A Composite Metric for Abstractive Summarization Evaluation: Integrating Lexical, Semantic, and Factual Dimensions

Thiago Ouverney, Valdecy Pereira

Universidade Federal Fluminense, Brazil
thiago.ouverney@id.uff.br
valdecy@ic.uff.br

Abstract. Automatic evaluation of abstractive summarization remains challenging because traditional lexical-overlap metrics often fail to capture semantic adequacy and factual consistency. This work proposes a composite evaluation metric integrating lexical, semantic, and factual dimensions. Metrics from each family are combined using a hierarchical weighted formulation, with both intra-family and inter-family weights optimized via Bayesian Optimization using the SummEval benchmark [Fabbri et al., 2021]. The resulting metric assigns 19% weight to lexical overlap, 58% to semantic similarity, and 23% to factuality, and achieves a Spearman correlation of 0.488 on the test set, outperforming the best individual metric, BERTScore-Precision (0.467). These results indicate that combining complementary evaluation dimensions yields a more robust and human-aligned signal for abstractive summarization.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Learning paradigms*; Supervised learning; • **Information systems** → *Data mining*.

Keywords: Automatic Text Summarization, Bayesian Optimization, Composite Metric, Evaluation Metrics, Natural Language Processing, Text Mining

1. INTRODUCTION

Automatic Text Summarization (ATS) has become increasingly relevant as the volume of digital information grows and users require concise and reliable representations of long documents. Among ATS approaches, abstractive summarization is particularly attractive because it rewrites and compresses information in a more natural manner than extractive systems. However, evaluating abstractive summaries remains an open challenge [Fabbri et al. 2021].

Traditional automatic metrics such as ROUGE [Lin 2004], BLEU [Papineni et al. 2002], and METEOR [Banerjee and Lavie 2005] were designed for settings in which surface overlap is a useful proxy for quality. These metrics remain common in summarization research, but their limitations are well documented in abstractive settings. Lexical similarity alone is often insufficient to capture semantic adequacy, factual consistency, or the broader notion of summary quality as reflected in human evaluation.

More recent work has introduced semantic and factual metrics to address part of this limitation. Semantic approaches improve sensitivity to paraphrasing, while factual metrics help identify whether the generated content is supported by the source document. Even so, the literature still presents these dimensions largely as separate signals, and no single metric has shown consistently strong agreement with human judgment across the main evaluation criteria.

No specific funding was reported for this study.

This work addresses that gap by proposing a composite evaluation metric that integrates lexical, semantic, and factual signals into a unified and learnable formulation. The central hypothesis is that these dimensions provide complementary information and that a calibrated combination can better approximate expert human evaluation than any single metric.

To examine this question, we propose a composite metric organized in two levels. First, metrics are combined within each family. Then, the family-level scores are combined into a final score. This hierarchical design makes it possible to estimate both the relative contribution of each family and the role of each metric inside it, while keeping the final formulation interpretable.

The main contributions of this article are threefold: (i) a composite metric for abstractive summarization evaluation that integrates lexical, semantic, and factual dimensions; (ii) an extension of the pyAutoSummarizer library [Pereira 2023] with a unified evaluation interface; and (iii) an empirical analysis of the proposed metric on the SummEval benchmark [Fabbri et al. 2021].

2. RELATED WORK

Research on summarization evaluation has shifted from lexical overlap metrics to semantic and factual approaches. This shift reflects the limitations of surface matching when the target summaries are abstractive.

Early evaluation methods are largely based on lexical overlap between system summaries and reference summaries. ROUGE [Lin 2004] became the standard metric in summarization by measuring overlap through n-grams and longest common subsequences. BLEU [Papineni et al. 2002] uses modified n-gram precision with a brevity penalty, while METEOR [Banerjee and Lavie 2005] extends this line of work by incorporating alignment and synonym matching. These metrics remain useful, but they are less adequate when summaries contain paraphrases that preserve meaning without preserving wording.

To address this issue, semantic metrics based on contextual embeddings were proposed. Sentence-BERT [Reimers and Gurevych 2019] enables sentence-level similarity comparisons, and BERTScore [Zhang et al. 2020] compares candidate and reference texts through contextual token alignments. These metrics tend to correlate better with human judgment than lexical metrics, but they do not directly assess factual consistency.

A separate line of work focuses on factuality through models such as FactCC [Kryściński et al. 2020] and QuestEval [Scialom et al. 2021]. These methods address unsupported statements, but they are usually applied as standalone metrics. Benchmarks such as SummEval [Fabbri et al. 2021] show that no single metric performs best in every dimension.

This work proposes a hierarchical composition in which both intra-family and inter-family weights are learned via optimization. Rather than treating metrics as flat features, the approach explicitly models lexical, semantic, and factual dimensions as structured components, enabling a more interpretable, data-driven integration that aligns with human judgment.

3. METHODOLOGY

This work proposes a hierarchical composite metric for abstractive summarization evaluation. The main idea is to combine lexical, semantic, and factual signals into a single automatic score, while learning the relative contribution of each component from human judgments.

Let each metric family $F_k \in \{F_{lex}, F_{sem}, F_{fact}\}$ contain a set of individual metrics $m_{k,i}$. The score of each family is defined as

$$M_k = \sum_i w_{k,i} \cdot m_{k,i} \tag{1}$$

where $w_{k,i}$ are non-negative weights normalized such that $\sum_i w_{k,i} = 1$.

The final composite metric is defined as

$$M = \sum_k \alpha_k \cdot M_k \tag{2}$$

where α_k represents the contribution of each family and $\sum_k \alpha_k = 1$.

This formulation makes it possible to learn both the relative importance of the metrics within each family and each family's contribution to the final score.

The lexical family includes ROUGE-1, ROUGE-L, BLEU, and METEOR. The semantic family includes BERTScore-Precision, BERTScore-Recall, Sentence-BERT average similarity, and Sentence-BERT maximum similarity. The factual family includes QuestEval and FactCC. Lexical and semantic metrics are reference-based, comparing a generated summary against human reference summaries, whereas factual metrics assess the consistency of the generated summary with the source document.

The evaluation unit in this study is the pair (d, s), where d denotes a source document and s denotes a summary produced by a given summarization system. Thus, a single source document may be associated with multiple generated summaries produced by different systems, and each of these outputs is evaluated independently. For reference-based metrics, the generated summary is compared against all available human reference summaries for document d, and the resulting values are aggregated at the (d, s) level.

In the experimental setting adopted in this study, the human target is derived from expert annotations in SummEval [Fabbri et al. 2021]. For each pair (d, s), the overall expert score is computed as the arithmetic mean of the judgments for coherence, consistency, fluency, and relevance:

$$y_{d,s} = \frac{1}{4}\left(y_{d,s}^{\text{coh}} + y_{d,s}^{\text{cons}} + y_{d,s}^{\text{flu}} + y_{d,s}^{\text{rel}}\right) \tag{3}$$

This aggregated target is adopted to provide a single supervision signal aligned with overall summary quality, while preserving the multidimensional nature of the expert evaluation. At the same time, this choice imposes an equal-weight definition of overall quality across the four human evaluation dimensions.

Let $M_{d,s}(\Theta)$ denote the final composite score assigned to the summary generated by system s for document d, where Θ denotes the full set of intra-family and inter-family weights. All weights are jointly optimized with Optuna, using a TPE-based Bayesian optimization strategy [Akiba et al. 2019]. In this sense, the weight search can be viewed as learning a metric combination from labeled examples, aligning the formulation with meta-learning approaches to evaluation design. The objective is to maximize the Spearman rank correlation between the composite metric and the aggregated expert score:

$$\mathcal{J}(\Theta) = \rho_s\left(\{M_{d,s}(\Theta)\}, \{y_{d,s}\}\right) \tag{4}$$

where $\rho_s(\cdot, \cdot)$ denotes Spearman's rank correlation coefficient [Spearman 1904]. The validation protocol used to estimate this objective is described in the experimental setup.

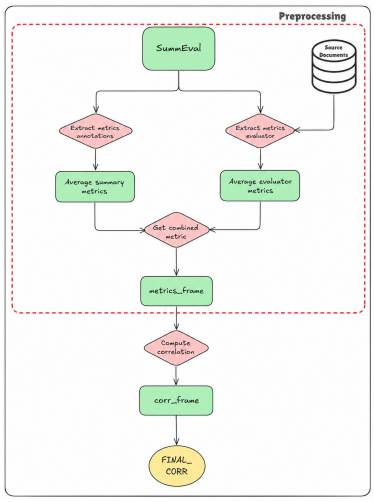

Fig. 1. Overview of the evaluation pipeline, from SummEval ingestion to final correlation computation.

Bayesian optimization with Optuna was adopted because the objective is rank-based and non-differentiable, and each candidate configuration must be evaluated through cross-validation. The search includes 13 continuous parameters representing the weights of the individual metrics and metric families. An exhaustive grid search would become costly as the numerical resolution increased. Optuna's Tree-structured Parzen Estimator provides a practical strategy for exploring this space under a fixed computational budget, while all candidate weights are normalized before evaluation.

3.1 Experimental Setup

The experiments use the SummEval benchmark [Fabbri et al. 2021], which contains 100 news articles, 16 summarization systems, and 1600 document-summary pairs. This study uses only expert judgments as the optimization target. For lexical metrics, punctuation and stop words are removed from generated and reference summaries; semantic and factual metrics use the original text.

Since multiple system outputs are associated with the same source document, all splits are performed at the document level to avoid leakage across partitions (Figure 1). Weight optimization uses 4-fold Group K-Fold cross-validation on the training portion, while a document-level holdout test set remains untouched until final evaluation (Figure 2).

Robustness on the held-out test set was additionally assessed through a paired document-cluster bootstrap with 20,000 resamples, keeping the learned weights fixed. Details are reported in Section 4.6.

4. RESULTS

4.1 Performance of Individual Metrics

Table I reports the correlation between individual metrics and expert judgments. Lexical metrics present lower correlations than the best semantic metric. ROUGE-1 reaches 0.24, ROUGE-L 0.17, BLEU 0.18, and METEOR 0.22. BERTScore-Precision achieves 0.47, the highest among the individual metrics considered. QuestEval reaches 0.30, while FactCC shows a correlation close to zero in this experimental setting.

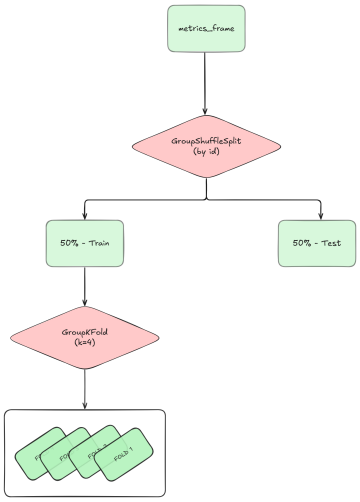

Fig. 2. Data splitting strategy using GroupShuffleSplit and GroupKFold to avoid document-level leakage.

Metric	Spearman correlation
ROUGE-1	0.24
ROUGE-L	0.17
BLEU	0.18
METEOR	0.22
BERTScore-Precision	0.47
BERTScore-Recall	0.31
SBERT Avg	0.14
SBERT Max	0.08
QuestEval	0.30
FactCC	-0.01

Table I. Spearman correlation between individual automatic metrics and expert judgments.

Component	Weight
Lexical family	0.19
Semantic family	0.58
Factual family	0.23

Table II. Global weights learned for the composite metric.

4.2 Learned Composition Weights

Table II presents the global weights learned during optimization: 19% lexical, 58% semantic, and 23% factual. Internally, BLEU and ROUGE-1 dominate the lexical family, BERTScore-Precision dominates the semantic family, and QuestEval receives almost all factual weight.

Figure 3 shows the optimization process. The weights move toward a stable configuration, and the correlation values improve in the initial trials before leveling off.

4.3 Final Comparison with Baselines

Table III presents the final comparison on the test set. The composite metric reaches 0.488, compared with 0.467 for BERTScore-Precision, 0.301 for QuestEval, and 0.238 for ROUGE-1.

The improvement over BERTScore-Precision is consistent with the final comparison reported. This

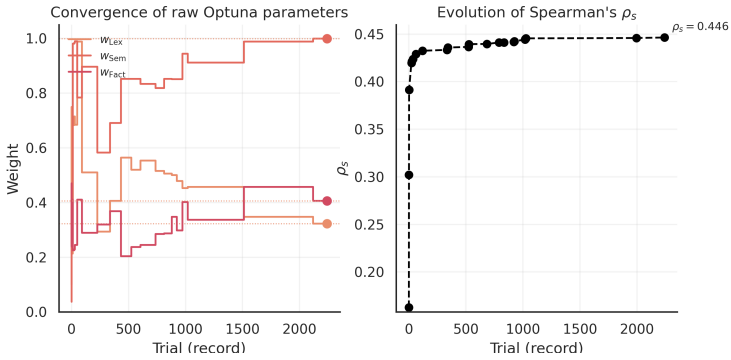

Fig. 3. Evolution of optimization trials. Left: convergence of component weights. Right: improvement of Spearman correlation across trials.

Metric	Spearman correlation
ROUGE-1	0.238
QuestEval	0.301
BERTScore-Precision	0.467
Composite metric	**0.488**

Table III. Final comparison between the composite metric and selected baselines on the test set.

Metric	Coherence	Consistency	Fluency	Relevance	Overall
Composite metric	**0.409**	**0.193**	**0.243**	**0.483**	**0.488**
BERTScore-Precision	0.389	0.156	0.214	0.501	0.467
QuestEval	0.241	0.160	0.217	0.225	0.301
ROUGE-1	0.132	0.168	0.127	0.262	0.238

Table IV. Spearman correlation with human judgments across evaluation dimensions.

suggests that the combination of lexical, semantic, and factual information can provide a more effective evaluation signal than the best individual baseline in this experimental setting.

4.4 Performance by Human Evaluation Dimension

Table IV reports performance by evaluation dimension. The composite metric outperforms BERTScore-Precision on coherence, consistency, and fluency, while remaining slightly below it on relevance; the largest gain appears in consistency.

4.5 Abstractive and Extractive Systems

To examine performance across system types, the composite metric was evaluated separately on abstractive and extractive systems, reaching Spearman correlations of 0.474 and 0.554, respectively. These results show that the metric is not limited to a single type of summarization system, while suggesting that the balance among lexical, semantic, and factual signals may vary with the summarization strategy.

4.6 Bootstrap Robustness Analysis

To assess the stability of the observed advantage without modifying the learned composition, we performed a paired document-cluster bootstrap on the held-out test split. This partition contains 50

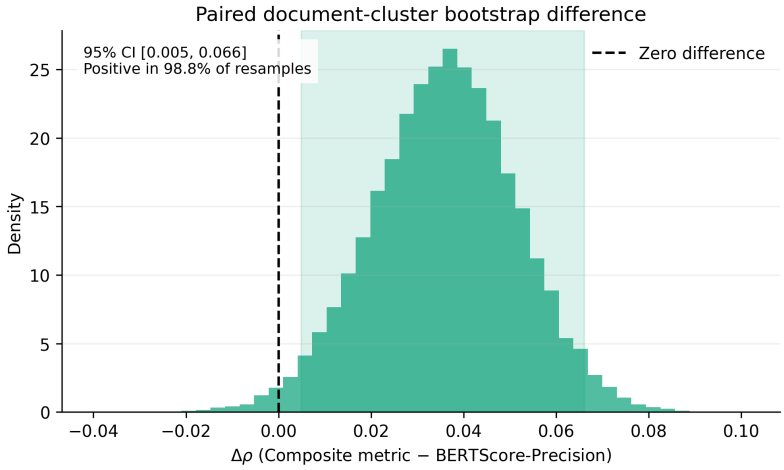

Fig. 4. Distribution of the paired Spearman correlation difference $\Delta\rho = \rho_{Composite} - \rho_{BERTScore-P}$ across 20,000 document-cluster bootstrap resamples.

source articles and 800 document-summary pairs. The composition weights were kept fixed throughout the analysis.

We generated 20,000 bootstrap resamples by sampling source articles with replacement and retaining all 16 system outputs associated with each selected article. This procedure preserves the dependence among summaries generated from the same source document. For each resample, we computed the difference $\Delta\rho = \rho_{Composite} - \rho_{BERTScore-P}$.

As shown in Figure 4, the percentile 95% bootstrap confidence interval for the paired difference is [0.005, 0.066], with the entire interval above zero. The composite metric achieves a higher correlation than BERTScore-Precision in 98.8% of the bootstrap resamples.

5. DISCUSSION

BERTScore-Precision remains the strongest standalone baseline, confirming the importance of semantic similarity on SummEval. Nevertheless, the composite metric achieves higher overall correlation and improves coherence, consistency, and fluency, suggesting that lexical and factual signals provide complementary information. The semantic family receives most of the learned weight, while lexical and factual components retain smaller contributions.

The dimension-level results reinforce this point. The largest gain appears in consistency, which is compatible with the role of factual metrics in capturing whether the generated content is supported by the source document. The smaller gains in coherence and fluency suggest that lexical and semantic components may help stabilize the metric across dimensions even when they do not dominate the final score. In contrast, the slight drop in relevance indicates that the proposed composition does not uniformly improve every aspect of human evaluation.

The higher correlation observed for extractive systems should be read conservatively: system type changes the balance among lexical, semantic, and factual cues, so a single set of learned weights may transfer unevenly across strata. Likewise, the optimization target is the arithmetic mean of coherence, consistency, fluency, and relevance, which provides a practical supervision signal but imposes a specific notion of overall summary quality. The learned weights should therefore be interpreted as benchmark-aligned rather than universal.

Within the factual family, the learned contribution is almost entirely associated with QuestEval,

while FactCC receives negligible weight and presents near-zero correlation with the human target. This result does not imply that FactCC is ineffective in general. The original model was trained using synthetically transformed source sentences, and the resulting error patterns may not fully represent the outputs or the aggregated quality target considered in this study. A possible direction is to adapt FactCC using factuality examples that better reflect the target summarization setting. Such adaptation should be performed only on training documents and evaluated on a separate holdout before its contribution to the composite metric is reassessed.

6. CONCLUSION

This work proposed a hierarchical metric combining lexical, semantic, and factual signals for abstractive summarization evaluation. On SummEval, the composite metric achieved a higher held-out Spearman correlation than BERTScore-Precision (0.488 vs. 0.467), while the document-cluster bootstrap yielded a positive 95% confidence interval for the paired difference and favored the composite metric in 98.8% of resamples. The learned weights indicate a dominant semantic contribution, complemented by lexical and factual signals. Future work will evaluate transfer to external benchmarks, alternative human-quality targets, and factuality-model adaptation.

ACKNOWLEDGMENTS

This article is derived from the undergraduate thesis of T. Ouverney, supervised by V. Pereira. Code and experiments are available at `https://github.com/thiago-ouverney/pyAutoSummarizer`.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, AK, USA, pp. 2623–2631, 2019.

Banerjee, S. and Lavie, A. 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 Summarization*. Ann Arbor, MI, USA, pp. 65–72, 2005.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. R. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* vol. 9, pp. 391–416, 2021.

Kryściński, W., McCann, B., Xiong, C., and Socher, R. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Online, pp. 9332–9346, 2020.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out*. Barcelona, Spain, pp. 74–81, 2004.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA, pp. 311–318, 2002.

Pereira, V. pyautosummarizer. `https://github.com/Valdecy/pyAutoSummarizer`, 2023.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. Hong Kong, China, pp. 3982–3992, 2019.

Scialom, T., Dray, R., Delbé, C., Galerne, A., Cohen, S., Dragoni, F., Boucheron, J., Déloison, L., Galley, M., Sène, K., Sekhari, A., and Teton, F. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic, pp. 659–670, 2021.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* 15 (1): 72–101, 1904.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*. Online, 2020.