Knowledge and Information Systems (2024) 66:7717–7738
https://doi.org/10.1007/s10115-024-02217-0

REGULAR PAPER

Evaluation metrics on text summarization: comprehensive survey

Ensieh Davoodijam[1] · Mohsen Alambardar Meybodi[2]

Received: 10 August 2023 / Revised: 30 March 2024 / Accepted: 19 August 2024 /
Published online: 31 August 2024

Abstract

Automatic text summarization is the process of shortening a large document into a summary text that preserves the main concepts and key points of the original document. Due to the wide applications of text summarization, many studies have been conducted on it, but evaluating the quality of generated summaries poses significant challenges. Selecting the appropriate evaluation metrics to capture various aspects of summarization quality, including content, structure, coherence, readability, novelty, and semantic relevance, plays a crucial role in text summarization application. To address this challenge, the main focus of this study is on gathering and investigating a comprehensive set of evaluation metrics. Analysis of various metrics can enhance the understanding of the evaluation method and leads to select appropriate evaluation text summarization systems in the future. After a short review of various automatic text summarization methods, we thoroughly analyze 42 prominent metrics, categorizing them into six distinct categories to provide insights into their strengths, limitations, and applicability.

Keywords Evaluation metric · Automatic summarization · Machine translation metric

1 Introduction

Text summarization is the process of condensing a text document into a shorter version while preserving its key points and overall context. This process relies on the utilization of natural language processing (NLP) techniques, which enable the identification of important information, semantic understanding, and linguistic analysis. NLP can automatically comprehend and extract relevant content from large volumes of text, facilitating efficient information retrieval and knowledge extraction. It plays a crucial role in automating the summarization

✉ Mohsen Alambardar Meybodi
m.alambardar@sci.ui.ac.ir

Ensieh Davoodijam
n.davodijam@gmail.com

[1] Department of Software Engineering, University of Kashan, Kashan, Iran

[2] Department of Applied Mathematics and Computer Science, Faculty of Mathematics and Statistics, University of Isfahan, Isfahan 81746-73441, Iran

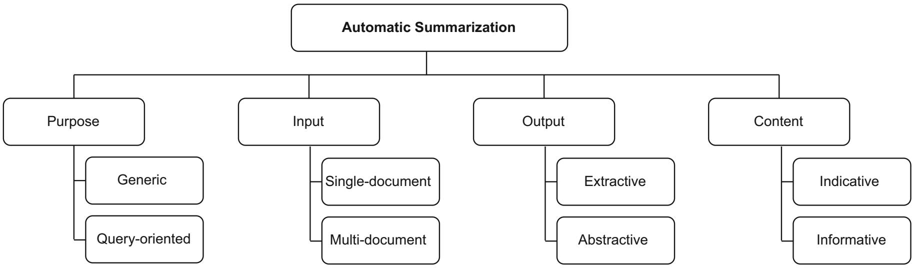

Fig. 1 Classification of Text Summarization Methods

process and generating concise summaries that capture the essence of the original text [1, 2]. Text summarization methods can be categorized into different approaches based on their characteristics and techniques. In the sequel, we will explore the classification of text summarization methods to gain a better understanding of the diverse strategies employed, their strengths, limitations, and applicability, which can aid in the selection of appropriate methods for text summarization tasks. V Noarious categorizations of text summarization approaches, taking into account factors such as the purpose, input, output, and content, are illustrated in Fig. 1. In the sequel, we explain in more detail.

- Purpose: Summarization methods can be classified into generic summarization and query-oriented summarization, which serve different purposes. In general, generic summarization provides a general overview of the main content, while query-oriented summarization generates summaries according to specific questions. The query-oriented method creates outlines corresponding to the query, providing more detailed information [3, 4].
- Input: Summarization methods are categorized as single-document summarization [5] or multi-document summarization [6–8]. During single-document summarization, information is condensed from one source document into a summary, while in multi-document summarization, information is extracted from multiple source documents to create a comprehensive summary [3, 4].
- Output: Summaries can be separated into extractive summaries and abstractive summaries [9]. Extractive summarization involves selecting the most important sentences or passages from the source documents and assembling them into a summary. In contrast, abstractive summarization goes beyond mere extraction and generates new sentences that may not exist in the original documents, creating summaries with a more natural language flow [2, 4].
- Content: Summaries can be indicative or informative, depending on their content. Informative summaries include sufficient content to convey the key information, enabling users to understand the main points without referring back to the original input. Indicative summaries provide a glimpse or overview of the content, serving as an indicator for users to decide whether they need to review the original document in more detail [3, 4].

The evaluation of summaries involves assessing their quality, effectiveness, and relevance [3, 4]. It encompasses measuring how well a summary captures the main points, maintains essential information, and conveys the intended meaning of the original text. Evaluating summaries is crucial for comparing different algorithms or approaches to text summarization and determining their performance. However, there are significant challenges in establishing a universally accepted evaluation standard for summarizing. Due to the subjective nature

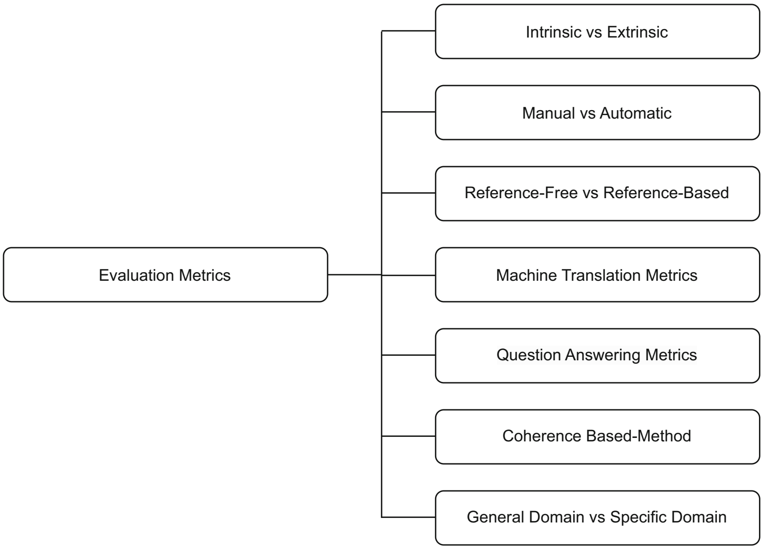
Evaluation Metrics
Intrinsic vs Extrinsic
Manual vs Automatic
Reference-Free vs Reference-Based
Machine Translation Metrics
Question Answering Metrics
Coherence Based-Method
General Domain vs Specific Domain

Fig. 2 Approaches for Evaluation Metrics in Text Summarization

of summarization, there is no clear definition of what constitutes a "good" summary. It is imperative to develop robust evaluations and metrics that accurately capture the essence of high-quality summaries in order to address the challenges associated with evaluating text summarization systems [10]. This paper provides an overview of text summarization evaluation methods. Through a comprehensive analysis of the current methods, prominent trends and challenges in evaluation are identified. Further research is also recommended, along with valuable insights and recommendations that will assist in the development of effective evaluation methodologies and metrics for this field. Hence, researchers, developers, and practitioners will gain a deeper understanding of the techniques used in text summarization evaluation.

2 Classification of evaluation metrics

Evaluation criteria in text summarization can be classified into different categories; each targets specific aspects of summarization quality and offers valuable insights into the efficacy of various metrics (see Fig. 2). These categories can be presented in detail as follows:

2.1 Intrinsic versus extrinsic

Intrinsic evaluation involves assessing the quality of the generated summaries by evaluating them based on a set of criteria, including relevance, comprehensiveness, informativeness, and accuracy [11]. It aims to measure how well a summary captures the main concepts, key information, and overall coherence. Two most famous criteria of this category, described in the sequel, are ROUGE [12] and BLEU [13]. They compare the generated summary against one or more reference summaries and provide scores based on various linguistic features. Extrinsic methods measure summary quality through a task-based performance such as the information retrieval-oriented task. Defining a meaningful extrinsic evaluation metric has been challenging for the summarization community for many years [3, 4, 14, 15]. In this approach, the quality of summaries is measured based on their utility and performance in real-world scenarios.

2.2 Manual versus automatic

Manual evaluation involves human judges who read both the reference summaries and the generated summaries and provide evaluations based on predefined criteria. This approach is considered reliable and accurate as human judges can apply their expertise, linguistic knowledge, and contextual understanding to the evaluation process. It is often considered the gold standard for evaluation and provides valuable insights for system development. However, manual evaluation is time-consuming, and expensive, and may introduce biases due to individual preferences and interpretations. In contrast, automatic evaluation relies on computational metrics and algorithms to evaluate summaries without human intervention. Automatic evaluation metrics utilize linguistic features, statistical analysis, or machine learning algorithms to assess the quality of summaries. They are convenient, fast, and can process a large volume of data. However, automatic evaluation metrics may not fully capture the nuances of a good summary and may not align with human judgments. They are often based on surface-level features and do not consider deeper semantic understanding [3, 4, 14].

2.2.1 Manual

Manual evaluations based on the reference are usually conducted by the following:

1. **Pyramid** [16, 17] relies on experts to identify occurrences of the same unit of meaning known as Summary Content Units (SCUs) in the reference summary and the candidate summary. A candidate summary that is similar to the reference summary and has many matches is more likely to be effective.
2. **Lightweight Pyramid** [18] utilizes the reference summaries, just as the original Pyramid did, and bases the score on less subjective SCU judgments to offers a more efficient and cost-effective approach to evaluation. The Lightweight Pyramid employs statistical sampling rather than exhaustive SCU extraction and testing, resulting in a reduced overall cost. It offers a practical alternative for evaluating summaries, particularly in scenarios where time and resources are limited.

2.2.2 Automatic

Automatic evaluation metrics can be classified into five primary categories: Pyramid-based, N-gram-based, N-gram graph-based, Basic element n-gram, and BERT-based methods. Each category offers a unique view to assess the quality and effectiveness of text summarization, as explained below:

2.2.2.1 Pyramid based Pyramid-based metrics aim to provide a quantitative assessment of summary quality by incorporating the principles and techniques of the original pyramid method. By automating the evaluation process, these metrics offer a more efficient and scalable approach for evaluating summaries.

3. **PEAK(Pyramid Evaluation via Automated Knowledge Extraction)** [19] employs open information extraction to identify subject–predicate–object triples and a graph is

created using these triples in order to identify and weigh salient triples. This algorithm uses the Munkres–Kuhn bipartite graph algorithms [20] to find the optimal assignment of model SCUs to target summaries.

4. **PyrEval** [21] is an automated tool designed to implement the manual pyramid method for content analysis in the evaluation of automatic summarization. It employs low-dimensional distributional semantics and the EDUA algorithm, Emergent Discovery of Units of Attraction [22], to build a content model using vectorized phrases. PyrEval enables efficient application of pyramid content evaluation without the need for retraining and has been extensively tested on various datasets containing human-written and machine-generated summaries, demonstrating good performance in both scenarios.

2.2.2.2 N-gram based N-gram based evaluates the quality of a summary by comparing the n-grams, which are consecutive sequences of n words, between the reference summary and the generated summary. In the following, some methods that utilize n-grams are listed.

5. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** [12] measures the overlap between the system-generated summaries and reference summaries produced by humans. There are multiple versions of ROUGE, each designed to evaluate different aspects of summaries. The selection of a particular version depends on the specific requirements of the task and the aspects of the summaries that need to be assessed:

   - **ROUGE-N** is a metric used to evaluate the similarity between the summary created by a system and the reference summary by measuring the overlap of n-grams. The value of n can be set to any integer, with $n = 1$ being the most commonly used value. In spite of ROUGE-N's simplicity and effectiveness, it does not take into account word order or semantic similarity [12].
   - **ROUGE-L** calculates the longest common subsequence (LCS) between the system-generated summary and the reference summary. It captures the similarity in terms of the longest sequence of words that appears in both summaries, regardless of their order or position. It focuses on evaluating the content overlap and can provide insights into the complete coverage of the generated summary compared to the reference [12].
   - **ROUGE-W** follows ROUGE-L's methods, but it gives each word a weight based on its position in the summary. Words at the start or end of the summary receive more weight than words in the middle [12].
   - **ROUGE-S** measures the skip-bigram overlap between the system-generated summary and the reference summary. A skip-bigram is a pair of words that are not adjacent but are separated by no more than k words. ROUGE-S is designed to capture semantic similarity, making it more effective than ROUGE-N for summaries that use different words to express the same ideas [12].
   - **ROUGE-SU** is similar to a ROUGE-S, but includes matches for unigrams in addition to skip-bigrams, which can be useful for longer summaries [12].
   - **ROUGE-X** is a generalization of the ROUGE-N metric which measures overlaps in terms of the number of n-grams of length X [12].
   - **ROUGE-WE (R-WE)** utilizes a more flexible approach based on the cosine similarity of word embeddings rather than relying on the hard lexical matching of bigrams [23].

When using n-gram-based evaluation metrics in text, there are two main issues. These metrics need one or more reference references, which can be hard and expensive to acquire, especially for tasks like summarizations or dialogues, where there are a variety of acceptable outputs.

In addition, n-gram-based approaches are insensitive to minor changes because, without references, they weigh all parts of the text equally, making them insensitive to errors. The n-gram-based metrics may have difficulties in detecting inconsistencies between two sentences that have opposite meanings, such as "I am writing my paper in Vancouver." and "I am not writing my paper in Vancouver" [24].

2.2.2.3 Basic element N-gram Similarities with ROUGE can be observed in the evaluation metrics known as Basic Elements [25] and BEwT-E [26]. These metrics involve the comparison of "basic elements," which are groups of semantic units extracted from the sentence structure of the summaries, rather than evaluating individual words.

6. **Basic Elements(BE)** [25] addresses some of the limitations of n-grams by using extremely small units of content, called Basic Elements. Three subproblems are required to assess the content of a summary, corresponding to the three modules in the BE Package. Preparation and scoring are both phases of the BE Package. In the preparation phase, the first module identifies reference BEs from the reference summaries, the second module merges similar reference BEs, and the third module scores each reference BE. During the second phase, the summary is broken up into separate lists of BEs; the first module compares each BE to the reference list of BEs; the second module assigns a score to each BE to be rated and computes an overall score for all BEs contained in the summary.
7. **BEwT-E (Basic Elements with Transformations for Evaluation)** [26] is referred to as the new and improved implementation of the BE method, known as BE with Transformations for Evaluation. The process is divided into three steps: Extracting BEs, Weighing BEs, and Definition of Transformations. In its original form, BE matched primarily based on lexical identity, and a machine translation program expanded it by paraphrasing through a large list of alternative phrases. In contrast, BEwTE employs transformations to match BEs that share similar semantic content, but differ in lexical structure.

2.2.2.4 N-gram graph based N-gram graph based represents summaries as graphs based on character n-grams and compute a similarity score between them. The following explanation provides a more detailed description of this process:

8. **AutoSummENG(AUTOmatic SUMMary Evaluation based on N-gram Graphs)** [27] compares texts with gold-standard summaries using n-gram graphs and other histograms. Each text has an n-gram graph and each model summary has another n-gram graph. A Value Similarity (VS) and a Cooccurrence Similarity (CS) measure the similarity between the evaluated text and each model summary. Based on these similarities, an overall performance is calculated.
9. **MeMoG (Merged Model Graph)** [28] uses the n-gram graph framework to create a single, "centroid" graph representation of a set of documents. Instead of averaging over individual model texts, all model texts are merged into a single graph and compared to the evaluated summary graph.
10. **NPowER: (N-Gram Graph Powered Evaluation via Regression)** [29] combines n-gram graph approaches with machine learning principles to provide statistically and language-independent n-gram graph-based approaches (AutoSummENG and MeMoG). Regression is used to model how n-gram graph evaluations can be consolidated to form a final grade for a summary using individual evaluations as features. By analyzing the primary evaluation scores of different methods, it develops a second-level grade estimator that is built as a regression problem, and based on that estimate, a target grade for example, responsiveness or Pyramid score.

2.2.2.5 LLMs based Large Language Models (LLMs) are pre-trained models that were developed by Google in 2017. LLMs adopt deep learning methods and transformer architecture on massive datasets. There are different types of LLMs such as GPT [30] and BERT. We introduce evaluation metrics based on these models as follows:

- **GPT Based** GPT (Generative Pretrained Transformers) is a family of deep learning models developed by OpenAI. It includes GPT-1, GPT-2, GPT-3, and GPT-4 [30]. The following refers to some metrics that are based on GPT:

  11. **G-EVAL** [31] is a prompt-base metric that employs Chain-of-Thoughts (CoT) and GPT$-$3.5 and GPT-4. It estimates the probabilities of the token for text summarization and dialogue generation.
  12. **GPTScore (generative pre-trained models to score generated text)** [32] calculates the score with zero-shot instruction of different pre-trained models from FLAN-T5-small to GPT-3.

- **BERT based** The Bidirectional Encoder Representations from Transformers (BERT) language model has become a popular language model developed by Google. Using Transformer architecture, it analyzes both the left and right context of words in a sentence to understand their meaning and context. By training on a large amount of text data, BERT can be fine-tuned for various natural language processing tasks, including text classification, named entity recognition, question answering, etc. The BERT algorithm is capable of understanding the semantic meaning of sentences and capturing contextual representations of words [33]. Some metrics utilizing this approach are mentioned below:

  13. **BERTScore** [34, 35] generates embeddings for both candidate (machine-generated) and reference texts using this metric. In addition to the position of the word in the sentence, these embeddings capture the contextual and semantic meaning of the word. BERTScore uses a variant of the Word Mover's Distance algorithm to compare embeddings that consider not only lexical similarity but also contextual and positional differences. By comparing the candidate text with the reference text, BERTScore provides a score ranging from 0 to 1, where a score of 1 indicates that the candidate text and reference text are perfectly in line.
  14. **MoverScore** [36] is a modified version of the BERTScore metric that combines contextualized representations with a distance measure to create a cost matrix. This metric is designed to capture semantic similarity between tokens and is found to be highly correlated with human judgments.
  15. **InfoLM** [37] is a new metric designed to assess text summarization and data-to-text generation without training. Using the candidate and reference sentences, a pre-trained masked language model (PMLM) computes discrete probability distributions over the vocabulary. Unlike other BERT-based metrics, InfoLM directly relies on the PMLM's probability distributions, avoiding layer selection and aggregation techniques. By considering synonyms, exact string matching, and distant dependencies, it overcomes the limitations of string-based metrics.

2.3 Reference-free versus reference based

A key consideration is whether or not a reference summary is required for automatic evaluation metrics, which are categorized similarly to manual evaluation metrics. A reference summary identifies key information found in an input document and is compared to a candidate summary to determine if they are similar. As a consequence, reference-free evaluations create a model of important content in the document that can be used either directly or indirectly in evaluating the candidate summary. A number of proposed methods exist in academic literature that model the salient content themselves, but they are not as widely used as reference-based metrics [38]. This approach is utilized in some metrics in the following section:

16. **HIGHRES (HIGHlight-based Reference-less Evaluation of Summarization)** [39] offers a highlight-based evaluation method that provides absolute evaluation and avoids reference bias. Highlight-based evaluation is particularly suitable for abstractive summarization tasks and can be efficiently crowdsourced without the need for expert annotations. There are three main components of the HIGHRES framework: document highlight annotation, content evaluation using highlight annotation, and clarity evaluation using highlight annotation. It surpasses limitations imposed by n-gram overlap and provides an effective approach for evaluating the quality of summaries.
17. **BLANC** [40] is primarily concerned with retrieving masked tokens, or the Cloze task [41], in which a model reconstructs obscured text spans using the BERT to determine which masked text tokens to retrieve. As BLANC emphasizes the masked token task rather than human-written reference summaries for evaluation, it allows for an objective and automated assessment that does not rely on human-written references. Studies evaluating BLANC have shown that it exhibits a similar correlation to human evaluations as the widely used ROUGE metrics.
18. **SUPERT (SUmmarization evaluation with Pseudo references and bERT)** [42] provides a method for evaluating the summarization of multiple documents without the need for human annotations or reference summaries. SUPERT utilizes contextualized embedding and soft token alignment techniques to evaluate how similar a summary is to a pseudo-reference summary, a set of selected salient sentences from the source documents. There are two steps in the SUPERT process: first, important information is selected from input documents to create a pseudo-reference summary, and then, the overlap between the pseudo-reference summary and the summary being evaluated is calculated.

2.4 Machine translation metrics

Machine translation metrics are measures designed to evaluate the quality of machine-generated translations. These metrics provide quantitative scores that indicate the similarity or adequacy of the translated output compared to one or more references. Machine translation metrics can be adapted and used for text summarization evaluation by treating the summarization task as a form of translation [43, 44]. The following are some of the most important machine translation metrics that can be used for text summarization evaluation.

19. **BLEU (Bilingual Evaluation Understudy)** [13] compares overlaps based on n-gram precision. By using a numerical score, the metric determines how close a machine translation is to a reference translation. This metric was developed by modifying the word error rate metric used in speech recognition to accommodate multiple reference translations and variations in word choice and order. Different weighting schemes are considered when BLEU calculates the weighted average of variable-length phrase matches.
20. **chrF** [45] uses character n-grams to calculate the similarity between reference and candidate sentences. The chrF algorithm applies character n-grams as opposed to word n-grams as in BLEU and ROUGE. By combining the precision and recall values generated by

various values of n (up to 6) using arithmetic averaging, the precision (chrP) and recall (chrR) scores can be determined.

21. **BLEURT** [46] is based on BERT and uses a few thousand potentially biased training examples to model human judgments. In order to train the system, BLEURT uses a unique pretraining scheme that uses random perturbations of Wikipedia sentences combined with a wide variety of lexical and semantic supervision signals.
22. **METEOR (Metric for Evaluation of Translation with Explicit Ordering)** [47] addresses the limitations of the BLEU metric, which matches unigrams between machine-generated and human-produced translations. This score is determined by combining unigram precision, unigram recall, and fragmentation. Various versions of this metric have been introduced to enhance its performance [47–51].
23. **Meteor Universal** [52] is an extension of the Meteor metric that allows for language-specific evaluation in any target language by utilizing linguistic resources. Based on exact, stem, synonym, and paraphrase matching, the Meteor metric scores are provided for English, Czech, German, French, Spanish, and Arabic.
24. **Prism-src (Probability is metric)** [53] is used to evaluate translated texts without references. All possible paraphrases of a sentence are represented and the similarity between the two texts can be measured without the use of human quality judgments. It uses a sentential sequence-to-sequence paraphraser.
25. **COMET-QE (Crosslingual Optimized Metric for Evaluation of Translation- Quality Estimation)** [54] is an altered version of COMET [55] that uses cross-lingual pre-trained language modeling for training machine translation evaluation models. With COMET, segment-level representations are encoded using the cross-lingual, pre-trained model XLM-LMRoBERTa [56] differs from reference-based COMET in that there is no reference, and, therefore, the combinations of features used for input to the feed-forward regressor differ as well.
26. **SUM-QE (Summary Quality Estimation)** [57] is a new quality estimation model for summarization, based on BERT. To predict linguistic quality scores, this model adds a task-specific layer to a pre-trained BERT model. An advantage of this model is that it evaluates aspects of linguistic quality that are not directly addressed by content-based methods of summarizing results. References are not required, and summaries are evaluated according to five specific linguistic qualities: grammar, non-redundancy, referential clarity, focus, structure, and coherence.
27. **BARTSCORE** [58] uses BART [59], as a sequence-to-sequence pre-trained model and measures information, fluency, and factual text. It provides an evaluation of sentence generation by the weights of BART. Weights are used to place different emphasis on tokens, which can be implemented in various ways, such as Inverse Document Frequency (IDF). BARTSCORE increases in computational complexity as all token probabilities and weights are calculated.
28. **RUSE (Regression-based Unsupervised Sentence Embeddings)** [60] combines three pre-trained sentence embeddings: InferSent, Quick-Thought, and Universal Sentence Encoders in a supervised regression model. Rather than focusing on local features such as characters or words n-grams, these sentence embeddings capture global information about sentences. A Multilayer Perceptron (MLP) and Support Vector Regression (SVR) regressor are used to calculate the RUSE score by using both hypothesis (system-generated summary) and reference (gold standard summary) embeddings.
29. **YiSi** [61] is a metric for evaluating neural machine translation quality. There are three variants of the YISI metric: YISI-0, YISI-1, and YISI-2. It is designed to assess machine translation quality in low-resource languages, monolingual languages, and cross-lingual

languages. With YISI-0, resources are free, and the longest common character substring accuracy and word frequency inverse document are used. The YISI-1 variant is monolingual and evaluates lexical semantic similarity using an embedding model. YISI-2 is a cross-lingual variant of YISI, requiring cross-lingual embeddings to determine cross-lingual lexical semantic similarity as well as calculating cosine similarity for cross-lingual lexical representations.

2.5 Question answering metrics

In a question answering (QA) evaluation framework, metrics are applied to evaluate a reference summary by transforming it into a set of question answer pairs. After determining the proportion of questions that can be answered correctly by the candidate summary, these metrics evaluate the candidate summary's ability to include the information contained in the reference summary. A candidate summary's quality is directly measured by QA-based metrics due to the fact that accurate answers are a function of the information contained in it, thus providing a signal of its quality that is not adequately captured by metrics based solely on the textual overlap [62].

30. **APES (Answering Performance for Evaluation of Summaries)** [63] evaluates summaries based on learned reading comprehension. Named entities are removed from reference summaries and predicted from candidate summaries. In order to evaluate summaries, it is beneficial to reduce them to an extrinsic task such as answering questions. APES program begins by receiving news article summaries, question-and-answer pairs relevant to the central information in the article, and an automatic QA system. After that, use this QA system to determine how many questions were answered correctly.
31. **SummaQA** [64] presents a question answering metric that does not use human annotation. It applies F1 scores and QA confidence metrics at the document level as well. Based on the APES approach described above, it proposes two unsupervised QA metrics, QAfscore(unsup) and QAconf(unsup), that account for both the quality and informativeness of the generated summary.
32. **QAEval** [62] is a reference-based metric that represents reference summary information through a set of questions and answers that are generated automatically from the reference. The process consists of four steps: Answer Selection, Question Generation, Question Answering, Answer Verification, and Scoring. In QAEval, questions are asked about noun phrases, while in APES, questions are asked about named entities only.
33. **FEQA (Faithfulness Evaluation with Question Answering)** [65] addresses the issue of evaluating the faithfulness of generated summaries. Using a learned model, it generates a set of "ground truth" QA pairs. In the next step, off-the-shelf reading comprehension models are evaluated using the answer spans extracted from the source documents. In terms of factual consistency, high accuracy indicates that the summary and the source document produce similar answers.
34. **QAGS (Question Answering and Generation for Summarization)** [24] measures the factual consistency of abstractive summaries. In order to generate the final QAGS score, questions based on the summary are generated. By utilizing both the source and the summary, the current framework designed for answering questions can be applied to other conditional text generation tasks like image captioning or machine translation.
35. **Q-Metrics** [66] proposes modifying existing metrics to emphasize answerability and N-gram similarity. Based on the task type (document quality assurance, knowledge base quality assurance, visual quality assurance), the weights for answerability and N-gram

similarity can be adjusted accordingly. In order to capture the answerability of the question, additional weights have been proposed for question types, content words, functions, and named entities. For the determination of these weights, a small amount of human-annotated data can be utilized; however, the weights may vary from task to task.

36. **QuestEval** [67] consists of Question Answering (QA) component and Question Generation (QG) component that evaluates factual consistency and relevance without any ground-truth reference. The framework combines previous QA approaches and extends them with question weighting, ensuring that factual consistency, relevance, and information selection are all taken into account.

2.6 Coherence-based method

Coherence occurs when sentences are interconnected and convey a unified meaning, while random sequences of sentences lack this cohesion and can be difficult to understand. The context of a sentence can often be helpful in understanding the meaning of its words. Coherence models differentiate between coherent and incoherent texts and can be applied to text generation, summarization, and scoring [68]. There are two widely recognized tasks in coherence modeling: sentence ordering (SO) and summary coherence rating (SCR). This section, however, focuses specifically on SCR as a metric to evaluate the coherence of summaries. In fact, incorporating the SCR metric as an evaluation criterion provides valuable insight into the coherence quality of automatic summarization systems. Researchers and developers can evaluate the generated summaries based on their coherence, which ensures logical connections, smooth transitions, and overall flow of information. As a result of formal discourse theories, several coherence models have been proposed. These models can be categorized into separate categories [69].

2.6.1 Entity grid approach

37. **Modeling Local Coherence: An Entity-Based Approach** [70] is an entity grid representation of discourse, suitable for the ranking-based generation and text classification tasks. For each text, an entity grid is employed, which consists of rows corresponding to sentences and columns corresponding to discourse entities. Discourse entities are groups of coreferent noun phrases. In addition to abstracting a text into entity transition sequences, it also collects distributional, syntactic, and referential information. In summarization evaluation, it compares the rankings produced by the model to human coherence judgments.
38. **A Neural Local Coherence Model** [71] identifies entity transitions and arbitrary entity-specific features. A distributed representation of entity transitions and entity features is used to achieve generalization. Also presented is an end-to-end approach to learning task-specific features. This architecture employs a convolutional neural network (CNN) to represent entity transitions and features.

2.6.2 Entity graph approach

There are limitations to grid-based approaches, despite their advantages, including data sparsity, dependence on specific domains, and computational complexity [72]. It is suggested to model local coherence using a graph-based approach for representing entities, followed by applying centrality measures to nodes within the graph:

39. **Graph-based Local Coherence Modeling** [73] provides an unsupervised, computationally efficient method of modeling local coherence based on graphs. Entity grids are used as the incidence matrix of bipartite graphs capturing text structure. This bipartite graph contains all the information needed to calculate local coherence, so there is no need for separate learning phases or feature vectors.
40. **A Neural Graph-based Local Coherence Model** [68] explores the advantages of encoding entity graphs with Relational Graph Convolutional Networks (RGCNs) for determining local coherence. It proposes a neural graph-based model that constructs a graph of relationships among sentences in a given text based on entity-based and linear relationships. Furthermore, this model uses convolutional networks to extract coherence features from these graphs.

2.7 General domain versus specific domain

The performance of a candidate summary depends on the domain in which it is used. A number of approaches have proved effective in evaluating the quality of summaries in the biomedical domain, such as SERA [74]. In contrast, SummTriver [75] another evaluation tool may be more accurate when assessing the coherence and relevance of summaries in the news domain. To ensure appropriate evaluation of automatic summarization systems, it is necessary to consider the specific characteristics and demands of the given domain during the decision-making process.

41. **SERA (Summarization Evaluation by Relevance Analysis)** [74] evaluates summaries by comparing the relevance between a machine-generated summary and its corresponding human-written version. A content quality score can be underestimated if only lexical overlaps are used to describe the same concepts. SERA proposes an approach based on the premise that concepts have meanings based on their context and that related concepts frequently co-occur.
42. **SummTriver** [75] evaluates trivergence between the distributions of events in the candidate summary, the distributions of events in the document generated from the collection of candidate summaries, and the distributions of events in the document itself as part of the method by considering multiple candidate summaries. To evaluate the trivergence, the system employs several compositions of divergences between these probability distributions. It provides a comprehensive analysis of trivergence between candidate summaries and the underlying documents, which is important to the evaluation of summarization systems.

3 Challenge

The main challenges related to text summarization evaluation categorized as follows:

- Subjectivity: There are different opinions on what constitutes a good summary, and summarizing text is a subjective task. The lack of a universal standard makes it difficult to evaluate the effectiveness of summarization algorithms [10].
- Domain-specificity: Depending on the topic or domain of the text, text summarization algorithms perform differently. For example, a summarization algorithm for scientific papers may not provide a satisfactory result for news articles [76].
- Length: The correct length of a summary can be a challenge. A too short summary may miss important points, but a too long summary may make it difficult to understand [4].

- Multimodal text: Multimodal texts, by definition, contain text as well as images, videos, and other multimedia elements. To summarize these types of texts, advanced algorithms must be used to analyze different types of data [4].

As a result, it is a complex task to evaluate text summarization algorithms that requires carefully considering these challenges and developing standardized evaluation methods to accommodate a variety of texts and tasks. All the metrics and their corresponding features are listed in Table 1. Note that the mentioned metrics may belong to different categories.

4 Comparison of metrics

4.1 Datasets

A shared task dataset was originally used to develop and test the current evaluation metrics. A discussion of some of the indicators associated with commonly used standard datasets is presented in the following section:

- **Document Understanding Conference (DUC)**[1]: Annual workshops on automatic document summarization were held by DUC from 2001 to 2007. As part of the workshops, participants were asked to create an automatic summary of a set of documents. Various metrics, such as ROUGE and BLEU, were then used to evaluate and compare the generated summaries with human-generated summaries. Most of the current evaluation metrics were developed and tested on DUC datasets, which have become a standard benchmark for evaluating summarization algorithms [77].
- **Text Analysis Conference (TAC)**[2]: NIAT (National Institute of Standards and Technology) organizes TAC, an annual conference that promotes research in natural language processing. An opportunity to develop and test algorithmic solutions to various NLP tasks, including summarization, machine translation, and question answering, using standardized datasets is provided for researchers in this conference.
- **CNN/DailyMail:** Current summarization research relies heavily on the CNN/DailyMail [78]. The task involves summarizing a single document on a large scale that differs from the datasets used in DUC and TAC. CNN/DailyMail references have shorter summaries (about 50 tokens) than TAC references, and they tend to be more extractive [79].
- **Biomedical Semantic Question Answering (BioASQ):** The University of Athens, Greece, organizes an annual challenge to develop and evaluate systems that can automatically answer biomedical questions. By providing a platform for researchers and industry professionals to evaluate and share their systems, this challenge aims to advance the state of the art in biomedical question answering. An extensive dataset of biomedical questions and related documents is provided to participants, who are asked to build systems that answer them automatically. BioASQ challenges include retrieval of biomedical information and question answering, as well as summarizing biomedical documents. It provides a series of benchmark datasets for researchers to evaluate and compare their models. Participants are provided with a set of biomedical articles and are asked to summarize them [80, 81].
- **Newsroom:** The NEWSROOM dataset was compiled utilizing social media and search engine metadata to compile 1.3 million summaries. There were more than 100 million

[1] https://duc.nist.gov/.

[2] https://tac.nist.gov/.

Table 1 Evaluation Metrics—M/A denotes manual or automatic, RB/RF indicates references base or references-free, MT refers to machine translation based, QA represents question answering based, CO represents coherence based, and G/S represents general or specific domain

Metrics	A/M	RB/RF	MT	QA	CO	G/S
BLEU (2001)	A	RB	*			G
Pyramid (2004)	M	RB				G
ROUGE (2004)	A	RB				G
METEOR (2005)	A	RB	*			G
BE (2006)	A	RB				G
Entity-Based Approach (2008)	A	RB			*	G
AutoSummENG (2008)	A	RB				G
BEwT-E (2008)	A	RB				G
MeMoG (2010)	A	RB				G
Graph-based Local Coherence Modeling (2013)	A	RB			*	G
NPowER (2013)	A	RB				G
Meteor Universal (2014)	A	RB	*			G
chrF (2015)	A	RB	*			G
PEAK (2016)	A					G
SERA (2016)	A	RB				S(NEWS)
Neural Local Coherence Model (2017)	A	RB			*	G
RUSE (2018)	A	RB	*			G
Q-Metrics (2018)	A	RB		*		G
SummTriver (2018)	A	RB				S(Biomedical)
Lightweight Pyramid (2019)	M	RB				G
SUM-QE (2019)	A	RF	*			G
YiSi (2019)	A	RF/RB	*			G
APES (2019)	A	RB		*		G
SummaQA (2019)	A	RF		*		G
PyrEval (2019)	A	RF/RB				G
MoverScore (2019)	A	RB				G
BERTSCORE (2019)	A	RB				G
QAEval (2020)	A	RB		*		G
FEQA (2020)	A	RB		*		G
QAGS (2020)	A	RB		*		G
HIGHRES (2020)	A	RF				G
BLANC (2020)	A	RF				G
SUPERT (2020)	A	RF				G
BLEURT (2020)	A	RB	*			G
Prism-src (2020)	A	RF	*			G
Neural Graph-based Local Coherence Model (2021)	A	RB			*	G
BARTSCORE (2021)	A	RB	*			G
QuestEval (2021)	A	RF		*		G

Table 1 continued

Metrics	A/M	RB/RF	MT	QA	CO	G/S
COMET-QE (2021)	A	RF	*			G
InfoLM (2022)	A	RB				G
G-EVAL (2023)	A	RB				G
GPTScore (2023)	A	RB	*			G

web pages crawled by various online publishers as part of the dataset creation process. This study aimed to identify newswire articles and extract summaries from their HTML metadata. It was originally created for use in search engine results as well as social media. In order to explore and analyze news-related content in a comprehensive manner, researchers and practitioners can access a wealth of information from diverse online sources with this dataset [79].

4.2 Comparative analysis of evaluation metrics

Numerous studies have examined and assessed various criteria in the evaluation domain. In these studies, specific metrics are examined and compared in order to shed light on their effectiveness, aiming to gain a thorough understanding of the subject. In addition to contributing to the advancement of evaluation methodologies, these studies pave the way for further research by looking at evaluation from different perspectives. Table 2 presents a collection of selected evaluation studies that assess text summarization from different perspectives. The following paragraphs describe a few selected studies:

1. Lin et al. have found that using multiple references for DUC tasks results in a higher correlation score with human judgment. According to the study, the ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S metrics were effective for summarizing single documents. For very short summaries, the ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4, and ROUGE-SU9 metrics showed good performance. Obtaining a high correlation score for multi-document summarization tasks proved challenging. A number of metrics, such as ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-S9, ROUGE-SU4, and ROUGE-SU9, performed reasonably well in evaluating the summaries when stop words were excluded from the matching process. However, a single reference summary with sufficient samples was considered a valid alternative to employing multiple references. It was observed that using multiple references enhanced the correlation with human judgments. In order to achieve statistical significance, a critical number of samples were necessary for evaluations to be stable and reliable [12].
2. Rankel et al. claimed that certain ROUGE variants can be combined with other ROUGE metrics to produce highly competitive evaluation metrics. Statistical significance should be included when reporting differences in ROUGE scores in order to enhance accuracy and credibility. The authors also suggested that combining ROUGE-BE/BETw-E with ROUGE-1, ROUGE-2, and ROUGE-4 metrics would result in summaries that were judged as significantly better by human evaluators [82].
3. Through comparing different ROUGE metric variants to BLEU metrics using DUC datasets, Graham conducted an extensive study in 2015. In the study, 192 ROUGE variants and BLEU were evaluated along with summarization metrics, and a detailed

Table 2 Comparison of Evaluation Metrics on Diverse Text Summarization Datasets

Refs.	Year	Data sets	Metrics	Subject
[12]	2004	DUC	ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S	Comparing Different Metric
[82]	2013	TAC	ROUGE-BE/BETw-E, ROUGE-1, 2, and 4	Comparing Different Metric
[83]	2015	TAC	ROUGE-BLEU	Comparing Different Metric
[84]	2019	TAC	ROUGE-2 (R-2), ROUGE-L (R-L), ROUGE-WE (R-WE) and S3	Comparing Different Metric
[85]	2021	CNN/DailyMail	ROUGE, ROUGE-WE, S3, BertScore, MoverScore, SMS, SummaQA, BLANC, SUPERT, BLEU, CHRF, METEOR, CIDEr	Comparing Different Metric
[38]	2022	–	–	Analyze the Limitations of Reference-Free Evaluation
[10]	2022	REALSumm, SummEval	ROUGE, BERTScore, AQEval	Re-Examining System-Level Correlations of Automatic Summarization
[76]	2022	Wos,IEEE and ACM	Rouge and Human Evaluation	Systematic Review on Text Summarization of Biomedical
[86]	2023	Newsroom, SummEval	ROUGE-1, ROUGE-2 and ROUGE-L, BERTScore, PRISM, BARTScore, ChatGPT	Comparing Different Metric

analysis of them was provided. Additionally, although BLEU exhibited the highest correlation with human assessment, it did not demonstrate a significant advantage over the best-performing variant of the ROUGE metric [83].

4. Peyrard et al., a study on evaluation metric performance, found that summaries tend to score within the typical range of TAC summaries. For summaries generated by current models, however, popular metrics such as ROUGE-2 (R-2), ROUGE-L (R-L), and ROUGE-WE (R-WE) show significant discrepancies and may not accurately reflect human judgment for higher-scoring summaries. The authors proposed collecting human judgments for summaries in this high-scoring range to address this problem [84].
5. A review of various methods and their limitations by Deutsch and his colleagues suggested that reference-free metrics are limited when evaluating generated text. As these metrics are similar to the models for generating text, they can be improved during testing and are more likely to favor outputs from similar models. The researchers suggested using reference-free metrics for analyzing model behavior instead of measuring progress on tasks. This is because reference-free metrics may not accurately assess higher-quality texts like human-written references [38].
6. Fabbri et al., in [85] re-evaluated for 14 automatic evaluation metrics using neural summarization model outputs as well as expert and crowdsourced human annotations. This research assembled the largest number of summaries generated by CNN/DailyMail. The SummarEval database, which provides a range of summarization resources, was also introduced in this research.
7. In order to determine the accuracy of automatic summarization evaluation metrics in replicating human judgments, system-level correlations are used. It has been shown by Deutsch and colleagues that the correlation of the ROUGE metric to human judgments is near zero when realistic scenarios are considered. Rather than just reporting system-level correlations, the authors recommend that new evaluation metrics include correlations between system pairs with scores that differ from one another to give users a better understanding of the reliability of observed improvements. SummEval [85] and REALSumm [87] datasets are used to analyze summary quality, both of which were collected from CNN/DailyMail. Based on the value of an observed improvement, users could determine the likelihood of humans agreeing that the improvement is real [10].
8. An overview of recent biomedical text summarization studies is presented in this article, which emphasizes the techniques used, the types of input data, and the metrics used for evaluating the systems. Search strategies developed by NLP experts and prior systematic reviews were used to search digital libraries such as WoS, IEEE, and ACM from January 1, 2014, to March 15, 2022. Data collected in the article are divided into two categories: Evaluation Metrics (Rouge, Rouge, and others) and Human Evaluation versus No Human Evaluation. The vast majority of performance evaluations were based on intrinsic or automatic metrics, such as Rouge [76].
9. Wang et al. evaluated ChatGPT for natural language generation (NLG) tasks such as summarization to measure the quality of the result. They performed ChatGPT as a metric on datasets that are introduced on SummEval [85], and NewsRoom [79] and compared with other baselines like ROUGE-1, ROUGE-2, and ROUGE-L, BERTScore, PRISM, BARTScore. The authors have considered different scenarios for a reference-free metric and reference-based metric. ChatGBT has shown better results than other methods in the dataset on SummEval for text summarization [86].

5 Conclusions

A comprehensive analysis of evaluation metrics for automated summarization systems is presented in this paper. It examines a wide range of evaluation metrics and categorizes them based on various viewpoints, discussing their advantages and disadvantages. There is no single metric that can adequately assess the quality of summary systems from all angles, so a combination of metrics is required to assess their quality in order to receive the most comprehensive assessment. The purpose of this survey is to provide a comprehensive review of 42 evaluation metrics that can be used to evaluate text summarization and serve as a valuable resource for both researchers and practitioners.

Author Contributions Both authors contributed equally to this work in writing, editing and....

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Lloret E, Palomar M (2012) Text summarisation in progress: a literature review. Artif Intell Rev 37:1–41
2. Widyassari AP, Rustad S, Shidik GF, Noersasongko E, Syukur A, Affandy A et al (2022) Review of automatic text summarization techniques & methods. J King Saud Univ Comput Inf Sci 34(4):1029–1046
3. El-Kassas Wafaa S, Salama Cherif R, Rafea Ahmed A, Mohamed Hoda K (2021) Automatic text summarization: a comprehensive survey. Expert Syst Appl 165:113679
4. Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey. Artif Intell Rev 47(1):1–66
5. Radev DR, Blair-Goldensohn S, Zhang Z (2001) Experiments in single and multidocument summarization using mead. In: First document understanding conference, pp 1–7
6. Qiang J-P, Chen P, Ding W, Xie F, Xindong W (2016) Multi-document summarization using closed patterns. Knowl-Based Syst 99:28–38
7. John A, Premjith PS, Wilscy M (2017) Extractive multi-document summarization using population-based multicriteria optimization. Expert Syst Appl 86:385–397
8. Widjanarko A, Kusumaningrum R, Surarso B (2018) Multi document summarization for the Indonesian language based on latent Dirichlet allocation and significance sentence. In: 2018 International conference on information and communications technology (ICOIACT). IEEE, pp 520–524
9. Khan A, Salim N (2014) A review on abstractive summarization methods. J Theor Appl Inf Technol 59(1):64–72
10. Deutsch D, Dror R, Roth D (2022) Re-examining system-level correlations of automatic summarization evaluation metrics. arXiv preprint arXiv:2204.10216
11. Lin J, Demner-Fushman D (2005) Evaluating summaries and answers: two sides of the same coin? In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp 41–48
12. Lin C-Y (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
13. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318
14. Lloret E, Plaza L, Aker A (2018) The challenging task of summary evaluation: an overview. Lang Resour Eval 52:101–148
15. Jones KS, Galliers JR (1995) Evaluating natural language processing systems: an analysis and review. Lecture Notes in Artificial Intelligence. Springer
16. Nenkova A, Passonneau RJ (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, pp 145–152

17. Nenkova A, Passonneau R, McKeown K (2007) The pyramid method: incorporating human content selection variation in summarization evaluation. ACM Trans Speech Lang Process 4(2):4-es
18. Shapira O, Gabay D, Gao Y, Ronen H, Pasunuru R, Bansal M, Amsterdamer Y, Dagan I (2019) Crowdsourcing lightweight pyramids for manual summary evaluation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (Long and Short Papers), pp 682–687
19. Yang Q, Passonneau R, De Melo G (2016) Peak: pyramid evaluation via automated knowledge extraction. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
20. Weisstein EW (2011) Hungarian maximum matching algorithm. https://mathworld.wolfram.com/
21. Gao Y, Sun C, Passonneau RJ (2019) Automated pyramid summarization evaluation. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL)
22. Zhang S, Zhang J, Zhang C (2007) Edua: an efficient algorithm for dynamic database mining. Inf Sci 177(13):2756–2767
23. Ng J-P, Abrecht V (2015) Better summarization evaluation with word embeddings for rouge. arXiv preprint arXiv:1508.06034
24. Wang A, Cho K, Lewis M (2020) Asking and answering questions to evaluate the factual consistency of summaries. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5008–5020
25. Hovy EH, Lin C-Y, Zhou L, Fukumoto J (2006) Automated summarization evaluation with basic elements. In: LREC, vol 6, pp 604–611
26. Tratz S, Hovy E (2008) Bewte: basic elements with transformations for evaluation. In: TAC 2008 workshop
27. Giannakopoulos G, Karkaletsis V, Vouros G, Stamatopoulos P (2008) Summarization system evaluation revisited: N-gram graphs. ACM Trans Speech Lang Process 5(3):1–39
28. Giannakopoulos G, Karkaletsis V (2010) Summarization system evaluation variations based on n-gram graphs. In: TAC, Citeseer
29. Giannakopoulos G, Karkaletsis V (2013) Summary evaluation: together we stand npower-ed. In: International conference on intelligent text processing and computational linguistics. Springer, pp 436–450
30. Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia JS, Parke R, Mwavu R, Martinez N, Gichoya JW, Ghassemi M, Demner-Fushman D et al (2024) Peer review of gpt-4 technical report and systems card. PLoS Digit Health 3(1):e0000417
31. Fu J, Ng S-K, Jiang Z, Liu P (2023) Gptscore: evaluate as you desire. arXiv preprint arXiv:2302.04166
32. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023) Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634
33. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
34. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2019) Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675
35. Liu Y (2019) Fine-tune bert for extractive summarization. arXiv preprint arXiv:1903.10318
36. Zhao W, Peyrard M, Liu F, Gao Y, Meyer CM, Eger S (2019) Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv preprint arXiv:1909.02622
37. Colombo PJA, Clavel C, Piantanida P (2022) Infolm: a new metric to evaluate summarization & data2text generation. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 10554–10562
38. Deutsch D, Dror R, Roth D (2022) On the limitations of reference-free evaluations of generated text. arXiv preprint arXiv:2210.12563
39. Narayan S, Vlachos A, et al (2019) Highres: Highlight-based reference-less evaluation of summarization. arXiv preprint arXiv:1906.01361
40. Vasilyev O, Dharnidharka V, Bohannon J (2020) Fill in the blanc: human-free quality estimation of document summaries. arXiv preprint arXiv:2002.09836
41. Taylor WL (1953) “Cloze procedure”: a new tool for measuring readability. Journal Q 30(4):415–433
42. Gao Y, Zhao W, Eger S (2020) Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. arXiv preprint arXiv:2005.03724
43. Turian J, Shen L, Melamed ID (2003) Evaluation of machine translation and its evaluation. In: Proceedings of machine translation summit IX: papers
44. Lee S, Lee J, Moon H, Park C, Seo J, Eo S, Koo S, Lim H (2023) A survey on evaluation metrics for machine translation. Mathematics 11(4):1006
45. Popović M (2015) chrf: character n-gram f-score for automatic mt evaluation. In: Proceedings of the tenth workshop on statistical machine translation, pp 392–395
46. Sellam T, Das D, Parikh AP (2020) Bleurt: learning robust metrics for text generation. arXiv preprint arXiv:2004.04696

47. Banerjee S, Lavie A (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, pp 65–72
48. Denkowski M, Lavie A (2010) Meteor-next and the meteor paraphrase tables: improved evaluation support for five target languages. In: Proceedings of the joint fifth workshop on statistical machine translation and MetricsMATR, pp 339–342
49. Agarwal A, Lavie A (2008) Meteor, m-bleu and m-ter: evaluation metrics for high-correlation with human rankings of machine translation output. In: Proceedings of the third workshop on statistical machine translation, pp 115–118
50. Denkowski M, Lavie A (2010) Extending the meteor machine translation evaluation metric to the phrase level. In: Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, pp 250–253
51. Denkowski M, Lavie A (2011) Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation, pp 85–91
52. Denkowski M, Lavie A (2014) Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation, pp 376–380
53. Thompson B, Post M (2020) Automatic machine translation evaluation in many languages via zero-shot paraphrasing. arXiv preprint arXiv:2004.14564
54. Rei R, Farinha AC, Zerva C, van Stigt D, Stewart C, Ramos P, Glushkova T, Martins AFT, Lavie A (2021) Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In: Proceedings of the sixth conference on machine translation, pp 1030–1040
55. Rei R, Stewart C, Farinha AC, Lavie A (2020) Comet: a neural framework for mt evaluation. *arXiv preprint* arXiv:2009.09025
56. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek, G Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116
57. Xenouleas S, Malakasiotis P, Apidianaki M, Androutsopoulos I (2019) Sumqe: a bert-based summary quality estimation model. arXiv preprint arXiv:1909.00578
58. Yuan W, Neubig G, Liu P (2021) Bartscore: Evaluating generated text as text generation. Adv Neural Inf Process Syst 34:27263–27277
59. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461
60. Shimanaka H, Kajiwara T, Komachi M (2018) Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In: Proceedings of the third conference on machine translation: shared task papers, pp 751–758
61. Lo C (2019) Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In: Proceedings of the fourth conference on machine translation (volume 2: shared task papers, Day 1), pp 507–513
62. Deutsch D, Bedrax-Weiss T, Roth D (2021) Towards question-answering as an automatic metric for evaluating the content quality of a summary. Trans Assoc Comput Linguist 9:774–789
63. Eyal M, Baumel T, Elhadad M (2019) Question answering as an automatic evaluation metric for news article summarization. arXiv preprint arXiv:1906.00318
64. Scialom T, Lamprier S, Piwowarski B, Staiano J (2019) Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint* arXiv:1909.01610
65. Durmus E, He H, Diab M (2020) Feqa: a question answering evaluation framework for faithfulness assessment in abstractive summarization. arXiv preprint arXiv:2005.03754
66. Nema P, Khapra MM (2018) Towards a better metric for evaluating question generation systems. arXiv preprint arXiv:1808.10192
67. Scialom T, Dray P-A, Gallinari P, Lamprier S, Piwowarski B, Staiano J, Wang A (2021) Questeval: Summarization asks for fact-based evaluation. arXiv preprint arXiv:2103.12693
68. Mesgar M, Ribeiro LFR, Gurevych I (2021) A neural graph-based local coherence model. In: Findings of the association for computational linguistics: EMNLP 2021, pp 2316–2321
69. Mohiuddin T, Joty S, Nguyen DT (2018) Coherence modeling of asynchronous conversations: a neural entity grid approach. arXiv preprint arXiv:1805.02275
70. Barzilay R, Lapata M (2008) Modeling local coherence: an entity-based approach. Comput Linguist 34(1):1–34
71. Nguyen DT, Joty S (2017) A neural local coherence model. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1320–1330

72. Elsner M, Charniak E (2011) Extending the entity grid with entity-specific features. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 125–129
73. Guinaudeau C, Strube M (2013) Graph-based local coherence modeling. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers), pp 93–103
74. Cohan A, Goharian N (2016) Revisiting summarization evaluation for scientific articles. arXiv preprint arXiv:1604.00400
75. Cabrera-Diego LA, Torres-Moreno J-M (2018) Summtriver: a new trivergent model to evaluate summaries automatically without human references. Data Knowl Eng 113:184–197
76. Chaves A, Kesiku C, Garcia-Zapirain B (2022) Automatic text summarization of biomedical text data: a systematic review. Information 13(8):393
77. Dang HT, Owczarzak K et al (2008) Overview of the tac 2008 update summarization task. In: TAC
78. Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. In: Advances in neural information processing systems, vol 28
79. Grusky M, Naaman M, Artzi Y (2018) Newsroom: a dataset of 1.3 million summaries with diverse extractive strategies. arXiv preprint arXiv:1804.11283
80. Krithara A, Nentidis A, Bougiatiotis K, Paliouras G (2023) Bioasq-qa: a manually curated corpus for biomedical question answering. Sci Data 10(1):170
81. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, Weissenborn D, Krithara A, Petridis S, Polychronopoulos D et al (2015) An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinform 16(1):1–28
82. Rankel PA, Conroy J, Dang HT, Nenkova A (2013) A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: short papers), pp 131–136
83. Graham Y (2015) Re-evaluating automatic summarization with bleu and 192 shades of rouge. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 128–137
84. Peyrard M (2019) Studying summarization evaluation metrics in the appropriate scoring range. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 5093–5100
85. Fabbri AR, Kryściński W, McCann B, Xiong C, Socher R, Radev D (2021) Summeval: re-evaluating summarization evaluation. Trans Assoc Comput Linguist 9:391–409
86. Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, Xu J, Qu J, Zhou J (2023). Is chatgpt a good nlg evaluator? A preliminary study. arXiv preprint arXiv:2303.04048
87. Bhandari M, Gour P, Ashfaq A, Liu P, Neubig G (2020) Re-evaluating evaluation in text summarization. arXiv preprint arXiv:2010.07100

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ensieh Davoodijam received the BS degree in computer engineering from Arak University, Arak, Iran in 2009, the MS degree from University of Isfahan, Isfahan, Iran in 2012 and the PhD from Isfahan University of Technology, Isfahan, Iran in 2021. Her main interests include data mining and text mining.

Mohsen Alambardar Meybodi received the BS, MS and PhD degrees in computer science from Yazd University, Iran in 2008, 2010 and 2019 respectively. He is currently an assistant professor with the Department of Applied Mathematics and Computer Science, University of Isfahan, Iran. His research interests include parameterized algorithms, computational complexity and machine learning.