Neural CRF Model for Sentence Alignment in Text Simplification

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

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.
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**.
**Weakness of Previous Work on Sentence Alignment**

<table>
<thead>
<tr>
<th>Method</th>
<th>Similarity metric</th>
<th>Alignment strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>JaccardAlign (Xu et al., 2015)</td>
<td>Jaccard similarity</td>
<td>Greedy</td>
</tr>
<tr>
<td>MASSAlign (Paetzold et al., 2017)</td>
<td>TF-IDF</td>
<td>Dynamic programming</td>
</tr>
<tr>
<td>CATS (Štajner et al., 2018)</td>
<td>Lexical-similarities</td>
<td>Greedy</td>
</tr>
</tbody>
</table>

**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$_{\text{finetune}}$ → A neural CRF alignment model.

<table>
<thead>
<tr>
<th>aligned + partial vs. others$^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
</tr>
</tbody>
</table>

$^*$ 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\(_{\text{finetune}}\) → A neural CRF alignment model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>JaccardAlign (Xu et al., 2015)</td>
<td>98.66</td>
<td>67.58</td>
<td>80.22</td>
</tr>
<tr>
<td>MASSAlign (Paetzold et al., 2017)</td>
<td>95.49</td>
<td>82.27</td>
<td>88.39</td>
</tr>
<tr>
<td>CATS (Štajner et al., 2018)</td>
<td>88.56</td>
<td>91.31</td>
<td>89.92</td>
</tr>
</tbody>
</table>

* 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\textit{finetune} → A neural CRF alignment model.

<table>
<thead>
<tr>
<th></th>
<th>aligned + partial vs. others*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
</tr>
<tr>
<td>JaccardAlign (Xu et al., 2015)</td>
<td>98.66</td>
</tr>
<tr>
<td>MASSAlign (Paetzold et al., 2017)</td>
<td>95.49</td>
</tr>
<tr>
<td>CATS (Štajner et al., 2018)</td>
<td>88.56</td>
</tr>
<tr>
<td>BERT \textit{finetune}</td>
<td>94.99</td>
</tr>
</tbody>
</table>

* 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\textsubscript{finetune} \rightarrow A neural CRF alignment model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>JaccardAlign (Xu et al., 2015)</td>
<td>98.66</td>
<td>67.58</td>
<td>80.22</td>
</tr>
<tr>
<td>MASSAlign (Paetzold et al., 2017)</td>
<td>95.49</td>
<td>82.27</td>
<td>88.39</td>
</tr>
<tr>
<td>CATS (Štajner et al., 2018)</td>
<td>88.56</td>
<td>91.31</td>
<td>89.92</td>
</tr>
<tr>
<td>BERT\textsubscript{finetune}</td>
<td>94.99</td>
<td>89.62</td>
<td>92.22</td>
</tr>
<tr>
<td>BERT\textsubscript{finetune} + paragraph alignment</td>
<td>98.05</td>
<td>88.63</td>
<td>93.10</td>
</tr>
</tbody>
</table>

* 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\textit{finetune} → A neural CRF alignment model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>JaccardAlign (Xu et al., 2015)</td>
<td>98.66</td>
<td>67.58</td>
<td>80.22</td>
</tr>
<tr>
<td>MASSAlign (Paetzold et al., 2017)</td>
<td>95.49</td>
<td>82.27</td>
<td>88.39</td>
</tr>
<tr>
<td>CATS (Štajner et al., 2018)</td>
<td>88.56</td>
<td>91.31</td>
<td>89.92</td>
</tr>
<tr>
<td>BERT\textit{finetune}</td>
<td>94.99</td>
<td>89.62</td>
<td>92.22</td>
</tr>
<tr>
<td>BERT\textit{finetune} + paragraph alignment</td>
<td>98.05</td>
<td>88.63</td>
<td>93.10</td>
</tr>
<tr>
<td>Ours CRF aligner</td>
<td>97.86</td>
<td>91.31</td>
<td>95.59</td>
</tr>
</tbody>
</table>

* 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)

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

Neural CRF alignment model

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

Seq2Seq generation models for text simplification

Train / evaluate

SOTA
Our Work

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

Neural CRF alignment model (SOTA)

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

Seq2Seq generation models for text simplification (SOTA)

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

Train / evaluate

Sentence Alignment

Text Simplification
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 \textit{aligned} / \textit{partially aligned} / \textit{not aligned}
  - Inter-annotator agreement: 0.807 (Cohen Kappa)

Step 3: Verify the crowdsourcing labels by \( \text{\# annotators} \times 4 \)

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

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)

Neural CRF alignment model

Train / evaluate

Seq2Seq generation models for text simplification

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

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate
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: \( \text{sim}P \in \mathbb{R}^{2 \times k \times l} \) to \( \mathbb{R}^{k \times l} \)
for \( i \leftarrow 1 \) to \( k \) do
  for \( j \leftarrow 1 \) to \( l \) do
    \( \text{sim}P[1, i, j] = \text{avg}_{s_p \in S_i} \left( \max_{c_q \in C_j} \text{simSent}(s_p, c_q) \right) \)
    \( \text{sim}P[2, i, j] = \max_{s_p \in S_i, c_q \in C_j} \text{simSent}(s_p, c_q) \)
  end
end
return \( \text{sim}P \)
```

```
Algorithm 2: Paragraph Alignment Algorithm

Input: \( \text{sim}P \in \mathbb{R}^{2 \times k \times l} \)
Initialize: \( \text{align}P \in \mathbb{R}^{k \times l} \)
for \( i \leftarrow 1 \) to \( k \) do
  \( j_{\text{max}} = \arg \max_j \text{sim}P[1, i, j] \)
  if \( \text{sim}P[1, i, j_{\text{max}}] > \tau_1 \) and \( \text{d}(i, j_{\text{max}}) < \tau_2 \) then
    \( \text{align}P[i, j_{\text{max}}] = 1 \)
  end
for \( j \leftarrow 1 \) to \( l \) do
  if \( \text{sim}P[2, i, j] > \tau_3 \) then
    \( \text{align}P[i, j] = 1 \)
  end
  if \( j > 1 \) \&\& \( \text{sim}P[2, i, j] > \tau_4 \) \&\& \( \text{sim}P[2, i, j - 1] > \tau_4 \) \&\& \( \text{d}(i, j - 1) < \tau_5 \) then
    \( \text{align}P[i, j] = 1 \)
    \( \text{align}P[i, j - 1] = 1 \)
  end
end
return \( \text{align}P \)
```
Neural CRF Sentence Alignment Model

Label $a$

Simple Paragraph $S$

Complex Paragraph $C$

$A_1 = 1$

$A_2 = 1$

$A_3 = 4$

$A_4 = 5$

$A_5 = 3$

$A_6 = 0$

Semantic Similarity

Alignment Label Transition

Linear-chain CRF

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

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

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>JaccardAlign (Xu et al., 2015)</td>
<td>98.66</td>
<td>67.58</td>
<td>80.22</td>
</tr>
<tr>
<td>MASSAlign (Paetzold et al., 2017)</td>
<td>95.49</td>
<td>82.27</td>
<td>88.39</td>
</tr>
<tr>
<td>CATS (Štajner et al., 2018)</td>
<td>88.56</td>
<td>91.31</td>
<td>89.92</td>
</tr>
<tr>
<td>BERT finetune</td>
<td>94.99</td>
<td>89.62</td>
<td>92.22</td>
</tr>
<tr>
<td>BERT finetune + paragraph alignment</td>
<td>98.05</td>
<td>88.63</td>
<td>93.10</td>
</tr>
<tr>
<td>Ours CRF aligner</td>
<td>97.86</td>
<td>91.31</td>
<td>95.59</td>
</tr>
</tbody>
</table>

* 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)
  - Neural CRF alignment model
  - Train / evaluate

- Seq2Seq generation models for text simplification
  - Train / evaluate

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

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate
New Corpora Contain Way Fewer Errors*

Wiki-Large
(Zhang and Lapata, 2017)

 alignment error 30%
real simplification 52%
not simpler 18%

Wiki-Auto (this work)
1.6 times larger

 alignment error 4%
not simpler 8%
real simplification 88%

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*

Newsgala (Xu et al., 2015)

- splitting: 16%
- deletion: 40%
- paraphrase: 27%
- deletion+paraphrase: 12%

Newsela-Auto (this work)

- splitting: 21%
- deletion: 9%
- paraphrase: 17%
- deletion+paraphrase: 15%
- splitting+paraphrase: 38%

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

Seq2Seq generation models for text simplification

Train / evaluate

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

Apply the trained alignment model to the entire Newsela and Wikipedia corpora to generate
Experiments on Text Simplification

- Transformer\textsubscript{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*

- LSTM: 35.8 + 0.2
- EditNTS: 35.8 + 0.3
- Transformer<sub>BERT</sub>: 36.6 + 2.2

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

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

Trained on old Newsela (Xu et al., 2015) - Red
Trained on Newsela-Auto (this work) - Green

SOTA
Human Evaluation on Text Simplification*

Trained on old Newsela (Xu et al., 2015)  Trained on Newsela-Auto (this work)

Fluency: 2.91 (Trained on old Newsela) vs. 3.76 (Trained on Newsela-Auto) + 0.85

Adequacy: 2.56 (Trained on old Newsela) vs. 3.21 (Trained on Newsela-Auto) + 0.65

Simplicity: 2.67 (Trained on old Newsela) vs. 3.18 (Trained on Newsela-Auto) + 0.51

Transformer$_{BERT}$ model

(In 5-point Likert scale)

* Evaluate on the Newsela-Auto (this work) test set.
Human Evaluation on Text Simplification*

Transformer\textsubscript{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 and B are partially overlapped**
  - A and B are partially overlapped (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.