An Open-Resource Knowledge Augmentation for Biomedical Lay Summarization

João Pedro Veloso, Evelin Amorim
INESC TEC
{joao.p.veloso, evelin.f.amorim}@inesctec.pt

Abstract

Automatic summarization aims to generate concise versions of texts while retaining relevant information. Summaries can be either extractive, using direct excerpts, or abstractive, rephrasing content to convey the same meaning. Lay summarization applies abstractive techniques to simplify complex texts, such as scientific literature, for broader audiences, thereby promoting public understanding of specialized knowledge. Prior work shows that knowledge augmentation improves lay summarization. Still, biomedical applications often rely on closed resources like the Unified Medical Language System (UMLS), which require expert curation and are costly to scale. We propose a four-step approach that leverages keyword extraction and DBpedia, an open general domain knowledge base, ideal to bridge the gap between expert and lay knowledge. First, we extract keywords from biomedical texts using YAKE!, a well-established unsupervised method. Second, we query DBpedia using these keywords to retrieve relevant concept entries. Third, we construct a graph of concepts for each document based on cosine similarity between DBpedia entries. Finally, we combine each graph with the original abstract to train a summarization model. Our method achieves competitive performance compared to UMLS-based systems in the eLife dataset (ROUGE-1: 58.44 vs. 60.26, ROUGE-L: 43.45 vs. 45.45), demonstrating that open-resource approaches can provide viable alternatives to licensed knowledge bases while maintaining accessibility for resource-constrained organizations.

Keywords: Biomedical Lay Summarization, Knowledge Graphs, DBpedia, Open Resources

1. Introduction

Scientific communication has accelerated rapidly due to the digitization of journals, pre-print servers, and online news (Bornmann and Mutz, 2015; Fraser et al., 2021). This shift has generated an unprecedented volume and velocity of biomedical findings entering the public domain (Landhuis, 2016), necessitating lay summaries to bridge the gap between experts and the general public (Goldstein and Krukowski, 2023). Within Natural Language Processing (NLP), automatic summarization has emerged as a key strategy for condensing complex documents while preserving salient information (Xie et al., 2023; Chaves et al., 2022; Wang et al., 2021). However, lay summarization of biomedical text still faces significant hurdles, including the need to provide sufficient methodological context to prevent model hallucinations and ensure the rigorous preservation of scientific facts (Xiao et al., 2025).

Several approaches have been developed to mitigate hallucinations in biomedical lay summarization. For instance, Goldsack et al. (2023) proposed representing biomedical papers through a combination of text embeddings and concept graphs extracted from the Unified Medical Language System (UMLS)[1]. Tang et al. (2023) also used graph-based methods, but of citation papers, to diminish the hallucinations. However, the use of citation papers provides limited utility for linguistic simplification. Furthermore, reliance on licensed knowledge bases like UMLS restricts the accessibility of these techniques for many organizations. Although UMLS is free for research use, access is not entirely straightforward: it requires registration, periodic license renewal, and typically some form of institutional affiliation. In our experience, this alone can already be a barrier for independent researchers or smaller teams. A more limiting issue is the license itself. UMLS explicitly forbids the redistribution of both the original resource and derived data products. This makes it difficult to share processed datasets or to release systems that depend on it in a fully reproducible way. For example, even when a model can be distributed, anyone attempting to use it must first obtain and configure UMLS independently, which is a non-trivial step. In practice, this makes it difficult to share and reproduce results in line with open research principles. This tension is one of the reasons we explore open-access alternatives, aiming to retain competitive performance while lowering the barrier to entry and enabling broader dissemination. We propose a new technique that employs an open resource for summarization and that presents competitive results with a state-of-the-art licensed technique (Goldsack et al., 2023).

Our proposed technique consists of three stages: (1) automatic keyword extraction from the source text using YAKE! (Campos et al., 2020); (2) retrieval of these keywords from DBpedia and subsequent classification into biomedical and non-biomedical

[1] https://uts.nlm.nih.gov/uts/umls/home

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 50–60

concepts; and (3) the construction of a concept graph integrated with text embeddings for summary generation. The hypothesis in using keywords is the identification of relevant concepts of a given paper more efficiently, discarding biomedical concepts that are not central to the whole idea of the research. To validate this, we conducted a manual evaluation of 500 extracted keywords (10 per document) to assess their ability to represent the paper's core medical meaning. Also, by using DBPedia, a general domain knowledge base, we aim to bridge the gap between technical and lay vocabulary. To diminish the noise from DBPedia, we apply a biomedical text classification to remove irrelevant concepts. Finally, we connect the concepts by using cosine similarity, and use the built graph to enrich the text representation to generate the lay summary. While UMLS-based approaches have achieved strong performance in biomedical lay summarization, their reliance on licensed resources creates barriers for many organizations, particularly those in low-resource settings or with limited budgets. Our work demonstrates that competitive performance (97% of UMLS-based ROUGE-1 scores) can be achieved using only open resources, thereby democratizing access to biomedical lay summarization technology. To our knowledge, this is the first study demonstrating that open general-domain knowledge bases can substitute licensed biomedical ontologies for lay summarization with minimal performance loss.

Our work aims to answer the following research questions:

RQ1: To what extent can a curated subset of automatically extracted keywords from biomedical papers represent central biomedical concepts?

RQ2: Can a lay summarization technique leveraging open-source knowledge bases compete with approaches using licensed ontologies?

Our contributions are two-fold: (i) we explore the feasibility of using keyword-based representations to preserve medical meaning, validated through a targeted manual assessment; and (ii) we establish that open-source knowledge bases provide a viable and competitive alternative to licensed proprietary databases in the biomedical domain. In the next sections, we describe similar work to ours, emphasizing our contributions, and then describe our method and its results. We release our code to facilitate reproducibility and further research[2].

[2] https://github.com/evelinamorim/openbiolaysumm

2. Related Work

With the introduction of neural and transformer-based architectures, significant improvements in performance have been made in the summarization of texts. In particular, for biomedical summarization such as (Luo et al., 2023), there has been the exploration of readability-controllable summarization. The authors use transformer models, adapting biomedical content to various levels of readability, which is crucial for bridging expert and lay audiences. In particular, a Longformer-Encoder-Decoder (LED) model is fine-tuned as a baseline model, with additional strategies built as variations upon it. They apply the proposed technique to a new dataset composed of 28,124 articles published in PLOS journals[3], where PLOS is a nonprofit open-access publisher covering a wide range of scientific disciplines, including medicine and the life sciences. The authors evaluated their results using ROUGE scores, achieving peaks of 49.87 for ROUGE-1 and 46.02 for ROUGE-L. While our strategy also utilizes the LED architecture, our proposal incorporates an external knowledge base in addition to the source text and summaries used by Luo et al. (2023).

Another approach that incorporates external knowledge is KeBioSum (Xie et al., 2022). The enriched content is based on the PICO framework (Population, Intervention, Comparison, and Outcome) (Huang et al., 2006), which is a structure designed to represent medical knowledge in queries. The method extracts named entities based on the PICO framework using fine-tuned models. Then, the identified elements are combined with the text document to produce the summaries. The authors used three datasets, ORD19 (Wang et al., 2020), PubMed (Cohan et al., 2018), and S2ORC (Lo et al., 2020) to evaluate the effectiveness of their approach. The ROUGE-L values for these datasets were respectively 29.10, 32.28, and 34.08. Despite enriching the content to improve summarization, the exploitation of external concepts is limited only to the identification of relevant terms.

Goldsack et al. (2023), similar to ours, uses the concepts from an external knowledge base, UMLS, and their relations to aid the learning of the biomedical lay summary model. Like our proposal: extraction of relevant concepts, building on the graph of the concepts, and learning the summaries. However, they use a UMLS-based library to identify concepts and their relations. They use two datasets, one composed of papers collected from PLOS journals and another from the eLife journal (Goldsack et al., 2022). The results reported in a combination of the datasets were 48.58 for ROUGE-1 and 45.71 for ROUGE-L. The use of UMLS can limit the usage by organizations, which is not a limitation of

[3] https://journals.plos.org/plosone/

our proposal that uses only open resources (YAKE! and DBPedia) to enrich the input content.

Xie et al. (2024) also employ UMLS to build word and sentence correlation graphs, which are used for topic modeling. The discovery of the topics and summarization are done simultaneously by a neural network, whose inputs are the graphs, the text embeddings, and the gold summary. Like (Xie et al., 2022), in this work, the datasets used are ORD19, PubMed, and S2ORC to evaluate the effectiveness of their approach. The proposed approach achieved ROUGE-1 values between 42.16 and 46.16. Again, the use of a licensed thesaurus limits the use of this approach by organizations.

In the survey by Bhowmik et al. (2025), one challenge highlighted is the need to adopt an open research culture in the biomedical lay summarization task. The reason is that many of the current competitive approaches use the licensed thesaurus UMLS. Our proposal tries to break this barrier with a competitive method based on open-resources.

3. Material and Methods

In this section, we describe the dataset used in our experiments and the main methods we used, including our proposed approach.

3.1. Dataset

Due to limitations of our computational resources, we used a subset of the eLife dataset proposed by Goldsack et al. (2022). The eLife dataset consists of biomedical research articles published in eLife, a peer-reviewed open-access journal in life sciences and biomedicine. Each article includes both the technical scientific abstract and an accompanying "eLife digest", which is a lay summary written to explain the research to general audiences. Our task is to generate these lay summaries from the technical abstracts, requiring both content selection and linguistic simplification. The selected subset is in our repository, and its statistics are described in Table 1.

Split	# Docs	# Words	Avg. Words
Train	1,582	44.38M	10,212.6 (±3,479.8)
Test	241	2.41M	10,009.0 (±3,324.7)
Validation	241	2.42M	10,043.5 (±3,285.9)

Table 1: Statistics of the eLife dataset subset. (M denotes millions).

3.2. Methods

We evaluated four distinct methods. The first is a BART model (Lewis et al., 2020) fine-tuned on the summaries dataset. The second method appends keywords extracted from YAKE! to the article text and feeds this to a BART fine-tuning process. The third is a reproduction of the approach by Goldsack et al. (2023), hereafter referred to as the UMLS-based method. Finally, we propose OpenBioLaySumm, an approach based on YAKE! (Campos et al., 2020) and DBpedia. In our approach, we make a distinction between keywords and concepts. YAKE! first extracts directly from the abstract keywords—n-gram text spans such as "neuron" or "cell membrane". We then map these keywords to concepts, i.e., entries in DBpedia obtained via the Lookup API. Each concept is associated with a URI, a textual description, and additional semantic context. For instance, the keyword "neuron" is mapped to the DBpedia resource `http://dbpedia.org/resource/Neuron` , along with its natural language description.

We implement the doc-enhance architecture from Goldsack et al. (2023), which combines textual and graph-based representations through a multi-component encoder-decoder framework. The pipeline consists of four main steps (Figure 1):

1. **Concept Extraction:** In the UMLS-based variant of our approach, biomedical concepts are identified through mappings to the UMLS Metathesaurus. Goldsack et al. (2023) originally relied on MetaMap for this purpose. However, following recent changes in the UMLS distribution (notably the discontinuation of the MetaMap NLP Content View in the 2025AB release[4]), we instead implement concept recognition using QuickUMLS (Soldaini and Goharian, 2016), which enables approximate string matching over the Metathesaurus and can be readily integrated with custom subsets. By contrast, OpenBioLaySumm does not directly rely on UMLS mappings for initial concept detection. Instead, it applies YAKE! (Campos et al., 2020) to extract the top-ranked keywords from each biomedical abstract. These keywords are subsequently used to query the DBpedia Lookup API[5] , which returns a ranked list of candidate entities for each query term, yielding up to 100 entity candidates per document. Because DBpedia includes entities from multiple domains, an additional filtering step is required to retain only biomedical concepts. To this end, candidate entities are automatically classified as biomedical or

[4] https://www.nlm.nih.gov/pubs/techbull/nd25/nd25_umls_release_ab.html

[5] https://lookup.dbpedia.org/

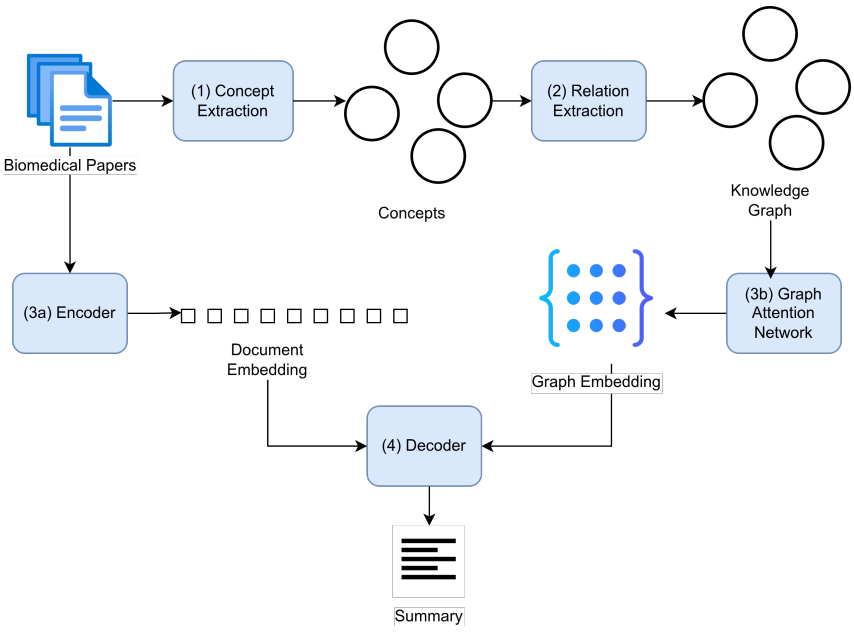

Figure 1: Pipeline for both summarization approaches (UMLS-based and OpenBioLaySumm), following the doc-enhance architecture from Goldsack et al. (2023). Both approaches: (1) extract concepts from biomedical abstracts, (2) construct concept graphs: the UMLS-based approach uses relations provided by the UMLS Metathesaurus, while OpenBioLaySumm computes edge weights via SciBERT cosine similarity > 0.7, (3a) encode the document with LED, (3b) process the graph with a 3-layer GAT, then concatenate both representations through an additional encoder layer, and (4) generate the lay summary with the LED decoder.

non-biomedical using a zero-shot large language model based on their textual descriptions. The results were classified using the description of the entry and DeepSeek-R1:7b as biomedical or non-biomedical concepts with the prompt: "Classify if the following term is related to the biomedical/biological domain: concept description. Answer yes if it's related to biomedical/biological, no if not.". Concepts not assigned to the biomedical domain are discarded prior to graph construction. A manual validation of these LLM classifications was performed by one of the authors to ensure quality (described in Section 4.1).

2. **Relation Extraction:** The two approaches differ mainly in how the concept graphs are constructed. In the UMLS-based setting, we rely on the semantic relations already defined in the UMLS Metathesaurus (e.g., is_a, part_of, treats), which are extracted together with the concepts using QuickUMLS. This results in graphs that reflect curated, domain-specific knowledge. By contrast, OpenBioLaySumm does not assume a fixed set of relations. Instead, relationships are inferred dynamically from semantic similarity. To do this, we encode the DBpedia description of each concept using SciBERT (Beltagy et al., 2019), truncating inputs to 250 tokens. We then apply mean pooling over the token embeddings to obtain a 768-dimensional vector for each concept. To make this step more efficient, we reduce the representations to 50 dimensions using PCA fitted on the training data. Once the representations are obtained, we compute cosine similarity between all pairs of concepts within a document. Edges are added between concepts whose similarity exceeds 0.7, yielding a relatively sparse graph in which each concept is linked to a set of semantically related neighbours rather than to a predefined relation type.

3. **Graph and Document Encoding**: Following the doc-enhance architecture, we encode information from both textual and structured modalities. The document text is processed using the Longformer-Encoder-Decoder (LED) model (Beltagy et al., 2020), which produces document-level representations suitable for long biomedical inputs (approximately 10,000 words on average in our dataset) (Figure 1, 3a). In parallel, the corresponding concept graph is encoded with a 3-layer Graph Attention Network (GAT) (hidden size 768)(Figure 1, 3b). ELU activations are applied in the first two layers, followed by a linear projection in the final layer. To combine both modalities, we concatenate the document embedding obtained from

LED with the graph-level representation produced by the GAT. The resulting representation is then passed through an additional encoder layer, allowing interactions between the textual and graph-based features prior to decoding.

4. **Summary Generation:** We use the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) as our base model, where the encoder processes the concatenated document and graph representations, and the decoder generates the lay summary. In the training, the summary is fed to the decoder to learn such an output.

Our implementation builds on the publicly available code provided by ~ Goldsack et al. (2023), while preserving the original doc-enhance architecture. We made the following changes to support our experimental setup: (1) we replaced MetaMap with QuickUMLS for UMLS concept extraction following recent changes in UMLS content views, (2) we incorporated a concept extraction step based on YAKE! and DBpedia for OpenBioLaySumm, and (3) we included a SciBERT-based similarity module to construct the concept graph during training.

4. Experiments and Results

We did three experiments regarding our method: (1) the extraction of the concepts, (2) the classification of the concepts in biomedical entities, and (3) the lay summarization process. Next, we detail each one of them and its results.

4.1. Extraction of the Concepts

This stage was performed by the UMLS-based method and OpenBioLaySumm. For the UMLS-based method, we adapted the code made available for the paper (Goldsack et al., 2023) in a GitHub repository [6]. The adaptation is because the tool used by the authors, MetaMap, was not maintained anymore[7]. Thus, we use an alternative, the QuickUMLS library which presents the same functionality of extracting UMLS concepts (Soldaini and Goharian, 2016). In addition to the extraction, this library also identifies the class of the concept and the relations between the concepts. Therefore, the next stage, classification of concepts, is not necessary for the UMLS-based method.

For OpenBioLaySumm, we exploit the keyword extraction called Yake! proposed by Campos et al. (2020), which uses a statistical technique to identify the main keywords in a given text. We apply Yake! for each document with the number of terms (top) set to 10, n-grams of up to 2 words(n = 2), which means only unigrams and bigrams, and a variable deduplication limit (dedupLim) set to 0.5. Deduplication merges similar keywords to avoid redundancy, and a value of 0.5 merges similar keywords with 50% similarity according to the Levenshtein distance. Examples of biomedical keywords extracted from the test set are: “neck region”, “cell membrane”, “cell growth”, and “endocytic events”. Some non-biomedical keywords are also extracted, but this is not relevant to the generation of the lay summary; that is why we discard these concepts using an automatic classification step, which is described in Section 4.2. Examples of non-biomedical words in the test set are: “figure supplement”, “Supplementary file”, “Time constant”, and “label set”.

To assess whether Yake! is able to extract terms that express biomedical concepts, one of the authors of this paper evaluated the extracted keywords. For this evaluation, we randomly selected 50 papers from the test set, and the 10 keywords extracted for each paper were evaluated as biomedical and non-biomedical. The total of evaluated keywords was 500 terms from the test set. In this set, 151 keywords were not considered biomedical, while 349 keywords were considered biomedical. Thus, 70% of the keywords extracted by Yake! were biomedical.

4.2. Classification of the Concepts

As described in Section 3.2, UMLS-based methods do not require domain classification since the UMLS Metathesaurus is curated by specialists. However, OpenBioLaySumm requires filtering since YAKE! extracts keywords from all domains without biomedical discrimination. This section presents the experimental validation of our zero-shot LLM classification approach.

To assess the capabilities of the model to classify a given concept in a zero-shot manner into biomedical or non-biomedical, one of the authors evaluated the output of the model. Similarly to the keywords evaluation, 10 keywords and their respective DBPedia entries of a sample of 50 documents were analyzed, for a total of 500 DBPedia entries. Cohen’s Kappa agreement between the LLM classifier and a human evaluator was computed per document, and the average of the values was 0.3324. Half of the agreements were poor (22, $\kappa < 0.20$), a few presented a slight (4, $0.20 \leq \kappa < 0.40$) , and the rest were moderate to almost perfect (24, $\kappa \geq 0.40$). In a more detailed analysis, it was possible to observe that the LLM classified a lot of the concepts as biomedical, probably because it is a general domain LLM, and it needed examples to understand

[6] https://github.com/TGoldsack1/Enhancing_Biomedical_Lay_Summarisation_with_External_Knowledge_Graphs

[7] https://wayback.archive-it.org/7867/20241213191952/https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html

what a biomedical concept is. However, since the majority of the concepts extracted are biomedical (70%), the influence of the mistakes of the classification is probably small in the summarization result. We leave as future work the application of a few-shot approach at this stage of the proposal.

4.3. Lay Summarization

For the OpenBioLaySumm, it is necessary to relate each concept to the others. In this stage, we use SciBERT and compute the cosine similarity between the concepts. Concept pairs with cosine similarity above 0.7 are connected by edges in the graph. The representation of the graph is learned by the Graph Attention Network (GAT) and concatenated with the LED text embedding of the document, and fed to a Decoder, which is then responsible for producing the summary. We trained all four models (BART baseline, BART+YAKE!, UMLS reproduction, and OpenBioLaySumm) using a learning rate 2e-6, and a batch size 2-4. All models used identical decoding parameters: beam search with 4 beams, minimum length 100 tokens, length penalty 2.0, and no-repeat-ngram size 3. However, due to limitations in our computational resources, the UMLS (Goldsack) strategy was trained only with 5 epochs, while BART, BART+YAKE!, and OpenBioLaySumm (YAKE!+DBPedia) were run with 20 epochs since they all required less GPU memory than the UMLS approach. All models were trained under an identical four-hour GPU budget on an NVIDIA A100 (40GB). Due to the higher memory footprint and preprocessing overhead of the UMLS-based pipeline, it completed five training epochs within this time constraint, while the other models completed approximately twenty epochs. Under the four-hour fixed GPU budget, the UMLS-based model validation performance peaked at the first epoch and did not show consistent improvement in subsequent epochs, suggesting early convergence under the given setup. Early stopping was applied uniformly across models.

Table 2 depicts the results of the four approaches. Observe that among the open-based approaches (BART, BART+YAKE!, and OpenBioLaySumm), the best results in three metrics are from OpenBioLaySumm. Another interesting observation is that BART+YAKE! improved over the baseline, demonstrating that keyword augmentation alone provides value even without graph construction. The ROUGE-Lsum for the BART+YAKE! approach is also the second-best result, which shows a better summary-level sequential coherence. This is expected: prepending keywords maintains a simple linear structure, while our GAT-encoded graph introduces richer but more complex concept relationships that may occasionally reorder information. However, the gain in the three other metrics achieved by OpenBioLaySumm is more significant compared to the gain of ROUGE-LSum.

The lay summarization task requires simplification of the text. Hence, readability measurements are a usual way to inspect the level of simplification of the results. Figure 2 shows a comparison between the gold summaries (reference) and three of the approaches tested. The Flesch score is a readability metric based on which the higher the value, the simpler the text. All the methods were able to simplify the text beyond the gold summaries. The UMLS (Goldsack) approach presented the best simplification results. Regarding the Coleman-Liau and Dale-Chall scores, the smaller the value, the simpler the text. These metrics indicate the same; the UMLS (Goldsack) presented the lowest values, while OpenBioLaySum (YAKE!+DBPedia) presented the second lowest ones. The remaining measures, average sentence length, average word length, and type-token ratio, shows similar results. One hypothesis for OpenBioLaySumm (YAKE!+DBPedia) presents this type of result, which is the type of text that DBPedia entries have compared to the text of the UMLS definitions. As a general domain knowledge base, DBPedia probably adds a more diversified vocabulary to the final summary compared to a specialist-curated knowledge base, such as UMLS.

In this section, we presented quantitative measures of the quality of the summaries. However, we acknowledge ROUGE's limitations: it measures lexical overlap rather than semantic fidelity, factual correctness, or accessibility for lay readers. Qualitative analysis is essential in those cases, although it would require expensive expert work. To mitigate the quantitative metrics limitations, and give a more qualitative view of the results we provide in the section, an analysis through two illustrative examples: a success case where OpenBioLaySumm (YAKE!+DBpedia) outperforms the UMLS (Goldsack) approach, and a failure case where it achieves significantly lower BLEU scores.

5. Discussion

We will divide our discussion into two parts. The first will discuss the quantitative results we presented in the last section, and the second part will present two examples and qualitatively discuss the UMLS-based approach (Goldsack) and the OpenBioLaySumm (YAKE! + DBPedia). The goal is to deepen the understanding of the enrichment role in the lay summarization task.

Quantitative results. We evaluated the three main stages of our pipeline: the keyword extraction, the classification of concepts in the biomedical domain, and the final lay summarization. The first evaluation helps to answer our first research question:

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
BART Baseline	45.84	11.85	20.92	42.55
BART+YAKE!	47.53	12.97	21.63	44.04
OpenBioLaySumm (YAKE!+DBpedia)	58.44	20.65	43.45	43.47
UMLS (Goldsack et al.)†	**60.26**	**22.55**	**45.45**	**45.44**

Table 2: ROUGE scores on the test split. †Our reproduction using QuickUMLS and 5 training epochs. Goldsack et al. (2023) reported ROUGE-1: 48.58 and ROUGE-L: 45.71 on PLOS+eLife combined dataset using MetaMap and full training. Direct comparison is limited due to dataset and implementation differences.

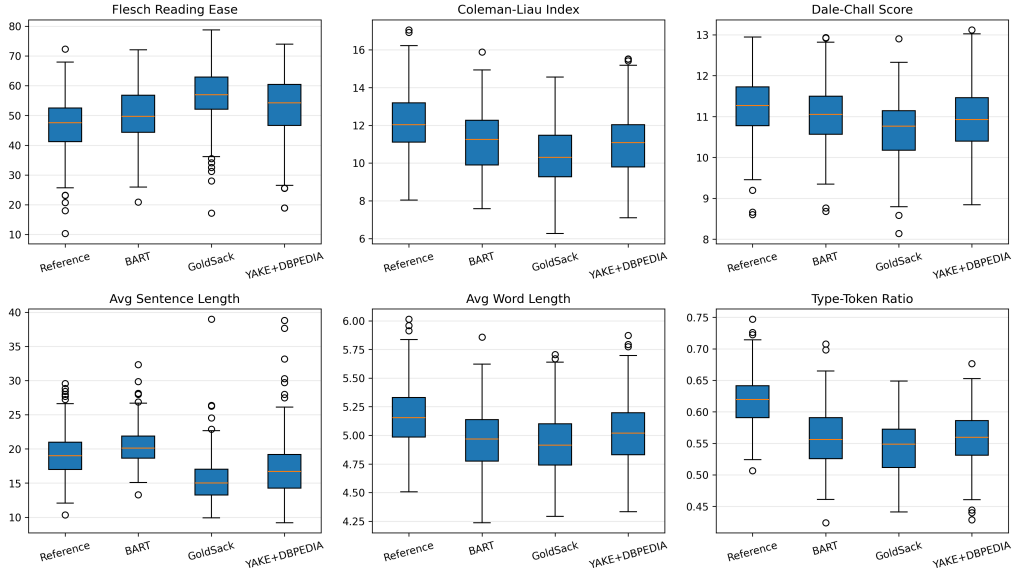

Figure 2: Readability measurements for Gold Summary, BART (baseline), OpenBioLaySumm(YAKE!+DBPedia), and UMLS (Goldsack)

RQ1) To what extent can a curated subset of automatically extracted keywords from biomedical articles represent central biomedical concepts? The results of the assessment indicate that Yake! effectively retrieves biomedical meaning from biomedical papers. In an evaluation of 50 random documents from our test set, we assessed that approximately 70% of the keywords extracted from the biomedical papers are domain-related. The minimum of biomedical terms in a document was 2, and the maximum was 10. Despite the minimum of 2 keywords, the average was 6.96 ± 1.79 and the mode was 8. These numbers suggest that Yake! can extract medical terms from the papers. This is an intuitive result since scientific papers usually follow a writing template, and the authors usually convey only the essential information to the readers, avoiding excess that can confuse.

The LLM classification results presented, most of the time, low agreement scores with the human evaluation. In this case, we also evaluated 50 documents, and 26 presented Cohen's kappa lower than 0.40. An inspection of the LLM results revealed that the LLM alone is not capable of being aware of what a biomedical concept is. We use a zero-shot approach, which can be the reason for such a low performance. The LLM classifier identified 444 of the retrieved concepts as biomedical, while the human evaluator identified only 349 of the 500 concepts. One possible solution for improvement is to develop a domain classifier, which could take advantage of the biomedical words to sieve the relevant entries. An only-encoder model could be fine-tuned to distinguish biomedical and non-biomedical terms based on their collected DBPedia entries. A domain-specific classifier could sieve the concepts more effectively, reducing the noise of out-of-domain concepts in the GAT. This could improve the results for an approach based on the YAKE! and DBPedia.

Finally, the results of the summarization show that the OpenBioLaySumm achieved 96.97%, 91.57%, 95.59%, and 95.96% of the performance in the respective ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores of the UMLS-based approach (Goldsack). In a close look at the readability metrics, we can observe that the average sentence length of the OpenBioLaySumm is greater

than that of the UMLS one (Goldsack). DBPedia uses an encyclopedic style in the writing, and compared to a more specialized thesaurus like UMLS, it probably contains more lengthy sentences and less objective explanations. The results, therefore, are a reflection of the DBPedia writing style. The same influence of the DBPedia style is possible to observe in the Type-Token Ratio (TTR) boxplot. This metric measures the number of unique tokens divided by the total number of tokens. The boxplot shows that the average of TTR is slightly higher than the BART approach, indicating that DBPedia introduced a more diverse vocabulary. The noise from the non-biomedical entries of the DBPedia would probably even worsen this effect. We also computed the BLEU scores for the summaries. Figure 3 presents the overall BLEU scores, in which the UMLS (Goldsack) presented a slightly superior advantage. The analysis of these results answers our second research question, **RQ2)**Can a lay summarization technique leveraging open-source knowledge bases compete with approaches using licensed ontologies? The answer to this question is positive since our preliminary results with an open-resource approach reached between 91.57% and 96.97% in the ROUGE metrics.

Qualitative results. Regarding the extracted keywords, it is possible to observe that some of the identified terms that are not biomedical are very repetitive. For instance, the word “figure” appears in 72 of the extracted keywords, and “supplement” appears in 43. Differently, the most frequent biomedical keywords present much lower frequency values. For instance, “cells“, “dna” and “cell” appear in 18, 16, and 15 keywords. This is probably because the articles in this particular dataset cover a broad number of topics within the biomedical domain, and the discovered keywords are meaningful for summarization.

The LLM classification, despite its poor results, was able to filter a very common non-biomedical keyword, “figure supplement”. However, the LLM made simple mistakes like classifying the concept *Politician*[8] as biomedical. This type of error caused a lot of noise in the final result.

For the summarization results, we will look at two examples. The first example is depicted in Table 3, the OpenBioLaySumm shows better performance than the UMLS (Goldsack) approach. Observe that in this example, both have low BLEU values, but OpenBioLaySumm presented a better phrasing, and the concepts in this abstract (“electrical impulse”, “vesicles”, “membrane”, etc) are very different, which favors the DBPedia approach.

We examined failure cases to understand when graph-based augmentation degrades performance.

[8] http://dbpedia.org/resource/Politician

Success Case: Neuroscience Abstract (Synaptic Transmission) **BLEU**_UMLS = 8.804; **BLEU**_Ours = 15.906 (+81%)
UMLS: Neurons communicate with each another ... a electrical signal is between a nerve ... vesicles are travel to each on the membrane...
OpenBioLaySumm: Neurons communicate with one another ... a electrical impulse reaches from a neuron ... vesicles are leave to receptors on the surface...
Green = Better phrasing; Red = Errors in UMLS

Table 3: OpenBioLaySumm achieves 81% higher BLEU by better capturing key biological relationships (receptors, synapses, vesicles) through DBpedia’s general-domain knowledge graph.

Table 4 shows an example where OpenBioLaySumm achieves only 34.34% of UMLS’s BLEU score (7.016 vs. 20.432) on a text about Notch signaling. The failure exhibits two issues. First, both methods struggle to correctly tokenize and preserve protein family names (Delta, Serrate, Jagged), producing fragments like "Seragged", "Lagagged", and repeated single-letter tokens "D D D". Second, OpenBioLaySumm’s graph structure amplifies these errors: the high semantic similarity between protein names (Notch, DLL1, DLL4, Jagged) creates densely connected subgraphs. This causes the GAT attention mechanism to cycle through similar fragmented tokens, leading to severe repetition patterns like "of of both of some of" that further degrade BLEU scores. Despite this type of case, our method’s overall strong performance (ROUGE-1: 58.44, 96.97% of UMLS) indicates this is a bounded problem affecting a small subset of highly technical abstracts.

6. Conclusion

In this work, we propose an open-resource approach to lay summarization of biomedical papers. We showed that the performance of the proposed approach is competitive, and information extraction can be achieved in different ways other than exclusively named entities, like YAKE!. Finally, we showed examples of the test set that demonstrate the strengths and weaknesses of our method.

However, this work has several limitations. One of them is that we only tested one LLM for the classification, and nowadays there is a diverse set of models that could be tested. Another one is that, due to constraints in computational resources, we did not

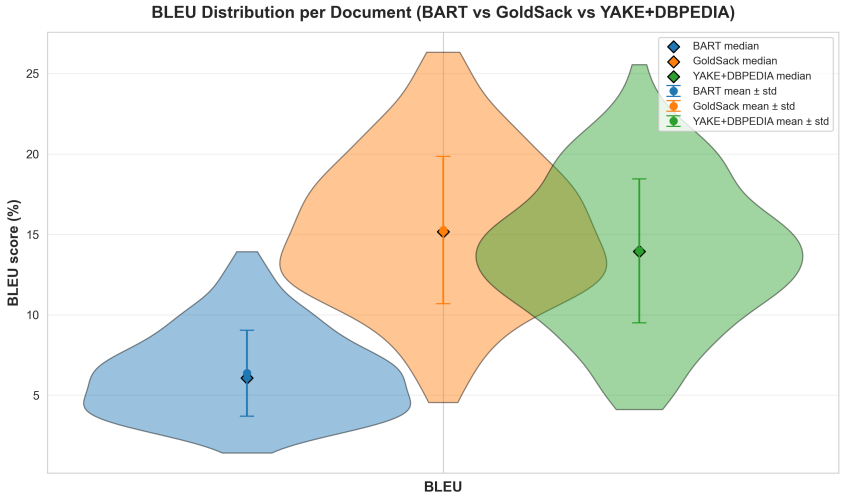

Figure 3: The BLEU scores for the UMLS (Goldsack) and the OpenBioLaySumm (YAKE!+DBPedia)

UMLS (Goldsack) BLEU = 20.432	**OpenBioLaySumm** BLEU = 7.016 (-65.66%)
Protein name fragmentation:	
receptors lig Seragged ligands bind...	to a on lig- Lagagged ligands bind...
Mammals have Not of Not ligands, D of of Jagged...	Theals have Not of lig ligands, D that of Jagged...
DLL1 and DLL4 are activate in in cells cells , of, D D some tissues D D D ...	DLL1 and DLL4 are bind on on the cells , of of both of some of have D type...
Structural errors:	
Onealing are between and forth cells on on and off ...	Oneals from from and forth the as and on on and off ...

Table 4: Failure case showing severe degradation with protein-heavy terminology. Both methods fragment protein names (Serrate/Jagged → "Seragged"/"Lagagged", Delta → "D"), but OpenBioLaySumm shows more repetition ("of of both of some of", 66% BLEU drop). The dense graph of similar protein concepts amplifies tokenization errors through attention cycles.

test several different parameters for the model, including YAKE! parameters. The prompt is also very simple, and a few-shot approach could improve the results. Also, the evaluation was conducted by only one person, who is also one of the authors of the paper. An additional human evaluation would strengthen the conclusions of the proposed approach. Our approach struggles with abstracts containing many similar, interconnected technical entities (e.g., protein families with multiple related members). The graph structure, while beneficial for most documents, can amplify confusion in such cases by creating densely connected concept clusters that lead to repetitive generation. Also, due to computational constraints, we reduced SciBERT embeddings from 768 to 50 dimensions via PCA. While this enabled processing within our GPU budget, we did not systematically evaluate the impact of this reduction on final performance. Future work should explore the trade-off between dimensionality, computational cost, and summarization quality. Finally, the classification could benefit from a standard approach with a BERT-based classifier trained on general data combined with biomedical data. This could be an efficient and effective approach compared to an in-context learning approach based on LLMs.

Future work should explore adaptive graph pruning that reduces edge density when concepts exceed a similarity threshold in localized graph regions. Also, we plan an ablation study with the used parameters and do a comparison between our GAT-based approach against simpler aggregation methods (e.g., mean-pooled concept embeddings) to isolate the contribution of graph-based reasoning. This would clarify whether performance gains stem from DBpedia's knowledge content or from the graph structure itself. Finally, we plan to conduct a more extensive evaluation of the results with biomedical experts.

7. Acknowledgements

We thank Tomas Goldsack for making the code from Goldsack et al. (2023) publicly available. Our implementation builds upon their doc-enhance architecture, with modifications for QuickUMLS integration and the YAKE!+DBpedia pipeline. The authors acknowledge the computational resources provided by the Deucalion supercomputer at the Minho Advanced Computing Centre (MACC), University of Minho, Guimarães, Portugal, through projects 2025.00013.AIvLAB.DEUCALION and 2025.00017.AIvLAB.DEUCALION. Deucalion is funded by the EuroHPC Joint Undertaking and the Portuguese Foundation for Science and Technology (FCT). Also, this publication is based upon work from COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

8. Bibliographical References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Souramita Bhowmik, Anupam Jamatia, Dwijen Rudrapal, and Kunal Chakma. 2025. Biomedical lay summarization: Research progress and challenges. In *2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 464–471. IEEE.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.*, 66(11):2215–2222.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. 2022. Automatic Text Summarization of Biomedical Text Data: A Systematic Review. *Information*, 13(8):393. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Nicholas Fraser, Liam Brierley, Gautam Dey, Jessica K. Polka, Máté Pálfy, Federico Nanni, and Jonathon Alexis Coates. 2021. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, 19(4):1–28. Publisher: Public Library of Science.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. Enhancing biomedical lay summarisation with external knowledge graphs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Carly M Goldstein and Rebecca A Krukowski. 2023. The importance of lay summaries for improving science communication. *Annals of Behavioral Medicine*, 57(7):509–510.

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4969–4983.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Readability Controllable Biomedical

Document Summarization. ArXiv:2210.04705 [cs].

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Chen Tang, Shun Wang, Tomas Goldsack, and Chenghua Lin. 2023. Improving biomedical abstractive summarisation with knowledge aggregation from citation papers. *arXiv preprint arXiv:2310.15684*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker, and Javed Mostafa. 2021. A systematic review of automatic text summarization for biomedical literature and EHRs. *Journal of the American Medical Informatics Association : JAMIA*, 28(10):2287–2297.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William K Cheung, et al. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 365–377.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models. ArXiv:2304.08763 [cs].

Qianqian Xie, Prayag Tiwari, and Sophia Ananiadou. 2024. Knowledge-Enhanced Graph Topic Transformer for Explainable Biomedical Text Summarization. *IEEE Journal of Biomedical and Health Informatics*, 28(4):1836–1847.