Is Summary Useful or Not? An Extrinsic Human Evaluation of Text Summaries on Downstream Tasks

Xiao Pu, Mingqi Gao, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
State Key Laboratory of Media Convergence Production Technology and Systems
puxiao@stu.pku.edu.cn, {gaomingqi, wanxiaojun}@pku.edu.cn

Abstract

Research on automated text summarization typically uses human and automatic evaluation methods. While most recent studies focus on intrinsic evaluation, which assesses the general quality of summaries, e.g. coherence and informativeness, we concentrate on task-based extrinsic evaluation to determine the usefulness of summaries. We incorporate three downstream tasks, namely question answering, text classification, and text similarity assessment, and measure the usefulness of summaries for these tasks by several metrics. Our findings reveal that summaries are generally useful in tasks that require a comprehensive grasp of the text but are less useful in tasks requiring a more specific understanding of the text. We also analyze the usefulness and inherent properties of summaries from different models, and find that fine-tuned models consistently produce more useful summaries across all three tasks. In contrast, zero-shot models tend to lean towards text classification and similarity assessment, providing more general and less detailed summaries. Additionally, we assess the correlation between 14 intrinsic automatic metrics and human judgments. Intrinsic metrics perform well in evaluating summaries for question answering but are less effective in the other two tasks. This highlights the limitations of relying solely on intrinsic metrics for assessing summary performance and usefulness.

Keywords: NLG evaluation, extrinsic evaluation, text summarization

1. Introduction

Automated text summarization condenses longer documents into concise versions, making it a valuable tool for quick understanding. However, evaluating the quality of generated summaries poses challenges. Evaluation methods can be categorized as automatic metrics or human judgments. While automatic metrics offer cost-effective evaluation, they may not be perfect substitutes for human annotation. Human evaluation, divided into intrinsic and extrinsic evaluation, provides more reliable performance assessment.

While intrinsic evaluation of text summarization focuses on assessing the quality of generated summaries, e.g. coherence, fluency, and informativeness (Fabbri et al., 2021a; Bhandari et al., 2020a), extrinsic evaluation, also known as task-based evaluation, explore the usefulness or helpfulness of text summaries in other tasks (Dorr et al., 2005). It is more objective and spontaneous because it evaluates human performance in a realistic usage scenario and is less demanding on the annotators.

Gillick and Liu (2010) have shown that when conducting intrinsic human evaluation experiments, the reliability of non-expert ratings is significantly lower compared to expert annotators, indicating that intrinsic evaluation methods place high demands on annotators. On the other hand, extrinsic evaluation methods is less demanding and more straightforward to annotators in that they do not have to directly rate the summaries but treat them as tools to accomplish other tasks. As a result, we can design simpler tasks that are more closely aligned with real-life scenarios, reducing the difficulty for annotators in the evaluation process, therefore obtaining reliable results on the usefulness of the summaries. Additionally, by avoiding direct scoring of the summaries, extrinsic evaluation methods provide a more objective assessment approach.

Previous studies on extrinsic evaluation of summarization models have utilized methods such as cross-comprehension tests (Kolluru and Gotoh, 2005), relevance judgment (Dorr et al., 2005), and question answering (Hirao et al., 2001). However, these studies are dated. In recent years, neural summarization systems, especially those based on pre-trained language models have made great strides in intrinsic evaluation (Fabbri et al., 2021a; Bhandari et al., 2020a). However, to the best of our knowledge, no work has investigated the usefulness of these approaches from the perspective of extrinsic evaluation. Additionally, they often rely on a single evaluation method or have limited human experimentation (Hovy and Lin, 1998). In light of these limitations, we propose a comprehensive extrinsic evaluation method, conducting human experiments on three designed tasks to assess the usefulness of text summaries. We also construct a trustworthy human-evaluated corpus for three downstream tasks. Our study addresses research questions on summary usefulness, task-

Code and datasets will be available at https://github.com/SophiaPx/extrinsic_eval.

LREC-COLING 2024, pages 9389–9404
20-25 May, 2024.

specific utility, preferred summary characteristics, and correlation between automatic metrics and human judgments. The contributions of our work are summarized as follows:

- We introduce an extrinsic evaluation framework to assess the usefulness of text summaries on three downstream tasks. We also build a web-based platform to facilitate the annotated data collection.
- We annotate and construct a reliable human extrinsic evaluation dataset of 4,000 texts, including 400 source texts, 400 human summaries, and 3,200 summaries generated by eight different text summarization systems.
- We analyze the usefulness of summaries on downstream tasks, and find that summaries are generally useful in tasks that require a complete understanding of the text but less useful in tasks requiring a more specific understanding of the text. We also explore the connection between the usefulness and intrinsic properties of summaries.
- We re-evaluate 14 intrinsic automatic metrics through our proposed criteria and discover that most of them fail to reflect the extrinsic metrics in classification and similarity tasks.

2. Methodology

This work aims to provide a comprehensive assessment of the usefulness of summaries in downstream tasks. Participants are asked to complete three tasks using source articles and summaries, and their performance is measured to determine the usefulness of summaries.

2.1. Measures of usefulness

In our study, we consider a summary to be useful (or helpful) if it is able to facilitate users to complete a task. A useful summary should help users save time by being shorter than the source text, while also providing them with the information they need to complete the task. Therefore, to assess the usefulness of the summaries, we decide to compare on two dimensions: time and correctness. Time refers to the amount of time it takes the participant to complete the task using either the source text or the summary. Correctness refers to the accuracy of the participant's response and is measured using different metrics for each task. A web-based platform is developed and deployed for this study, to automatically record the completion time and submitted answers by participants for each task.

2.2. Design of downstream tasks

In the course of devising specific tasks for this research, our objective is to emulate and represent diverse real-world applications of summaries. After thorough deliberation, we have elected to construct three distinct tasks: question answering, classification, and similarity assessment.

Question answering. Summaries frequently serve as concise substitutes for original texts, allowing users to access key information expeditiously. To address this particular use case, we have designed the QA task, in which participants are asked to answer questions based on the information provided in the source text or the summary. To evaluate the participant's accuracy, we use two commonly used evaluation metrics in QA systems to calculate the overlap between the answers submitted by the participant and the ground true answers. Additionally, we also propose a distinguished metric to reflect on the probability of the participants' answer attempts. By evaluating their performance in the QA task, we are able to determine the amount of useful information contained in the summary.

Classification. In specific scenarios, users may engage with summaries to swiftly identify content that aligns with their specific interests. We have integrated the classification task to cater to this real-world application. In this task, participants are asked to select one or more tags based on the article or summary they see. The accuracy of their choices is calculated as a way of determining whether different types of summaries are useful in helping people make an overall judgment about the article.

Similarity assessment. In situations where individuals encounter news articles of interest, they often rely on summaries to determine if the content is related to prior reports or covers similar subject matter. We have devised the similarity assessment task to explore this scenario, where participants are asked to take into account various factors such as the topic, event field, writing style, tone, etc. of the two articles to make a comprehensive judgment, and then score the similarity of the two articles or summaries on a scale of 1 to 4. By calculating how similar their scores are to the ground truth scores, we can determine how useful the summaries are for similarity judgments.

3. Experimental Settings

In this section, we present the construction and annotation of three datasets for use and the design of our user study for extrinsic evaluation. Specifically, we focus on three downstream tasks: question answering (QA), text classification, and text similarity assessment. We then propose extrinsic metrics reflecting on the usefulness of summaries, and

introduce the intrinsic metrics we use to make a comparison with our proposed extrinsic metrics.

3.1. Data Preparation

Processing and annotating datasets. For our proposed QA, classification, and similarity tasks, we sample, reprocess and manually annotate the following datasets in our study:

CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016) is a widely used benchmark for text summarization, which includes a collection of news articles and their corresponding reference summaries that are typically 3-4 sentences in length. We use it for building dataset for our question answering task: We select 100 pairs of source text and reference summary randomly from the test set and annotate two datasets: QA-ref and QA-source. For QA-ref, we manually write four questions and their corresponding answers according to reference summaries, while for QA-source, questions and answers are written according to source articles. Multiple correct answers may exist for each question in both QA datasets.

New York Times Annotated Corpus (Sandhaus, 2008) contains a set of news articles along with human-written summaries. Each article is associated with multiple tags or labels. For the classification task, we randomly sample 100 news articles from the test set and categorize them into 11 tags after having carefully removed some redundant tags.

The SemEval-2022 Task 8 dataset (Chen et al., 2022) is a multilingual collection of the URLs of news articles that have been paired and annotated for their similarity level, therefore we utilize this for our similarity task. We crawl 300 pairs of news pages according to the links provided by this dataset and then extract titles, descriptions, and body parts of each article by data cleaning, resulting in 100 pairs of news articles with corresponding summaries and similarity scores.

Generating summaries for different systems. We select 8 representative summarization systems to generate automatic summaries, including 6 abstractive summarization models, namely BART (Lewis et al., 2019), Pegasus (Zhang et al., 2020), BRIO (Liu et al., 2022), T5 (Raffel et al., 2020), T0 (Sanh et al., 2021) and GPT3 (Brown et al., 2020)[1] and 2 simple extractive summarization models, namely Lead-n [2] and Lexrank (Erkan and Radev, 2004). To ensure fairness in comparing summaries across different systems, we generate summaries of similar lengths for each task. More details regarding this process are shown in Appendix A.

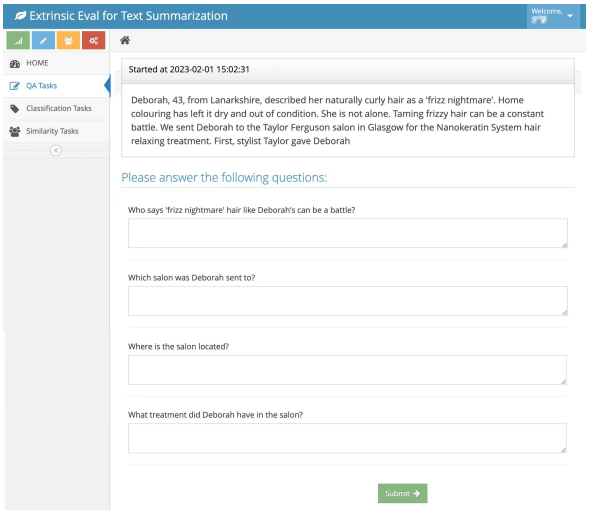


Figure 1: A screenshot of the answering page for the QA task. The user information on the platform has been anonymized.

3.2. Web-based Platform for Evaluation

We implement a web-based platform (as shown in Figure 1 and Appendix B) to facilitate users' participation in the tasks and the acquisition of experiment data, which includes responses and completion time for each question. To guarantee impartiality, the platform is designed to prohibit the utilization of the copy-paste/search functionality. Furthermore, the website offers guidance information and exemplar answers to assist participants to fully understand the tasks.

3.3. Experimental Details

We recruit 20 university students to participate in the experiment and they are required to possess a high level of proficiency in English. The numbers of male and female participants are the same. The initial ten participants complete the QA-ref, classification and similarity tasks, while the latter ten participants complete the QA-source task.

To ensure that the participants' responses are only based on the content of the text currently being viewed and to minimize the influence of individual differences, a method for distributing the texts is devised. The following considerations are taken into account: (1) To prevent them from having an advantage due to prior exposure to a similar text, each person is allowed to see only one text (either source text or summary) from the same source. (2) To ensure fairness and minimize the influence of individual differences, each person must be exposed to the same number of texts from each system,

[1]Among these models, BART, Pegasus, BRIO and T0 have been finetuned on the CNN/DailyMail Dataset.

[2]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

regardless of their proficiency level.

The distribution method is as follows: One source text is associated with nine summaries (including reference summary), resulting in ten texts (including source text) originating from the same source text.

First, all summaries are aligned with the source text, then different systems are arranged in the following order: [Source, Human, BART, Pegasus, Lexrank, Lead-n, BRIO, T5, T0, GPT3]. After that, all texts are numbered, with text_id (0-999) as their unique identifier. Therefore, the hundreds place indicates the system corresponding to the text, and the tens place and the individual place indicate the corresponding source text.

The texts are assigned to different participants according to the system it belongs to and the corresponding source text. Each participant is assigned to a user_id and the correspondence between texts and participants is established by the following formula:

$$y = \lfloor \frac{text\_id - \lfloor \frac{text\_id}{100} \rfloor \times 100}{10} \rfloor - \lfloor \frac{text\_id}{100} \rfloor$$

$$\text{user\_id}(y) = \begin{cases} y, & \text{if } y \geq 0 \\ 10 + y, & \text{if } y < 0 \end{cases}$$

3.4. Proposed Extrinsic Metrics

Based on the three downstream tasks, we propose the following extrinsic metrics to evaluate the usefulness of the summaries.

For the QA task, let y_n^k denote the participant's answer to the k-th question of n-th article, and $\hat{y}_n^{ki}$ denote the i-th key answer to the k-th question of n-th article. N represents the number of summaries of each system, which equals 100, and K represents the number of questions for each article, which equals 4 in this case, and the three metrics are calculated as follows.

- Answerable measures the proportion of questions that can be answered according to the text.
- Exact Match Ratio (EM), which counts the overall accuracy rate of the answers. EM of each system is calculated as:

$$EM = \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} MAX_i(I(y_n^k == \hat{y}_n^{ki}))$$

$$\text{with } I(y_n^k = \hat{y_n}^{ki}) = \begin{cases} 1, & \text{if } y_n^k = \hat{y_n}^{ki} \\ 0, & \text{if } y_n^k \neq \hat{y_n}^{ki} \end{cases}$$

- F1 is a looser measure of the average overlap between the prediction and ground truth answer. When calculating F1, both y_n^k and $\hat{y_n}^k$ are tokenized into sets of words. F1 is calculated as

$$F_1 = \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} MAX_i \frac{2|y_n^k \cap \hat{y_n}^{ki}|}{|y_n^k| + |\hat{y_n}^{ki}|}$$

For the classification task, we use EM and F1 as extrinsic metrics, which are commonly used in multi-class classification tasks.

For the similarity task, we use the following metrics:

- Mean Squared Error (MSE), which indicates the extent to which the participant's answer deviates from the standard answer.
- Spearman's ρ, a measure of the correlation between the participant's judgment and the true similarity. It can only be used for system-level analysis because it cannot be calculated using separate texts.

4. Evaluating Summaries' Usefulness

In this section, we compare the performance of different summarization systems by means of the proposed extrinsic evaluation method (as shown in Table 1) and try to answer some questions regarding the usefulness of summaries.

4.1. How useful are text summaries compared to source articles?

Results from three downstream tasks demonstrate that the use of summaries significantly reduces the time required for task completion. Specifically, compared to the source articles, the average time participants spent using summaries to complete QA tasks drops by 61-62% (as shown in Table 2). Similar results can also be observed in the classification and similarity tasks, with the time-saving percentages of 59% and 42%, respectively (as shown in Table 3).

We also find that **summaries are particularly useful in classification and similarity tasks**. In the QA task, source texts outperform summaries on average, while in the classification and similarity tasks, participants spend less time as well as perform better with summaries. This may be due to the fact that making an overall judgment about the text, such as classification or similarity assessment, does not require as much information as answering specific questions. As a result, the excess information in the long source text may not aid in decision-making and even interfere with human

system	QA (ref-based)				QA (source-based)				Classification			Similarity		
	answerable	EM	F1	time(seconds)	answerable	EM	F1	time(seconds)	EM	F1	time(seconds)	MSE	ρ	time(seconds)
source	0.8550	0.3225	0.5077	280.04	**0.8875**	**0.5050**	**0.6796**	211.64	0.8827	0.8951	72.97	0.9136	0.6184	37.74
reference	**0.8875**	**0.5400**	**0.7535**	**93.94**	0.5375	0.2725	0.3746	83.3	**0.9127**	**0.9156**	34.37	0.7736	0.7060	**19.92**
bart	0.4975	0.2400	0.3240	108.37	0.4900	0.2325	0.3197	**83.05**	0.8964	0.9015	**25.43**	0.9803	0.6085	21.94
pegasus	0.5475	0.2100	0.3222	112.55	0.5125	0.2825	0.3662	89.66	0.8900	0.8942	29.88	0.9836	0.6014	23.93
lexrank	0.3625	0.0900	0.1631	111.78	0.3775	0.1500	0.2291	92.01	0.9000	0.9017	29.88	1.2403	0.5323	23.77
Lead-n	0.4175	0.1600	0.2483	110.78	0.4775	0.2475	0.3342	84.33	0.8773	0.8792	31.29	1.4336	0.4536	23.42
BRIO	0.5825	0.2350	0.3598	104.21	0.5425	0.3075	0.4040	90.15	0.9000	0.9036	25.9	**0.7569**	0.6998	21.07
t5	0.4400	0.1600	0.2416	106.07	0.4375	0.2075	0.2861	86.57	0.8791	0.8814	34.86	1.3736	0.4699	20.17
t0	0.5350	0.1875	0.3003	107.21	0.5100	0.2600	0.3530	98.6	0.8864	0.8889	28.57	0.7669	**0.7087**	20.96
gpt3	0.4200	0.1575	0.2338	100.02	0.4500	0.1975	0.2855	83.74	0.9036	0.9068	29.11	0.8469	0.6741	20.66

Table 1: Usefulness of different systems on downstream tasks, including the average time taken by participants to complete tasks with different system outputs and results of extrinsic metrics based on user performance.

	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%

Table 2: Summaries compared to source texts in the QA tasks. The green percentages indicate that summaries are more useful compared to the source text, i.e. participants take less time or perform better in completing the task. The green ones indicate less useful. Though all summaries represent a significant time saving, participants perform worse in QA tasks using the summaries compared to source texts.

judgments. This is supported by observed people's tendencies in the classification task, where they tend to assign more tags to longer source articles, potentially leading to a higher recall but lower precision in comparison to the human summaries.

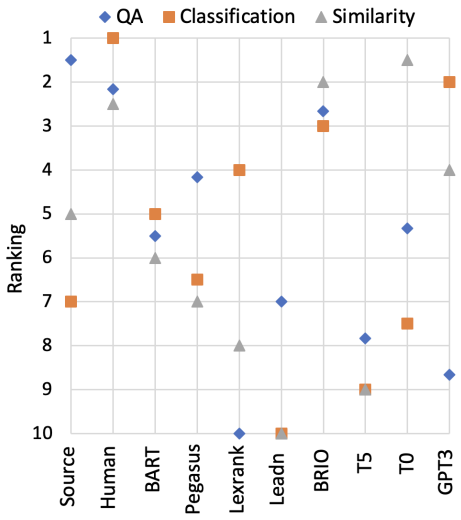


Figure 2: Average ranking of different systems on three different tasks. Each ranking is calculated by averaging the rankings over extrinsic metrics for the same task.

4.2. Which summarization systems are more useful?

We divide all the automated summaries into three categories based on the model used to generate them: fine-tuned, zero-shot, and simple extractive. A question we want to know is, how stable or consistent is the usefulness level of summaries across different downstream tasks? By analyzing rankings of the source text and summaries in the three tasks, as is shown in Figure 2, we find that: The summaries generated by fine-tuned models have higher consistency in usefulness across different tasks, such as those generated by BART, Pegasus, and BRIO, with a stable ranking similar to that of the human summaries. This suggests that summaries generated by fine-tuned models are insensitive to differences between tasks. The summaries generated by simple extractive models and models in the zero-shot setting exhibit a varying ranking across tasks. For example, both zero-shot GPT3 summaries and simple extractive Lexrank summaries show high or above average rankings in the classification task, medium rankings in the similarity task, and very low rankings in the QA task.

4.3. Go deeper: what kind of summaries are more useful?

Furthermore, we would like to know what leads to differences in the usefulness of summaries from different systems. We start by analyzing different features of summaries based on metrics and our case study C, and then explore the relationship between these features and usefulness on downstream tasks.

Summary Style: To quantify the summary style, namely abstractive or extractive, we follow Grusky et al. (2018) to employ the Ext-cvg (Extractive Fragment Coverage) metric, which assesses the extent to which a summary is extractive. As shown in the Table 4, traditional extractive models like Lead-n

	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%

Table 3: Summaries compared to source texts in the classification and similarity tasks. The green percentages indicate that summaries are more useful compared to the source text, i.e. participants take less time or perform better in completing the task. The green ones indicate less useful. It shows that summary serves about the same function as the source text in these two tasks, and even helps participants to do tasks better.

	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

Table 4: Intrinsic features of summaries from different summaries.

	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

Table 5: System-level pearson correlation between the metrics reflecting on intrinsic features of summaries and the extrinsic metrics.

and Lexrank exhibit an Ext-Cvg of 100%. According to our case study, they contain relatively less important information in a limited space. Interestingly, the reference summaries written by humans score notably lower at 87.51%, indicating that human-written reference summaries creatively incorporate expressions beyond those present in the source text. The zero-shot GPT3 ranks just below the reference summaries, surpassing summaries generated by all fine-tuned models. This quantitative result aligns with the observation in our case study, that summaries generated by fine-tuned models tend to be more informative and specific, including more factual details such as times, places, and numbers [3], while summaries generated by models in the zero-shot setting seem more abstractive and general. Compared to them, simple extractive summaries are more coarse-grained and less useful.

As shown in Table 5, the system-level Pearson correlation between the extrinsic evaluation metrics for the three tasks and the aforementioned intrinsic features is presented. We observe that the summary's Ext-Cvg and the extrinsic metrics for classification and similarity tasks exhibit a moderate negative correlation of approximately -0.60. This implies that more abstractive summaries tend to be more useful for classification and similarity tasks. This makes perfectly sense that those summaries without some details may be easier for people grasp the whole story and are therefore naturally better suited for tasks involving overall judgments, such as classification and similarity tasks.

Grammaticality: Following Lee et al. (2022), we employ the languagetool website[4] to identify grammar errors in the summaries. Less grammar errors per word denote greater grammaticality. Surprisingly, we discover that human-written text does not consistently surpass machine-generated one in terms of grammaticality. Notably, the reference summaries exhibit the highest proportion of grammar errors, and even summaries generated by Lead-n and Lexrank (constructed directly from source text sentences) show nearly higher error rates compared to other summaries. As shown in Table 5, grammar errors ratio demonstrates a moderate correlation (-60% to -40%) with extrinsic metrics for QA and similarity tasks, while displaying minimal correlation with extrinsic metrics for classification tasks. This indicates that more grammatical summaries are more beneficial for QA and similarity tasks, whereas this is less crucial for classification tasks. This intuitively makes sense, as tasks like classification often require grasping keywords to label topics, whereas tasks that involve evaluating the overall similarity between two texts are more intricate. These tasks necessitate a comprehensive judgment incorporating various textual aspects such as specific events, writing styles, and the tone of the articles, which places a demand on the grammaticality of summaries.

[3] It's important to note that this observation is only based on the summaries fine-tuned on the CNN/DailyMail dataset. Fine-tuning on other datasets may produce different results and therefore cannot be generalized as all summaries generated by fine-tuned models. When referring to summaries generated by fine-tuned models, it should only be understood as those fine-tuned on the CNN/DailyMail dataset.

[4] The link to this website is: https://languagetool.org/

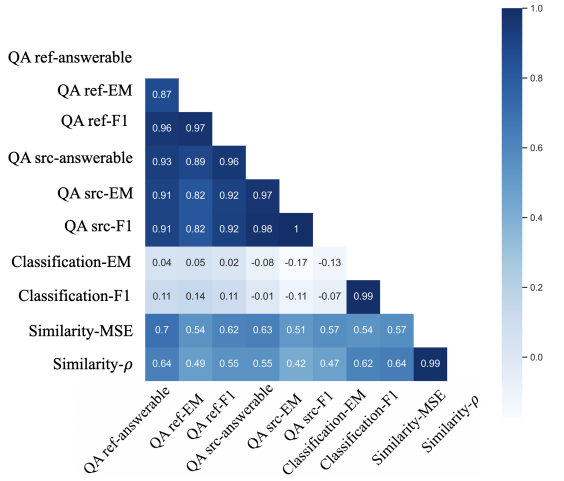


Figure 3: System-level Pearson correlation of all extrinsic metrics.

Sentence length: We measure average number of words per sentence of summaries. Our analysis reveals that traditional extractive summaries display significantly longer average sentence lengths, with approximately 30 words per sentence. This aligns with the inherent characteristic of the source text's sentence length. In contrast, human reference summaries average less than 17 words per sentence. Fine-tuned summaries approximate the reference summaries' lengths, while GPT-3's zero-shot generated summaries feature relatively longer sentences, averaging around 24 words per sentence. For QA and similarity tasks, the average sentence length of the summaries demonstrates a moderate correlation with their usefulness, whereas the correlation is notably low for classification tasks.

5. Correlation between Metrics

5.1. Analyzing Our Extrinsic Metrics

1 In this section, we study the relationship between our proposed extrinsic metrics. We compute system-level correlations of all the extrinsic metrics (as shown in Figure 3).

According to the Pearson's *r*, extrinsic metrics of the same downstream task are highly correlated, ranging from 0.8 to 1. QA-ref and QA-source are highly correlated at system level, with Pearson's *r* above 0.8 and Kendall's τ above 0.69. This suggests that there is little difference in the relative performance of the systems on QA-ref and QA-source, although they differ in the way the dataset is constructed. Comparing the metrics of the different downstream tasks, we find that the QA task and the classification task are poorly correlated, with Pearson's *r* ranging from -0.2 to 0.2. Whereas the similarity task is moderately correlated with both the other two tasks, with Pearson's *r* ranging from 0.4 to 0.7. This indicates that the QA task and the classification task are very divergent, while the similarity task can be considered as a point of comparison between QA and classification tasks. Overall, moderate to weak correlations illustrate that our experiment involves three tasks of different perspectives to measure the usefulness of the summary.

5.2. Evaluating Intrinsic Automatic Metrics

We perform a meta-evaluation using Pearson's *r* and Kendall's τ to compare various intrinsic automatic metrics with our extrinsic metrics. The intrinsic automatic metrics include n-gram overlap-based measures such as ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and CHRF (Popović, 2017). For metrics based on word embeddings, we report BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), Rouge-we (Ng and Abrecht, 2015), Embedding average (Landauer and Dumais, 1997), Vector extrema (Forgues et al., 2014), Greedy matching (Rus and Lintean, 2012). Furthermore, we also include a model-based metric SummaQA (Scialom et al., 2019) in our evaluation. All scores are reported in the range of 0-1. These scores will be compared with our extrinsic human evaluation results.

Our analysis reveals that there is a high correlation between extrinsic metrics in the QA task and intrinsic automatic metrics, as shown in Table 6, with Pearson's *r* values ranging between 0.7 and 1. Additionally, we find that there is little difference between the performance of different intrinsic automatic metrics, indicating that they are able to evaluate the QA task relatively well.

On the other hand, we observe that extrinsic metrics in classification and similarity tasks have low to moderate correlation with most intrinsic automatic metrics. The Embedding Average metric is found to be strongly correlated with the extrinsic metrics for the classification task (statistically significant at $p < 0.01$) and show a moderate correlation for the similarity task. Other word embedding-based metrics such as Greedy Matching, Rouge-we, BERTScore and MOVERScore also show moderate correlation with extrinsic metrics in classification and similarity tasks.

In terms of the best and worst intrinsic automatic metrics, we find that no single metric consistently performs the best across all tasks. However, two intrinsic automatic metrics that are closest to the extrinsic metrics are Rouge-1 (better in the QA task) and Embedding Average (better in the similarity and classification tasks). On the other hand, CIDEr is found to be least correlated with the extrinsic metrics, and show little relevance for the similarity

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

Table 6: 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.

and classification tasks.

We further evaluate the reliability of intrinsic automatic metrics in quantifying differences between systems with competitive performance,i.e., top-k system analysis. As illustrated in Figure 4, k systems are ranked based on different extrinsic metrics. We observe that for the QA-ref answerable metric and QA-source F1 and answerable metrics, the correlation between automatic and extrinsic metrics decreases slightly as the number of systems increases from 3, then increases when the number of systems reaches 5. A similar trend is also observed in the plot of the F1 indicator for the classification task, but with more noticeable fluctuations. However, we find a significant decline in the correlation between extrinsic and intrinsic automatic metrics of the similarity task as k increased, which suggests that intrinsic automatic metrics should not be used to compare systems with substantial differences in usefulness in this task. While the correlation between the QA-ref answerable metric and intrinsic automatic metrics remains stable at a high level even as k changed, we find that most intrinsic automatic metrics may not consistently and reliably quantify differences of usefulness between systems.

6. Related Work

We now discuss the literature most related to our work, and defer a more complete review to Appendix D.

Kolluru and Gotoh (2005) have acknowledged the human's subjectivity in evaluating summaries, and attempted to alleviate this through the use of cross-comprehension tests. Usefulness of summaries has also been evaluated through a single extrinsic task, i.e. relevance judgment (Dorr et al., 2005) and question answering (Hirao et al., 2001). While Hovy and Lin (1998) have proposed a set of tasks to measure the information content of full text and summaries, including a Shannon Game, a Question Game, and a Classification Game, finding that different extrinsic evaluation methods rate summaries differently, the scale of the experiments was too small to draw statistically significant conclusions. Bhandari et al. (2020b) and Fabbri et al. (2021b) have investigated the relationship between intrinsic automatic metrics and intrinsic human judgments in the field of text summarization. Goyal et al. (2022a) have examined the consistency between intrinsic automatic metrics and human preferences for different types of summaries and found that intrinsic automatic metrics cannot reliably evaluate summaries generated by models in the zero-shot setting.

7. Conclusions

In this work, we employ an extrinsic evaluation approach to assess the usefulness of summaries across three downstream tasks, including question answering, classification, and similarity assessment. These tasks are designed to represent and simulate real-world scenarios where summaries are used. A specially-developed web platform enables us to collect annotators' feedback, including their completion times and accuracies, allowing us to analyze usefulness from two different angles. Regarding the usefulness of text summaries, our main findings include:

1. Summaries are particularly useful in classification and similarity assessment tasks while being less useful for the QA task.

2. According to the systems rankings of usefulness, fine-tuned summaries exhibit consistent relative usefulness across various tasks. Conversely, zero-shot and simple extractive summaries demonstrate varying rankings across tasks.

3. Further exploration reveals that various features of summaries, such as style, grammatical correctness, and sentence length, reflect differences among types of summary systems, which therefore impact their usefulness to varying degrees in downstream tasks.

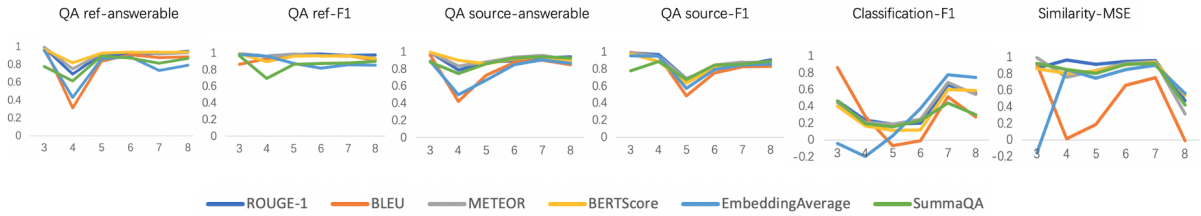


Figure 4: System-level Pearson correlations between intrinsic automatic metrics and proposed extrinsic metrics on top-k systems.

4. Intrinsic automatic metrics can reflect the usefulness of summaries in the question answering task, but their correlation with usefulness is limited when it comes to tasks requiring people to make an overall judgment about the text, namely classification and similarity tasks.

Limitations

In our selection of extractive text summarization models, we are focusing on simple extractive text summarization models and not including more advanced deep learning-based models. Therefore, although the experimental results may show that the performance of the two extractive models is not as good as that of generative models, this cannot be taken as conclusive evidence that generative text summarization models are inherently superior.

Ethics Statement

To ensure ethical practices, all annotators involved in our study are remunerated for their time and effort. We compensate them at a rate of 10 USD per hour, which exceeds the local minimum wage requirements. By providing a fair and competitive compensation package, we aim to acknowledge the valuable contributions of our annotators and uphold ethical standards in research.

Acknowledgements

This work was supported by Beijing Science and Technology Program (Z231100007423011), National Key R&D Program of China (2021YFF0901502), National Science Foundation of China (No. 62161160339) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

8. Bibliographical References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020a. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020b. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3110–3120, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Bonnie Dorr, Christof Monz, Richard Schwartz, and David Zajic. 2005. A methodology for extrinsic evaluation of text summarization: does rouge

correlate? In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 1–8.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Rosa Gaudio, Aljoscha Burchardt, and António Branco. 2016. Evaluating machine translation in a usage scenario. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1–8, Portorož, Slovenia. European Language Resources Association (ELRA).

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022b. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Tsutomu Hirao, Yutaka Sasaki, and Hideki Isozaki. 2001. An extrinsic evaluation for question-biased text summarization on qa tasks. In *Proc. of the NAACL 2001 Workshop on Automatic Summarization*, pages 61–68.

Eduard Hovy and Chin-Yew Lin. 1998. Automated text summarization and the summarist system. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

BalaKrishna Kolluru and Yoshihiko Gotoh. 2005. On the subjectivity of human authored summaries. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 9–16.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami, Pooya Khosravyan Dehkordy, and Asghar Tajoddin. 2008. Optimizing text summarization based on fuzzy logic. In *Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)*, pages 347–352. IEEE.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Igor V Mashechkin, MI Petrovskiy, DS Popov, and Dmitry V Tsarev. 2011. Automatic text summarization using latent semantic analysis. *Programming and Computer Software*, 37:299–305.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Makbule Gulcin Ozsoy, Ferda Nur Alpaslan, and Ilyas Cicekli. 2011. Text summarization using latent semantic analysis. *Journal of Information Science*, 37(4):405–417.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *American statistician*, pages 59–66.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. *arXiv preprint arXiv:1904.05929*.

Ladda Suanmali, Mohammed Salem Binwahlan, and Naomie Salim. 2009. Sentence features fusion for text summarization using fuzzy logic. In *2009 Ninth International Conference on Hybrid Intelligent Systems*, volume 1, pages 142–146. IEEE.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Sukriti Verma and Vagisha Nidhi. 2017. Extractive summarization using deep learning. *arXiv preprint arXiv:1708.04439*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

9. Language Resource References

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.

Evan Sandhaus. 2008. The new york times annotated corpus.

Supplemental Materials

A. Details of Controlling the Length of Summaries

In order to ensure a fair comparison among summaries generated by different systems, we aim to maintain similar lengths across the summaries.

We first calculate the upper and lower limits of the ideal summary length in terms of token count by allowing a deviation of 20 tokens from the average length of the reference summaries For fine-tuned models such as BART and Pegasus, during the summary generation process, we set the "min_new_tokens" parameter to the lower limit. This ensured that the generated summaries would have a minimum token count as specified.

As for lexrank and leadn, since we can only set the number of sentences to be generated, we start with one sentence and gradually increase the number of generated sentences. After each sentence addition, we calculate the token count of the summary and compare it to our predetermined range. We stop adding sentences when the token count was within the desired range or closest to it. For prompt-based GPT-3 and T0, we employ specific prompts to guide the generation of summaries with a desired length. For instance, for the question-answering task, we use prompts such as "Please summarize this in about 50 words" to provide explicit instructions for generating summaries of the specified length. After generating the summaries, we then perform sentence-level truncation on summaries that exceeded the desired word count range, only if after truncation, it will be within or closer to our desired range.

This approach allowed us to control the length of the generated summaries for different systems while adhering to the predefined length boundaries. Figure 5 shows the length of generated summaries in the three tasks. Consequently, we can infer that the information content of the summaries is not significantly different among the various systems. The factor of summary length will not interfere with the results of our extrinsic human experiments.

B. Details of our evaluation platform

Here we provide some screenshots of our web-based platform for evaluation. The annotators are first required to read through the guidelines, as shown in Figure 6 and Figure 7. After a QA session to ensure all annotators fully comprehend the annotation process, rating criteria and any other details, they are directed to the list page (Figure 8) and then the answering page (Figure 9).

C. A Case Study of Summary Style

By looking at the source text and the summaries generated with different models, as shown in Figure 10, we find that the zero-shot GPT3 summary tends to paraphrase the news in a more general way, making it easier for readers to capture the main point, but often omitting detailed information. Instead,

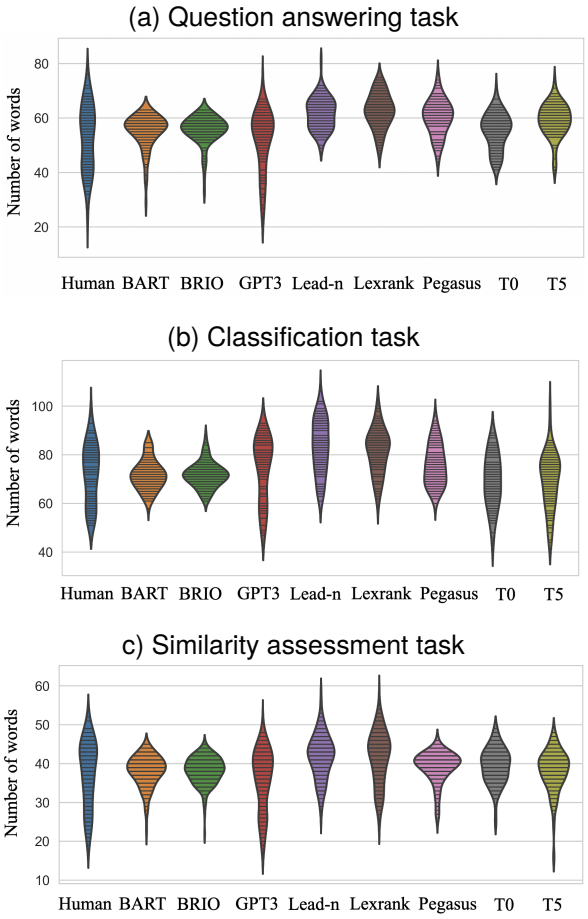


Figure 5: Length of summaries from different systems in three tasks.

summaries of fine-tuned BRIO and T0 models contain more detailed information, making it more suitable for QA tasks. The coherence between sentences in the extractive Lexrank summary is poor, causing difficulty in reading.

D. Related work

D.1. Intrinsic Evaluation for Summarization

Past works that have assessed the quality of summaries through intrinsic evaluation methods can be classified into two main categories: intrinsic automatic metrics and intrinsic human evaluation. Some researchers evaluate summaries by computing the n-gram word overlap between reference summaries and generated summaries, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which have been proven to be relatively effective over time. With the development of representation learning, researchers have proposed new intrinsic automatic metrics based on word embeddings, such as Greedy Matching (Rus and Lintean,

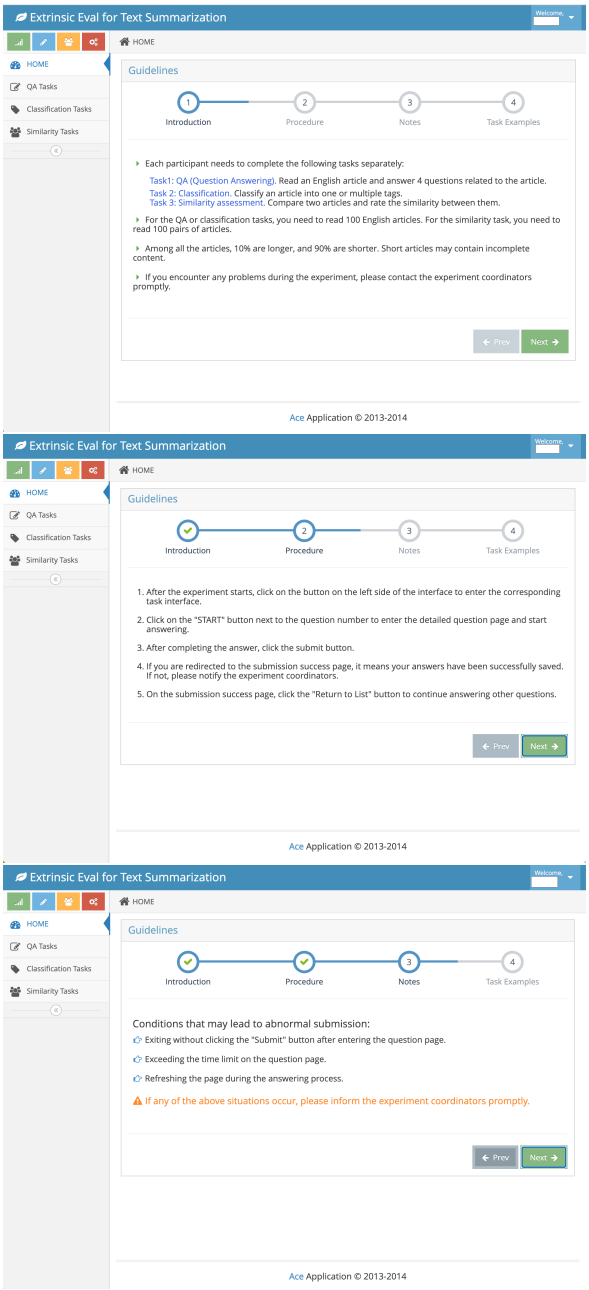
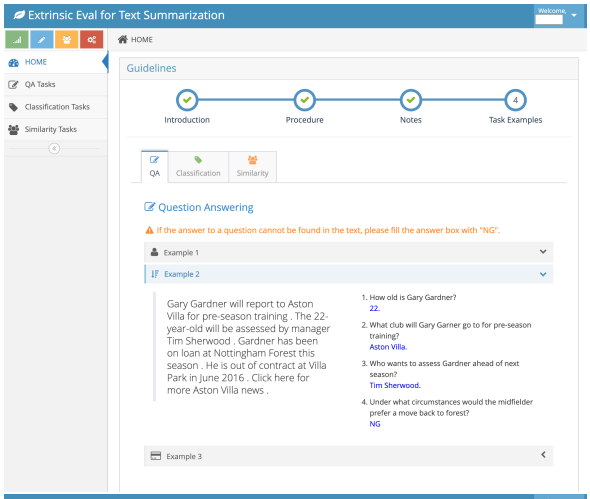
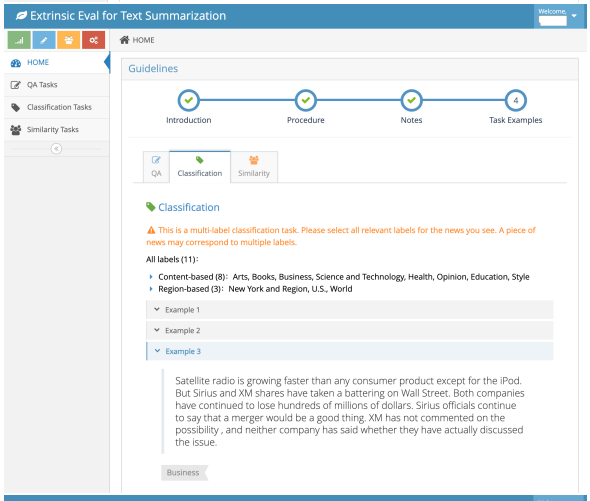


Figure 6: Instruction and guidelines.

2012) and SMS (Clark et al., 2019), which compute the similarity of word embeddings between reference summaries and generated summaries. Additionally, automatic metrics based on question-answering (Scialom et al., 2019) and entailment classification (Kryscinski et al., 2020) have also been proposed. Human evaluation, on the other hand, is considered as the gold standard for evaluating generated summaries. The Pyramid method (Nenkova and Passonneau, 2004) serves as a viable framework for human evaluation, which has been further improved into a crowdsourcing method (Shapira et al., 2019). Previous research has also investigated the relationship between intrinsic auto-





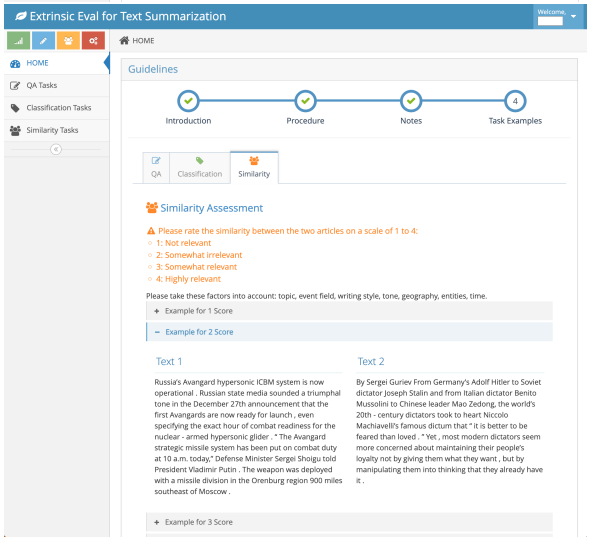


Figure 7: Examples for each task to help the annotators understand the rating criteria.

matic metrics and intrinsic human judgments in the field of text summarization. A common approach to conduct meta-evaluation is to have annotators score the quality of summaries by Pyramid method (Bhandari et al., 2020a) or on multiple dimensions (Fabbri et al., 2021a) such as coherence, consistency, relevance, and fluency, and compute the correlation coefficient between the output scores of automatic evaluation metrics and human judgments. Gillick and Liu (2010) have shown significant differences in the performance of experts and non-experts in scoring summaries. Goyal et al. (2022b) have examined the consistency between intrinsic automatic metrics and human preferences for different types of summaries and found that intrinsic automatic metrics cannot reliably evaluate summaries generated by models in the zero-shot setting, while our work investigates the correlation between intrinsic automatic metrics and extrinsic human judgments.

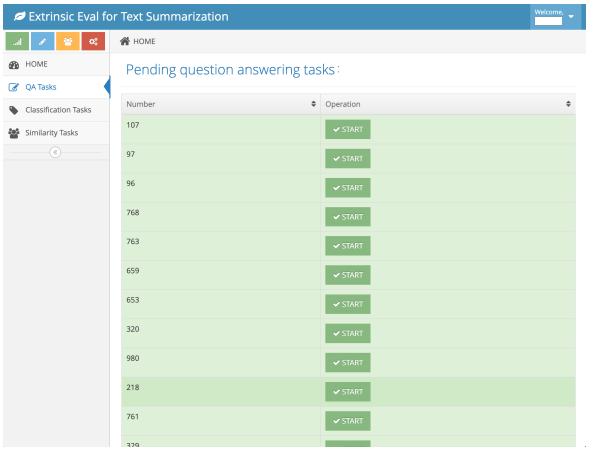
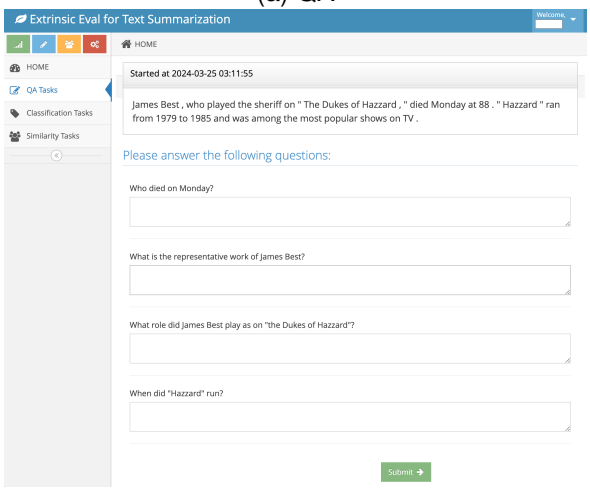


Figure 8: The list interface displays the tasks assigned to each annotator.

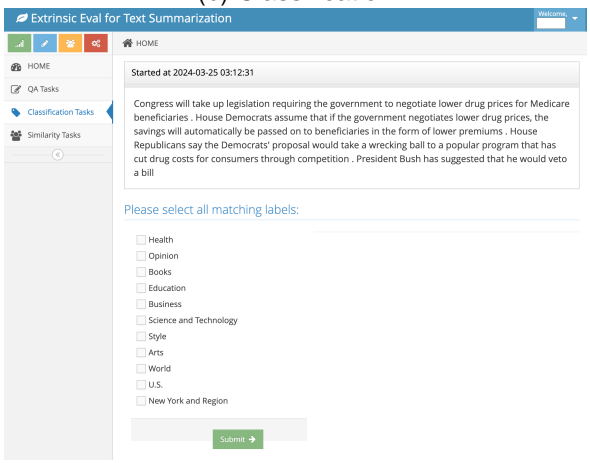
D.2. Extrinsic Evaluation for Summarization

Kolluru and Gotoh (2005) have acknowledged the human's subjectivity in evaluating summaries, and has attempted to alleviate this through the use of cross-comprehension tests. Usefulness of summaries has also been evaluated through a single extrinsic task, i.e. relevance judgment (Dorr et al., 2005) and question answering (Hirao et al., 2001). While Hovy and Lin (1998) have proposed a set of tasks to measure the information content of full text and summaries, including a Shannon Game, a Question Game, and a Classification Game, finding that different extrinsic evaluation methods rate summaries differently, the scale of the experiments was too small to draw statistically significant conclusions. We design three distinct extrinsic evaluation tasks with a larger scale of human judgments and evaluates the summaries generated by the recently proposed summarization approaches.

(a) QA



(b) Classification



c) Similarity

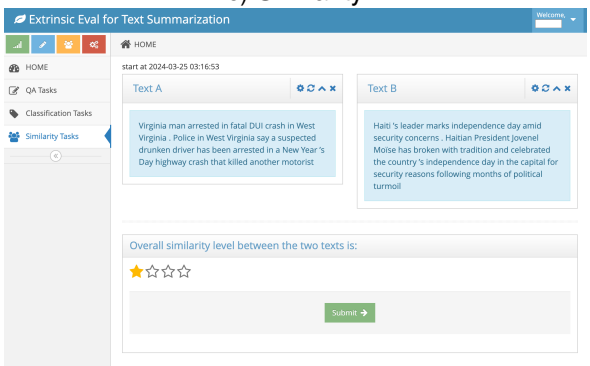


Figure 9: The answering page for annotators.

D.3. Summarization Models

Summarization models can be broadly categorized into two groups: extractive and abstractive. Extractive models directly identify and extract the most important sentences or words from the source text as the summary. Non-neural models, such as graph-based models, fuzzy logic-based models, and latent semantic analysis have been proposed and investigated (Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Suanmali et al., 2009; Kyoomarsi et al., 2008; Ozsoy et al., 2011; Mashechkin et al., 2011). Additionally, extractive summarization based on neural network models have also been explored (Nallapati et al., 2017; Verma and Nidhi, 2017; Narayan et al., 2018; Liu, 2019). On the other hand, abstractive models generate a summary text that is not necessarily a direct extraction of the source text. In recent years, abstractive summarization models based on neural networks have been advancing and become dominant in the summarization field. A common paradigm is pre-training and fine-tuning (Liu and Lapata, 2019; Lewis et al., 2019; Zhang et al., 2020). Additionally, some prompt-based approaches have been proposed (Brown et al., 2020; Sanh et al., 2021), enabling summarization models to learn from specific task instructions.

Source text:
A heartbroken pensioner is believed to have killed himself six days after his wife 's death by jumping from a bridge at their ' special place ' where they used to take romantic walks together. [...] Today officers confirmed a body pulled from the River Trent on April 15 by a specialist underwater search unit was sadly that of the missing pensioner. [...] June tragically died on March 31 , eight hours after collapsing suddenly from what doctors at the Queen 's Medical Centre in Nottingham described as a ' catastrophic bleed ' to the brain. [...]

GPT3 summary:
A man is believed to have killed himself by jumping from a bridge at a picturesque spot where he and his wife used to take romantic walks together, six days after she died from a brain hemorrhage.

BRIO summary:
John Lord , 86 , went missing from his home on April 6 less than a week after his beloved wife June , 81, died from a ' catastrophic bleed ' to the brain. The body of the pensioner was recovered from the River Trent on April 15. His family believe he may have jumped from a bridge at the picturesque beauty [...]

T0 summary:
John Lord, 86, went missing from his home on April 6. His wife June, 81, died from a 'catastrophic bleed' to the brain. Family feared the worst after finding a note describing how much he missed her. Mr Lord's body was pulled from the River Trent on April 15.

Lexrank summary:
Mr Lord , 86 , went missing from his home in St Ann 's on Monday, 6 April . ' A Nottinghamshire Police spokesperson said : ' The body of a man found in the River Trent on April 15 , 2015 , has been confirmed as that of missing John Lord . ' Message :Mr Lord 's daughter Alison said her father was grieving and had left a heart- breaking note signed [...]

Figure 10: A case study to illustrate the difference of summary style.