
Liam Cripwell, Joel Legrand, Claire Gardent
Existing Metrics for Simplification Models

Reference-based Metrics

Most popular evaluation metrics require *multiple high-quality references*

- something not readily available for simplification
- makes it difficult to evaluate on unseen domains.
Existing Metrics for Simplification Models

Reference-based Metrics

Most popular evaluation metrics require *multiple high-quality references*

- something not readily available for simplification
- makes it difficult to evaluate on unseen domains.

Single-Score Metrics

Most popular metrics use a *single score* that aims to quantify simplicity, meaning preservation and fluency (e.g. SARI, LENS)

- Inverse correlation between meaning preservation and simplicity.
- High scores might mean high faithfulness but low simplicity or vice-versa
We Evaluate

- Document level Simplification Models
- Meaning Preservation and Simplification
- In- and Out-of-Domain
Outline

- Models

- Reference Less metrics for Simplicity and Meaning Preservation

- Data

- Results
  - In domain
  - Out of domain
  - Human Evaluation

- Summary and Open Challenges
Models
Models

One Text-Only Model

- LED_{para}

  Paragraph-level input,
  Longformer

3 Plan-Guided Models conditioned on a simplification plan

- LED_{para}+Plan

  Paragraph-level input,
  Longformer

- PGDyn

  Sentence-level input, BART

- ConBART

  PGDyn conditioned on document context

<table>
<thead>
<tr>
<th>Model</th>
<th>Plan</th>
<th>Input</th>
<th>Document Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>LED_{para}</td>
<td>No</td>
<td>Paragraph</td>
<td>No</td>
</tr>
<tr>
<td>LED_{para}+Plan</td>
<td>Yes</td>
<td>Paragraph</td>
<td>No</td>
</tr>
<tr>
<td>PGDyn</td>
<td>Yes</td>
<td>Sentence</td>
<td>No</td>
</tr>
<tr>
<td>ConBART</td>
<td>Yes</td>
<td>Sentence</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Metrics
Evaluating Meaning Preservation

SummaC

- an NLI entailment-based metric
- compute an NLI entailment matrix between input and output sentences.
- compute score for each output (P) or input (R) sentence
- Sentence scores are then averaged.

QAFactEval

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

Entity Matching between input and output

- R, P and F1
Evaluating Conservativity

- BLEU with respect to the input
- Average lengths of outputs (nb of tokens and sentences)
Evaluating Simplicity

FKGL

- Average length of sentences and syllable count of words in the document

$\epsilon_{SLE_{doc}}$

- Uses a RoBERTa-based simplicity scoring model
- Computes the absolute error of predicted scores compared to target simplicity level
- Average scores over a document's sentences.

Cripwell et al. 2023
Data
Simplification Datasets

Newsela

- High quality
- 1,130 English news articles manually rewritten at five different reading levels (0-4)

→ Training and ID Testing

English Wikipedia

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

→ OOD evaluation
In Domain Evaluation
In Domain Performance - References

- References have highest simplicity (lowest FKGL and best $\epsilon$SLE$_{doc}$)
- All models have higher meaning preservation scores than the references

Models under-simplify and are overly conservative
The End-to-End model (LEDpara, No planning)

- is more meaning preserving
- has worst simplicity performance
- has highest BLEU ('conservativity')
- produces longer outputs than the references

Plan-guidance helps reduce conservativity.
In Domain Performance - Best Models

The best models are plan-based and use a window context to plan (PGdyn, ConBART) and to generate (ConBART)
Out of Domain Evaluation

Training on Newsela, testing on Wiki-Auto
End-to-End Model (no planning) produces very short texts

- different from In-Domain Results (less meaning preserving)
- Could be a result of over-fitting (i.e. being biased towards Newsela paragraph lengths).
- Could also be a result of over-deletion due to a lack of plan-guidance.
Paragraph models produce texts with fewer sentences

- This could indicate less sentence splitting, or an over-deletion of sentences.
Sentence-level models achieve better simplicity and are less meaning preserving than paragraph-based models.

- Mirror ID performance
Human Evaluation
Human Evaluation

- At the paragraph-level

- Evaluators are then asked to judge whether the generated text is fluent, consistent with, and simpler than the input (binary yes/no).

- Sample 250 paragraphs from the test set that contain between 3-6 sentences.

- The proportion of positive ratings is used as the final score.
Same best models as for ID Evaluation

- Plan-based models with window context
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

- Less Meaning Preserving
- Simplify: Length and BLEU close to reference
- Still Conservative; higher faithfulness scores than the references
Brief Summary of Out-Of-Domain Results

Text Only Model (No Planning)

- produces very short texts
- different from In-Domain Results
- overfits to Newsela text length

Plan-Guided models

- have good simplicity and meaning preservation scores
Brief Summary of Human Evaluation

Text Only Models

- underperforms on meaning preservation and simplicity

Plan-Guided Models

- are better overall
Conclusion
Open Challenges for Simplification Evaluation

Trade-off Meaning Preservation / Conservativity / Simplicity

→ Can we define a metric which correctly capture this trade-off?
Open Challenges for Simplification Evaluation

Trade-off Meaning Preservation / Conservativity / Simplicity

→ Can we define a metric which correctly capture this trade-off?

Out-of Domain Evaluation

→ Can we make this metric reference-less?
Open Challenges for Simplification Evaluation

Trade-off Meaning Preservation / Conservativity / Simplicity

→ Can we define a metric which correctly capture this trade-off?

Out-of Domain Evaluation

→ Can we make this metric reference-less?

Multilinguality

→ Can we make this metric multilingual?
Thank You