# Summarization of Biomedical Texts using Keyword Extraction, Pretrained Models and Ontological Evaluation

Dimitrios A. Koutsomitropoulos, *member IEEE*, Athanasios N. Pyrgas and Andreas D. Andriopoulos
Department of Computer Engineering and Informatics, University of Patras, Greece
Email: {koutsomi, pyrgas, andriopa}@ceid.upatras.gr

***Abstract*—The rapid expansion of biomedical literature necessitates efficient methods for generating concise and accurate summaries. This paper investigates transformer-based models for biomedical text summarization and generation, integrating keyword extraction with a fine-tuning pipeline. We evaluate BioBART v2 and BioT5, across summarization tasks for biomedical abstracts. We suggest an automated process based on named entity extraction to enable large-scale dataset creation without manual annotation. A multi-dimensional evaluation framework is introduced, combining lexical, semantic, and concept-level metrics based on UMLS. Authoritative ontological concepts can offer ground in terminology and better capture the preservation of scientific information. Results show that BioBART v2 achieves superior performance in semantic alignment and concept retention.**

***Index Terms*—Biomedical NLP, Text Summarization, Transformers, UMLS, BioBART, BioT5**

## I. Introduction

The exponential growth of biomedical literature presents a major challenge for knowledge extraction. Databases such as PubMed [1] contain millions of scientific articles, making manual analysis impractical. Biomedical text differs from general-domain text due to its high information density, specialized terminology, and strict requirement for factual accuracy. Errors such as hallucinations can lead to misleading scientific conclusions.

The biomedical summarization task presents particular difficulties due to the high density of information and the need to preserve scientifically accurate concepts. Within this context, modern language models may exhibit hallucination phenomena [2], especially in open generative settings. This makes it necessary not only to develop robust models but also to adopt reliable evaluation methods that go beyond traditional lexical metrics.

In this paper we focus on the comparative evaluation of two domain-adapted encoder–decoder models, BioBART v2 [3] and BioT5 [4], selected for their suitability for conditional text generation and their ability to be fine-tuned on biomedical data.

The evaluation is based on a multidimensional framework that includes lexical metrics (BLEU, ROUGE), semantic metrics using pretrained models (BERTScore) and concept-level evaluation through biomedical entity recognition using SciSpaCy [5] and mapping to UMLS (Unified Medical Language System) [6]. This approach enables assessing the quality of generated texts not only in terms of linguistic similarity but also in terms of preservation of scientific information.

Initially, a thematically focused dataset is developed, based on real biomedical abstracts from the PubMed database, with emphasis on selected sub-areas of oncology. This dataset is organized into input–target pairs and supports both the training and the systematic evaluation of text-generation models in a setting characterized by high terminological specialization.

We suggest a text condensation pipeline by means of the YAKE algorithm [7] and named entity extraction, in order to achieve compression and enable efficient processing and evaluation with large-volume datasets. Subsequently, a comparative study is conducted on BioBART v2 and BioT5, enabling analysis of how architecture and pretraining strategy influence the quality of generated biomedical text.

A central contribution of this work is the development of a multidimensional evaluation framework that combines lexical metrics, semantic metrics and concept-level evaluation based on UMLS. The integration of conceptual criteria allows assessment of the quality of generated texts at the level of biomedical ontology concepts, addressing a critical gap in traditional evaluation methods.

Experimental source code, datasets and results are openly available on Github[1]. The rest of this paper is organized as follows: In Section II we review related work towards summarization of scientific texts, standard evaluation metrics and their existing limitations. Section III presents our methodology and pipeline from dataset instantiation to text condensation to models' finetuning. The actual evaluation procedure, its metrics and conceptual-level evaluation are detailed in Section IV. Section V presents our results and performs a discussion about models performance. Finally, Section VI summarizes our conclusions and future work.

## II. Related Work

Automatic summarization of scientific texts is one of the most active research areas in natural language processing, especially after the widespread adoption of pretrained models and the availability of large datasets. In the biomedical domain,

[1] https://github.com/swigroup/bioSum

the need for effective summarization is stressed by the volume and complexity of available scientific information.

Early approaches relied mainly on extractive methods, which selected sentences from the original text based on statistical or semantic criteria. Large datasets and benchmarks such as the BioASQ challenge [8] have significantly contributed to progress in the field by providing a common reference point for evaluating different methods. These works focus primarily on extracting and organizing information from large-scale biomedical data, while also incorporating elements of summarization and question answering. However, such methods often fail to produce coherent and natural language, which led to a gradual shift toward abstractive approaches that use neural models to generate new text and allow greater flexibility in rephrasing information.

With the introduction of transformer models, a substantial improvement in the quality of generated summaries has been observed. Studies leveraging architectures such as BART [9] and T5 [10], as well as their biomedical adaptations, have shown that these models can produce more coherent and information-dense summaries compared to earlier methods. For example, the work of Cohan et al. [11] proposes an abstractive summarization model for long scientific documents that takes discourse structure into account, highlighting the importance of exploiting the structure of scientific articles.

At the same time, there is growing interest in more open-ended biomedical text generation scenarios that go beyond the classical summarization problem. This includes large-scale autoregressive models such as BioGPT [12] and Med-PaLM [13], which are capable of generating extensive and linguistically coherent scientific text based on limited or even incomplete input. These models shift the problem from compression towards knowledge-driven generation, introducing new challenges regarding the reliability and interpretability of the produced information.

However, many approaches rely either on general-domain datasets or on relatively small datasets with annotated summaries, which restricts the ability of models to generalize across different biomedical subdomains. Moreover, the quality of available gold-standard summaries often exhibits inconsistencies, affecting the reliability of both training and evaluation.

One of the key issues concerns evaluation methods. Established metrics such as BLEU [14] and ROUGE [15] rely mainly on surface-level word or phrase overlap and do not adequately capture the semantic quality or scientific accuracy of generated texts. The introduction of metrics such as BERTScore [16] attempts to address this by leveraging contextual representations, yet it still does not provide a complete picture of conceptual correctness.

In addition, the phenomenon of hallucinations [2] remains one of the most significant limitations of modern text-generation models. Models may produce linguistically convincing text that contains inaccurate or nonexistent information, which is particularly problematic in the biomedical domain, where accurate information is critical. This issue becomes more pronounced in open generative settings, where the output is not strictly constrained by the input.

Furthermore, recent studies on large-scale systems such as Med-PaLM have shown that, despite strong performance on medical knowledge tasks, issues related to transparency, controllability, and verifiability of generated information persist. At the same time, studies such as SummEval [17] highlight that automatic metrics often do not fully correlate with human judgment, making it necessary to explore alternative evaluation approaches.

In this context, the present work aims to contribute to the investigation of the above issues, focusing particularly on the gap concerning the reliable evaluation of generated information in biomedical settings. Specifically, work in this paper does not merely rely on traditional text-comparison metrics but adopts a multidimensional evaluation approach that combines lexical, semantic and conceptual metrics, based on biomedical entities through UMLS.

## III. Methodology

### A. Dataset

For the purposes of this work, a specialized dataset of biomedical abstracts was constructed from PubMed. PubMed was selected due to its central role in disseminating biomedical knowledge and its extensive coverage of scientific publications. The dataset was constructed from 18,885 PubMed abstracts collected using 20 manually selected oncology-related search terms covering major cancer types, treatment modalities, biomarkers, and cancer biology topics.

Each dataset instance consists of an *input* i.e., the full scientific abstract and a *target* that is, the shortened version of the abstract. Targets were generated automatically rather than manually. The YAKE algorithm was used to extract key terms based on statistical features such as term frequency, position, and co-occurrence. Sentences containing the most important terms were then selected and combined to form the condensed target summary.

This automated process enables large-scale dataset creation without manual annotation, which would be impractical in such a specialized domain. However, it should be noted that the quality of the generated targets depends on the performance of the keyword-extraction method and may not fully capture the semantic structure of the original text.

For the quantitative assessment of the condensation process, basic statistical measures were computed on the test set (Table I). This process yields a compression ratio of approximately 0.44, meaning that the target texts are on average less than half the length of the original abstracts. The dataset is then split into training (70%), validation (10%), and test (20%) sets for experimental purposes: the training set was used for fine-tuning the models, the validation set for monitoring the training progress and adjusting hyperparameters, and the final evaluation was conducted on a shared test set consisting of 3,777 examples.

TABLE I
STATISTICS OF THE TEST SET

| Metric | Value |
|---|---|
| Mean input length | 226.59 words |
| Median input length | 227 words |
| Mean target length | 99.93 words |
| Median target length | 76 words |

### B. Pipeline

The creation of the input–target pairs used for training the summarization models was carried out through an automated processing pipeline applied to the original abstracts. This pipeline was designed to produce condensed versions of the texts while preserving the most important elements of the scientific content.

As shown in Fig. 1 the overall procedure follows a sequential workflow, beginning with preprocessing of the abstracts and ending with the generation of the final target texts. Initially, the abstracts undergo basic preprocessing, including text cleaning, removal of non-informative characters, and sentence segmentation.

Next, the YAKE algorithm is applied to extract keywords from each abstract. YAKE assigns an importance score to each candidate keyword based on statistical features such as term frequency, position within the document, and co-occurrence patterns. Since YAKE relies exclusively on statistical features, it may highlight terms that lack genuine biomedical relevance.

YAKE was configured for English text (lan="en") with a maximum n-gram length of 10 (n=10) and the extraction of the top 20 candidate keywords (top=20). To improve biomedical relevance, candidate keywords were further filtered based on Named Entity Recognition (NER) using biomedical entities identified by SciSpaCy. Specifically, only sentences containing at least one recognized biomedical entity are retained and then combined to form the final condensed version of the abstract, which serves as the training target.

### C. NER

For biomedical entity extraction, the SciSpaCy library was used. SciSpaCy extends spaCy and provides models specifically adapted for processing scientific and biomedical text. These models are trained on specialized corpora and can recognize entities related to diseases, drugs, genes, proteins, and broader biological or clinical concepts.

Entity recognition is applied both to the reference texts and to the model-generated summaries to ensure that results are comparable at the same conceptual level. SciSpaCy models rely on neural architectures that use contextual embeddings, enabling more effective entity detection in domain-specific texts.

For each document, the system identifies lexical mentions corresponding to candidate biomedical entities and produces an initial set of conceptual units for further processing. In this work, the `en_core_sci_lg` model was used, offering broad coverage of biomedical terminology.

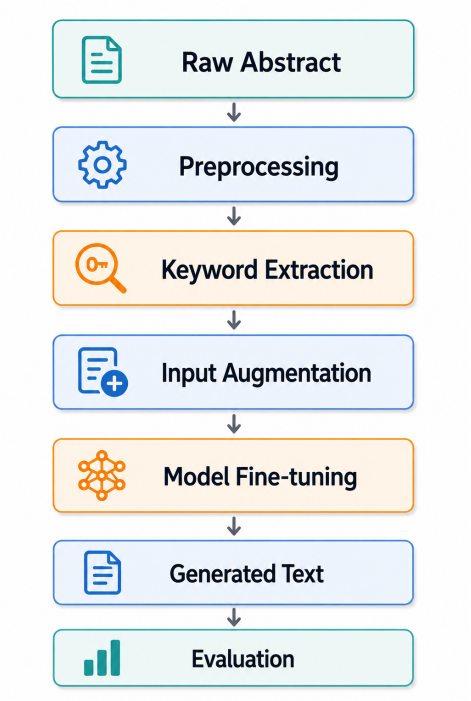


Fig. 1. Medical Summarization Pipeline Overview

Beyond basic NER, SciSpaCy also supports entity linking, enabling the alignment of recognized terms with corresponding concepts in external knowledge bases such as the UMLS. This capability is particularly important, as it serves as the intermediate step between surface-level text analysis and the ontological normalization described in the next Section IV.

### D. Models training and finetuning

The training of the models was carried out through a fine-tuning process applied to pretrained, transformer-based models, using the dataset constructed in the previous stage. Specifically, fine-tuning was performed on the BioBART v2 and BioT5 models, both of which have been pretrained on large-scale biomedical corpora.

*BioBART v2* is an adaptation of the BART model to the biomedical domain through additional pretraining on scientific corpora. It follows an encoder–decoder architecture and is designed for text-to-text tasks such as summarization and translation.

Its pretraining relies on denoising autoencoding objectives, where the model is required to reconstruct the original text from a corrupted version. Typical corruptions include word deletion, sentence shuffling, and noise injection. This process enables the model to learn strong representations that preserve textual coherence and structural consistency during generation. Due to this training strategy, BioBART performs particularly well on summarization tasks, as it tends to reproduce the structure and content of the input more faithfully, producing stable and coherent summaries.

*BioT5* is an adaptation of the T5 model for the biomedical domain and follows the text-to-text philosophy, in which all

natural-language processing tasks are formulated as text transformation problems. Its pretraining is based on span-corruption objectives, where spans of text are replaced with special tokens and the model is trained to reconstruct them. This approach encourages the model to generate more flexible and abstractive formulations.

As a result, BioT5 exhibits stronger paraphrasing and generalization capabilities, making it suitable for producing more free-form text. However, this increased flexibility may come at the cost of reduced faithfulness to the original content, especially in tasks requiring strict preservation of scientific information.

Using pretrained models enables the exploitation of rich linguistic and semantic representations learned during pretraining. These representations capture both general language knowledge and domain-specific biomedical information, making the models particularly suitable for scientific text processing.

Fine-tuning consists of further training the models on the summarization task using the constructed input–target pairs. The models were trained using a shared set of hyperparameters to ensure a fair comparison, with differences in performance due primarily to architectural and pretraining factors. Training employed a learning rate of $1 \times 10^{-5}$, a batch size of 16 with 2 gradient-accumulation steps, maximum input and output lengths of 512 and 256 tokens respectively, the Adam optimizer with weight decay set to 0.01, and up to 100 epochs with early stopping based on ROUGE-L performance on the validation set.

### *E. Text generation*

After completion of the fine-tuning process, trained models are used to generate summaries for each abstract in the test set. This corresponds to the inference stage, during which the model is applied to unseen data in order to evaluate its generalization ability.

Output generation is performed through auto-regressive decoding, where the model produces the output sequence step by step, predicting the next token based on the previously generated tokens and the encoded representation of the input. This process continues until a special end-of-sequence token is produced or the maximum allowed sequence length is reached.

A decoding strategy is used to control the generation process, determining how the next token is selected. In this work, beam search (with a beam width of 4) is applied, allowing the model to explore multiple candidate sequences simultaneously and select the one with the highest overall probability. Beam search improves the coherence and quality of the generated text, although it may reduce linguistic diversity. The same decoding strategy is applied to both models to ensure comparability of results.

The generated summaries are then stored and used for the evaluation of the models, enabling a systematic comparison of their performance on real-world data. A sample of the generated text for a particular input and target is shown in Table II.

TABLE II
EXAMPLE OF MODEL OUTPUTS FOR A BIOMEDICAL ABSTRACT

| | |
|---|---|
| **Input Abstract (excerpt)** | Previous studies have demonstrated corilagin's inhibitory effects on various cancer cells. This study investigated the effects of corilagin on apoptosis in A2780 ovarian cancer cells and explored the underlying molecular mechanisms. |
| **Target Summary (excerpt)** | Corilagin inhibits A2780 ovarian cancer cell proliferation and induces apoptosis through cell cycle arrest and activation of the PI3K/AKT-p53 pathway. |
| **BioBART v2 Output (excerpt)** | Corilagin induces apoptosis in A2780 ovarian cancer cells through the PI3K/AKT pathway and mitochondrial dysfunction. |
| **BioT5 Output (excerpt)** | Corrugoervel Cinn-Exhibits A2780 A2780 Cell Adenocarcinoma Cellular Motm Growth. Previous studies have demonstrated corilagin's inhibitory effects on the growth of various cancer cells. |

## IV. EVALUATION

### *A. Procedure*

After completing the fine-tuning process, the models were used to generate summaries for every instance in the test set. The resulting summaries were then compared with their corresponding reference summaries through a multi-level evaluation framework.

First, lexical and structural similarity was examined using the BLEU and ROUGE metrics, which quantify the degree of overlap between generated and expected summaries. Next, semantic proximity was evaluated using BERTScore, which leverages contextual token embeddings to identify meaning-preserving expressions even when surface forms differ substantially. Finally, an additional biomedical, concept–level evaluation was introduced through entity extraction and normalization within the UMLS framework, enabling an assessment of how well each model preserves domain-specific scientific information. The various levels of the evaluation procedure are shown in Fig. 2.

### *B. Metrics*

*1) BLEU:* The BLEU metric (Bilingual Evaluation Understudy) is used to measure n-gram overlap between the generated summary and the target summary, with higher scores indicating stronger lexical correspondence and therefore greater similarity in local linguistic structure. In this work, BLEU was computed at the example level for all instances in the test set, and the final score for each model corresponds to the average across all samples. During training, the evolution of BLEU was also monitored across epochs on the validation set to track the models' improvement trajectory and assess whether training led to consistent gains in lexical alignment.

*2) ROUGE-L:* The ROUGE-L metric is widely used in the evaluation of automatic summarization systems, as it measures the longest common subsequence (LCS) between two texts. This allows the metric to capture not only content preservation but also the structural ordering of information between the

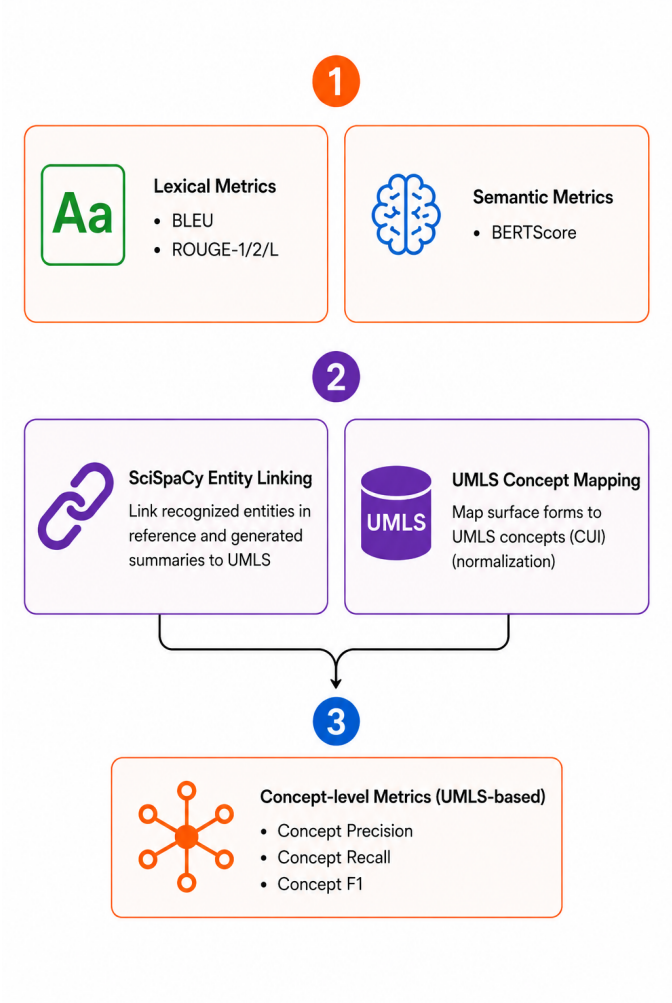


Fig. 2. The multi-level evaluation framework.

generated summary and the target text. In contrast to BLEU, which emphasizes local n-gram overlap, ROUGE-L is better suited for summarization tasks because it more effectively reflects global coherence and information flow. For this reason, ROUGE-L serves as the primary evaluation metric in this work and is used as the main criterion for the early stopping procedure.

*3) BERTScore:* Although BLEU and ROUGE provide useful indications of lexical and structural similarity between generated summaries and target texts, they do not fully capture semantic equivalence across different phrasings. In practice, two summaries may convey the same information using different wording or syntactic constructions, leading surface-overlap metrics to underestimate their true quality.

To address this limitation, the BERTScore metric is employed, leveraging contextual embeddings produced by pre-trained language models. Unlike traditional n-gram–based metrics, BERTScore computes semantic similarity between tokens in the embedding space, enabling it to detect meaning-preserving expressions even when the lexical form differs substantially.

The metric reports three quantities: *Precision*, reflecting how well the generated summary's tokens semantically match those of the reference; *Recall*, indicating how much of the reference information is captured; and the *F1-score*, the harmonic mean of the two, used as the overall semantic similarity indicator. In this paper, BERTScore was computed across the entire test set by comparing each model's generated summaries with their corresponding target summaries.

### C. Concept-level evaluation

Although the BLEU, ROUGE, and BERTScore metrics provide useful indications of lexical and semantic similarity between generated summaries and target texts, they do not directly assess whether the critical biomedical concepts of the original text are preserved. These metrics operate primarily at the level of words or distributed representations and may therefore fail to fully capture the correctness of the underlying scientific information.

In the biomedical domain, maintaining the meaning of concepts is particularly important, as even small deviations in terminology (for example, distinctions between related but non-equivalent concepts) can lead to substantial distortion of the conveyed information. For this reason, we have included an additional concept-level evaluation based on the UMLS (Unified Medical Language System). UMLS integrates a wide range of medical and biological ontologies and taxonomies, providing for each concept a unique identifier known as the *Concept Unique Identifier (CUI).*

Our approach enables the mapping of both reference texts and generated summaries to CUIs, which serve as normalized representations of biomedical concepts. In this way, the comparison between texts can be performed at the level of semantic units, independently of their specific lexical realization.

The evaluation follows the procedure below, applied uniformly to all examples in the test set:

1) *Named Entity Recognition* on both target texts and generated summaries using the SciSpaCy library.
2) *Normalization* of the extracted entities to CUIs through matching with the UMLS lexicon.
3) *Comparison* of the resulting concept sets between generated and target summaries to estimate the degree of scientific information preservation.

This comparison is based on the precision, recall, and F1-score metrics, which are computed over sets of extracted concepts. *Precision* expresses the proportion of concepts in the generated summary that correspond to concepts in the reference text, serving as an indicator of the "purity" of the generated information. *Recall* reflects the proportion of reference concepts that are covered by the generated summary, functioning as a measure of completeness. The *F1-score*, as the harmonic mean of the two, provides an overall indicator of concept-level alignment by balancing these complementary dimensions.

## V. Results and Discussion

### A. Main Results

Table III presents the results for the BLEU and ROUGE-L metrics across the models evaluated. As a means of ablation, results also include the corresponding performance of the general-purpose models (BART and T5), allowing to illustrate the impact of domain adaptation on model behavior. As observed, the biomedical models consistently exhibit improved performance both in quality metrics such as BLEU and ROUGE, as well as throughout the training process, confirming the importance of specialized pretraining for domain-specific tasks.

TABLE III
BLEU and ROUGE-L scores for the Summarization Task

| Model | BLEU | ROUGE-L |
|---|---|---|
| BioBART v2 | 0.6433 | 0.7205 |
| BART | 0.4735 | 0.6043 |
| BioT5 | 0.5357 | 0.6168 |
| T5 | 0.4406 | 0.5419 |

The superiority of BioBART v2 with respect to the BLEU metric is particularly indicative of the model's tendency to preserve a higher degree of lexical fidelity to the reference text. This finding is consistent with the model's pretraining strategy (denoising autoencoding), which encourages reconstruction and retention of the input structure.

It is also important to note that the reference summaries were produced mostly through an extractive process, a factor that may favor models exhibiting greater lexical overlap with the original text. Consequently, the higher BLEU performance of BioBART v2 may reflect not only better pretraining alignment but also better compatibility with NER and keyword extraction.

BioT5 converges to slightly lower ROUGE-L values, indicating that despite its general ability to produce coherent text, it does not necessarily maintain a high degree of structural alignment with the reference summary. This finding suggests that a model's capacity to generate more "free-form" and abstractive text does not inherently translate into better performance on metrics that rely on structural correspondence, such as ROUGE-L.

Table IV shows the BERTScore results between the two models. The superiority of BioBART v2 in terms of BERTScore indicates that the model does not merely preserve the surface form of the text but also captures the underlying semantic content more faithfully. This observation aligns with its denoising-based pretraining strategy, which strengthens the model's ability to maintain the coherence and semantic structure of the input.

### B. Concept-level evaluation results

The primary comparison for evaluating the models is performed between the reference summaries (target) and the generated summaries, as this directly reflects the extent to which the models succeed in preserving the critical biomedical concepts, ensuring conceptual preservation (Tables V & VI).

TABLE IV
BERTScore results for the Summarization Task

| Model | Mean Precision | Mean Recall | Mean F1-score |
|---|---|---|---|
| BioBART v2 | 0.9636 | 0.9616 | 0.9623 |
| BioT5 | 0.9532 | 0.9467 | 0.9496 |

TABLE V
Macro-level concept metrics (Target vs. Generated)

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BioBART v2 | 0.7778 | 0.7624 | 0.7434 |
| BioT5 | 0.7553 | 0.6633 | 0.6754 |

TABLE VI
Micro-level concept metrics (Target vs. Generated)

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BioBART v2 | 0.7551 | 0.7198 | 0.7370 |
| BioT5 | 0.7505 | 0.6226 | 0.6806 |

The results show that BioBART v2 consistently outperforms BioT5 across all concept-level metrics, both under macro and micro averaging. This superiority indicates that the model preserves a larger proportion of the biomedical concepts present in the reference summaries.

The difference in recall is particularly noteworthy, as it is substantially higher for BioBART v2 (0.7624 versus 0.6633 at the macro level). This finding suggests that BioBART v2 captures a greater share of the concepts contained in the reference summaries, whereas BioT5 tends to lose information.

The difference in precision is smaller, indicating that both models maintain a similar level of purity in the concepts they generate and largely avoid introducing irrelevant or incorrect concepts. However, the lower recall of BioT5 leads to a significant reduction in overall performance, as reflected in the F1-score.

### C. Discussion

The BLEU and ROUGE-L results show that BioBART v2 consistently outperforms BioT5 in lexical and structural alignment with the reference summaries. The gains are evident for both BLEU (0.6433 vs. 0.5357) and ROUGE-L (0.7205 vs. 0.6168), indicating stronger preservation of reference wording and content ordering. At the semantic level, BioBART v2 also achieves higher BERTScore precision, recall, and F1 values. Together, these results suggest that the denoising-based BioBART v2 architecture is better aligned with the keyword-guided, extractive reference summaries used in this study.

At a semantic level, as measured by BERTScore, BioBART v2 is also better, even if only slightly, with both models achieving relatively high F1-scores. This difference highlights a critical limitation of surface-overlap metrics: lexical similarity does not necessarily imply semantic equivalence.

At the concept level, BioBART v2 achieves higher F1 scores, with absolute improvements of 0.068 under macro

averaging and 0.056 under micro averaging. The focus is now on the preservation of biomedical concepts regardless of the linguistic form in which they are expressed. This approach enables a direct assessment of the scientific information retained in the generated summaries, offering a level of analysis that is not accessible through traditional metrics.

The comparison of results shows that strong performance in one category of metrics does not necessarily imply comparable performance in another. A model may achieve high BLEU or ROUGE scores by reproducing surface-level elements of the text, while still failing to preserve the full set of critical concepts. Similarly, even high semantic similarity, as captured by BERTScore, does not guarantee complete conceptual coverage, particularly in cases of aggressive compression.

## VI. Conclusions and Future Work

Evaluating natural language generation systems is one of the most complex and open problems in NLP, as there is no single metric capable of fully capturing the quality of the generated text. The findings of this paper confirm that the use of multiple complementary metrics is a necessary condition for reliably assessing model performance.

The BLEU and ROUGE-L metrics provide an initial estimate of the lexical and structural similarity between generated texts and reference texts. The use of BERTScore then enables a shift from lexical to semantic evaluation, as it relies on contextualized vector representations.

Concept-level evaluation through UMLS allows for the comparison of texts at the level of biomedical concepts (CUIs), providing a more direct assessment of the preservation of scientific information. This underscores the importance of a multidimensional evaluation framework, especially in biomedical NLP applications, where the accuracy and completeness of scientific information are critical factors.

The suggestion of using keyword extraction (YAKE) is shown to allow for efficient scaling and creation of larger evaluation datasets. Still, the use of human-annotated summaries or the creation of multi-reference datasets, where multiple acceptable summaries correspond to the same source text, could lead to more realistic training and more reliable evaluation, particularly in open generative settings.

Finally, future work can investigate hybrid approaches that combine summarization and text generation; as a result, models could dynamically adjust the degree of information compression or expansion depending on the available context.

## References


[1] National Library of Medicine, "PubMed: Biomedical Literature Database," https://pubmed.ncbi.nlm.nih.gov/, U.S. National Institutes of Health, Bethesda, MD.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[3] Q. Yuan, Z. Yuan, H. Li, X. Wang, S. Wu, and J. Zhang, "BioBART: Pretraining and Evaluation of Biomedical Generative Language Models," in *Proceedings of the BioNLP Workshop at ACL 2022*, 2022, pp. 1–10.

[4] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian, G. Altan-Bonnet, K. Van Auken, and W. Yoon, "BioT5: Enriching Cross-Task Biomedical Text Generation with Pre-trained Language Models," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 4235–4245.

[5] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 319–327.

[6] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology," *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D267–D270, 2004.

[7] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword Extraction from Single Documents Using Multiple Local Features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[8] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, and I. A. Kakadiaris, "BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering," *BMC Bioinformatics*, vol. 22, pp. 1–30, 2021.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of ACL 2020*, 2020, pp. 7871–7880.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[11] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents," in *Proceedings of NAACL-HLT 2018*, 2018, pp. 615–621.

[12] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022.

[13] K. Singhal, S. Azizi, T. Tu, S. Mahdavi, J. Wei, H. W. Chung, N. Scales *et al.*, "Large Language Models Encode Clinical Knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of ACL 2002*, 2002, pp. 311–318.

[15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[16] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations (ICLR)*, 2020.

[17] K. Krishna, F. Schmidt, B. Dhingra *et al.*, "SummEval: Re-evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021.