DocAsRef: An Empirical Study on Repurposing Reference-based Summary Quality Metrics as Reference-free Metrics

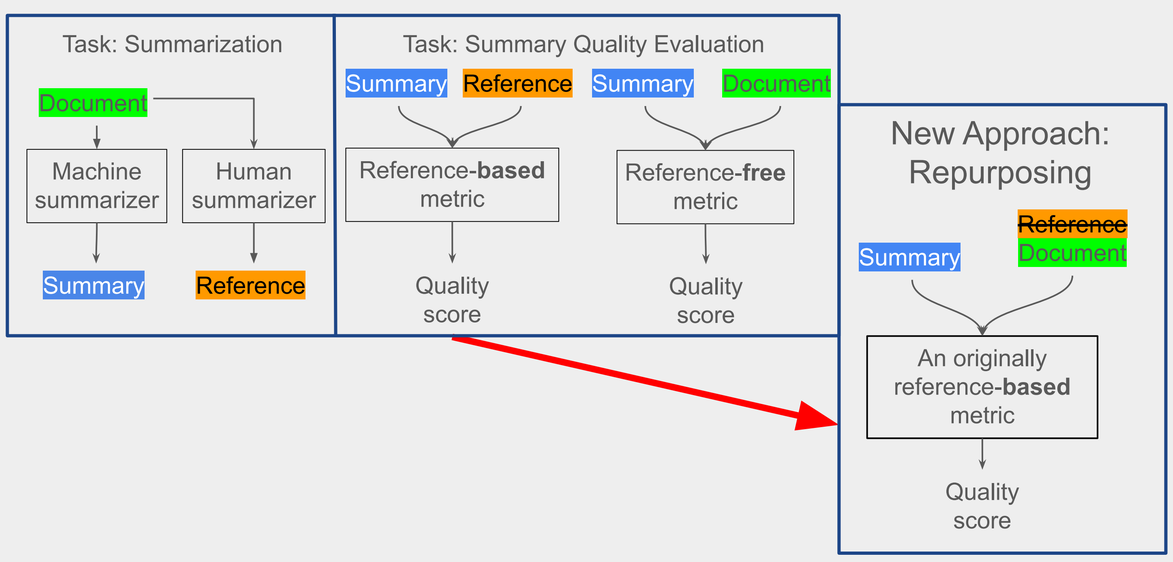
Forrest Sheng Bao[▷◁◇], Ruixuan Tu[▷◁♡], Ge Luo[◇], Yinfei Yang[♣], Hebi Li[◇], Minghui Qiu[♠], Youbiao He[◇], and Cen Chen[¶]

[◇]Department of Computer Science, Iowa State University, Ames, IA, USA
[♡]Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA
[♣]Sunnyvale, CA, USA [♠]ByteDance, China
[¶]School of Data Science and Engineering, East China Normal University, Shanghai, China
[▷◁]Equal contribution, forrest.bao@gmail.com, ruixuan@cs.wisc.edu

SigmaWe/DocAsRef

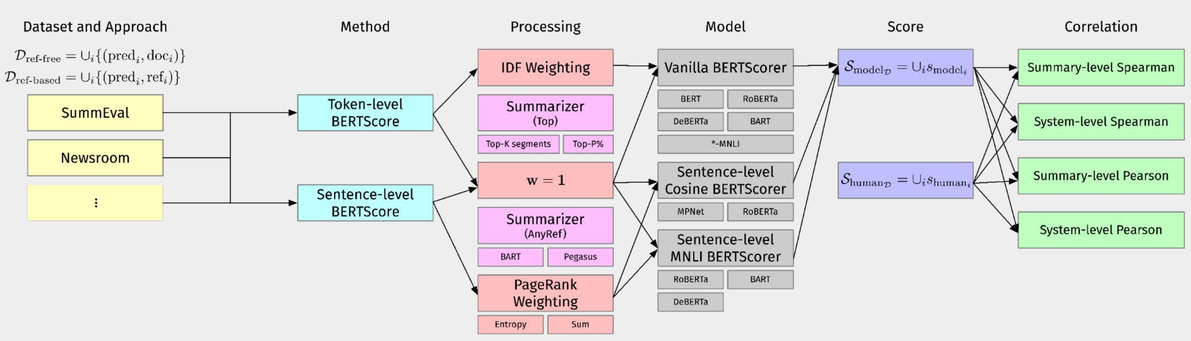
Motivation

- Motivation
  - Reference-free metrics: simple but not scalable.
  - Reference-based metrics: complex but not scalable.
  - **Q:** Can we bring the best of both worlds together?
- Idea (Repurposing): f(system summary, reference) → f(system summary, document) (use **doc**ument **as ref**erence)



Settings

- Tables: report Spearman's correlation coefficients at summary-level; font in tables: best in each column in **bold**, 2nd best underlined.
- Tweaks for BERTScore
  - try -base and -large LMs (RoBERTa, DeBERTa, BART), for both versions of pre-trained and fine-tuned for MNLI dataset
  - expanding BERTScore to sentence level by similarity between sentences instead of tokens (perform worse than token-level)
- Aspects to calculate correlation coefficients
  - SummEval: CONsistency, RELevance, COHerence, FLUency
  - Newsroom: INFormativeness, RELevance, COHerence, FLUency



Single-document Summarization

BERTScore: Before and After Re-purposing

- SummEval: coverage of 23 modern summarizers, many of which exhibit highly similar behavior
- Newsroom: coverage of 7 systems with distinct performances

Upper: reference-free; Lower: reference-based; **U > L**
Transformer-based reference-free metrics (BERTScore, MoverScore, BLEURT) become **more accurate after repurposing**.

	SummEval				Newsroom			
	CON	REL	COH	FLU	INF	REL	COH	FLU
BERTScore P	**0.318**	0.375	**0.471**	**0.265**	0.611	0.591	0.633	0.591
BERTScore R	0.235	0.343	0.258	0.162	**0.750**	**0.658**	0.659	0.590
BERTScore F	0.308	**0.401**	0.416	0.241	0.689	0.617	**0.663**	**0.618**
MoverScore	0.180	0.245	0.138	0.093	0.695	0.615	0.589	0.537
ROUGE-1 R	0.145	0.128	0.002	0.067	0.744	0.639	0.564	0.476
ROUGE-2 R	0.262	0.155	0.049	0.163	0.746	0.648	0.591	0.511
ROUGE-L R	0.289	0.187	0.106	0.183	0.746	0.641	0.591	0.515
BLEURT	0.221	0.252	0.336	0.172	0.549	0.507	0.596	0.562
BERTScore P	0.008	0.208	0.275	0.083	-0.034	0.012	0.044	0.045
BERTScore R	0.158	0.355	0.284	0.148	0.315	0.294	0.311	0.320
BERTScore F	0.088	0.301	0.321	0.139	0.149	0.171	0.185	0.187
MoverScore	0.129	0.238	0.088	0.096	0.136	0.153	0.112	0.077
ROUGE-1 R	0.148	0.250	0.117	0.109	0.105	0.128	0.071	0.073
ROUGE-2 R	0.166	0.194	0.109	0.102	0.069	0.087	0.016	0.037
ROUGE-L R	0.123	0.205	0.146	0.099	0.035	0.063	0.016	0.025
BLEURT	0.048	0.215	0.174	0.087	0.154	0.140	0.071	0.075

BERTScore vs. Baselines

SummEval

Repurposed BERTScore **>> all non-GPT baselines**, **comparable to GPT3.5-based baselines** on RELevance and COHerence. **>** one GPT-3.5-based approach on CONsistency.

	CON	REL	COH	FLU
BERTScore, repurposed, using respective LMs below				
RoBERTa-large P	0.318	0.375	**0.471**	0.265
RoBERTa-large F	0.308	0.401	0.416	0.241
RoBERTa-large-MNLI P	0.387	0.358	0.438	0.287
RoBERTa-large-MNLI F	0.357	0.382	0.373	0.241
DeBERTa-large P	0.338	0.341	0.418	0.280
DeBERTa-large F	0.289	0.357	0.315	0.211
DeBERTa-large-MNLI P	0.399	0.293	0.351	0.303
DeBERTa-large-MNLI F	0.344	0.333	0.291	0.239
Best of Repurposed BERTScore	0.399	0.401	0.471	0.303
Baselines, reference-free				
Blanc	0.244	0.197	0.089	0.132
SummaQA-F1	0.197	0.165	0.123	0.140
SUPERT	0.330	0.216	0.120	0.230
SueNes	0.190	0.177	0.167	0.228
ChatGPT (Wang et al., 2023)	**0.432**	**0.428**	0.470	0.353
G-Eval (GPT-3.5) (Liu et al., 2023)	0.386	0.385	0.440	**0.424**
Best of Baselines	0.432	0.428	0.470	0.424

Newsroom

Repurposed BERTScore **> all baselines except SueNes**, which is fine-tuned using data explicitly augmented for the summary evaluation task.

	INF	REL	COH	FLU
BERTScore, repurposed, using respective LMs below				
RoBERTa-large R	0.750	**0.658**	0.659	0.590
RoBERTa-large F	0.689	0.617	0.663	0.618
RoBERTa-large-MNLI R	0.737	0.621	0.632	0.550
RoBERTa-large-MNLI F	0.680	0.582	0.641	0.563
DeBERTa-large R	0.747	0.646	**0.669**	0.604
DeBERTa-large F	0.720	0.625	0.676	0.613
DeBERTa-large-MNLI R	0.748	0.629	0.668	0.583
DeBERTa-large-MNLI F	0.739	0.635	0.674	0.595
Best of Repurposed BERTScore	0.750	0.658	0.669	0.618
Baselines, reference-free				
Blanc	0.688	0.608	0.586	0.531
SummaQA-F1	0.569	0.516	0.490	0.466
SUPERT	0.693	0.605	0.617	0.539
SueNes	**0.753**	0.647	**0.669**	**0.674**
ChatGPT (Gao et al., 2023)	0.521	0.524	0.484	0.480
ChatGPT (Wang et al., 2023)	0.578	0.461	0.469	0.507
Best of Baselines	0.753	0.647	0.669	0.674

Multi-document Summarization

TAC2010 provides 10 source docs $d_1, \ldots, d_{10}$ for generating a system summary s. Heuristic given s and single-doc metric f:

$$\text{score}(s) = \sum_{i \in [1..10]} f(d_i, s)$$

	Pyramid	Linguistic	Overall
BERTScore, repurposed, using respective LMs below			
DeBERTa-large-MNLI P	0.496	0.401	0.455
DeBERTa-large-MNLI R	0.526	0.405	0.492
DeBERTa-large-MNLI F	**0.539**	0.422	**0.500**
BART-large-MNLI P	0.471	0.272	0.415
BART-large-MNLI R	0.422	0.202	0.380
BART-large-MNLI F	0.481	0.245	0.426
RoBERTa-large-MNLI P	0.469	0.306	0.418
RoBERTa-large-MNLI R	0.481	0.340	0.450
RoBERTa-large-MNLI F	0.509	0.356	0.464
Baselines, reference-free			
Blanc	0.427	0.294	0.397
SummaQA-F1	0.301	0.243	0.286
SUPERT	0.479	0.324	0.427
SueNes	0.492	**0.460**	0.470

> all baselines except Linguistic aspect

Ablation Study

- IDF makes a very small impact
- In many cases, IDF even decreases the performance

Table 5: The performance of BERTScore-P with and without IDF. Summary-level Spearman's correlation coefficients in comparison. Model size: base. Yellow cells are when using IDF is worse than without IDF and green cells are for the opposite.

		SummEval				Newsroom			
		CON	REL	COH	FLU	INF	REL	COH	FLU
IDF on	RoBERTa	0.295	0.284	0.381	0.228	0.627	0.579	0.589	0.536
	BART	0.279	0.283	0.359	0.208	0.673	0.631	0.664	0.620
	DeBERTa	0.262	0.252	0.316	0.206	0.614	0.556	0.613	0.544
IDF off	RoBERTa	0.307	0.315	0.408	0.240	0.597	0.551	0.579	0.531
	BART	0.291	0.322	0.390	0.233	0.675	0.650	0.661	0.610
	DeBERTa	0.281	0.276	0.345	0.221	0.628	0.587	0.631	0.586

arXiv abs/2212.10013