The University Of Sheffield.

Imperial College London

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

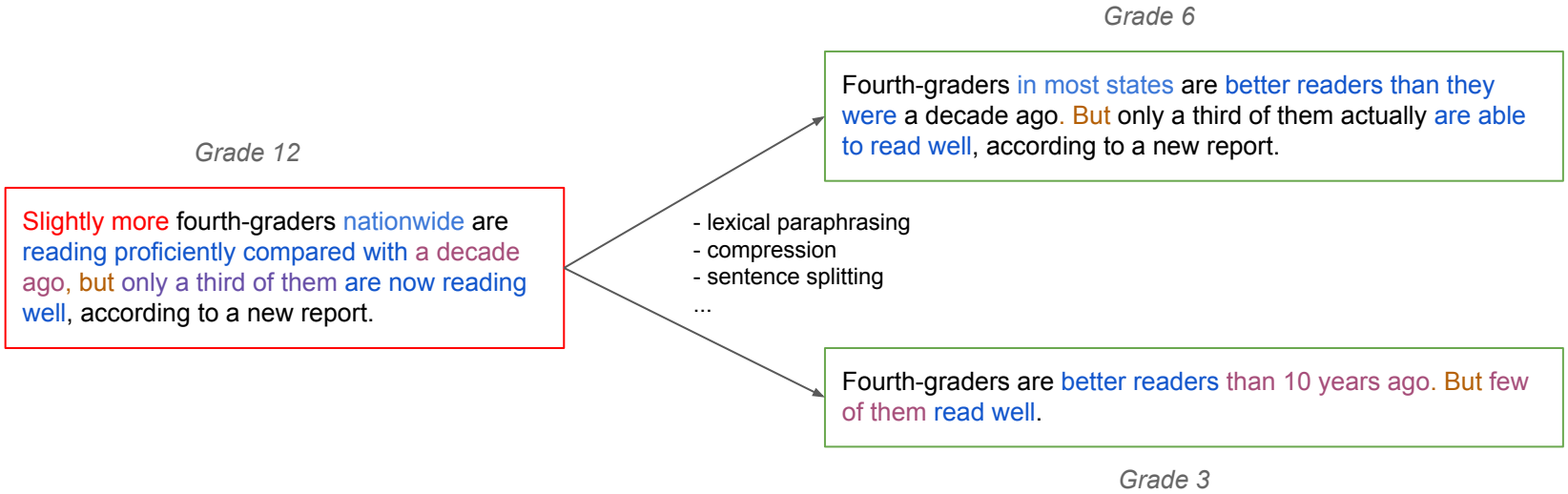
Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia

EMNLP 2021

7-11 November 2021

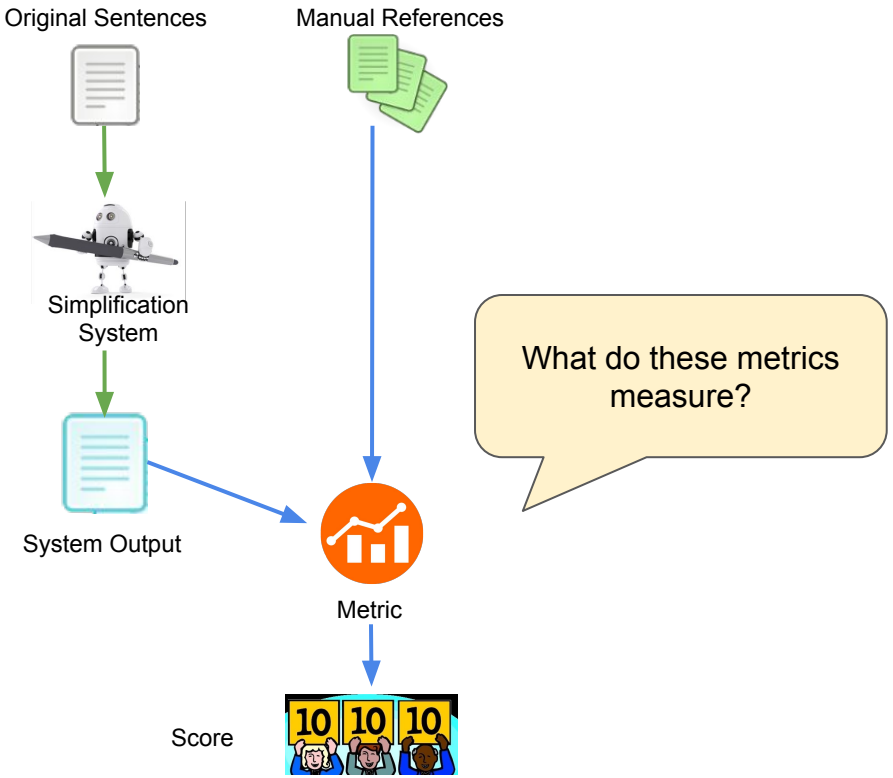
What is Text Simplification?

Modify the content and structure of a text so that it is **easier to understand** while preserving its original meaning



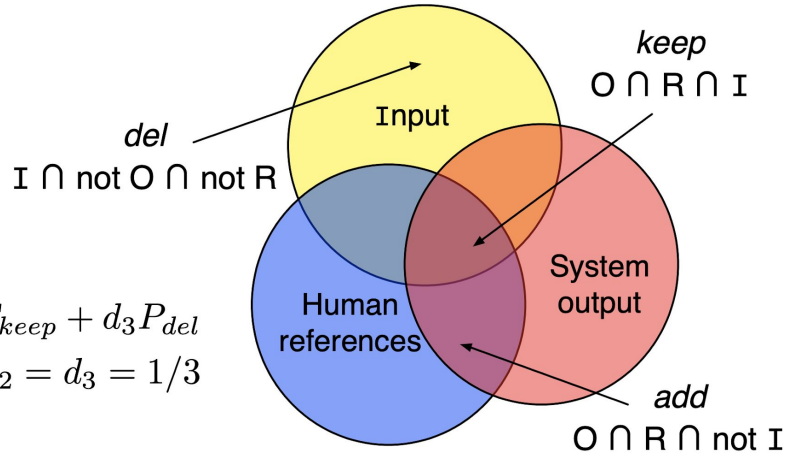
Examples from: Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. *Problems in Current Text Simplification Research: New Data Can Help.*Transactions of the Association for Computational Linguistics, 3:283–297.

Standard Automatic Evaluation Pipeline



SARI (Xu et al., 2016)

Lexical Paraphrasing



$$\text{SARI} = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$
$$d_1 = d_2 = d_3 = 1/3$$

Input: About 95 species are currently accepted.

REF-1: About 95 species are currently known .
REF-2: About 95 species are now accepted .
REF-3: 95 species are now accepted .

Output-1: About 95 you now get in . → 0.2683
Output-2: About 95 species are now agreed . → 0.7594
Output-3: About 95 species are currently agreed. → 0.5890

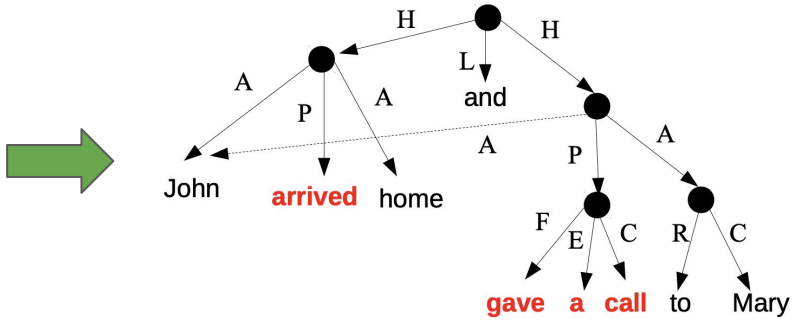
SAMSA (Sulem et al., 2018)

Sentence Splitting

Assumption: In an ideal simplification each event is placed in a different sentence.

Original Sentence:

John arrived home and gave a call to Mary.



System Output:

John arrived home. John called Mary.

Score:
1.0

Readability Indices

- **Flesch Reading Ease** (Flesch,1948)

$$FRE = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$$

- **Flesch-Kincaid Grade Level** (Kincaid et al., 1975)

$$FKGL = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$$

Metrics used in Machine Translation

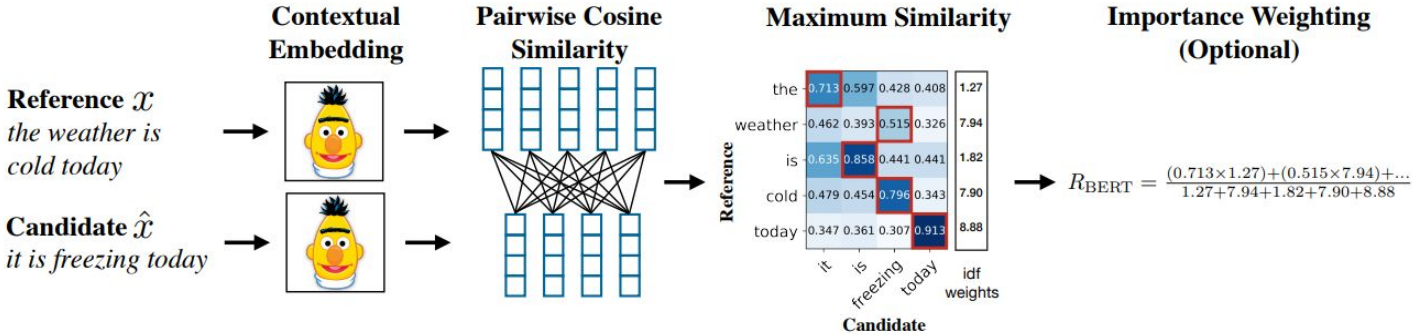
- **BLEU** (Papineni et al., 2002)

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

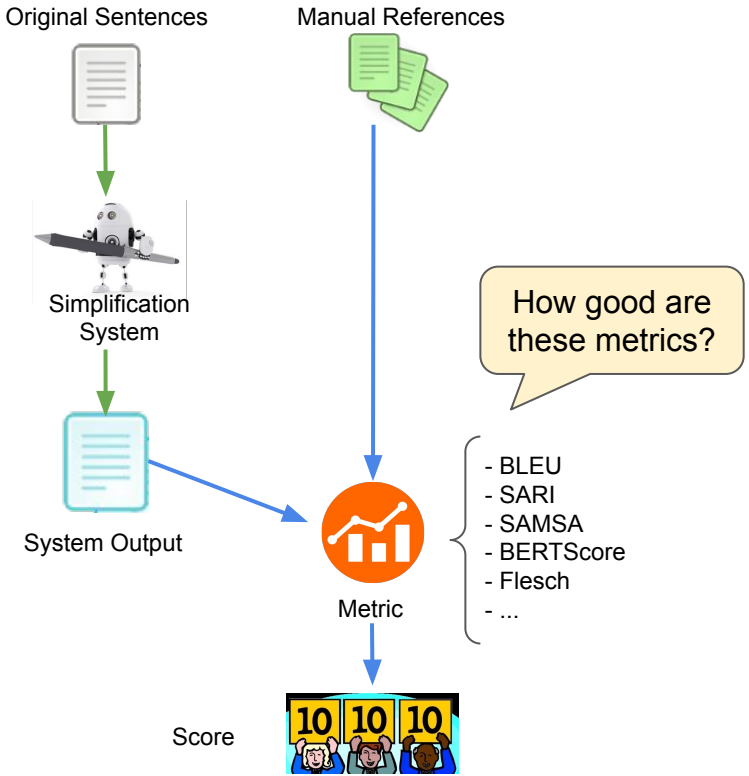
$$BP = \begin{cases} 1 & if \ \ c > r \\ e^{1-\frac{r}{c}} & if \ \ c \le r \end{cases}$$

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

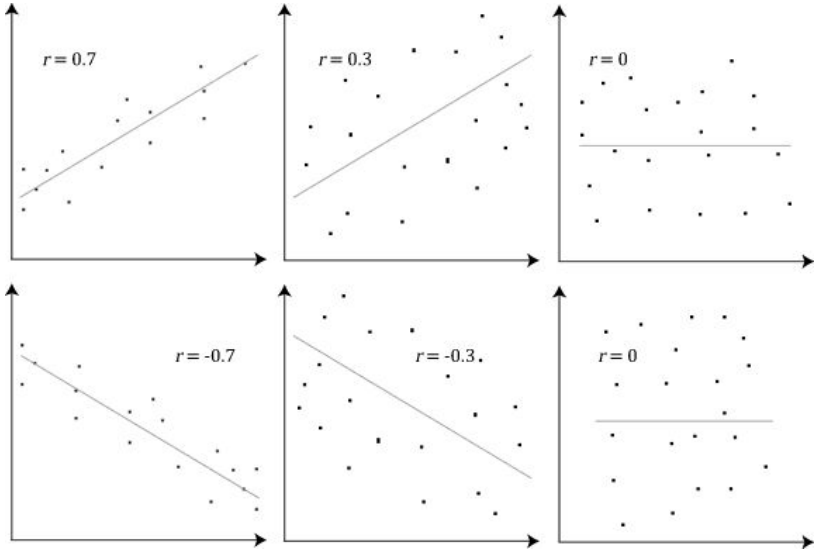
- **BERTScore** (Zhang et al., 2020)



High Correlation = “Good” Metric?



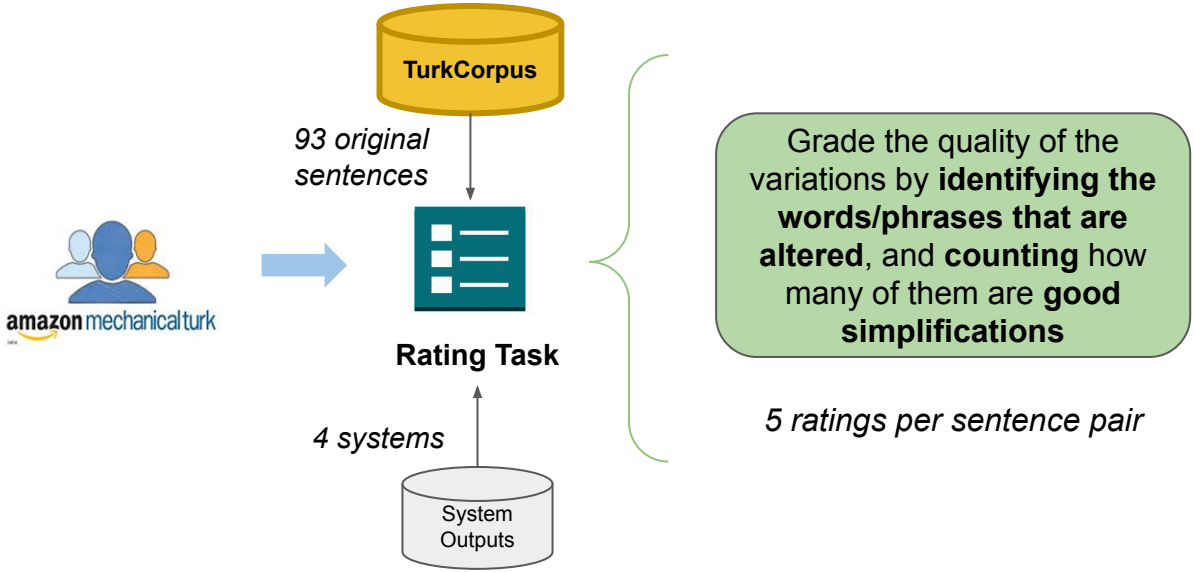
Correlation between Metrics and Human Judgements



Simplicity Gain

Lexical Paraphrasing

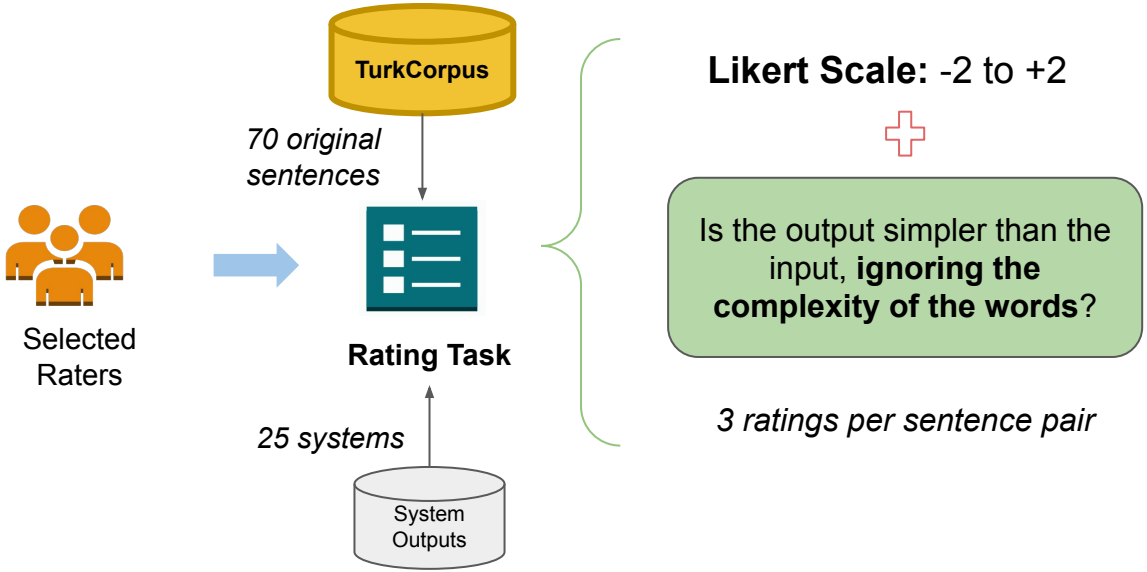
SARI



Structural Simplicity

Sentence Splitting

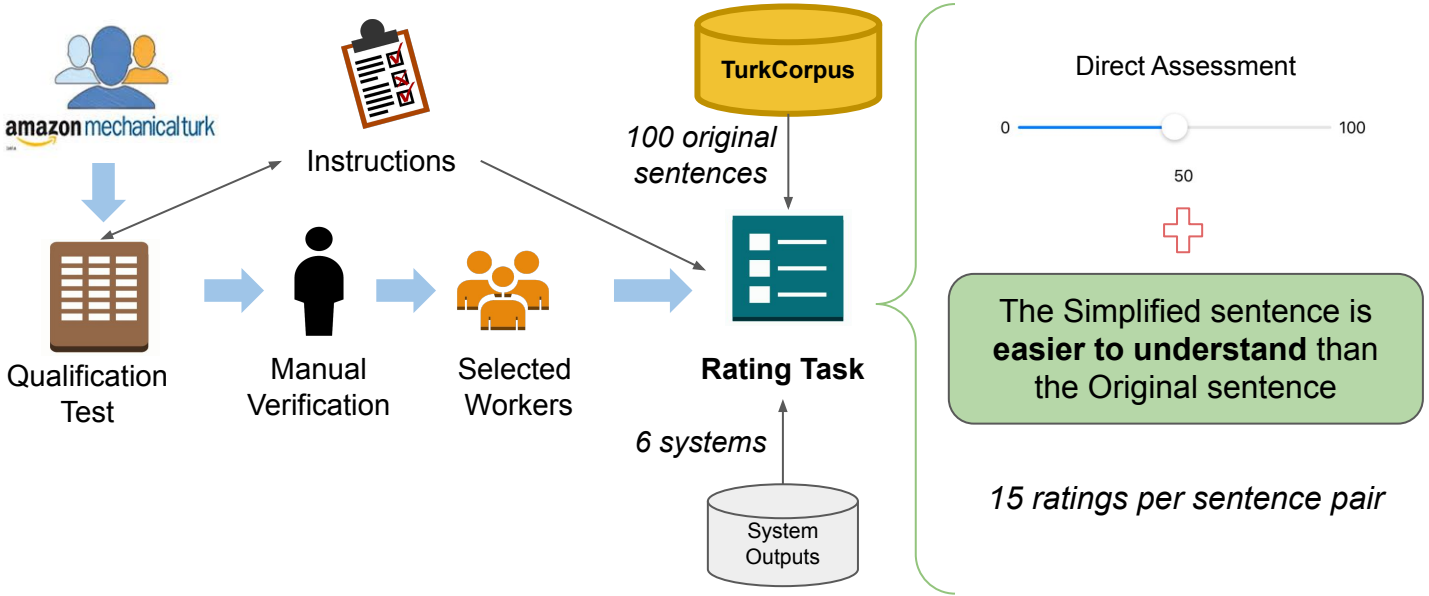
SAMSA



Simplicity-DA

NEW!

General Simplicity



Meta-Evaluation of Automatic Metrics

Experimental Setting

- Study the behaviour of automatic metrics at the **sentence-level**
- Focused on metrics that measure (some form of) **simplicity**
- Analyse the variation of correlation w.r.t.
  a. Simplicity levels
  b. System type
  c. Set of manual references
- Metrics
  a. SARI, SAMSA, FKGL, BLEU, BERTScore
  b. Averages of BLEU, SARI, SAMSA

Metrics across Simplicity Levels

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

Simplicity-DA

	Metric	Low (N = 300)	High (N=300)	All (N=600)
Reference-based (using ASSET)	$BERTScore_P$	0.512	0.287	**0.617**
	$BERTScore_{F1}$	0.518	0.224	0.573
	BLEU-SARI (AM)	0.417	0.239	0.503
	$BERTScore_R$	0.471	0.172	0.500
	BLEU	0.405	0.235	0.496
	BLEU-SARI (GM)	0.408	0.215	0.476
	SARI	0.336	0.139	0.359
Non-Reference-based	FKGL	0.272	0.093	0.117
	SAMSA	0.103	0.010	0.058

Metrics across Simplicity Levels

Differences are not as considerable as observed for Simplicity-DA

Simplicity Gain

	Metric	Low (N = 186)	High (N=186)	All (N=372)
Reference-based (using TurkCorpus)	$BERTScore_P$	0.209	0.231	0.241
	$BERTScore_{F1}$	0.215	0.236	0.247
	BLEU-SARI (AM)	0.223	0.172	0.187
	$BERTScore_R$	0.221	0.217	0.241
	BLEU	0.178	0.132	0.123
	BLEU-SARI (GM)	0.246	0.177	0.214
	SARI	0.292	0.240	**0.331**
Non-Reference-based	FKGL	0.045	0.101	0.147
	SAMSA	0.120	0.042	0.013

Metrics across Simplicity Levels

BERTScore$_P$ is only the best when scoring "low" quality simplifications

Structural Simplicity

	Metric	Low (N = 875)	High (N=875)	All (N=1750)
Reference-based (using HSplit)	BERTScore$_P$	**0.552**	0.310	0.090
	BERTScore$_{F1}$	0.483	0.529	0.325
	BLEU-SARI (AM)	0.346	0.599	0.431
	BERTScore$_R$	0.411	0.601	0.430
	BLEU	0.421	**0.643**	0.443
	BLEU-SARI (GM)	0.329	0.589	0.438
	SARI	0.137	0.418	0.313
Non-Reference-based	FKGL	0.070	0.165	0.228
	SAMSA	0.103	0.431	0.284

Metrics across System Types

Encouraging results considering the current trend in simplification models

Simplicity-DA

	Metric	SBMT (N = 100)	PBMT (N=100)	NMT (N=300)	Sem+PBMT (N=100)
Reference-based (using ASSET)	$BERTScore_P$	0.537	0.459	**0.650**	**0.624**
	$BERTScore_{F1}$	0.528	0.400	0.588	0.568
	BLEU-SARI (AM)	0.315	0.336	0.536	0.335
	$BERTScore_R$	0.527	0.375	0.484	0.470
	BLEU	0.295	0.347	0.546	0.333
	BLEU-SARI (GM)	0.298	0.320	0.508	0.308
	SARI	0.228	0.173	0.310	0.240
Non-Reference-based	FKGL	0.055	0.063	0.104	0.062
	SAMSA	0.184	0.067	0.126	0.248

Effect of Simplification References

All metrics (but SARI) improve their correlations

Simplicity-DA

	ASSET (10 references)			ASSET + TurkCorpus + HSplit (22 references)			Selected References (Different refs. per instance according to the operations performed)		
Metric	**Low**	**High**	**All**	**Low**	**High**	**All**	**Low**	**High**	**All**
$BERTScore_P$	0.512	0.287	**0.617**	0.541	0.280	**0.629**	0.543	0.276	**0.635**
$BERTScore_{F1}$	0.518	0.224	0.573	0.530	0.202	0.576	0.534	0.202	0.584
BLEU-SARI (AM)	0.417	0.239	0.503	0.418	0.218	0.519	0.418	0.221	0.523
$BERTScore_R$	0.471	0.172	0.500	0.476	0.165	0.506	0.479	0.165	0.511
BLEU	0.405	0.235	0.496	0.404	0.230	0.526	0.402	0.223	0.525
BLEU-SARI (GM)	0.408	0.215	0.476	0.410	0.195	0.490	0.410	0.205	0.496
SARI	0.336	0.139	0.359	0.366	0.097	0.353	0.352	0.115	0.350

Recommendations for Automatic Evaluation

- **Which automatic metric(s) should be used?**
  - Use multiple metrics, and mainly BERTScore_P
- **Which manual references should the metric(s) compare against?**
  - References in ASSET seem to be enough
- **How should the automatic scores be interpreted?**
  - First, use BERTScore_P to ensure that the output is of high quality
  - Then use SARI and/or SAMSA to verify specific gains
  - However, human evaluation should be preferred for final conclusions
    - Task-based evaluation?

Check the paper for recommendations on development of new metrics

Development of New Metrics

Is the way we evaluate simplicity adequate for the goals of the task?

- **Collecting More Human Judgements**
  - Simplicity-DA offers flexibility but is more subjective
  - Simplicity Gain and Structural Simplicity require more quality control
- **Combining the best characteristics of current ones**
  - Similarity based on contextual word embeddings, as in BERTScore
  - Take the input sentence into account, as in SARI and SAMSA
- **Enrich manual references**
  - Inform of the simplicity level of the references
  - Identify (manually) the operations that were performed

Contributions

- A **new dataset for evaluation of automatic metrics** following the Direct Assessment methodology
- The **first meta-evaluation** of Sentence Simplification metrics
  - Metrics can more reliably score low-quality simplifications
  - Correlations change depending on system type
  - More references could improve correlations (better to be smart with selecting references!)
- **Recommendations for automatic evaluation** of current simplification models

Thanks!

Datasets and scripts available in: https://github.com/feralvam/metaeval-simplification

Fernando Alva-Manchego

@feralvam

https://feralvam.github.io/

Datasets with Human Judgements on 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
System Types	PBMT SBMT	PBMT SBMT NMT Sem Sem+PBMT Sem+NMT	PBMT SBMT NMT Sem+PBMT
ICC	0.176	0.465	0.386
Spearman's *p*	0.299	0.508	0.607

Includes SotA