LREC-COLING 2024

THE 2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION

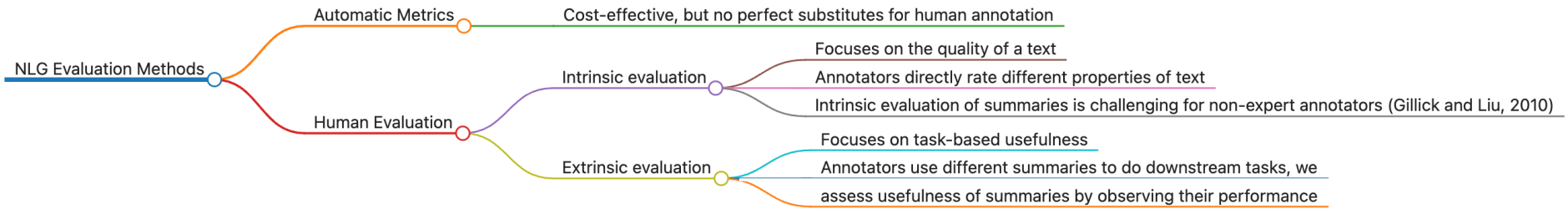
Is Summary Useful or Not?

An Extrinsic Human Evaluation of Text Summaries on Downstream Tasks

Xiao Pu Mingqi Gao Xiaojun Wan

北京大學
PEKING UNIVERSITY

Motivation


NLG Evaluation Methods
Automatic Metrics
Cost-effective, but no perfect substitutes for human annotation
Human Evaluation
Intrinsic evaluation
Focuses on the quality of a text
Annotators directly rate different properties of text
Intrinsic evaluation of summaries is challenging for non-expert annotators (Gillick and Liu, 2010)
Extrinsic evaluation
Focuses on task-based usefulness
Annotators use different summaries to do downstream tasks, we
assess usefulness of summaries by observing their performance

Methodology

- A summary is useful if it can facilitate users to complete a task
- 2 dimensions of usefulness: time and correctness
- We design different downstream tasks to represent diverse real-world applications of summaries:

TASK	METRIC
Question answering	Answerable, EM, F1
Classification	EM, F1
Similarity assessment	MSE, Spearman's ρ

Experimental Setting -- Dataset

QA

randomly collect 100 pairs of source articles and reference summaries from CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016).

We then annotate two datasets: QA-ref (QA-pairs are written according to the reference summaries) and QA-src (QA-pairs are written according to source articles).

Classification

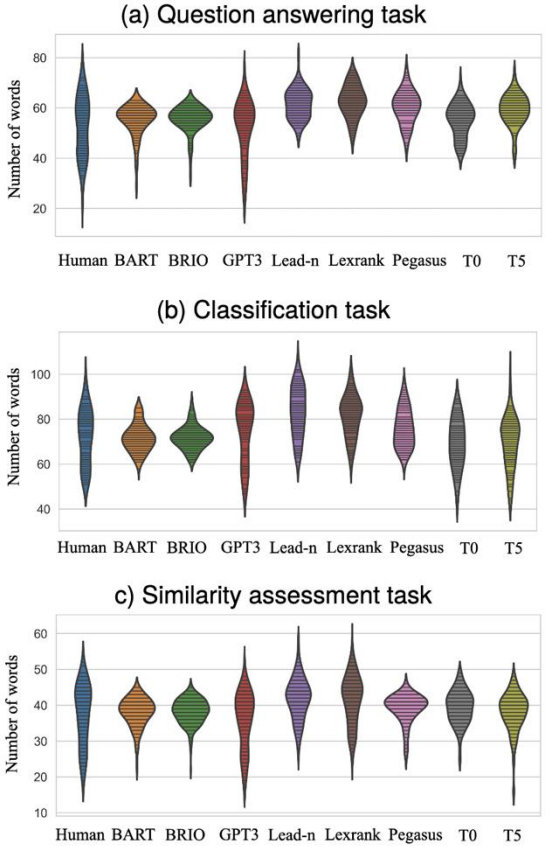
randomly sample 100 news articles from New York Times Annotated Corpus (Sandhaus, 2008). Each article is paired with one or multiple tags.

Similarity Assessment

we use the SemEval-2022 Task 8 dataset (Chen et al., 2022) to collect 100 pairs of news articles with summaries and similarity scores.

Experimental Setting – Models and Details

- Include 8 summarization models: BART, Pegasus, Lexrank, Lead-n*, BRIO, T5, T0, GPT3
- To ensure fairness in comparing summaries across different systems, we generate summaries of similar lengths for each task
- Include 20 university students proficient in English in the experiment



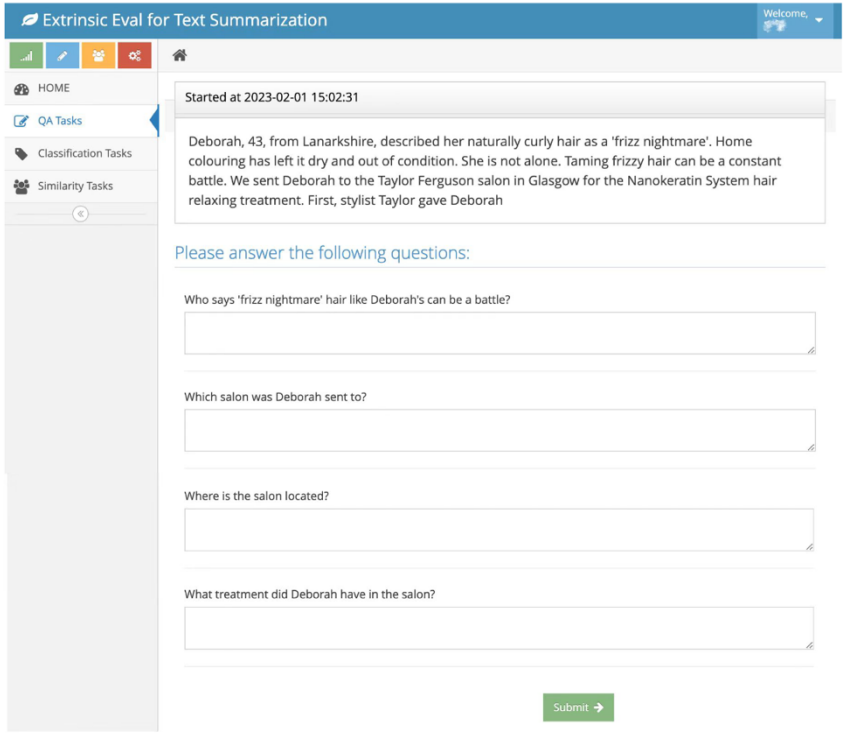
Length of summaries from different systems in three tasks

* We modify the Lead-3 setting and refer to it as the Lead-n model, which selects the first several sentences that are closest to the summary length we set.

Experimental Setting – Platform

A Web-based Platform for Evaluation

- Offers guidelines for annotators
- Collect experiment data, including the answers and completion time of each question
- prohibits the utilization of the copy-paste/search functionality to guarantee impartiality



A screenshot of our platform

Results: Evaluating Summaries' Usefulness

RQ1: How useful are text summaries compared to source articles?

- The use of summaries generally reduces the completion time.
- Summaries are particularly useful in classification and similarity tasks, with higher correctness on tasks.

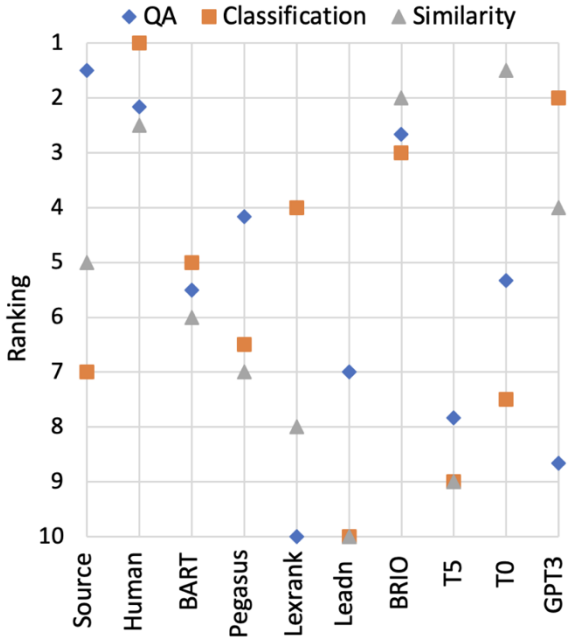
	QA (ref-based)								QA (source-based)							
	Answerable		EM		F1		Time(seconds)		Answerable		EM		F1		Time(seconds)	
Source	0.86		0.32		0.51		280		0.89		0.51		0.7		212	
Human Summaries	0.89	+4%	0.54	+67%	0.75	+48%	94	-66%	0.54	-39%	0.27	-46%	0.4	-45%	88	-58%
All Summaries	0.52	-39%	0.22	-32%	0.33	-36%	106	-62%	0.52	-41%	0.24	-53%	0.3	-52%	83	-61%

	Classification						Similarity					
	EM		F1		Time(seconds)		MSE		Spearman's ρ		Time(seconds)	
Source	0.88		0.90		73		0.91		0.6		38	
Human Summaries	0.91	+3%	0.92	+2%	34	-53%	0.77	-15%	0.7	+14%	20	-47%
All Summaries	0.89	+1%	0.90	-	30	-59%	1.02	+11%	0.6	-	22	-42%

Summaries compared to source texts in the downstream tasks. The green percentages indicate that summaries are more useful compared to the source text, i.e. participants take less time or perform better. The red ones indicate less useful.

RQ2: Which summarization systems are more useful?

- Divide the summarization models into fine-tuned, zero-shot, and simple extractive.
- Fine-tuned models have higher consistency in usefulness across different tasks, and are less sensitive to differences between tasks.
- Zero-shot and simple extractive methods exhibit a varying ranking across tasks.



Average ranking of different systems on three different tasks. Each ranking is calculated by averaging the rankings over extrinsic metrics for the same task

RQ3: What kind of summaries are more useful?

Explore how the inner features of summaries influence their usefulness in tasks

Inner properties and their metrics:

- Summary Style (abstractive or extractive): we employ the Ext-cvg (Extractive Fragment Coverage) to assess the extractiveness of summaries
- Grammaticality: the ratio of grammar errors in the summaries
- Sentence length: average number of words per sentence of summaries

Findings:

Abstractive summaries tend to be more useful for classification and similarity tasks. Grammatically correctness and shorter sentences contribute to more useful summaries in QA and similarity tasks.

	Ext-Cvg(%)	Errors(%)	Sent-Len
Ref	87.51	10.89	16.95
BART	98.83	5.13	17.91
BRIO	96.60	3.66	15.96
GPT3	93.01	2.78	24.12
Lead-n	100.00	9.09	28.24
Lexrank	100.00	10.82	29.10
Pegasus	98.96	5.28	17.83
T0	94.97	3.20	18.29
T5	96.44	9.17	16.18

intrinsic features of summaries from different systems

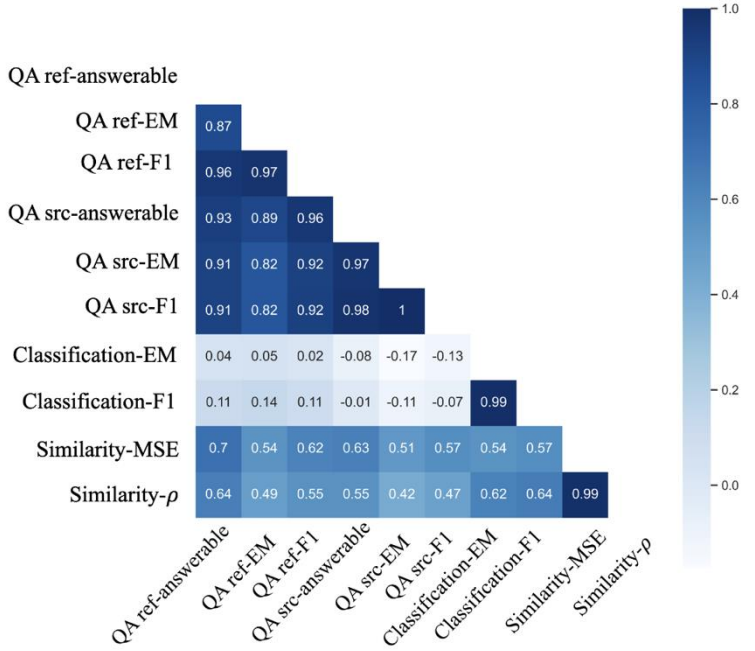
	Ext-Cvg	Errors	Sent-Len
qa_EM	-0.223	-0.440	-0.595
qa_F1	-0.291	-0.441	-0.597
cls_EM	-0.602	0.120	-0.090
cls_F1	-0.591	0.072	-0.143
sim_MSE	-0.641	-0.603	-0.551
sim_Spearman's ρ	-0.642	-0.597	-0.507

System-level pearson correlation between intrinsic features and our extrinsic metrics

Results: Correlation between Metrics

(1) Analyzing the relationships between our extrinsic metrics:

- Extrinsic metrics within the same task are highly correlated.
- There are only weak to moderate correlations among tasks, meaning that the tasks involved are diverse, reflecting different perspectives of usefulness



System-level Pearson correlation of our extrinsic metrics

(2) Evaluating automatic metrics using extrinsic criteria:

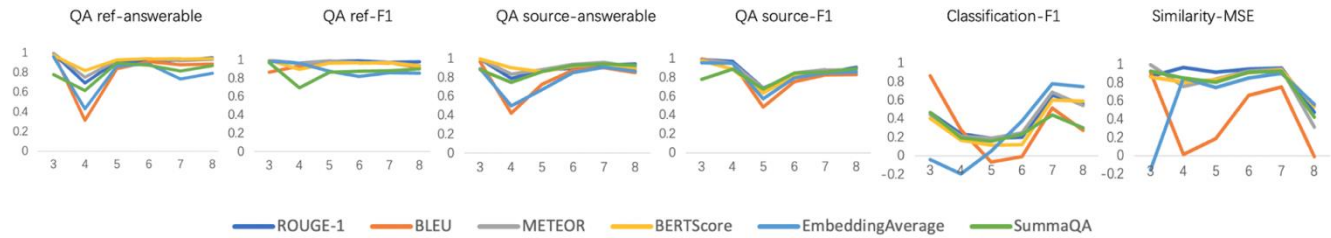
- Automatic metrics can well reflect the usefulness of summaries in the QA task, but their correlations with extrinsic metrics are generally low for classification and similarity tasks.

Extrinsic Criteria / Automatic Metrics	QA (ref-based)						QA (source-based)						Classification				Similarity			
	answerable		EM		F1		answerable		EM		F1		EM		F1		MSE		ρ	
	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ
ROUGE-1	0.95**	0.71*	0.94**	0.76**	0.98**	0.86**	0.95**	0.79**	0.89**	0.64*	0.91**	0.64*	0.51	0.50	0.56	0.50	0.48	0.43	0.40	0.36
ROUGE-2	0.97**	0.79**	0.94**	0.91**	0.98**	0.93**	0.92**	0.71*	0.89**	0.71*	0.89**	0.71*	0.23	0.21	0.29	0.21	0.18	0.29	0.10	0.36
ROUGE-L	0.99**	0.93**	0.93**	0.76**	0.97**	0.79**	0.91**	0.71*	0.87**	0.71*	0.87**	0.71*	0.33	0.43	0.40	0.43	0.29	0.29	0.22	0.36
BLEU	0.89**	0.64*	0.88**	0.84**	0.92**	0.93**	0.85**	0.71*	0.83*	0.71*	0.83*	0.71*	0.21	0.21	0.28	0.21	-0.01	0.14	-0.08	0.21
METEOR	0.93**	0.64*	0.88**	0.84**	0.94**	0.79**	0.91**	0.86**	0.87**	0.71*	0.89**	0.71*	0.49	0.50	0.54	0.50	0.31	0.36	0.24	0.29
CHRF	0.95**	0.64*	0.90**	0.84**	0.96**	0.93**	0.91**	0.71*	0.88**	0.71*	0.89**	0.71*	0.48	0.50	0.52	0.50	0.31	0.29	0.23	0.36
CIDEe	0.75*	0.50	0.83**	0.69*	0.85**	0.79**	0.82*	0.71*	0.82*	0.57	0.83*	0.57	0.12	0.00	0.20	0.00	-0.03	0.07	-0.09	0.00
BERTScore	0.94**	0.71*	0.87**	0.62*	0.93**	0.71*	0.89**	0.93**	0.85**	0.79**	0.86**	0.79**	0.54	0.43	0.59	0.43	0.54	0.43	0.48	0.36
MOVERScore	0.97**	0.79**	0.93**	0.69*	0.97**	0.79**	0.93**	0.86**	0.87**	0.71*	0.88**	0.71*	0.55	0.50	0.60	0.50	0.46	0.43	0.39	0.36
ROUGE-we	0.95**	0.71*	0.94**	0.76**	0.98**	0.86**	0.95**	0.79**	0.90**	0.64*	0.91**	0.64*	0.50	0.50	0.55	0.50	0.45	0.43	0.38	0.36
EmbeddingAverage	0.79*	0.50	0.82*	0.69*	0.86**	0.79**	0.87**	0.71*	0.85**	0.57	0.86**	0.57	0.71*	0.57	0.75	0.57	0.56	0.50	0.51	0.43
VectorExtrema	0.80*	0.57	0.80*	0.76**	0.86**	0.86**	0.82*	0.64*	0.84**	0.64*	0.84**	0.64*	0.37	0.21	0.42	0.21	0.40	0.36	0.33	0.29
GreedyMatching	0.89**	0.64*	0.80*	0.69*	0.88**	0.79**	0.85**	0.71*	0.85**	0.71*	0.86**	0.71*	0.60	0.50	0.64	0.50	0.43	0.50	0.36	0.43
SummaQA	0.87**	0.57	0.85**	0.62*	0.91**	0.71*	0.93**	0.79**	0.87**	0.64*	0.89**	0.64*	0.24	0.21	0.30	0.21	0.43	0.43	0.35	0.36

Pearson's r and Kendall's τ between intrinsic automatic metrics and extrinsic criteria. Significance is indicated by * for p-values less than or equal to 0.05 and ** for p-values less than or equal to 0.01

(2) Evaluating automatic metrics using extrinsic criteria:

- According to top-k system analysis, most automatic metrics fail to consistently and reliably quantify differences in usefulness between systems.



System-level Pearson correlations between intrinsic automatic metrics and proposed extrinsic metrics on top-k systems

Conclusions

- An extrinsic evaluation framework to assess the usefulness of text summaries with a web-based platform to facilitate the data collection.
- A new human extrinsic evaluation dataset with 4k annotated articles.
- We find that summaries are generally useful in tasks that require a comprehensive understanding of an article. We also explore the connection between the usefulness and intrinsic properties of summaries.
- We re-evaluate 14 automatic metrics and discover that most of them fail to reflect the extrinsic metrics in classification and similarity tasks.