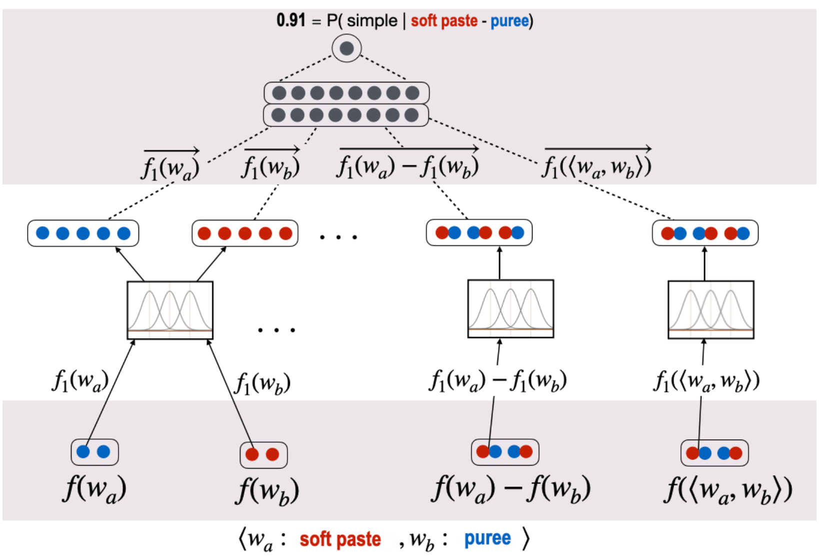
| Paraphrase Rule | | Score |
|---|---|---|
| | → *self-supporting* | 0.93 |
| *self-reliant* | → *self-sufficient* | 0.48 |
| | → *self-sustainable* | -0.60 |
| | → *possible* | 0.94 |
| *viable* | → *realistic* | 0.15 |
| | → *plausible* | -0.91 |
| | → *in-depth review* | 0.89 |
| *detailed assessement* | → *careful examination* | 0.28 |
| | → *comprehensive evaluation* | -0.87 |

complex

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" in EMNLP (2018)

# Open Source

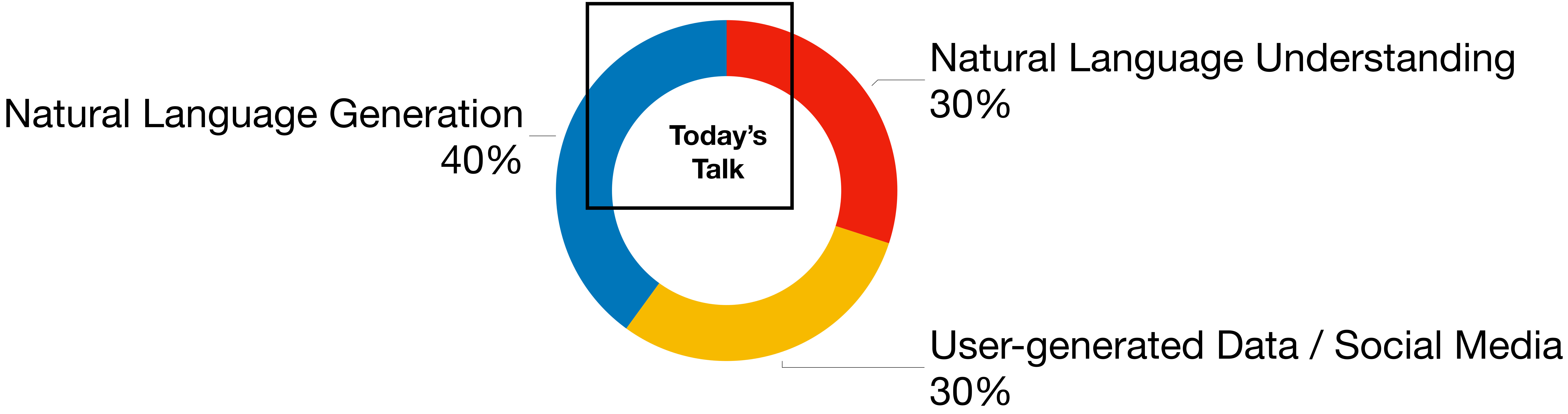
Code and data are available at - https://github.com/mounicam/lexical_simplification




**A Word-Complexity Lexicon and a Neural Readability Ranking Model for Lexical Simplification**

Maddela Mounica et al. (EMNLP 2018)

# My Research



Natural Language Generation
40%
Today's
Talk
Natural Language Understanding
30%
User-generated Data / Social Media
30%

# Thank you!

**https://cocoxu.github.io/**

thank u 4 ur time

thanku

I am grateful

thx

thanks a lot

thanking you

appreciate it

gratitude

gramercies

3x

tyvm

thanks

say thanks

thank you very much

thnx

thanks a ton

wawwww thankkkkkkkkkk you alottttttttttt!

I can no other answer make but thanks, and thanks, and ever thanks.

Ohio Supercomputer Center