Towards Interpretable Summary Evaluation via Allocation of Contextual Embeddings to Reference Text Topics

Ben Schaper  
Technical University of Munich  
ben.schaper@tum.de

Christopher Lohse  
Trinity College Dublin  
lohsec@tcd.ie

Marcell Streile  
IBM  
streile@de.ibm.com

Andrea Giovannini  
IBM  
agv@zurich.ibm.com

Richard Osuala  
Universitat de Barcelona  
richard.osuala@ub.edu

Abstract

Despite extensive recent advances in summary generation models, evaluation of auto-generated summaries still widely relies on single-score systems insufficient for transparent assessment and in-depth qualitative analysis. Towards bridging this gap, we propose the multifaceted interpretable summary evaluation method (MISEM), which is based on allocation of a summary’s contextual token embeddings to semantic topics identified in the reference text. We further contribute an interpretability toolbox for automated summary evaluation and interactive visual analysis of summary scoring, topic identification, and token-topic allocation. MISEM achieves a promising .404 Pearson correlation with human judgment on the TAC’08 dataset. Our code and toolbox are available at https://github.com/IBM/misem

1 Introduction

Auto-generated text summaries are becoming an increasingly mature, useful, and time-saving tool in research and industry with multiple applications such as email summary generation, summarizing research papers, and simplifying knowledge management and knowledge transfer in companies (El-Kassas et al., 2021). It is crucial for production-ready applications that auto-generated summaries do not omit critical information.

In this regard, currently used summary evaluation metrics have many known limitations, which are becoming even more apparent as natural language generator (NLG) models evolve (e.g., better paraphrasing) (Gehrmann et al., 2022). Therefore, the generated texts are becoming less assessable based on surface-level (e.g., n-gram overlap based) methods of older evaluation metrics (Gehrmann et al., 2022). Recent advances in summary evaluation leverage semantic similarity and often compute a single numerical evaluation score (Zheng and Lapata, 2020; Xenouleas et al., 2020; Gao et al., 2020). However, as noted by Gehrmann et al. (2022), a single numerical score alone is likely too narrow to reliably indicate the quality of a text summary.

Thus, there is a clear need for an NLG evaluation metric that provides a multifaceted and interpretable quality measure. Moreover, such a measure is needed as quality gate in industrial settings that NLG-generated summaries need to pass before being displayed to end-users.

Contributions Providing a multifaceted view of quality beyond a single score, our contributions are two-fold:

- **Evaluation Method:** We propose an interpretable summary evaluation method that identifies semantic topics present in the reference text, assigns summary text tokens to these topics, and measures the summary’s semantic coverage of each of these topics.

- **Interpretability Toolbox:** We provide an interactive interpretability toolbox that allows users to evaluate their summaries, adjust hyperparameters, explore topic-wise semantic overlap with reference texts, and detect missing parts of the summary.
2 Related Work

Evaluation based on Gold Standard This type of evaluation determines the relevance of a summary by comparing its overlapping words with a gold standard summary. Such evaluation is often labor-intensive, as it requires (multiple) human-written gold standard summaries. Baseline methods commonly used in gold standard evaluation are ROUGE (Lin, 2004) and BLEU (Papineni et al., 2001). More recent techniques enhance the ROUGE score with WordNet (ShafieiBavani et al., 2018), word embeddings (Ng and Abrecht, 2015) or use contextual embeddings (Zhang et al., 2019) in order to capture semantic similarity. In addition to that Mrabet and Demner-Fushman (2020) combine lexical similarities with BERT-Embeddings (Devlin et al., 2019).

Annotation-based Evaluation Annotation-based evaluation methods require manually annotated summary ratings following predefined guidelines. For example, the Pyramid method (Nenkova and Passonneau, 2004) works by annotating relevant topics in the source text and ranking the summaries accordingly. Böhm et al. (2020) use annotated texts to train a BERT-based evaluation model using rated summaries as training data.

Unsupervised Evaluation Unsupervised approaches infer a quality score of a summary based on its reference text without using a gold standard or manual annotations. Over the past few years, there have been multiple approaches to unsupervised methods for summary evaluation. For instance, many works explore BERT-Embeddings to detect semantic similarity between summary and reference text (Zheng and Lapata, 2020; Xenouleas et al., 2020; Gao et al., 2020). Zheng and Lapata (2020) propose a method for summary evaluation called Supert that combines BERT-embeddings with a directed graph in which each embedding is a node and node-wise similarity is computed based on graph positions. In order to rate a summary, SUPERT (Gao et al., 2020) uses contextual word embeddings and soft token alignment alignment techniques.

3 Methodology: Summary Evaluation

Our unsupervised evaluation method is based on the hypothesis that a generated summary can be evaluated by how much of the reference document’s semantic information it preserves. To increase the granularity and interpretability of our approach, we evaluate the preservation of content per semantic topic of the reference document. The semantic topics are identified by clustering the reference document’s sentence embeddings. Next, inspired by maximum mean discrepancy (MMD) (Gretton et al., 2012), our method measures the correspondence of the summary’s contextual token embeddings to each reference text topic centroid embedding. As each token is assigned to each topic, this assignment is weighted by the normalized cosine similarity between token and topic embedding. The pseudocode implementing our method is illustrated in Algorithm 1.

Encoding Following SUPERT (Gao et al., 2020), we first split both reference text and summary text into sentences. Then, the pre-trained Sentence-BERT model (Reimers and Gurevych, 2020) is used to encode the reference text into a set of sentence embeddings $R$ and the summary text into a set of contextual token embeddings $I$. In Algorithm 1, this encoding step is performed in lines 1-2.

Topic Clustering Our method requires a set of reference text topics defined as clusters $T$. In our experiments, $T$ are computed using the agglomerative clustering algorithm from Pedregosa et al. (2011). Line 3 of Algorithm 1 contains this step. Furthermore, our method requires topic centroids $T$, which are computed as $\bar{T}_t = \frac{1}{|T|} \sum_{i=1}^{|T|} t_i$, where $t$ represents one topic in $T$. Each topic centroid is calculated by taking the average of its associated reference text sentence embeddings. As the MISEM method assumes that the importance of a topic is determined to some degree by its length, a topic weight is calculated for each topic in $T$. The
Figure 2: Annotation Tool. By selecting one of the reference text topics (highlighted in yellow), the user can explore which summary text tokens best reflect that topic (in green). A cosine similarity threshold can be set to increase or decrease the number of highlighted summary text tokens. The example is based on text from TAC’09 (Dang and Owczarzak, 2008).

**Summary Text Correspondence** How well a topic is reflected by the summary text tokens is calculated using cosine similarity $C = \frac{\overline{T} \cdot \overline{I}}{\|\overline{T}\| \cdot \|\overline{I}\|}$, which computes the similarity between the topic centroid $\overline{T}$ and the summary token embeddings $\overline{I}$. Our experiments show that performance can be improved by normalizing the distribution of topic correspondence values for each token using the softmax function (Bridle, 1989). These two steps of calculating and normalizing the similarity scores $C$ can be found in steps 8-9 of Algorithm 1. For each topic $t$, the topic score is then calculated by taking the sum of its associated similarity scores $s_t = \sum_{i=1}^{n} C_{ti}$. Furthermore, we empirically find that applying the softmax normalization to $S$ improves the results. Algorithm 1 features the topic score calculation and subsequent normalization using the softmax function in lines 10-13.

**Weighted Aggregation** Finally, the summary score $m$ is calculated as the sum of the topic scores $S$ scaled by the topic weights $W$ as follows: $m = W \cdot S$. In Algorithm 1 this final step is performed in line 14.

**Interpretability** One concern with multiple summary scoring methods is that they only provide a single score and, as such, are limited in explaining why a particular summary was assigned a low or high score. The MISEM score addresses this problem by providing the summary text correspondence scores as part of the method. These values provide a means to examine how well a particular topic is represented by the summary, resulting in a more nuanced and transparent view of the final score. The following section introduces the interpretability toolbox, which leverages the inherent interpretability of the MISEM score to create insightful visualizations.

**4 Interpretability Toolbox**

We further contribute an interactive interpretability toolbox, which visually demonstrates the inherent interpretability of our evaluation method. As pointed out in Section 1, a single numerical score does not suffice to reliably evaluate a summary. Due to that our interpretability toolbox allows users to explore the consecutive calculations of the summary score of our method that make model behavior, various hyperparameters, and computed topic correspondence values intuitively interpretable.

We propose a workflow in which users (i) initiate the summary scoring, (ii) utilize the scatter plot (see Figure 3) to examine the relatedness of clustered sentences within different topics, and (iii) use the annotation tool (see Figure 2) to explore specific topics and summary text insufficiencies in-detail.

**Scatter Plot** The interactive scatter plot visualizes the reference text sentence embeddings clustered into topics. The embeddings are reduced to three dimensions using t-SNE (García-Alonso et al., 2008). This visualization empowers the user to explore the different topics covered in the reference text. Additionally, the user can hover individual points in the scatter plot to reveal their contents as depicted in Figure 3.

**Annotation Tool** The annotation tool allows the user to selectively explore individual reference text topics (see Figure 2). Leveraging the insights gained from the scatter plot, the annotation tool allows for further in-depth investigation. It shows the reference and summary texts displayed side-by-side to simplify the users’ diagnostic process. This
makes it a useful qualitative tool for users and domain experts alike to scrutinize the performance of NLG-generated summaries beyond the numerical score. Upon user selection of a topic and a similarity threshold, the summary’s numerical score for that topic is visually explained showing the topic’s allocated tokens with their score based on cosine similarity.

5 Experimental Setup and Results

Datasets We evaluate our method using TAC’08 and TAC’09 (Dang and Owczarzak, 2008), two multi-document summarization datasets available from the Text Analysis Conference (TAC)\(^1\) shared tasks. Following Louis and Nenkova (2013), we only use the initial summaries (the A parts). The TAC’08 and TAC’09 datasets include a total of 92 topics. Each topic has ten news articles, four reference summaries, and 57 (TAC’08) or 55 (TAC’09) machine-generated summaries. The average news article has 611 words and 24 sentences, while the average summary has 100 words. The Pyramid score is used as the ground-truth human rating in our experiments.

Related Methods We compare our method to two other unsupervised evaluation metrics, PacSum (Zheng and Lapata, 2020) and SUPERT (Gao et al., 2020), which are described in Section 2.

Note on Implementation Table 1 shows the best parameters derived from a grid search using TAC’08. We evaluated several clustering algorithms\(^2\) and text encoding models\(^3\) to determine the best combination. The best performance was achieved by combining S-BERT (all-mpnet-base-v2) (Reimers and Gurevych, 2020) contextual embeddings and agglomerative clustering from scikit-learn (Pedregosa et al., 2011). Sentence splitting was done using spaCy (Honnibal and Montani, 2017). The interpretability toolbox is based on streamlit\(^4\).

| Agglomerative Clustering | S-BERT-Model |
|--------------------------|--------------|
| Affinity                 | Linkage      | Distance | Model | normalize |
| Cosine                   | Complete     | 0.38     | all-mpnet-base-v2 | True |

Table 1: Optimal Hyperparameters

Results As summarized in Table 2, our interpretable method shows, e.g., with a Pearson’s correlation of .404 on TAC’08, promising correlation with human judgment compared to state-of-the-art methods (Zheng and Lapata, 2020; Gao et al., 2020).

| Method                | TAC’08 | TAC’09 |
|-----------------------|--------|--------|
|                        | r      | ρ      | τ     | r      | ρ      | τ     |
| (F1)-PacSumTopM       | .502   | .506   | .381  | .495   | .461   | .337  |
| (F β)-PacSumTopM      | .507   | .508   | .380  | .500   | .465   | .339  |
| MROBERTa (SUPERT)     | .366   | .526   | .235  | .357   | .316   | .229  |
| MSBERT (SUPERT)       | .466   | .428   | .311  | .436   | .435   | .320  |
| MISEM (ours)          | .404   | .375   | .270  | .349   | .376   | .274  |

Table 2: Main results on multi-document summarization datasets. Pearson’s r, Spearman’s ρ, and Kendall’s τ correlation with human scores are reported. All scores except for MISEM were taken from their respective publications.

6 Conclusion

In this paper, we present MISEM – an unsupervised method for interpretable summary evaluation based on semantic topics extracted from the reference text. Our method aims to provide a multifaceted view of a generated summary and allows for both quantitative as well as qualitative evaluation. Our experimental results show that our method achieves

\(^1\)https://tac.nist.gov/

\(^2\)e.g., DBSCAN, K-Means, Fast-Clustering

\(^3\)e.g., paraphrase-albert-small-v2, all-mpnet-base-v2

\(^4\)https://streamlit.io/
a competitive correlation with the human judgement Pyramid score on the TAC’08 and TAC’09 datasets. Beyond the MISEM method, we also provide an interpretability toolbox, which allows users to interactively and intuitively understand the summary evaluation, explore reference text topics, and analyze summary token to topic allocation.

**Limitations** We suggest the application of our method as summary quality gate in conjunction with other complementary methods to overcome several limitations. The latter include that our method does not incentivize conciseness of evaluated summaries due to not penalizing their length. Another limitation is an exclusive reliance on the semantic similarity of text embeddings. For instance, contradicting opinions (e.g., expressed via negation terms) in the summary can still yield a high semantic similarity with the respective reference text topic. Also, our method’s focus on semantics limits its capability of capturing morphological or syntactic errors that may also affect a summary’s quality. We further note that our method is constrained by the following three assumptions: (1) that topical diversity is vital for high-quality summaries, (2) that clustering of sentence embeddings provides a good representation of such topics, and (3) that computing, visualizing, and comparing the similarity between the contextual token embeddings of a summary and the centroids of clustered sentence embeddings of its reference text allows to meaningfully interpret the summary’s quality.

**Future Work** We suggest future work to test fine-tuning the encoding model, e.g., with the goal of optimizing the creation of sentence embedding clusters. Moreover, promising extensions of our method and toolbox are integrated scoring of summary morphology and syntax. Also, we motivate the validation in further domains (e.g. legal and medical texts) and on multilingual data using multilingual encoders. In addition to that, validating our approach on more recent datasets e.g. SummEval (Fabbri et al., 2021), or RealSumm (Bhandari et al., 2020) and Newsroom (Grusky et al., 2018) are suggested to improve the generalizability of our findings. The interpretability toolbox could potentially be enriched by the addition of an interactive Sankey diagram, as indicated in Figure 1. The latter further enables users to instantly and visually detect missing, under- and overrepresented topics in the summary.

**Closing Remarks** Our work calls for further research on interpretable and multifaceted summary evaluation methods and shows promising potential towards closing this research gap identified by Gehrmann et al. (2022).

**References**

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2020. Better rewards yield better summaries: Learning to summarise without references. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3110–3120. Association for Computational Linguistics.

John S. Bridle. 1989. *Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters*. In *Advances in Neural Information Processing Systems*, MI, pages 211–217.

Ht Dang and Karolina Owczarzak. 2008. *Overview of the TAC 2008 update summarization task*. *Theory and Applications of Categories*, pages 1–16.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, volume 1*, pages 4171–4186. Association for Computational Linguistics (ACL).

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. *Automatic text summarization: A comprehensive survey*. *Expert Systems with Applications*, 165.

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

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*.

Carlos R. García-Alonso, Leonor M. Pérez-Narango, Juan C. Fernández-Caballero, Laurens der Maaten,
and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11):187–202.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *arXiv preprint arXiv:2202.06935*.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

C Y Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)*, pages 25–26.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Yassine Mrabet and Dina Demner-Fushman. 2020. HOLMS: Alternative Summary Evaluation with Large Language Models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5679–5688.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *HLT-NAACL 2004 - Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, pages 145–152.

Jun Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, and Yorktown Heights. 2001. IBM Research Report Bleu : a method for Automatic Evaluation of Machine Translation. *Science*, 22176:1–10.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Nils Reimers and Iryna Gurevych. 2020. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, pages 905–914.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2020. SUMQE: A BERT-based summary quality estimation model. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 6005–6011.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*.

Hao Zheng and Mirella Lapata. 2020. Sentence centrality revisited for unsupervised summarization. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 6236–6247. Association for Computational Linguistics (ACL).