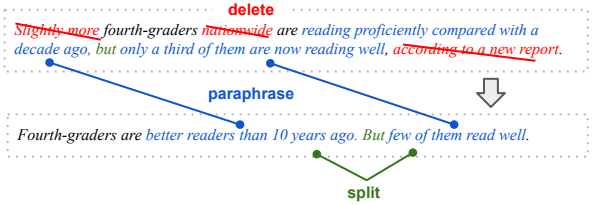
The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification

Fernando Alva-Manchego, Carolina Scarton, Lucia Specia

The University Of Sheffield. | Imperial College London

1. Automatic Text Simplification

Rewrite a text so that it is **easier to understand** while preserving its original meaning



2. Human Judgements of Simplicity

	Simplicity Gain (Xu et al., 2016)	Structural Simplicity (Sulem et al, 2018)	Simplicity-DA
Type of Rating	Discrete (count)	Discrete (Likert scale)	Continuous
Instances	372	1,750	600
ICC	0.176	0.465	0.386
Spearman's *p*	0.299	0.508	0.607

NEW!!

- direct assessments of **general simplicity** (vs. operation-specific)
- system outputs include **neural state-of-the-art** simplification models
- good **annotation reliability**

See the paper for details on the data collection and further comparisons between the datasets

3. Meta-Evaluation of Automatic Metrics

Find similar analysis with the other datasets in the paper

Simplicity-DA

Low scores indicate "bad" quality of a simplification, but high scores do not necessarily imply "good" quality

Type of Metric	Metric	ASSET (10 references) Low	ASSET (10 references) High	ASSET (10 references) All	ASSET + TurkCorpus + HSplit (22 references) Low	ASSET + TurkCorpus + HSplit (22 references) High	ASSET + TurkCorpus + HSplit (22 references) All	Selected References (Different refs. per instance acc. to the operations performed) Low	Selected References (Different refs. per instance acc. to the operations performed) High	Selected References (Different refs. per instance acc. to the operations performed) All
Reference-based	$BERTScore_P$	0.512	0.287	**0.617**	0.541	0.280	**0.629**	0.543	0.276	**0.635**
	BLEU-SARI (AM)	0.417	0.239	0.503	0.418	0.218	0.519	0.418	0.221	0.523
	BLEU	0.405	0.235	0.496	0.404	0.230	0.526	0.402	0.223	0.525
	SARI	0.336	0.139	0.359	0.366	0.097	0.353	0.352	0.115	0.350
Non-Reference-based	FKGL	0.272	0.093	0.117						
	SAMSA	0.103	0.010	0.058						

Having more references slightly improves correlations for most metrics

Selecting specific references for each system output could improve correlations

4. Recommendations

Which current metrics should be used?

1. Use BERTScore to ensure the output is of reasonable quality
2. Use SARI and SAMSA to verify specific types of gains
3. Prefer human evaluation for more accurate conclusions

Development of New Metrics

- Simplicity-DA vs Simplicity Gain vs Structural Simplicity → ***What are better ways to ask about simplicity?***
- Combine the best characteristics of current metrics → ***Take the original sentence into account!***
- Enrich manual references with meta-information
  - → ***How simple is each reference?***
  - → ***Which operations were applied to create each of them?***

https://github.com/feralvam/metaeval-simplification