Reference-Free Summarization Evaluation with Large Language Models

Abbas Akkasi
School of Computer Science, Carleton University
abbasakkasi@cunet.carleton.ca

Kathleen C. Fraser
National Research Council Canada and
School of Computer Science,
Carleton University
kathleen.fraser@nrc-cnrc.gc.ca

Majid Komeili
School of Computer Science,
Carleton University
majidkomeili@cunet.carleton.ca

Abstract

With the continuous advancement in unsupervised learning methodologies, text generation has become increasingly pervasive. However, the evaluation of the quality of the generated text remains challenging. Human annotations are expensive and often show high levels of disagreement, in particular for certain tasks characterized by inherent subjectivity, such as translation and summarization. Consequently, the demand for automated metrics that can reliably assess the quality of such generative systems and their outputs has grown more pronounced than ever. In 2023, Eval4NLP organized a shared task dedicated to the automatic evaluation of outputs from two specific categories of generative systems: machine translation and summarization. This evaluation was achieved through the utilization of prompts with Large Language Models. Participating in the summarization evaluation track, we propose an approach that involves prompting LLMs to evaluate six different latent dimensions of summarization quality. In contrast to many previous approaches to summarization assessments, which emphasize lexical overlap with reference text, this method surfaces the importance of correct syntax in summarization evaluation. Our method resulted in the second-highest performance in this shared task, demonstrating its effectiveness as a reference-free evaluation.

1 Introduction

Text summarization is a natural language processing (NLP) task that aims to condense a given text into a shorter version while retaining its most essential information. It plays a crucial role in information retrieval, content extraction, and document management. Automatic summarization systems, whether extractive (selecting and rearranging existing sentences) or abstractive (generating novel sentences), offer significant advantages in various domains such as news articles, legal documents, academic papers, and online content. The ability to generate concise and coherent summaries enhances information accessibility, facilitates quicker decision-making, and improves user experience in an era of information overload (Cajueiro et al., 2023).

A good summary plays a pivotal role in information processing and communication across various domains. It serves as a concise yet comprehensive representation of a larger body of text, distilling the core ideas, key information, and essential insights. The importance of a good summary lies in its ability to save time and effort for readers, enabling them to grasp the main points quickly and make informed decisions without delving into extensive documents or articles. A well-crafted summary is not merely a condensation of content; it is a bridge between complex information and its audience, ensuring that knowledge is accessible and actionable.

Evaluating the output of summarization systems is of paramount importance to ensure their effectiveness and utility. It involves assessing key factors like coherence, informativeness, and fluency. Adequate evaluation frameworks help researchers and practitioners to fine-tune algorithms, identify areas for improvement, and compare different summarization methods (Indu and Kavitha, 2016). A comprehensive evaluation not only facilitates the development of robust summarization algorithms but also guides their practical applications in real-world scenarios, addressing the increasing need for efficient content summarization in the digital age.

Numerous well-established evaluation metrics, as detailed in Section 2, are typically employed to assess the quality of generated summaries compared to reference summaries. These metrics include, but are not limited to, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), METEOR, BertScore, and MoverScore, among others. The majority of the metrics employed for the evaluation of generated summaries share a com-

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 193–201
November 1, 2023 ©2023 Association for Computational Linguistics

mon requisite; namely, the availability of reference summaries. Although reference-based evaluation methods can offer valuable insights into the performance of summarization systems, they come with inherent limitations. One significant challenge is the subjectivity of reference summaries. Summarization tasks often involve multiple valid ways to condense and express content, leading to diverse reference summaries for the same source text. Consequently, reliance on a limited set of references can introduce bias and fail to capture the full spectrum of acceptable summarization outputs (Steinberger and Ježek, 2009).

Another problem with reference-based evaluation is the issue of task-specific references. Creating reference summaries requires significant human effort, making it impractical to amass a large and diverse reference set for every possible source text. As a result, reference summaries may not adequately cover the variety of linguistic styles, domain-specific terminologies, or nuances in summarization needs, leading to biased evaluations that favor systems generating summaries similar to the available references.

Furthermore, most reference-based metrics primarily hinge on the presence or absence of specific words within generated summaries as the core element of their evaluation criteria. Nevertheless, other critical factors, such as coherence, readability, fluency, and consistency, among others, have been recognized as pivotal elements in the usability of text summaries (Fabbri et al., 2020). These essential aspects of summary assessment can be regarded as latent dimensions in the overall quality assessment.

To overcome the previously discussed challenges and in light of the recent advancements in Large Language Models (LLMs) and their widespread applicability, the Eval4NLP workshop organized a shared task. This task was specifically designed to investigate whether LLMs can be used to evaluate text summarizes, solely on the basis of the original text. With this aim in mind, the organizers provided a list of six LLMs sourced from Hugging Face, as outlined in Section 5. These models are diverse in their parameter counts and training data.

We participate in this challenge by designing different types of prompts focusing on the latent dimensions of the evaluation process. We conducted various experiments combining different prompts with the six available LLMs – including both large and small models – and evaluated the results on the training and validation sets to develop the final methodology. The final evaluation on the test set revealed that our best proposed prompt, coupled with a smaller LLM, achieved a notable Kendall τ correlation value of 0.49. This outcome positioned our system as the second-best performer in the competition.

The remainder of this paper is structured as follows: we commence with a review of related work in Section 2. Section 3 is dedicated to the dataset employed in our experiments, providing an overview of its characteristics. Subsequently, in Section 4, we delve into the solutions implemented. In Section 5, we elaborate on the experimental framework and present the results obtained. Lastly, we conclude the paper in Section 6 with a discussion of our findings and areas for future work.

2 Related Work

The quality evaluation of textual data generated in the era of natural language processing has always been seen as a difficult task because of the inherent complexity and diversity of textual data (Chen et al., 2023). The fact that a single idea can be expressed in multiple ways poses a challenge for reference-based methods, as they cannot cover all possible scenarios comprehensively, besides the costs of preparing the references for the evaluation. On the other hand, creating dependable reference-free metrics is not a straightforward endeavor and can be problematic as they must be able to correctly evaluate the different summaries generated from a same source text.. Traditional metrics of summarization quality have also failed to take into account important aspects such as coherence, fluency, and consistency (Zhang et al., 2019; Shen et al., 2022).

Various reference-based evaluation metrics are frequently used in text generation tasks. Some of the important ones are as follows: ROUGE stands as a widely adopted metric in the assessment of summarization quality. It quantifies the degree of overlap in n-grams between the generated summary and the reference summary. ROUGE is computed for various word n-gram sizes, such as 1-gram, 2-gram, and 3-gram, and the resulting scores are aggregated to produce a comprehensive evaluation score(Lin, 2004).

BLEU is another reference-based metric used to assess the quality of machine-generated text summaries by measuring how closely they match human-written reference summaries. It quantifies the precision of n-grams in the machine summary that also appear in the reference summary, providing a score that indicates the summary's accuracy and fluency (Papineni et al., 2002). Though BLEU and ROUGE both evaluate language quality, they diverge in their emphasis and methodology. BLEU places a primary focus on precision, whereas ROUGE prioritizes recall as its key metric.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a text summarization metric that evaluates the quality of machine-generated summaries by considering a variety of linguistic aspects, including unigram matching, stemming, synonyms, and word order. It provides a comprehensive measure of overall summary quality and can account for different ways of expressing the same information, making it a robust evaluation metric for text summarization (Banerjee and Lavie, 2005). MoverScore (Zhao et al., 2019) and CHRF (Popović, 2015a) assess the quality of generated summaries by comparing character n-grams between the generated summary and human reference summaries. CHRF accounts for both precision and recall and is particularly useful for languages with complex morphology and word forms(Popović, 2015b).

Moving away from the reference-based approach, Scialom et al. (2019) have introduced new metrics that rely on question-answering and demonstrated their positive outcomes when employed as rewards in a reinforcement learning setting. Importantly, these metrics do not depend on human references and can be computed directly from the text to be summarized. In another study by Chen et al. (2023), the authors explored the viability of LLMs, focusing on ChatGPT and the text-davinci series models, for reference-free text quality assessment. They conducted a comparative analysis of various techniques for evaluating text quality and identified the utilization of an explicit score generated by the GPT model as the most efficacious and consistent approach. They also discussed prompt design as an important factor influencing quality of scores generated by GPT model.

BertScore is another reference-free text summarization metric that leverages BERT (Bidirectional Encoder Representations from Transformers) embeddings to measure the similarity between the machine-generated summary and human reference summaries. It considers contextual information and semantic similarity, providing a more nuanced and accurate evaluation of summary quality(Zhang et al., 2019).

Chen and Eger (2023), introduces a novel approach by advocating the direct utilization of pre-trained Natural Language Inference (NLI) models as evaluation metrics. Furthermore, they developed a novel preference-based adversarial test suite for machine translation and summarization metrics. With this approach, there is no need for human annotators and it is particularly well-suited for reference-free evaluation. Additionally, their research findings indicate that NLI metrics exhibit strong performance in the context of summarization but yield results below the established standard metrics in the domain of machine translation. In the study conducted by (Kocmi and Federmann, 2023), GEMBA, an assessment method based on GPT technology, was introduced. The researchers conducted an evaluation of their metrics by comparing them to the metrics included in the WMT22 Metrics shared task. Remarkably, their approach demonstrated state-of-the-art performance on the MQM 2022 test set across three distinct language pairs: English to German, English to Russian, and Chinese to English.

Fernandes et al. (2023), did a comprehensive analysis of the potential of large language models in the context of machine translation evaluation through score prediction. They introduced a novel prompting technique known as AUTOMQM, which effectively harnesses the Multidimensional Quality Metrics (MQM) framework for the purpose of achieving interpretable machine translation (MT) evaluation using Large Language Models (LLMs).

A study by Goyal et al. (2022) aimed to assess the alignment of current reference-free evaluation metrics with human preferences when ranking summarization systems. They focused on two principal categories of metrics: *quality* and *factuality* metrics. Within the quality metrics, they examined SUPERT (Gao et al., 2020), which assesses the quality of generated summaries by contrasting them with automatically identified pivotal sentences from the input, along with BLANC (Vasilyev et al., 2020), which scrutinizes summaries via language understanding tasks. The second category of metrics is specifically designed to gauge the presence of inaccuracies in generated summaries concerning the source article.

Ermakova et al. (2019) provided a comprehensive overview of existing metrics for summary evaluation. They pointed out various limitations in these existing evaluation frameworks and introduced an automatic evaluation framework that eliminates the need for human annotations. They categorized the evaluation metrics into *informative* metrics like ROUGE and *readability* metrics including coherence, conciseness, content, grammar, recall, pithiness etc. Sai et al. (2022) conducted another extensive survey of the currently available automatic evaluation metrics in the domain of Natural Language Generation (NLG). They subsequently introduced a systematic taxonomy to categorize these evaluation metrics, with the categorization structured around the methodologies they employ.

Jain et al. (2023), showed that in-context learning can serve as a viable alternative to fine-tuned evaluation metrics for assessing NLG tasks. By employing a limited set of examples, in-context learning evaluators can achieve, and in some cases surpass, the current state-of-the-art performance in multi-dimensional evaluation. This approach's robustness is evident across various in-context examples. Furthermore, the research reveals a strong alignment between in-context learning evaluators and human judgments when evaluating summaries generated by GPT-3.

The present study shares similarities with the previously discussed reference-free evaluation metrics in that it operates without the need for reference summaries. However, unlike other approaches that entail intricate configurations, the model introduced here solely relies on straightforward prompts used with pre-trained LLMs.

3 Data and Evaluation

In the Eval4NLP 2023 shared task, the dataset provided for the summarization track comprises training and validation subsets, each containing source texts along with their corresponding summaries. These summaries have been generated by a summarization model that was trained on the CNN/DailyMail dataset, as documented by (Fabbri et al., 2020). Notably, the training dataset includes associated scores for each generated summary relative to its source text, which are intended for use in the system development process.

Furthermore, the organizers have also introduced a test set, which encompasses sentences and paragraphs extracted from English Wikipedia pages created subsequent to the date of July 15, 2023 (i.e., beyond the LlAMA2 training cutoff) (Leiter et al., 2023). For a comprehensive overview of the dataset, including key statistics, please refer to Table 1.

The validation and test data sets do not include explicit score annotations, necessitating participants to submit their results on the shared task page hosted on CodaBench [1]

The evaluation process in this study adheres to the metrics established in the WMT22 competition, as described by (Freitag et al., 2022), and employs segment-level Kendall correlation as the primary evaluation metric. In the realm of statistics, the Kendall rank correlation coefficient, commonly known as Kendall τ coefficient, is a statistical measure employed to assess the ordinal association between two measured variables. A τ test, which is a non-parametric hypothesis test used to determine statistical dependence based on the τ coefficient, is employed for this purpose. The ranking of systems in the shared task will be determined by their Kendall correlation scores on the test set, with the highest correlation indicating superior performance.

4 Solution

Irrespective of the type of summarization, whether it pertains to single or multi-document summarization or falls within the categories of abstractive or extractive summarization, certain fundamental criteria must be met by any generated summary. As highlighted by ter Hoeve et al. (2020), five of these criteria include: (1) *coherence* (does information flow logically from one sentence to the next?), (2) *completeness* (does the summary capture the most important information from the text?), (3) *conciseness* (is the summary brief and to the point?), (4) *consistency* (does the information in the summary align with that in the original text), and (5) *readability* (is the summary written in a clear and understandable manner?). Additionally, adhering to the conventions of correct language *syntax* stands as an imperative prerequisite, representing a sixth criterion complementing the other aforementioned factors for any text generated for various purposes.

In our approach to the Eval4NLP shared task, we devised straightforward prompts encompassing the six latent dimensions mentioned above. These

[1] https://www.codabench.org/competitions/1359/#/pages-tab

	Number of samples	Average length of the source text	Average length of the summary
train	320	361.56	62.08
validation	1280	358.77	63.21
test	825	199.57	38.55

Table 1: Statistics of data used for the experiments.

	Model Name
M1	Guanaco-65B-GPTQ
M2	Platypus2-70B-Instruct-GPTQ
M3	Nous-Hermes-13b
M4	OpenOrca-Platypus2-13B
M5	WizardLM-13B-V1.1-GPTQ
M6	orca_mini_v3_7b

Table 2: List of LLMs provided by task organizers

prompts were then input into the LLMs provided by the organizers, as detailed in Table 2. In Table 3, we present an overview of the prompts tailored to each of the evaluation factors. Our aim was to keep the prompts as simple as possible, instructing the LLMs to produce a score ranging from 0 to 100 for each pair of (source text, generated summary). Furthermore, we combined the prompt definitions for all these factors to create a single comprehensive prompt, denoted as "All."

Subsequently, we proceeded to assess the performance of the six varying-sized LLMs by employing all the prompts on both the training and validation datasets (see Section 5 for results). Following this evaluation, and guided by the outcomes obtained from the training and validation data, we selected the most promising prompt for application to the test dataset. Subsequently, we submitted the results for evaluation to CodaBench (Xu et al., 2022) to obtain the final scoring.

5 Experiments

In line with the prompt design outlined in Section 4, we leveraged the computational resources offered by the Canada Digital Alliance to apply the designated models with diverse prompts across both the training and validation datasets.

Table 4 presents the performance results in terms of Kendall τ on training and validation data. It is important to emphasize that the performance metrics for the training data were calculated using the available reference scores. However, for the validation data (which did not include reference scores), the performance metrics were computed by submitting the scores through the CodaLab page of the SharedTask.[2].

The organizers categorized models with parameters fewer than 25b as "small" and the rest as "large" models. We conducted experiments across all these models, and the performance variations, as indicated in Table 4, underscore how the model's effectiveness depends on the nature of the prompts they receive. Notably, it becomes evident that, in general, models M3 and M4 (both small models) consistently outperform the others across various prompt types. It is pertinent to observe that leveraging a prompt in conjunction with a specific model might yields superior results compared to other prompt-model combinations.

When evaluated on the training data, the best performance was achieved by the following prompts (in ranked order): P7, P2, P1, P5, P6, P4, and P3. In contrast, for the validation data, a slightly different order emerged, with P5, P2, P6, P7, P4, P1, and P3 being more effective. This variation is reasonable given that the source texts and generated summaries for the two datasets originate from different sources.

Subsequently, we proceeded to apply certain model-prompt combinations that had demonstrated promising results during the training and validation phases to the released test data. The performance of these selected model-prompt pairs, as evaluated by the organizers on the test data, is presented in Table 5.

Upon comparing the similarity between the results from the validation and test sets, it becomes evident that the test set exhibits greater similarity to the validation data rather than the training data. These results confirm that the utilization of large-scale language models (i.e. the LLMs with an extensive parameter count) without fine-tuning does not consistently yield high performance in the context of evaluation score generation tasks. In addition, the best results were achieved using the prompt for *syntax*, emphasizing the significance

[2] https://codalab.lisn.upsaclay.fr/competitions/15072#participate-submit_results

	Name	Prompt Definition
P1	**ALL**	The summary of a source text should be coherent and easy to understand', with a clear beginning, middle, and end.\n Summary completeness is a measure of how well a summary captures the most important information from the source text. \n A summary with high completeness will include all the key points and main ideas from the source text, while a summary with low completeness may omit or overlook important information.\nA summary is concise if it is brief and to the point, avoiding unnecessary details and using clear language to convey the main idea of the source text.\n A summary is readable if it is written in a clear and understandable manner. It should use simple language, concise sentences, and organized structure to effectively convey the main points of the source text.\n A summary is syntactically correct if it has proper sentence structure and arrangement of words. This includes using correct word order, subject-verb agreement, and appropriate use of phrases and clauses to convey the intended meaning accurately. \n Summary and the source text are consistent if summary accurately reflects the main ideas and key information of the source text without introducing new or conflicting information.\n The summary should align with the overall message, tone, and context of the original document to maintain coherence and reliability.\nGive a consistency score between 0 and 100 to the summary created from the source text.\n Zero means that 'summary and source text are not consistent, summary is not complete, coherent, readable, concise, and syntactically correct' at all and 100 means summary is 'fully consistent, coherent, readable, concise, complete and syntactic.'
P2	**Coherence**	The summary of a source text should be coherent and easy to understand', with a clear beginning, middle, and end.\n Give a coherence score for the given summary of the source text below on a continuous scale from 0 to 100, \n where a score of zero means 'no coherent' and score of one hundred means 'fully coherent'.
P3	**Completeness**	Summary completeness is a measure of how well a summary captures the most important information from the source text. \nA summary with high completeness will include all the key points and main ideas from the source text, while a summary with low completeness may omit or overlook important information.\nGive a completeness score between 0 and 100 to the summary created from the source text. \nZero means a 'very incomplete' and 100 means 'a complete summary.'
P4	**Conciseness**	A summary is concise if it is brief and to the point, avoiding unnecessary details and using clear language to convey the main idea of the source text.\nGive a conciseness score between 0 and 100 to the summary created from the source text. Zero means a 'inoncise' and 100 means a 'fully concise summary.'
P5	**Consistency**	Summary and the source text are consistent if summary accurately reflects the main ideas and key information of the source text without introducing new or conflicting information.\nThe summary should align with the overall message, tone, and context of the original document to maintain coherence and reliability.\n Give a consistency score between 0 and 100 to the summary created from the source text.\n Zero means that 'summary and source text are not consistent' at all and 100 means they are 'fully consistent.'
P6	**Readability**	A summary is readable if it is written in a clear and understandable manner. It should use simple language, concise sentences, and organized structure to effectively convey the main points of the source text."\n Give a readability score between 0 and 100 to the summary created from the source text.\n Zero means the 'summary is not readable' and 100 means summary is 'fully readable.'
P7	**Syntax**	A summary is syntactically correct if it has proper sentence structure and arrangement of words. This includes using correct word order, subject-verb agreement, and appropriate use of phrases and clauses to convey the intended meaning accurately. \n Give a syntax score between 0 and 100 to the summary created from the source text.\n Zero means a 'the syntax is completely unacceptable' and 100 means the syntax of summary is 'fully correct.'

Table 3: Prompts' Definition

		Train							**Validation**						
		P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Large Models	**M1**	0.36	0.44	0.25	0.41	**0.45**	**0.42**	0.40	0.35	**0.45**	0.26	**0.42**	0.42	**0.44**	0.42
	M2	0.24	0.25	0.21	0.22	0.24	0.11	0.22	0.27	0.22	0.21	0.24	0.26	0.12	0.22
Small Models	**M3**	**0.45**	**0.47**	**0.4**	**0.42**	0.41	0.41	**0.49**	**0.41**	0.22	**0.41**	0.41	**0.45**	0.4	**0.43**
	M4	**0.45**	**0.47**	**0.4**	**0.42**	0.41	0.41	**0.49**	**0.41**	0.44	**0.41**	0.41	**0.45**	0.4	**0.43**
	M5	0.17	0.18	0.12	0.27	0.33	0.26	0.26	0.23	0.18	0.17	0.28	0.31	0.28	0.28
	M6	0.36	0.32	0.32	0.35	0.39	0.31	0.38	0.36	0.36	0.35	0.33	0.37	0.33	0.37

Table 4: Performance of different models with different prompts in terms of Kendall τ. M1:Platypus2-70B-Instruct-GPTQ, M2:Guanaco-65B-GPTQ,M3:Nous-Hermes-13b, M4:OpenOrca-Platypus2-13B, M5:WizardLM-13B-V1.1-GPTQ, M6:orca_mini_v3_7b and P1:All Explained, P2: Coherence, P3: Completeness, P4:Conciseness, P5:Consistency, P6:Readability, P7:Syntax

		P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Large Models	**M1**	-	0.46	-	-	-	0.41	-
	M2	-	-	-	-	-	-	-
Small Models	**M3**	-	-	-	-	-	-	-
	M4	0.46	-	0.47	0.45	-	-	**0.49**
	M5	-	-	-	-	-	-	-
	M6	-	-	-	-	0.44	-	-

Table 5: Performance results on test data. M1:Platypus2-70B-Instruct-GPTQ, M2:Guanaco-65B-GPTQ,M3:Nous-Hermes-13b, M4:OpenOrca-Platypus2-13B, M5:WizardLM-13B-V1.1-GPTQ, M6:orca_mini_v3_7b and P1:All Explained, P2: Coherence, P3: Completeness, P4:Conciseness, P5:Consistency, P6:Readability, P7:Syntax

of this latent dimension in the quality of the generated summaries. Syntax is largely overlooked by reference-based metrics that focus on lexical overlap between the generated summary and a reference summary; however, our results suggest that it plays an important role in evaluation. The second-highest score was achieved using the prompt for *completeness*, consistent with the idea that a summary should include the most salient points from the original text.

It is worth highlighting that regulatory constraints imposed on participants prevented us from exploring the possibility of combining the scores from various prompts and models during our experimental phase. However, by employing a solitary model, we achieved a notable second-place ranking in the competition.

6 Conclusion

The assessment of summarization system outputs is vital to ascertain their efficiency and usefulness. Traditional approaches to summarization evaluation involve comparing the generated text with human-written reference summaries. However, the constraints associated with reference-based metrics encourage the researchers and practitioners to seek reference-free metrics for the evaluation and comparison of various summarization methods.

With the objective of formulating effective prompts for utilization along with LLMs, the Eval4NLP organized a collaborative initiative. The primary goal of this endeavor was to systematically examine the potential utility of LLMs in the evaluation of text summaries, relying exclusively on the source text. In this study, we actively engaged in the development of prompts tailored to each of the six latent dimensions (i.e. completeness, conciseness, readability, coherence, consistency and syntax) found to be relevant to summary evaluation. One specifically devised prompt, centered on the syntactic assessment of generated summaries, garnered a noteworthy score of 0.49 in terms of Kendall τ, thereby securing the second-highest position among performance evaluation systems.

Our primary focus in the present work involved the utilization of individual LLMs. Nevertheless, we acknowledge that the collaborative use of various models presents a promising avenue for potential performance enhancement, which we consider as a valuable direction for future investigations.

Acknowledgments

The authors would like to express their sincere appreciation to the Natural Sciences and Engineering Research Council of Canada (NSERC). We are also grateful for the assistance and resources provided by the Canada Digital Alliance, which significantly contributed to the success of this project.

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.

Daniel O Cajueiro, Arthur G Nery, Igor Tavares, Maísa K De Melo, Silvia A dos Reis, Li Weigang, and Victor RR Celestino. 2023. A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding. *arXiv preprint arXiv:2301.03403*.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information processing & management*, 56(5):1794–1814.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.

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

M Indu and KV Kavitha. 2016. Review on text summarization evaluation methods. In *2016 international conference on research advances in integrated navigation systems (RAINS)*, pages 1–4. IEEE.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multidimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison for NLP systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. 2015a. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2015b. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

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

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. *arXiv preprint arXiv:2202.08479*.

Josef Steinberger and Karel Ježek. 2009. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Maartje ter Hoeve, Julia Kiseleva, and Maarten de Rijke. 2020. What makes a good summary? reconsidering the focus of automatic summarization. *arXiv preprint arXiv:2012.07619*.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Zhen Xu, Sergio Escalera, Adrien Pavao, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7).

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*.

Appendix

In the course of this research, we utilized the subsequent modules:

1. PyTorch: 2.0.1+cu117
2. guidance: 0.0.64
3. transformers: 4.34.5
4. auto_gptq: 0.3.2