Lay Summarization of Biomedical Text Using Keywords and Knowledge Bases

João Pedro Brito Veloso
Master in Network and Information Systems Engineering
Department of Computer Science
Faculty of Sciences of the University of Porto
2026

Lay Summarization of Biomedical Text Using Keywords and Knowledge Bases

João Pedro Brito Veloso

Dissertation carried out as part of the Master in Network and Information a Systems Engineering

Department of Computer Science

2026

Supervisor

Evelin Carvalho Freire de Amorim, Assistant Researcher, INESC TEC

Acknowledgements

I would like to express my gratitude to all of those who have contributed to the conclusion of this thesis.

Firstly, I'd like to thank my family for their love, encouragement, and belief in my abilities. Their support, both emotional and financial, has been a fundamental source of support throughout my journey.

Secondly, I'd like to thank Professor Evelin Amorim for her guidance, comprehension, invaluable insights, and knowledge throughout this research.

Thirdly, I would like to thank the University of Porto; the INESTEC institution; the computational resources provided by the Deucalion supercomputer at the Minho Advanced Computing Centre (MACC); and the University of Minho, Guimarães, Portugal, through the projects 2025.00013.AIvLAB.DEUCALION and 2025.00017.AIvLAB.DEUCALION.

Finally, I want to thank my friends for always being there for me.

Thank you,

João Veloso

Abstract

Automatic summarization seeks to generate concise texts that preserve the key information of the source. In the biomedical domain, lay summarization is particularly important because it helps non-specialist audiences understand complex scientific findings. Although knowledge-augmented approaches improve performance, many biomedical systems depend on closed and curated resources, such as the Unified Medical Language System (UMLS), which are costly to maintain and difficult to scale in low-resource settings. This thesis proposes an open-resource biomedical lay summarization based on keyword extraction and DBpedia that follows a four-step path. First, keyword extraction from biomedical abstracts using YAKE!, a well-established unsupervised method. Second, retrieval of relevant concept entries using DBpedia queries containing the keywords extracted previously. Third, graph construction for each document using cosine similarity between relevant concepts. And finally, integration of information obtained from the graphs with the source abstract to train an abstractive summarization model. From the experiments, our method shows competitive performance against the UMLS-based systems, with results close to the closed-resource baseline (ROUGE-1: 58.44 vs. 60.26; ROUGE-L: 43.45 vs. 45.45). These findings reveal that open-resource approaches can provide viable alternatives to license knowledge bases while maintaining accessibility for resource-constrained organizations.

Keywords: Biomedical lay summarization, Automatic summarization, Language Models, DBpedia, Open-Resources

Resumo

Este trabalho tem como objetivo gerar automaticamente resumos leigos mais curtos e acessíveis de textos científicos, reduzindo a complexidade e a terminologia especializada. A proposta desta tese é baseada num modelo de transformer encoder-decoder enriquecido com conhecimento estrutural e externo. A informação estrutural é obtida a partir de grafos de conceitos construídos com palavras-chave extraídas através de um algoritmo de extração de keywords. A relevância biomédica destes conceitos é determinada por um modelo de classificação e representada por embeddings, de forma a captarem relações semânticas. O conhecimento adicional é encontrado na base de dados DBpedia, permitindo incorporar contexto e factos externos durante a geração de resumos. A proposta distingue-se de outras porque utilizamos recursos de acesso livre e representamos o conhecimento de forma mais compacta, já que o foco está apenas nos conceitos extraídos do texto.

Os resultados foram obtidos no conjunto de dados eLife, que contém artigos biomédicos acompanhados de resumos leigos produzidos por especialistas. No conjunto de teste, a abordagem baseada em DBpedia obteve um macro ROUGE-L de 43.45, com desempenho competitivo face à alternativa baseada em UMLS (45.45) e superior à baseline BART (ROUGE-L 21.75). Embora não supere o método UMLS, a nossa proposta usa um recurso aberto (DBpedia) comparado ao UMLS, que é um recurso fechado, e uma representação mais compacta, por apenas considerar as palavras-chave como relevantes.

Esta tese contribui para o campo da sumarização com o uso de palavras-chave e bases de conhecimento no domínio biomédico, ao propor uma metodologia que implementa os modelos de linguagem de grande escala com extração de palavras-chave e bases de conhecimento abertas, permitindo a incorporação de conhecimento externo de forma estruturada e compacta. Recursos abertos, como a DBpedia, junto com uma representação focada apenas em conceitos relevantes, resultam numa abordagem mais eficiente e acessível, mantendo desempenho competitivo na geração de resumos biomédicos para o público geral.

Palavras-chave: Sumarização Automática, modelos de linguagem, DBpedia, recursos abertos

Table of Contents

Acknowledgements i
Abstract ii
Resumo iii
List of Tables vi
List of Figures vii
Acronyms viii

1. Introduction 1
   1.1. Context 1
   1.2. Motivation 1
   1.3. The Main Goal 2
   1.4. Methodology 3
   1.5. Research Questions 3
   1.6. Organization 3
2. Background 5
   2.1. Large Language Models 5
      2.1.1. Type of LLMs 5
      2.1.2. How do LLMs work 6
   2.2. Summarization 8
      2.2.1. Types of automatic summarization 8
   2.3. Graph Neural Networks 9
      2.3.1. How GNNs Work 9
   2.4. Knowledge bases 10
      2.4.1. Types of KBs 10
      2.4.2. DBpedia and UMLS 11
   2.5. Keyword Extraction 12
3. Related Work 15
   3.1. Automatic summarization in Biomedical/Biological NLP 15
   3.2. Medical Knowledge Integration in NLP 16
   3.3. Large Language Models for Lay Summarization 17
   3.4. Benchmarks 19
   3.5. Recurring Challenges, Limitations and Research Gaps 19
4. Methodology 21
   4.1. Golsack et al. Method 21
   4.2. YAKE! Method with DBpedia 22

4.3. BART with YAKE! 24
4.4. BART 25
4.5. Classification Models 25
5. Experiments and Results 28
5.1. Datasets 28
5.1.1. eLife 28
5.1.2. News 29
5.1.3. eLife + AG News + 20 News Groups 29
5.1.4. Guidelines 29
5.2. Yake 30
5.3. LLM Classification 31
5.4. Summarization Results 33
5.4.1. Experiments Setup 33
5.4.2. Summarization Results 34
5.4.2.1. GoldSack method 34
5.4.2.2. YAKE+DBPEDIA 35
5.4.2.3. Comparison 36
5.5. Error Analysis 36
5.5.1. Readability 41
6. Conclusion 42

Bibliography 50

List of Tables

5.1. Word count statistics across dataset splits. 28
5.2. Word count statistics for the AG News and 20 Newsgroups datasets. 29
5.3. Word count statistics for the eLife + AG News + 20 News Groups dataset in section ”text” 29
5.4. Precision of LLM classification in 50 random articles of each split 31
5.5. Statistics for each file 33
5.6. Statistics from GoldSack model across all runs on the validation split 35
5.7. Statistics from our model across all runs on the validation split 35
5.8. ROUGE scores on the test split 36
5.9. Comparative Analysis: Notch Signaling Example (GoldSack Success) 38
5.10 Comparative Analysis: Neuron Synapses Example (YAKE Success) 39

1. Number of keywords that were considered relevant to biomedicine/biology by human evaluation that did not have DBpedia entries related to biomedicine/biology 44
2. Confusion matrix components (TP, FP, FN, and TN) for biomedical keyword classification across dataset splits 44
3. Performance metrics of the LLM on biomedical keyword classification 44
4. Validation results using XGBoost classifier using TF-IDF for downsampled train/test datasets (eLife articles + AG News + 20 Newsgroups) 44
5. Validation results using XGBoost classifier using TF-IDF for non-downsampled train/test datasets (eLife articles + AG News + 20 Newsgroups) 44
6. Comparative Analysis: Example 102 (GoldSack Success) 46
7. Comparative Analysis: Example 238 (YAKE Success) 47
8. Comparative Analysis: Example 26 (GoldSack $\approx$YAKE) 48
9. Comparative Analysis: Example 54 (Both Low) 49

List of Figures

4.1. The building of the graph of concepts based on keywords extracted by YAKE!.....24

5.1. Strength of agreement 32
5.2. GoldSack - Best Epoch Results (by ROUGE-L) on the validation split.....35
5.3. YAKE+DBPEDIA - Best Epoch Results (by ROUGE-L) on the validation split.....36
5.4. BLEU Distributions 37
5.5. Readability statistics.....41

1. GoldSack - All RUNS ROUGEL VALUES 45
2. YAKE+DBPEDIA - All RUNS ROUGEL VALUES 45

Acronyms

AI	Artificial Intelligence
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CUI	Concept Unique Identifier
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GPT	Generative Pre-trained Transformer
KB	Knowledge Base
LED	Longformer Encoder-Decoder
LLM	Large Language Model
MACC	Minho Advanced Computing Centre
NER	Named Entity Recognition
NLP	Natural Language Processing
RLHF	Reinforcement Learning from Human Feedback
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
T5	Text-to-Text Transfer Transformer
TTR	Type-Token Ratio
UMLS	Unified Medical Language System
UTS	UMLS Terminology Services
YAKE!	Yet Another Keyword Extractor

1. Introduction

In this introductory chapter, we present our motivation, the research problem, objectives, methodology, and structure of the thesis, as well as the expected contributions to the field of lay summarization.

1.1. Context

Scientific communication has accelerated thanks to the digitization of journals, preprint servers, and online news [1, 2]. This increase led to a much higher volume and velocity of biomedical findings reaching public channels [3].

Within natural language processing (NLP), summarization has surged as a key strategy to condense complex documents while preserving salient content and is commonly categorized into extractive and abstractive approaches.

The role of summarization is consistently highlighted by domain surveys in the address of information overload in biomedical literature and clinical texts, and it traces the transition from pre-trained encoders to large language models (LLMs) and retrieval-augmented pipelines tailored to domain constraints (e.g., terminology, factuality, and safety) [4–6].

In the specific area of biomedical lay summarization, recent benchmark efforts have improved readability and factual consistency for non-expert audiences by defining tasks, datasets, and evaluation protocols. This can be seen in the series BioLaySumm, especially between 2023 and 2024[7]. BioLaySumm shares tasks that contextualize the problem and reports system trends, including LLLM-based understanding. Representative 2024 systems include (i) LLM adaptation for lay summarization (HGP-NLP); (ii) readability-aware retrieval-augmented generation (Team YXZ); and (iii) extract-then-summarize pipelines enriched with Wikipedia knowledge (UIUC BioNLP).

These studies show a methodology that blends controllable generation, external knowledge sources, and task-specific constraints that aim to make it more accessible to lay readers [8, 9].

1.2. Motivation

Biomedical knowledge must be reliable, regardless of the time or place. Especially during public health crises, like the ”COVID-19” pandemic, where decisions must be made under uncertainty. The World Health Organization has emphasized the challenge of the

modern ”infodemic,” where misinformation and disinformation undermine evidence-based guidance and erode trust. In these contexts, high-quality lay summaries help the public understand better expert knowledge by being readable, accurate, and grounded in authoritative sources [10, 11].

The production of lay summaries in scale, especially manual production, is costly and slow. The automation of this pipeline is done through NLP systems that integrate domain knowledge bases and readability control. This offers the audience proper context without sacrificing fidelity. The recent BioLaySum systems showed this by combining LLMs with retrieval and explicit readability objectives. This improved accessibility while maintaining factual grounding. Advancing these lines—especially with biomedical ontologies/knowledge graphs and keyword-guided control—directly targets the dual need for clarity and credibility in an era of information disorder [8, 9].

In the biomedical domain, a large part of existing summarization and extraction systems relies heavily on closed or semi-closed ontologies, such as the Unified Medical Language System (UMLS). Although these resources present higher adoption and high quality, their restricted access, licensing constraints, and limited transparency result in challenges for the reproducibility, auditability, and long-term sustainability of NLP systems. And compared to the use of open and auditable knowledge bases, they present lower transparency, higher difficulty of integration across systems, and less accessibility for both researchers and practitioners. Therefore, exploring open knowledge sources as an alternative for biomedical text summarization is a necessary step to build systems that are more transparent, reproducible, and trustworthy for public-facing scientific communication.

1.3. The Main Goal

The main goal of this thesis is to develop and evaluate methods for generating lay summaries of biomedical texts by using a combination of keyword extraction and open external knowledge bases. The aim is to make non-expert audiences have more access to complex biomedical literature by producing concise, accurate, and easy-to-understand summaries. This achievement was done through the integration of advanced natural language processing techniques, such as keyword extraction and graph-based representations, with large pre-trained language models.

1.4. Methodology

In this work, we propose a summarization methodology for biomedical articles that leverages an encoder-decoder transformer architecture, specifically the Longformer Encoder-Decoder (LED) model, enhanced with graph-based knowledge integration. The system processes full-length biomedical texts, encodes them with the LED encoder, and incorporates structured biomedical knowledge from DBpedia through a custom Graph Attention Network (GAT) module. During training, the model learns to generate concise and coherent summaries by attending to both the textual content and the relevant graph-based features. The approach is evaluated using ROUGE metrics in a dedicated validation set, and the best-performing model is selected based on ROUGE-L scores. This methodology enables the generation of high-quality abstractive summaries that benefit from both deep contextual understanding and external biomedical knowledge.

1.5. Research Questions

In summary, this study aims to answer the following research questions:

1. Can keywords obtained from keyword extraction convey relevant biomedical concepts from scientific papers?

2. Is it possible to discard non-biomedical keyword concepts automatically using classifiers?

3. Does a DBpedia-based keyword/graph pipeline perform competitively with a UMLS-based pipeline for biomedical summarization?

The contributions of this thesis can be summarized as follows:

1. A lay summarization proposal that employs keyword extraction and open external knowledge bases;

2. A comparative analysis of the proposal using different representations of node characteristics, including the analysis of trade-offs between semantic-rich characteristics.

1.6. Organization

The remainder of this document consists of X sections that are structured as follows.

- **Chapter** 2 presents some theoretical background necessary to develop the investigation.

- **Chapter** 3 reviews some existing literature on automatic summarization, integration of medical knowledge in NLP, and Large Language Models for lay summarization. It discusses the approaches and techniques commonly used in this kind of problem.
- **Chapter** 4 describes the UMLS methodology used in [12] and the methodology proposed in this work.
- **Chapter** 5 presents the experiments and results obtained by the proposed methodology presented in the previous chapter. The results are interpreted and discussed with respect to the research objectives.
- **Chapter** 6 summarizes the main findings of the study and addresses their implications for the field of lay summaries of texts, especially biomedical texts. It also mentions the limitations of our research and areas for future improvement.
- **Chapters** Appendix A and Appendix B contains more detailed information about the results of our experiments.

2. Background

The main concepts related to this thesis are Large Language Models (LLMs), summarization, knowledge bases and keyword extraction. Next, we will briefly introduce them.

2.1. Large Language Models

Large Language Models (LLMs) are advanced artificial intelligence (AI) algorithms that model human natural language. Using deep learning techniques, these models learn how to represent words using large datasets. Using LLMs, several language-related tasks can be solved, from content generation to translation and analysis [13].

2.1.1 Type of LLMs

LLMs can be divided into three main types according to their underlying architecture: encoder-only, decoder-only, and encoder-decoder models. Each type has unique purposes and is optimized for specific tasks.

- *Encoder-only Models*

  Encoder-only models were designed to understand the input text in depth by extracting relevant information from the input text. With this extraction, an output is created that contains a continuous representation (embedding) of the input data text.

  As a result, encoder-only models excel at tasks that require deep language understanding, such as: Text classification, Named entity recognition (NER), Sentence similarity, and Question answering (with short answers). Good examples of these models are the BERT [14] and RoBERTa [15] models.

- *Decoder-only Models*

  Decoder-only models were designed to generate text. These models are masked, leading to these models maintaining an auto-regressive property. This ensures that the prediction of the next word is only based on previous words until that point and also ensures that the output tokens generated by the model are only generated one at a time. This property makes it easier for these models to generate text but not to understand text since they are less effective at bidirectional understanding compared to encoder-only models.

As a result, decoder-only models are highly effective in the following areas: text generation, code generation, conversational agent, and auto-completion.

A great example of these models is the GPT family [16]. A series of decoder-only models that are pre-trained on large-scale unsupervised text data and fine-tuned for specific tasks such as text classification, sentiment analysis, question-answering, and summarization.

- *Encoder-decoder Models*

  Encoder-decoder models are models that combine the previous two models. These models contain encoder-decoder transformers, where the encoder part is responsible for the creation of the embeddings that the decoder uses to generate an output sequence.

  These models are well-suited for: machine translation, summarization, question answering (with full-sentence answers), and text-to-text transformations.

  Good examples of these models are the BART [17] and T5 [18] models.

In this work, we used an encoder-decoder (sequence-to-sequence) architecture, specifically the Longformer Encoder-Decoder (LED) model, which leverages both an encoder to process the input text and a decoder to generate the output summary.

2.1.2 How do LLMs work

LLMs generally have the following three core areas of practical concern in the context of LLMs:

- **Development**

  As the name suggests, Development focuses on how LLMs are initially trained. Development begins by exposing the models to a huge amount of raw data. This stage corresponds to a large-scale *Pre-training*, where the models learn to predict missing or next tokens in a sentence, allowing them to understand language patterns and semantics. Then, these models may undergo further training on domain-specific data (e.g., legal, medical, financial). This is called *Domain-adaptive pre-training* (DAPT) that, even though it still uses unsupervised learning, it focuses on text from a particular domain. This allows the model to adapt to specialized vocabulary, stylistic conventions, and domain-specific knowledge in related contexts. Importantly, DAPT is not yet fine-tuning.

Together, these stages form the model's initial development, enabling it to acquire both general linguistic competence and domain-sensitive knowledge before the model is later refined for specific tasks through fine-tuning.

- **Adaptation**

  Adaptation is the area where pre-trained models are modified so they can perform better in specific tasks or align with human expectations. This includes:

  - *Instruction tuning* — Models are refined to follow better human instructions through the use of instruction-response datasets. We use this type of model in our research to classify text into biomedical and non-biomedical classes. We used the DeepSeep-R1, a decoder-only model [19], in our experiments. The reason for this choice is that it ensures consistency with broader focus on generative LLMs and allows us to explore how decoder-based architectures can be adapted to classifications tasks, even though they are not the most computationally efficient option.

  - *Alignment tuning* — Model outputs are aligned with human values, in most cases, using reinforcement learning from human feedback (RLHF).

  - *Full parameter fine-tuning* — All model parameters are updated to maximize performance on specialized datasets. We also experiment with and fine-tune a model to classify text into biomedical and non-biomedical classes. The type of LLM we employed was an encoder model of the family BERT (PubMed-BERT [20]).

- **Utilization**

  The utilization governs the area of how to interact with LLM so that we can take advantage of its advantages and capabilities. The two main strategies are as follows:

  - *In-context learning* - By providing examples directly in the prompt, LLMs are able to perform tasks without having to modify the model's parameters. The instructed tuned models are usually employed using this technique, as is our research.

  - *Chain-of-though learning* - LLMs generate responses between steps to explain the reasoning of the answer for complex, multi-step problems.

The main core area applied in this thesis was the ”Development” core area, since a model was built from the ground up to learn from a dataset.

2.2. Summarization

Summarization is the act of expressing the most important facts or ideas about something or someone in a short and clear form, or a text in which these facts or ideas are expressed [21]. Automatic summarization is a computational process that transforms large amounts of data into a concise summary, capturing the most important and relevant information from the original content. Automatic summarization has become incrementally more important in today’s information-rich environment, helping people efficiently consume and understand vast amounts of data.

2.2.1 Types of automatic summarization

Automatic summarization techniques can be broadly categorized into the following types [22]:

- *Extractive summarization*

  Key phrases are extracted directly from the source material (in most cases, it’s texts), which are combined to create summaries. In this way, the original content is preserved while maintaining high readability. This method relies on identifying and pulling the most informative portions.

- *Abstractive summarization*

  Compared to extractive summarization, abstractive summarization captures the core meaning of the source material by generating new sentences that are eventually all combined in a summary that goes beyond the mere extraction of the source material. These summaries are more fluent and natural because the content of the summaries needs to be expressed in a way that is not totally present in the source material. There are two main models that use this approach:

  - Specialized Summarization Models like BART [17] and PEGASUS [23];
  - Large Language Models like Amazon Titan [24], Claude [25], and OpenAI GPT [26].

- *Multi-level summarization*

Multi-level summarization is the combination of both the previous approaches to handle long documents [27]. There are two possible variants:

1. Extractive-Abstractive: First applies extractive summarization (e.g., BERT Extractive Summarizer) to select key sentences, then uses abstractive models (e.g., Amazon Titan, GPT-4, PEGASUS) to generate a coherent summary from the extracted content [27].

2. Abstractive - Abstractive: Uses techniques like Map-Reduce or Map-ReRank, where long documents are split into chunks, each chunk is summarized abstractively, and then these intermediate summaries are combined into a final summary using another abstractive model [28].

In this work, we use an abstractive summarization approach, in which the model generates novel summaries by understanding and rephrasing the input text, rather than simply extracting sentences. This is achieved using an encoder-decoder transformer architecture (LED), which allows flexible and coherent summary generation.

2.3. Graph Neural Networks

Graph Neural Networks (GNNs) are designed to handle complex data structured as graphs. Compared to traditional neural networks, which excel at processing Euclidean data such as text or images, GNNs process data with complex relationships and interdependence between objects.

A graph contains two fundamental components, Nodes, which represent entities or objects, and Edges, which define the relationships between nodes. With this structure, traditional machine learning approaches struggle with graph data because the number of neighbors per node can vary drastically; there is no fixed order for the nodes; topological relationships must be preserved, even if they are complex; and each instance depends on its connected neighbors.

2.3.1 How GNNs Work

There are three main steps the GNNs go through to process the data:

- Node Embeddings: First, the nodes are mapped to a lower-dimensional space while still preserving structural relationships.

This includes several approaches such as Graph Convolutional Networks (GCNs) (Graph Neural Network (GNN) Methods), Random walk-based approaches (Traditional Embedding Methods) or SciBERT (for text/graph hybrid approaches) (Text-Based Methods).

- Neighborhood Aggregation: In the second step, the information from connected nodes is combined.

  In this way, the local context is preserved while still maintaining the structure of the graph and learning complex patterns and relationships.

- Multi-layer Processing: Finally, in the third step the node representations are built using successive layers. Here, each layer transforms and processes the node features, allowing the model to capture more complex patterns in the graph structure.

In this work, Graph Neural Networks (GNNs) are used because biomedical data is better represented in a graph. Nodes represent biomedical entities (such as genes, proteins, diseases, drugs), while edges represent relationships such as co-occurrence, semantic relations, or hierarchical dependencies. This representation enables multi-hop reasoning across medical concepts, linking genes to diseases through protein pathways, and integrates external knowledge bases (such as UMLS or DBpedia), leading the abstractive summaries to have a higher credability thanks to the validated biomedical relationships.

2.4. Knowledge bases

A Knowledge Base (KB) is a centralized repository that stores, organizes, and shares information to help users find answers quickly and efficiently. It serves as a digital library of information about a company's products, services, or industry-related topics, designed to streamline information management and boost productivity.

2.4.1 Types of KBs

The types of knowledge bases is categorised by seeing their intended audience and access control mechanisms:

- Internal Knowledge Bases

  Knowledge Bases created for employees and internal stakeholders that contain HR policies, training materials, and technical documentation. It is used for the onboarding of employees and for the processing of documentation. To access these KBs, it must authenticate and authorize.

- External Knowledge Bases

  Knowledge Bases that have open access and are tailored to customers. It also contains product documentation, FAQs, and user manuals while also providing self-service support options. Compared to internal KBs, it may require minimal or no authentication.

2.4.2 DBpedia and UMLS

The two knowledge bases used in this thesis are DBpedia and UMLS. One is a more general KB, while the other is a more focused KB.

- DBpedia

  This KB is more general because it extracts information from Wikipedia, so it is a large-scale KB. The information is structured and made available on the Web. Because of that, DBpedia has open access and regular updates. It is a crucial component of the Linked Data movement because it provides Wikipedia's information in a structured kind of view [29]. Some key characteristics are the following:

  - Contains more than 6.7 million entities;
  - Links to other knowledge bases and datasets;
  - Provides data through SPARQL endpoints and dumps;
  - Available in about 140 languages (for version DBpedia Release 2025-06).

- UMLS

  This KB is designed for the domain of medicine and healthcare. Its source is 200 plus medical vocabularies [30]. This primary use is, of course, for healthcare systems and clinical support. Due to this specific domain, UMLS access is more restricted and involves the creation of a UTS account. Some key characteristics are the following:

  - Provides standardized medical concepts;
  - Enables healthcare system interoperability;
  - Supports clinical decision support systems;

  - Integrates more than 900,000 concepts and millions of term variants drawn from dozens of controlled vocabularies;
  - Cross-ontology mapping is enabled through unique concept identifiers (CUIs) and includes lexical tools to normalize terms across textual variants;
  - supports around 31 languages (version UMLS 2026AA).

UMLS (Unified Medical Language System) is a domain-specific biomedical knowledge base; DBpedia was still the one used as our primary resource for several reasons. Compared to UMLS, DBpedia is openly accessible and does not require a license to be used. This makes it easier to integrate it with the work that is being done. DBpedia also presents its data in a wider range of languages, which means an improvement in generalization and scalability of my approach. Finally, the regular updates and structured nature of DBpedia make it easier to retrieve and link biomedical entities, supporting a richer process in a widely applicable and more flexible manner. With this in mind, DBpedia was chosen as an alternative knowledge base to UMLS.

DBpedia was used for candidate resources and to fetch English abstracts, rerank candidates by semantic similarity to pick the best definition, and attach the chosen keyword-definition pairs to articles.

2.5. Keyword Extraction

Keyword extraction is the task where salient terms or phrases (unigrams to short n-grams) are identified to best represent the main topics and concepts of a document. For biomedical texts, domain terminology (e.g., gene names, diseases, methods) must be extracted effectively while handling multi-words expressions and redundancy reduction and still being robust to lexical variation. This task usually follows this pipeline:

- Preprocessing: lowercasing, tokenization, optional sentence segmentation; stopword and punctuation handling.
- Candidate generation: select n-grams (e.g., 1-3) excluding stopwords; allow domain terms and multi-word entities.
- Scoring: importance given based on local statistics (frequency, dispersion, position), co-occurrence/relatedness, casting, and context variability.

- Ranking and selection: produce a top-k list; optionally de-duplicate near-duplicates via lexical or embedding similarity; map to knowledge bases (e.g., DBpedia/UMLS) if needed.

This is important for our work because the extracted keywords we obtain are later embedded and used to construct concept graphs that support summarization. One well-know keyword strategy is ”YAKE!” (Yet Another Keyword Extractor), which is an unsupervised, domain-independent keyword extraction method designed to identify the most important words or phrases in a text. Works directly with the document used as input and makes it very lightweight, fast, and highly adaptable compared to other extractors [31].

For each candidate keyword, ”YAKE!” calculates a score based on several statistical features derived from the text itself, including:

- Term frequency - how often a term appears in the document
- Word position - whether the term appears early or later in the text
- Word relatedness - how often a word co-occurs with others
- Case sensitivity - whether a word appears capitalized
- Context variance - the diversity of contexts in which a word appears

All these features combined give weight to each keyword. The lower the score, the more relevant the keyword is considered to be. We use YAKE! because it helps with the identification of the keywords while also removing the less useful text. Since it is unsupervised, it works without labeled training data, which is useful in biomedical tasks where annotations are hard to get. All of this improves focus, reduces noise, and supports more relevant summaries.

Another strategy widely used is KeyBERT [32]. In this technique the entire document is transformed into an embedding representation, according to the chosen language model. Then, there is the selection of the candidates to be keywords, and they are also turned into embedding vectors. Finally, cosine similarity is computed for each candidate and the overall text. The most similar candidates are selected as keywords.

For the experiments in this thesis, we chose YAKE! instead of strategies based on language models because YAKE! is robust to rare expressions, which can happen in biomedical papers, and for its clear logic in computing the relevance score for each selected keyword.

3. Related Work

This chapter gives a review of previous research efforts in two main areas that are relevant for this work: automatic summarization and approaches that incorporate medical knowledge bases for NLP tasks, particularly in biomedical and biological contexts.

3.1. Automatic summarization in Biomedical/Biological NLP

Automatic summarization has experienced significant development in the past several decades, including a more specialized version of it, lay summarization. This task transforms complex technical or academic content into clear and easy-to-understand texts for non-experts in those areas. However, these models need to understand those same complex concepts and manage terminology simplification to preserve factual accuracy [33].

Older and more traditional methods include rule-based or statistical approaches but often lack the capacity to handle semantic complexity. With the introduction of neural and transformer-based architectures, significant improvements in performance have been made. Examples such as Luo et al. (2023) [34] explore 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 articles published in PLOS journals*, 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 [35], 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.

In similar manners, KeBioSum [36] enhances biomedical extractive summarization by integrating PICO domain knowledge into PLMs via lightweight adapters. It employs dual complementary adapters - a generative adapter that reconstructs masked PICO elements and a discriminative adapter that predicts PICO category labels- which are then infused into pre-trained models like BioBERT [37] and RoBERTa to improve extractive summarization of biomedical literature. The technique proposed by KeBioSum was applied to four

*https://journals.plos.org/plosone/

datasets composed of biomedical literature in CORD-19 journals* that contains articles related to COVID-19, long scientific documents from biomedical literature PubMed-Long, introduction sections only from PubMed documents (PubMed-Short), and EBM-NLP for PICO detection training. The authors evaluated their results using ROUGE scores, achieving peaks of 42.36 for ROUGE-1 and 38.66 for ROUGE-L. While KeBioSum uses structured PICO labels and is an extractive task (sentence selection), our approach uses continuous embeddings (DBpedia), and it's an abstractive task (generation).

Another line of work, such as DORIS [38], combines graph-based representations enriched with external biomedical knowledge (UMLS/DBpedia) with topic modeling to identify main themes and a transformer encoder-decoder for abstractive summarization. The Knowledge-Enhanced Graph Topic Transformer (KGTT) technique was applied to four datasets: CORD-19 journals, PubMed-Long, PubMed-Short, and S20RC containing random-sampled papers from biology and medical domains with human-written gold summaries. The authors evaluated their results using ROUGE scores, achieving peaks of 46.16 for ROUGE-1 and 42.16 for ROUGE-L. While KGTT uses a single unified graph-based approach, our work explores various knowledge integration strategies (UMLS, DBpedia, YAKE) across different model architectures (BART, UMLS variants). It also includes more extensive preprocessing and candidate generation mechanisms, where the focus is on comparative performance evaluation across different knowledge sources rather than a single optimized architecture.

Even with these advances, transformer-based summarization models (e.g., BART, T5) continue to struggle with biomedical texts when they do not have appropriate domain adaptation [33]. Challenges like handling technical terminology, assurance of readability for the public, and the preservation of factual accuracy remain.

3.2. Medical Knowledge Integration in NLP

To improve the quality and relevance of biomedical NLP tasks, several approaches have incorporated structured medical knowledge other than knowledge bases. Resources like:

- **MeSH (Medical Subject Headings)**: A hierarchical controlled vocabulary maintained by the National Library of Medicine, organizing medical terms into parent-child relationships. MeSH enables standardized vocabulary normalization and hierarchical entity disambiguation in biomedical text [39].

*https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244839

- **SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms)**: A comprehensive clinical ontology with 350k+ medical concepts connected by explicit semantic relationships. SNOMED CT provides fine-grained concept definitions and relationship types essential for capturing domain-specific entity interactions and clinical context [40].

These resources are the most used for entity recognition, disambiguation, relation extraction, and summarization. Thanks to these resources, models can work within specific domains, such as the biomedical domain, where they can identify relevant medical content, filter noise, and generate more accurate outputs.

3.3. Large Language Models for Lay Summarization

With new advancements in large language models (LLMs), zero-shot and few-shot summarizations opened up new opportunities. These models, such as GPT-4, PaLM, and Claude, are able to create coherent, context-sensitive summaries without the need for extensive task-specific training. For lay summarization, these capabilities are valuable because they translate complex, technical content into language that is accessible to non-experts.

Compared to traditional summarization models, LLMs can perform abstractive summarization that captures key ideas while rephrasing content in simpler terms. This is made possible through their exposure to vast and diverse corpora during pre-training, enabling them to generalize across a wide range of topics and writing styles.

Several studies were conducted where LLM's efficiency was evaluated, especially in domains such as the one we are also covering.

Several recent studies illustrate various degrees in strategies:

- Goldsack et al. (2023) [12] and Goldsack et al. (2022) [12] where LLMs were shown to generate lay summaries from biomedical abstracts that a human had evaluated before. In these studies, the zero-shot approach with the combination of question answering and summarization was used. Two novel datasets were used: PLOS, a large-scale dataset of biomedical journal articles with expert-written lay summaries that were selected from PLOS ONE publications, and eLife (Section 5.1).

- Ming, Guo, and Kilicoglu (2025) [41] improved the ability of LLMs for knowledge-guided biomedical lay summarization using MeSH terms that represent the articles'

main topics to enchance background information generation and a multi-turn dialogue framework employed during instruction-tuning to effectively incorporate the MeSH terms. This enables the LLMs to explain technical jargon and provide context beyond the original paper. This approach was evaluated on the eLife dataset, where the best model achieves, roughly, a 2% improvement on the ROUGE-1 score over the state-of-the-art on the eLife dataset. Compared to our approach, it utilizes a MeSH hierarchical vocabulary while using an LLM-based (instruction-tuning) architecture.

- WisPerMed team (BioLaySum 2024) [42] fine-tuned autoregressive LLMs (BioMistral, Llama3) for lay summarization in the BioLaySumm2024 shared task. This includes instruction tuning to guide model behavior, few-shot learning with well-crafted prompts to improve relevance and factuality, and prompt variations tailored with specific context. Furthermore, a Dynamic Expert Selection (DES) mechanism optimizes output selection based on readability and factuality metrics. This approach was evaluated on the BioLaySum2024 Shared Task dataset, which consists of biomedical scientific articles paired with lay summaries. From the 54 teams that participated in this task, the WisPerMed team achieved 4th place overall with only a 1.5% behind 1st place. Compared to our approach, it uses prompt engineering and few-shot learning for knowledge integration, while ours integrates structured knowledge graphs (UMLS, DBpedia) and uses fine-tuned autoregressive LLMs compared to LED encoder-decoder with graph neural networks used in our approach. This study was highlighted because it exemplifies a top-performing LLM-based pipeline in BioLaySumm 2024, relying solely on open models and prompting rather than closed biomedical ontologies. This makes it a representative case of modern lay-summarization approaches and complementary to other systems in this section that depend on structured resources such as MeSH or UMLS.

- Guo et al. (2021) [43] presented one of the earliest abstractive summarization approaches to convert technical biomedical reviews into lay-friendly language. The approach was evaluated on eLife and PLOS datasets and acheived peaks of 53.02 ± 0.48 for ROUGE-1 and 50.23 for ROUGE-L. Compared to our approach, it doesn't have structured knowledge, relying only on language model representations and optimizing for lay language clarity, while our approach maintains medical precision via knowledge-augmented decoding and uses a standard transformer compared to our

LED+GNN architecture that separates long-document encoding from knowledge-guide generation.

3.4. Benchmarks

Through the several shared tasks, some of those are now standard benchmarks:

- BioLaySum 2023 [44] and 2024 [45]: This shared task, through its own website, has been generating lay summaries from biomedical articles in both controllable and non-controllable settings - Various groups and teams have participated using different methods like prompt-engineered LLMs, fine-tuned domain models, and hybrid RAG-based pipelines.

- PLABA and PLOS/eLife datasets contain expert-level biomedical abstracts and their lay summary, offering valuable training and evaluation resources.

3.5. Recurring Challenges, Limitations and Research Gaps

Upon revision of the literature and benchmarks from Sections 3.1 to 3.4, recurring challenges gave more motives for our approach. The recurring challenges were:

- LLM's tendency to introduce factual inconsistencies: Models like BART, T5 and general LLM's tend to introduce factual errors while simplying biomedical terminology. Luo et al. (2023) show that LED improves factuality but still produces incorrect simplifications in biomedical terms. Goldsack et al. (2023) also reports that zero-shot LLM's generate plausible explanations but factually wrong. This demonstrates the necessity for explicit integration of structured knowledge, something our approach trys to resolve.

- Long biomedical documents are difficult to handle: LED and Longformer mitigate the problem, but models like BART and T5 still continue to truncate information. That is why we use external structures (keywords + graphs) to reduce textual load.

- Existence of a mismatch between human evaluation and automatic evaluation metrics: ROUGE and BLEU do not capture legibility nor clarity for the lay audience. BioLaySumm 2023-2024 reports clear discrepancies between automatics metrics and human evaluation. That is why we also compute the readability and qualitive analysis.

Even after recent advancements, some important research gaps still persistin the biomedical lay summarization area:

- Use of closed resources: A big part of the systems still utilizes closed knowledge bases

- Factual fidelity: ensure that simplified summaries remain accurate. Despite improvements in factual consistency, LLMs still produce simplifications that distort biomedical meaning. Prior work highlights that factual errors often arise when models attempt to reduce terminology complexity. This gap motivates the integration of explicit domain knowledge and keyword filtering in our pipeline, aiming to preserve factual accuracy while simplifying content.

- Readability control: tailoring outputs to specific non-expert audiences. Existing systems rarely provide mechanisms to adapt summaries to different lay audiences. Readability is typically treated as a secondary metric rather than a design constraint. Our methodology addresses this gap by incorporating readability scoring and qualitative analysis, ensuring that outputs are accessible to non-experts.

- Domain coverage: most datasets are limited to abstracts or narrow subfields. Most datasets used in biomedical lay summarization focus on abstracts or narrow subfields, restricting generalization. By evaluating our approach on eLife, AGNews and 20NewsGroups, we explore whether hybrid summarization techniques can generalize beyond biomedical abstracts and operate across broader domains.

- Evaluation methods: the current metrics used do not fully capture human perceptions of clarity and usefulness. Current evaluation practices rely heavily on automatic metrics such as ROUGE and BLEU, which do not capture clarity, usefulness or accessibility from a human perspective. This gap motivates the inclusion of readability metrics and human-oriented qualitative evaluation in our experimental design.

These gaps lead to the exploration of hybrid models that combine LLMs with domain-specific knowledge bases and the development of better evaluation frameworks for biomedical lay summarization.

4. Methodology

The approaches of this thesis are based on two methods; the first is based on the work of Goldsack et al. [12], which identifies biomedical concepts using UMLS, and the second, our proposal, is based on YAKE!, which identifies keywords as concepts in biomedical abstracts. Next, we detail both approaches.

4.1. Golsack et al. Method

This section describes how the work of Goldsack et al. implements UMLS to extract, filter, and standardize biomedical concepts from research articles for lay summarization. The main steps of the technique are described below.

1. *Concept Extraction and Mapping*. First, this work uses UMLS-based tools, such as MetaMap and term lookup utilities through the use of the UMLS API, to scan biomedical texts—titles, abstracts, and body— and and map phrase candidate windows (single and multi-word) to UMLS CUIs. This process includes:

    - Generation of permutations and right-truncated variants if no exact match is found.
    - Applying lexical normalization for punctuation, case, and hyphens. These techniques mirror those found in UMLS-Query tools, which support phrase-to-concept mapping and CUI retrieval.

2. *Filtering and Ranking* To ensure relevance and clarity:

    With curated UMLS ”content views” extracted concepts have ambiguous, overly general, or irrelevant terms filtered out.

    The remaining concepts are then ranked based on various criteria, such as frequency, their position in the document (e.g., if they appear in the title or early sentences), and semantic types (e.g., diseases, treatments). Concepts with higher contextualization and importance are prioritized by the summarization process.

3. *Lay Mapping and Summary Construction*

    If chosen, top-ranked concepts (CUIs) are mapped to simpler and more lay-friendly terms. This stage is crucial because it transforms the relational structure captured in

the lay mapping into a coherent narrative. By integrating node semantics with their contextual connections, the system can generate summaries that preserve both factual accuracy and the underlying relationships between biomedical entities.

.

4.2. YAKE! Method with DBpedia

This section describes our approach for extracting, structuring, and leveraging biomedical concepts from research articles using the Yet Another Keyword Extraction (”YAKE”) algorithm for summarization.

1. *YAKE! Overview*

   YAKE! is an unsupervised, domain-agnostic keyword extraction method that identifies important terms in a document based on statistical and positional features, without requiring external corpora or dictionaries. It’s particularly suitable for biomedical texts, where terminology can be highly variable and context-dependent.

2. *Concept Extraction and Embedding*

   With YAKE!, each biomedical abstract is processed to have a set of keywords or concepts extracted. They are intended to capture the core biomedical ideas present in the text. To ensure that only the most relatable biomedical concepts are retained, we use a DeepSeek language model (**DeepSeek-R1:7b**) that classifies whether each keyword is related to the biomedical domain while also removing keywords that often appear that the DeepSeek model had problems classifying (e.g., keywords like ”Figure Supplement”). This classification followed this prompt:

   > *“Classify if the following term is related to biomedical / biological domain: {keyword}. Answer yes if it is related to biomedical / biological, no if not.”*

   Then, each keyword classified that was related to the biomedical domain by the DeepSeek model is transformed into an ense vector representation (embedding) using a pretrained language model such as SciBERT. With this step, we are able to capture semantic relationships between concepts even if they are lexically different.

3. *Graph Construction*

This phase transforms the extracted and classified biomedical keywords into a structured knowledge representation that captures semantic relationships. This process is divided into various key steps:

(a) Similarity Computation: With the vector embeddings obtained in the previous phase, we compute pairwise similarities between all keywords within an abstract. This calculation is done through cosine similarity between embedding vectors, which measures the angular distance between concepts in the embedding space. This metric is used for this calculation because it is invariant to vector magnitude and it's effective at capturing semantic proximity. For each abstract with *n* keywords, this results in $\binom{n}{2}$ similarity scores.

(b) Edge Creation via Thresholding: Using a similarity threshold τ (above 0.5), a graph fully connected and overly dense was avoided. This means that only keyword pairs with similarity scores above the τ could be connected by edges in the resulting graph. This mechanism ensures that only meaningful relationships are preserved by filtering out weak or spurious connections and by creating a sparse graph structure that is computationally tractable while still being interpretable.

(c) Graph Structure: The resulting graph $G = (V, E)$ has the keyword as vertices V and similarity-based relationships as edges E. Each edge can optionally be weighted by the similarity score, so the summarization model can distinguish between strongly related concepts (concepts with an edge with a high similarity score) and weakly related concepts (concepts with an edge with low similarity score that still is above the threshold). This graph-based representation explicitly captures which biomedical concepts are semantically linked within the abstract, adding structured context that complements the original text.

(d) Semantic Interpretation: The graph serves as an intermediate knowledge representation that bridges the gap between unstructured text and the summarization model. Connected components or dense subgraphs often correspond to coherent thematic clusters within the abstract, which the model can leverage to generate more focused and accurate summaries.

4. *Integration with Summarization Model*

Our encoder-decoder summarization model receives the constructed graph and the original abstract embeddings as input and then trains to generate abstractive summaries that are informed by both the textual content and the structured knowledge captured in the concept graph.

The graph building process is described in Figure 4.1.

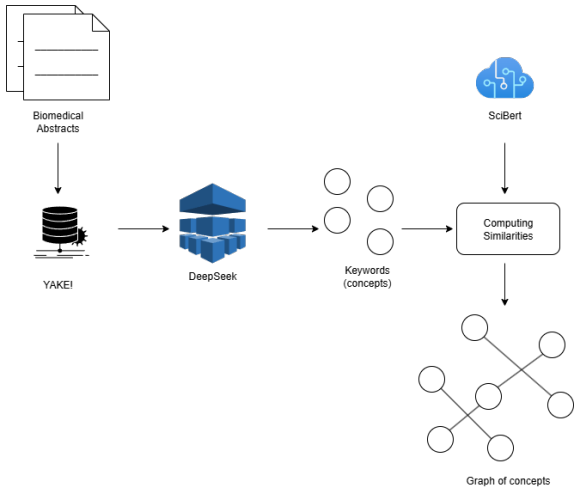

Figure 4.1: The building of the graph of concepts based on keywords extracted by YAKE!

.

4.3. BART with YAKE!

The third approach evaluated in this study combines the BART sequence-to-sequence model with unsupervised keyword extraction via YAKE to guide abstractive summarization.

1. *Keyword Extraction*: YAKE is applied to each article and *k* keywords are then extracted without requiring external knowledge resources or supervised training.

2. *Input Augmentation*: The list of keywords are then formatted as a structured concept block prepended to the original article text:

   augmented_input = keywords + original article text

3. *Model Training*: The standard BART model is fine-tuned on pairs of augmented articles and reference summaries, with no modifications to the base architecture.

4. *Inference & Evaluation*: During inference, BART generates summaries from augmented inputs. Using ROUGE metrics (ROUGE-1, ROUGE-L, etc.), evaluation is

done to each model checkpoint, where the best-performing epoch is retained based on ROUGE-L.

4.4. BART

The final approach is the baseline BART model trained for abstractive summarization, where the input is only the original article text. There is no external knowledge injected, no concept graphs, and no keyword prompts. Its role is to provide a clean refernece point for measuring the contribution of the augmentation strategies.

1. *Input construction*: each sample is represented as a source article and target summary.

2. *Tokenization and truncation*: source and target are tokenized to fixed maximum lengths.

3. *Training objective*: BART is fine-tuned with forced cross-entropy loss over target tokens.

4. *Generation and validation*: Summaries are generated with beam search and evaluated with ROUGE variants across epochs.

5. *Checkpointing*: the epoch checkpoint with best performance is saved using the validation metrics

4.5. Classification Models

Two classification models were implemented to ensure that our summarization pipeline only processed biomedical content: a traditional TF-IDF with XGBoost classifier and a fine-tuned transformer-based model (PubMedBERT).

1. *Task Definition*: The task is formulated as binary classification. Given a certain input text (title and/or abstract), the model will predict if the article belongs to the biomedical domain (label = 1) or if it belongs to the non-biomedical domain (label = 0).

2. *TF-IDF with XGBoost Classifier*: Term Frequency-Inverse Document Frequency (TF-IDF) [46] is a feature extraction method that represents documents as sparse vectors based on term importance.

    - Feature Extraction: For each article, TF-IDF scores are computed across the vocabulary. The TF component picks each term and measures how frequently

they appear in the article, and their local importance. The IDF component finds the most common terms (terms that appear across many documents) and penalizes them by reducing their weight. Mathematically, for term t in article *a* within corpus *D*:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log \frac{|D|}{|\{d' \in D \mid t \in d'\}|} \tag{4.1}$$

- Classification Model: The vectors created by TF-IDF serve as input features for an XGBoost (eXtreme Gradient Boosting) classifier. XGBoost [47] is an ensemble method that creates multiple decision trees sequentially, where each tree corrects errors made by previous trees. Compared to non-linear models, XGBoost can capture non-linear relationships and feature interactions, making it more powerful for complex classification tasks while remaining computationally efficient.

- Advantages and Limitations: This approach is computationally efficient, interpretable, and can handle high-dimensional sparse data well. Nevertheless, it still relies on lexical features and can't capture deep semantic meaning. For example, terms such as "heart disease" and "cardiac pathology" are treated as distinct features despite semantic similarity.

3. *PubMedBERT Classifier*: PubMedBERT is a domain-specific transformer model pre-tained on PubMed abstracts and full-text articles, making it inherently suited for biomedical text understanding.

   - Model Architecture: PubMedBERT is based on the BERT [14] (Bidirectional Encoder Representations from Transformers) architecture, consisting of 12 transformers encoder layers with multi-head self-attention mechanisms. For this classification, we added a linear classification head on top of the [CLS] token representation, producing a binary logit output.

   - Tokenization and Input Representation: The input text is tokenized using PubMedBERT's specialized vocabulary into subword tokens via WordPiece tokenization. Each sequence is prepended with [CLS] token and truncated to a fixed maximum length. This token sequence is then processed by the model and produces contextualized embeddings for each position.

- Semantic Understanding: Compared to TF-IDF, PubMedBeert captures deep semantic and contextual information. Meaning that it can spot biomedical content even when the terminology varies (using the given example from TF-IDF, it can understand that ”myocardial infarction” relates to ”heart attack”) because it has been exposed to a diverse number of biomedical expressions during the pretraining phase.

All models have their own purpose. The LLM prompting is to take advantage of the background knowledge of the large model; the TF-IDF classifier serves as a fast baseline, and PubMedBERT provides higher accuracy due to its semantic understanding.

5. Experiments and Results

In this chapter, we detail the datasets used in the experiments. For the biomedical experiments, we use a biomedical dataset, and for the classifiers, we employ a combination of two news datasets. We also describe the experiments we ran and the results they produced.

5.1. Datasets

The datasets used in this work were eLife articles from the eLife journal*, AG News articles† [48], and the 20 Newsgroups dataset, which contains nearly 20,000 newsgroup documents partitioned (nearly) evenly across 20 different newsgroups‡ [49].

The main dataset used was the eLife corpus, which contains complete biomedical research articles paired with expert-written lay summaries [50]. AG News and 20 Newsgroups, both non-biomedical news datasets, were used later to test new classifiers.

5.1.1 eLife

The dataset is divided into four splits: test, validation, training, and filtered training. The first two splits have 241 documents each, the train split has 4346 documents, and the *train filtered* split has 1582 documents. These 1582 documents are a subset of the original 4346 documents in the train split.

Table 5.1: Word count statistics across dataset splits.

Split	Section “Title”	Section “Abstract”
Val	12.25 ± 3.37	165.7 ± 19.4
Test	12.33 ± 3.45	166.28 ± 21.03
Train	12.30 ± 3.18	166.32 ± 21.63
Train filtered	12.38 ± 3.19	166.53 ± 21.04

A title example from an eLife article:

> National and regional seasonal dynamics of all-cause and cause-specific mortality in the USA from 1980 to 2016

*https://elifesciences.org/
†http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
‡http://qwone.com/~jason/20Newsgroups/

5.1.2 News

As the negative class, we used two news datasets: AG News [51] and 20 Newsgroups [52]. Table 5.2 summarizes the statistics for each dataset and the splits used to train the text classifiers.

Table 5.2: Word count statistics for the AG News and 20 Newsgroups datasets.

Dataset Split	Section "Text" (Mean $\pm$ SD)
AG News (All)	37.84 ± 10.09
20 Newsgroups (All)	181.64 ± 503.13

An example of an article from each news dataset:

> AG News: Kids Rule for Back-to-School The purchasing power of kids is a big part of why the back-to-school season has become such a huge marketing phenomenon.

> 20 News Group: AE is in Dallas...try 214/241-6060 or 214/241-0055. Tech support may be on their own line, but one of these should get you started.

5.1.3 eLife + AG News + 20 News Groups

This dataset combines all three datasets. Due to its size, we created two versions: one with downsampling and one without downsampling. Both versions were split into two files using an 80/20 split: `train.jsonl` and `test.jsonl`, respectively.

Table 5.3: Word count statistics for the eLife + AG News + 20 News Groups dataset in section "text"

Dataset Split	Train (Mean $\pm$ SD)	Test (Mean $\pm$ SD)
eLife + AG News + 20 News Groups (Downsample)	35.06 ± 107.44	30.88 ± 52.65
eLife + AG News + 20 News Groups (No Downsample)	55.71 ± 184.60	55.96 ± 187.49

Due to several time constraints and lesser good results, the work with the classification models, first mentioned in Section 4.5, was not explored further as it should have been. The objective was to create another method of comparsion, where the classification was done through a classifier like PubMedBert.

5.1.4 Guidelines

The eLife includes a plain-language summary in most research articles, known as an *eLife digest*. These digests are written by the eLife Features team—often with contributions from freelance writers—and are carefully edited before being sent to the authors for review to ensure clarity for non-specialist audiences [53].

According to the *eLife Features Team*, digests should:

- Briefly explain the background and significance of the paper in language that is accessible beyond the specialist community;
- Use clear, concise, and active sentence structures, avoiding jargon or defining any technical terms at first use; acronyms should be used sparingly (no more than three) [53];
- Be “clear, concise, and complete,” avoiding both overly technical language and exaggerated analogies or hype [54].

The data analyzed in this thesis were derived from eLife files used in *Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature* [12], which were originally downloaded from the GitHub repository *Corpora_for_Lay_Summarisation* [55].

5.2. Yake

For automatic keyword extraction, the *YAKE!* (Yet Another Keyword Extractor) algorithm, first mentioned in Section 2.5.1, was used. The implemented function used the following main parameters: the number of terms (*top*) set to 10, n-grams of up to 2 words ($n = 2$, i.e., only bigrams), and a variable deduplication limit (*dedupLim*) depending on the dataset split. For the validation and test files, *dedupLim* was set to 0.5, while for the training files, the value used was 0.9. This variation was configured directly in the function call.

Here are some examples of biomedical words found by YAKE!:

- action potential;
- ATP hydrolysis;
- spindle size;
- calcium transients;
- SNARE complex.

Here are some examples of non-biomedical words found by YAKE!:

- figure supplement;
- Figure;

- dashed line;
- Grand Island;
- data sources.

5.3. LLM Classification

The classification evaluation started with the random selection of 50 articles from the three eLife files, from which 10 keywords were extracted from each article using the YAKE algorithm, first mentioned in Section 4.2. DeepSeek-R1 then classified each keyword following the prompt first mentioned in Section 2, which states:

> *"Classify if the following term is related to biomedical / biological domain: {keyword}. Answer yes if it is related to biomedical / biological, no if not."*.

To provide some context, a brief description for each keyword was retrieved using DBpedia, first mentioned in Section 2.3.2, supplying the first description from its source index.

After the 50 articles in each file were analyzed by the LLM, we calculated the performance of DeepSeek-R1 using precision (other metrics found in Appendix A), defined by the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

Where:

- TP refers to true positives (relevant keywords correctly identified)
- FP refers to false positives (irrelevant keywords extracted)

The average precision was calculated using the macro-average, that is, the arithmetic mean of the individual precision scores obtained for each document.

File	Precision
Val	0.743
Test	0.821
Train	0.797

Table 5.4: Precision of LLM classification in 50 random articles of each split

These values represent the percentage of keywords that DeepSeek-R1 deemed relevant to biomedical/biological domains.

After that, for the same articles that DeepSeek-R1 classified, the classification was also done by me. The similarity was calculated using the Cohen Kappa formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5.2}$$

Where:

- p_o is the observed agreement. In other words, the number of times both raters agree divided by the total number of items,
- p_e is the expected agreement by chance.

The expected agreement is calculated as:

$$p_e = \sum_{i=1}^{k} p_{i1} \cdot p_{i2}$$

The final Cohen's kappa value indicates the strength of agreement between the two raters, in this case, the annotator and the LLM DeepSeek-R1.

Cohen's Kappa statistic (κ)	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.20–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Figure 5.1: Strength of agreement

In addition to Cohen Kappa, we also checked the standard statistics such as the median and standard deviation for each file:

Statistics	Val	Test	Train
Mean	0.328	0.341	0.434
Median	0.355	0.390	0.41
Standard Deviation	0.282	0.350	0.299
Minimum Cohen Kappa Value	-0.18	-0.15	-0.15
Maximum Cohen Kappa Value	1.00	1.00	1.00
Zero Values	16	19	13
Negative Values	3	1	3

Table 5.5: Statistics for each file

From Table 5.5, we can see that the strength of the agreement was fair in the val and test files, while it was moderate in the train file. This is due to the high percentage of articles that negatively affected the average Cohen Kappa value (around 39% for the validation file, 40% for the test file, and 33% for the train file) by decreasing its value or simply not changing the value. We can also see a clear difference between the use of the values 0.5 and 0.9 in the *DedupLim* variable due to the better average Cohen Kappa value obtained in the train file.

5.4. Summarization Results

This section presents the experiments and results from applying the various methodologies described in Chapter 4. The focus is on the two main approaches of this thesis: the ”GoldSack et al. Method” and the ”YAKE! Method with DBpedia”. We use tables and figures to discuss, analyze, and interpret the performance of each model and any significant findings or trends observed.

5.4.1 Experiments Setup

All experiments were carried out using the Deucalion high-performance computing service [56], where each script for each method was executed as a scheduled GPU job. This execution environment imposed several constraints on the experimental setup for each method. These constraints included limitations such as wall-clock execution time, maximum sequence lengths for the encoder and decoder, and batch size. This led to adjustments in the original configuration of some methods in order to ensure a feasible comparison.

This can be seen when comparing the original GoldSack configuration with the configurations used in our experiments. The original configuration had an encoder maximum length of 8192 tokens, a decoder maximum length of 2048 tokens, and a batch size of 8. However, with these values, memory and runtime constraints appeared on the Deucalion GPU nodes. This led to both GoldSack's method and our method (YAKE!+DBPEDIA) being trained with reduced maximum lengths: an encoder maximum length of 2048, a decoder maximum length of 512, and a batch size of 4. This explains the adjustments done to the original configuration of some methods. This mostly affected only the runtime of each script.

Each training run follows a standard supervised learning procedure. For each epoch in a run, the model is trained using the subset file described in Section 5.1.1, based on the eLife `train.json` split. After the training phase of each epoch, the model's performance is evaluated on the eLife validation set (`val.json`) using ROUGE metrics [35] (ROUGE-1, ROUGE-2, and ROUGE-L). The model with the best ROUGE-L score is then saved as the best model. The validation process also saves the predicted summaries for qualitative analysis. At the end, the best model is evaluated on the eLife test set to measure performance on unseen data.

Another important point is that all experiments and results were produced using QuickUMLS [57] instead of MetaMap, including the GoldSack results. This was due to setup complexity: initial development started on a newer Ubuntu release where MetaMap support was not available. Therefore, we adopted QuickUMLS. QuickUMLS is a tool for fast, unsupervised biomedical concept extraction from medical text. It scans text for candidate medical terms and matches them against a pre-indexed UMLS dictionary [58], returning concept identifiers (CUIs), semantic types, and scores.

5.4.2 Summarization Results

In this subsection, we summarize the results of the two main methods in this thesis using tables and figures.

5.4.2.1 GoldSack method

This table presents the statistics across all runs of the GoldSack method, while the figure shows the best ROUGE-L obtained in each run on the eLife validation split.

Metric	Mean	Std	Min	Max	Median
ROUGE1	29.99	6.61	24.58	46.42	27.49
ROUGE2	4.93	2.89	3.16	12.25	3.75
ROUGEL	17.02	2.39	13.63	21.09	17.23
ROUGELSUM	27.72	6.33	21.85	43.21	25.45

Table 5.6: Statistics from GoldSack model across all runs on the validation split

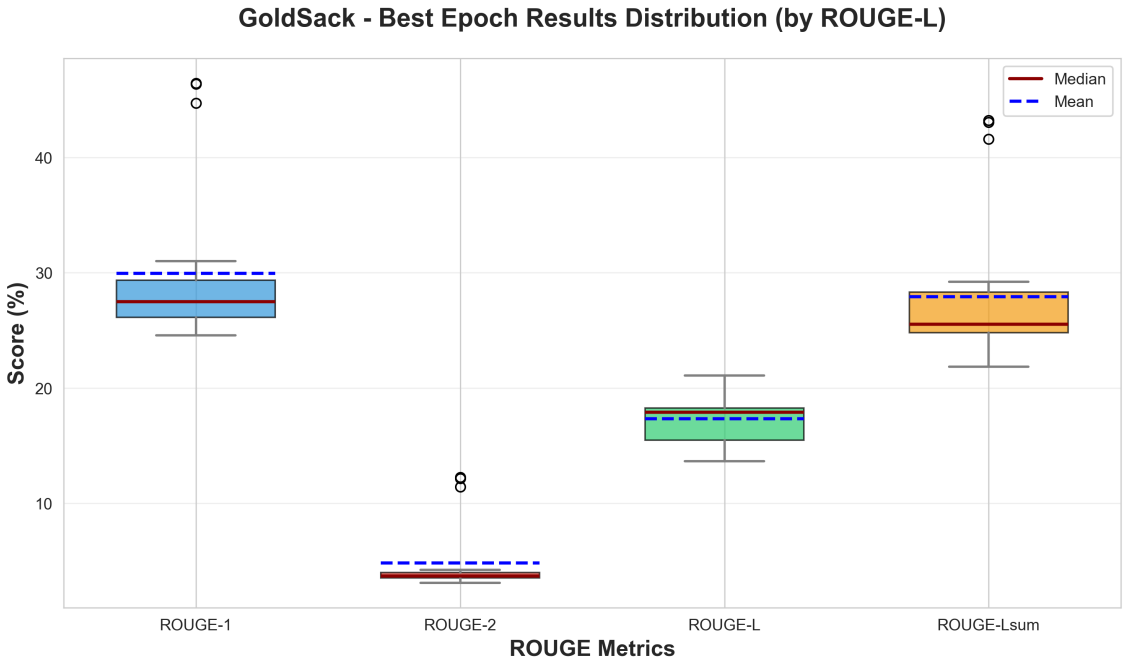

Figure 5.2: GoldSack - Best Epoch Results (by ROUGE-L) on the validation split

5.4.2.2 YAKE+DBPEDIA

This table presents the statistics across all runs of our method, while the figure shows the best ROUGE-L obtained in each run on the eLife validation split.

Metric	Mean	Std	Min	Max	Median
ROUGE1	37.08	25.71	0.16	59.11	57.83
ROUGE2	12.11	9.77	0.0	21.07	19.91
ROUGEL	27.45	18.18	0.16	43.41	42.13
ROUGELSUM	35.46	24.47	0.16	56.37	55.29

Table 5.7: Statistics from our model across all runs on the validation split

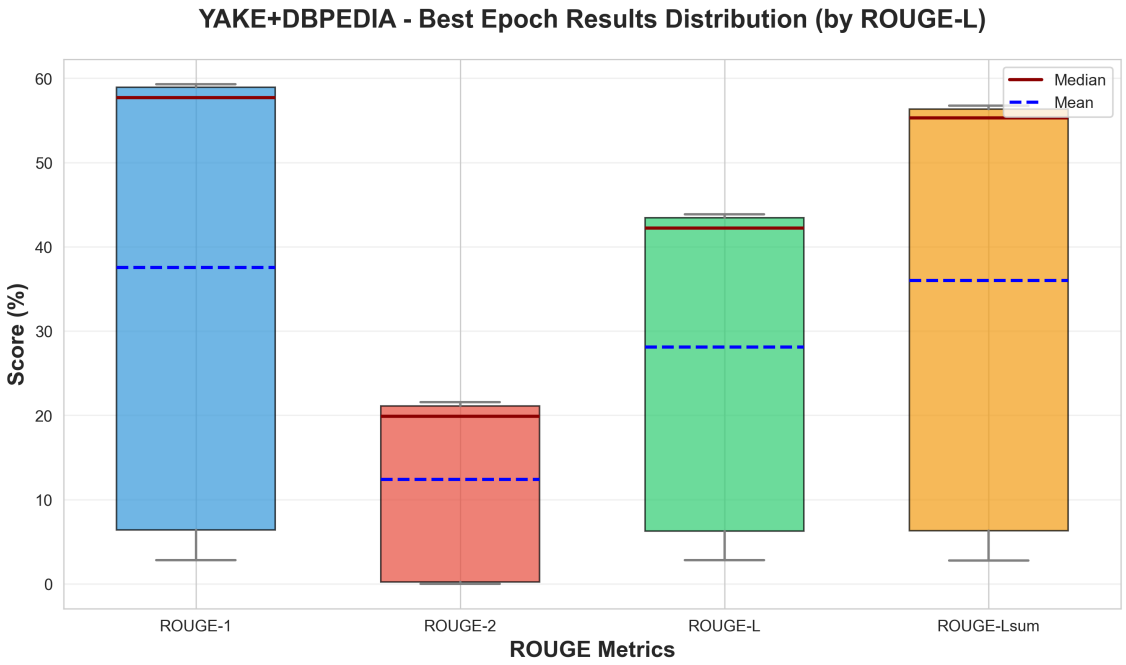

Figure 5.3: YAKE+DBPEDIA - Best Epoch Results (by ROUGE-L) on the validation split

5.4.2.3 Comparison

As can be seen in these tables and figures, our YAKE+DBPEDIA method demonstrates superior performance on the validation set, consistently achieving higher ROUGE scores across all metrics. However, even though it beats the BART baseline and the BART+YAKE! method on the unseen test set, the GoldSack method achieves better generalization performance. This suggests that although our method learns better representations during validation, GoldSack's simpler approach to knowledge integration generalizes better to unseen data. This is more effective for GoldSack's method because the ultimate goal of abstractive summarization is to produce high-quality summaries for real, unseen articles.

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
YAKE+DBpedia	58.44	20.65	43.45	43.47
GoldSack (UMLS)	**60.26**	**22.55**	**45.45**	**45.44**

Table 5.8: ROUGE scores on the test split

5.5. Error Analysis

To complement the aggregated ROUGE results, obtained in the previous section 5.4.2, we analyzed per-document BLEU distributions [59] and qualitative differences between the BART baseline, YAKE+DBPEDIA, and GoldSack methods. Figure 5.4 shows an overlapping violin plot of BLEU scores computed at the document level for BART, GoldSack, and

YAKE+DBPEDIA. Each violin contains the mean, median, and standard deviation, showing the central tendency and variability. From the graph, we can see that even though GoldSack reaches higher test-set generalization on average, YAKE+DBPEDIA exhibits competitive distributions and a wider tail in some documents, suggesting stronger performance in specific cases.

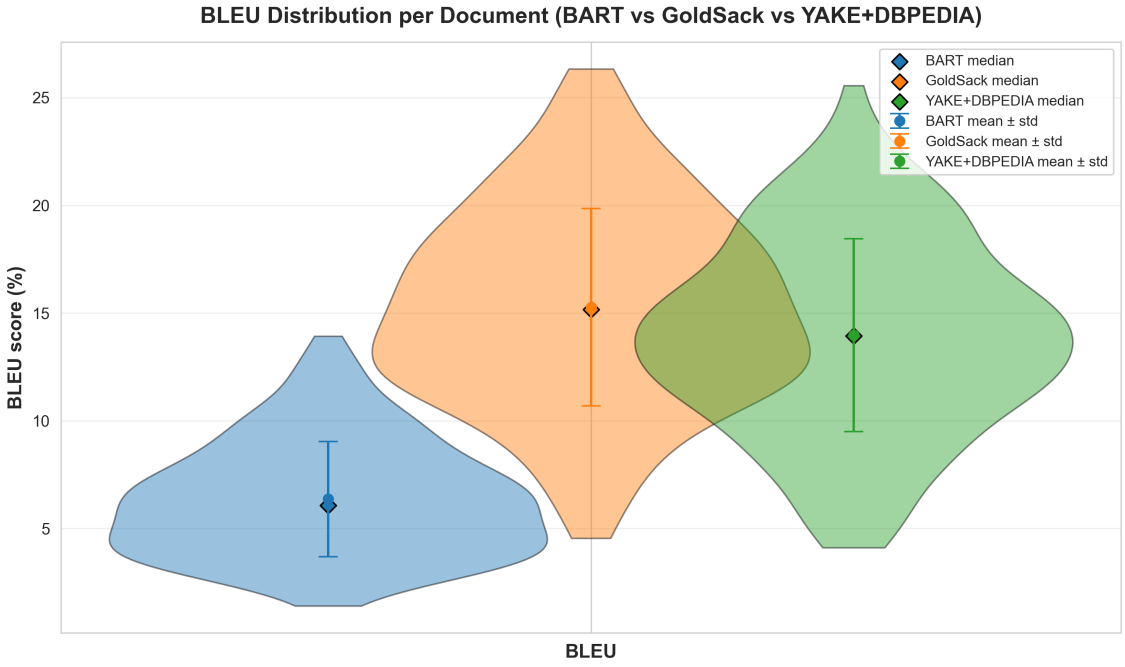

Figure 5.4: BLEU Distributions

We continued the analysis by selecting representative documents across four categories (with a detailed example for the first two categories):

(i) GoldSack > YAKE+DBPEDIA:

- **ID: 110** ($BLEU_G$ = 20.432; $BLEU_Y$ = 7.016) Subarea: ”Developmental Biology”
- **ID: 102** ($BLEU_G$ = 17.514; $BLEU_Y$ = 5.904) Subarea: ”Immunology”
- **ID: 56** ($BLEU_G$ = 24.476; $BLEU_Y$ = 14.337) Subarea: ”Immunology”

(ii) GoldSack < YAKE+DBPEDIA

- **ID: 164** ($BLEU_G$ = 8.804; $BLEU_Y$ = 15.906) Subarea: ”Neuro Science”
- **ID: 44** ($BLEU_G$ = 6.257; $BLEU_Y$ = 17.487) Subarea: ”Neuro Science”
- **ID: 238** ($BLEU_G$ = 5.512; $BLEU_Y$ = 12.527) Subarea: ”Genetics/Genomics”

Table 5.9: Comparative Analysis: Notch Signaling Example (GoldSack Success)

Success Case: Notch Signaling & DLL Ligands (GoldSack BLEU: 20.43, YAKE BLEU: 7.02)
Reference: A small number of signaling systems control how an animal develops from a single cell into a complex organism made up of many different cell types. Signals pass back and forth between cells, switching genes on and off to direct the development of tissues and organs. One of these signaling systems, called Notch, is so ancient that it appears in nearly all multicellular organisms. A cell sends a Notch signal using proteins called Delta or Jagged ligands that span membrane of the cell, so that part of the protein sits inside the cell and part remains outside. To change the behavior of another cell, the ligands bind to proteins called Notch receptors that span the membrane of the receiving cell. Mammals have two types of Delta ligand, two types of Jagged ligand and four types of Notch receptor...
GoldSack: protein protein of cells molecules are the cells animal develops, its single cell to a complex organism made up of many different cell types . Signals pass back and forth between cells , switching genes on and off to direct the development of tissues and organs . One of these signaling systems, called Notch, is so ancient that it appears in nearly all multicellular organisms . A cell sends a Notch signal using proteins called Delta or Jagged ligands that span the membrane of the cell , so that part of the protein sits inside the cell and part remains outside . To change the behavior of another cell, the ligands bind to proteins called Notch receptors that span the membrane of the receiving cell . Mammals have two types of Delta ligand, two types of Jagged ligand and four types of Notch receptor ...
YAKE: protein signal of cells molecules in the cells animal develops, its single fertil to a complex of of of many different types types . Oneals from from and forth between the as and on on and off , control the formation of different and organs . The of the signaling pathways is called Notch signaling is involved important that it is in the all multicellular organisms . The receptor detects a signalch signal to a on lig-Lagagged ligands , bind the on the receptor. and they they of the receptor on on the membrane membrane allows of outside . The bind the activity of the lig, the ligands must to the on Notch receptors on are the surface of the receptor cell . Theals have two Not of lig ligands, D that of Jagged ligand and one activating of Lagch receptor ...

Table 5.10: Comparative Analysis: Neuron Synapses Example (YAKE Success)

Success Case: Neuron Synapses (YAKE BLEU: 15.91, GoldSack BLEU: 8.80)
Reference: Neurons communicate with one another at junctions called synapses. When an electrical signal travels along a neuron and arrives at a synapse, vesicles filled with small neurotransmitter molecules fuse with the cell membrane and release the neurotransmitter. These chemicals rapidly bind to receptors on the downstream neuron that induce an electrical response in that cell. Vesicles can be consumed at prodigious rates, up to 500 a second, so the cell must recover the membrane rapidly and regenerate more vesicles filled with neurotransmitter. Experiments in the 1970s and 1980s suggested that when vesicles empty their contents into the synapse, they fuse completely with the membrane and are lost. To recover the membrane, the cell forms 'pits', by means of a coat protein called clathrin...
GoldSack: Neurons communicate with each another by junctions called synapses. When an electrical signal is between a nerve, reaches at a synapse, vesicles called with small neurotransmitter molecules are with the membrane membrane and release the neurotransmitter . These chemicals are rapidly bind to receptors on the downstream neuron and induce an electrical response in that cell . Vesicles can be consumed by synapses at prodigious rates , up to 100 times a second, and it recycling must be them vesicles. with neurotransmitter . Experiments in frog 1970s and 1980s suggested that when vesicles empty their, they are fused with the cell and recycled . However test the membrane, vesicles must form kissockets that which of a coat protein called clathrin ...
YAKE: Neurons communicate with one another at junctions called synapses . When an electrical signal travels from a neuron, arrives at a synapse, vesicles containing with small neurotransmitter molecules fuse with the cell membrane and release the neurotransmitter . These chemicals rapidly bind to receptors on the downstream neuron and induce an electrical response in that cell . Vesicles can be consumed by synapses at prodigious rates up to 100 times a second and it neurons needs be them vesicles. with neurotransmitter . Experiments in frogs laboratory and 1980s suggested that when vesicles empty their they are fused with the membrane. are recycled . However test the membrane, vesicles must form kissou' that which of a coat protein called clathrin ...

- **Green:** Accurate preservation matching reference
- **Blue:** Acceptable rephrasing (different wording, correct meaning)
- **Yellow:** Grammatical/syntax errors (spacing, word repetition, wrong articles)
- **Orange:** Entity fragmentation (incomplete entity names, split tokens, corrupted biological terms)
- **Red:** Semantic loss or critical meaning errors

(iii) GoldSack ≈ YAKE+DBPEDIA

- **ID: 2** ($BLEU_G$ = 15.990; $BLEU_Y$ = 16.369) Subarea: ”Neuro Science”
- **ID: 13** ($BLEU_G$ = 14.774; $BLEU_Y$ = 15.514) Subarea: ”Immunology”
- **ID: 26** ($BLEU_G$ = 17.613; $BLEU_Y$ = 17.118) Subarea: ”Genetics/Genomics”

(iv) Both methods fail (low BLEU scores in both methods)

- **ID: 5** ($BLEU_G$ = 9.543; $BLEU_Y$ = 7.767) Subarea: ”Neuro Science”
- **ID: 52** ($BLEU_G$ = 7.430; $BLEU_Y$ = 7.429) Subarea: ”Genetics/Genomics”
- **ID: 54** ($BLEU_G$ = 6.176; $BLEU_Y$ = 7.328) Subarea: ”Neuro Science”

Where $_G$ and $_Y$ represent the GoldSack and YAKE+DBPEDIA methods, respectively.

For cases where YAKE+DBPEDIA underperforms against GoldSack, the outputs tend to indicate that it preserves specialized biomedical entities better, which may be related to UMLS-driven concept grounding. In the opposite category, where YAKE+DBPEDIA outperforms GoldSack, YAKE+DBPEDIA tends to perform better on documents where the summary benefits from high-level paraphrasing and general-audience phrasing. This is most likely due to the descriptions obtained from the DBpedia encyclopedia. For cases where BLEU scores are almost equal, both methods show a similar degree of capture of the main content of the document. The minimal differences between methods are most likely stylistic (sentence structure, phrasing, or ordering) rather than substantive content choices. Cases where both methods underperform often involve abstracts with highly technical terminology or dense experimental protocols, making it difficult for the methods

to compress these abstracts into lay summaries without losing meaning. We can also see that subareas such as ”Immunology” favor the GoldSack method over YAKE+DBPEDIA.

5.5.1 Readability

Finally, we computed readability statistics for the reference summaries and each system output (Table 5.5): Flesch Reading Ease, average sentence length (words), average word length (characters), and type-token ratio.

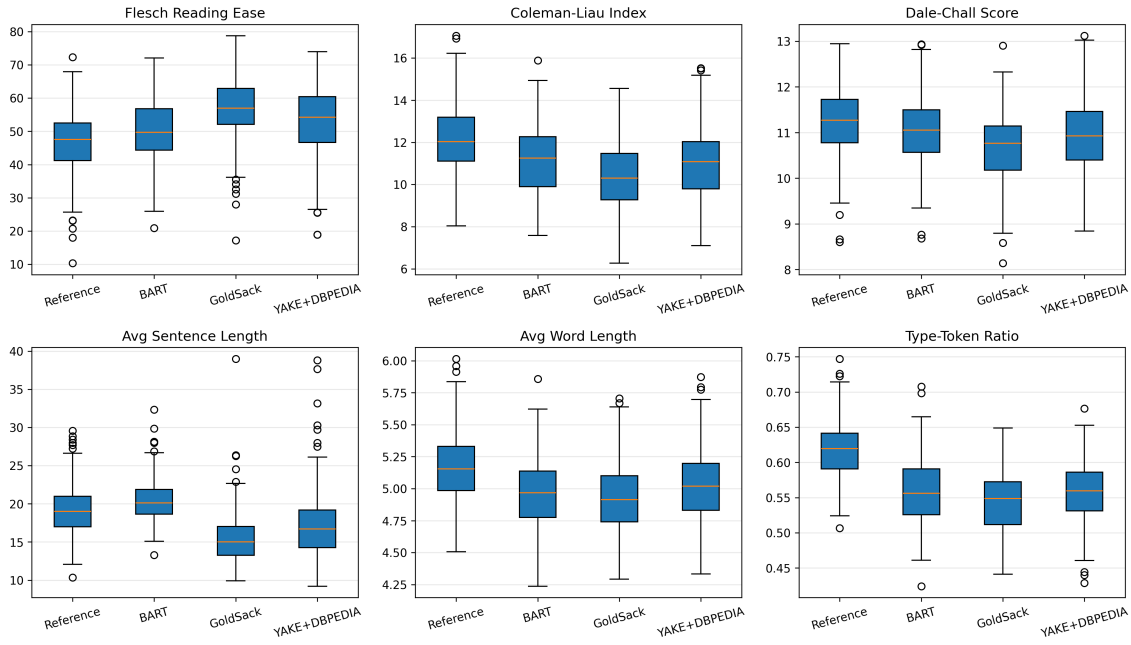

Figure 5.5: Readability statistics

From Figure 5.5, GoldSack presents the highest Flesch score (56.77) and the shortest mean sentence length (15.37 words), indicating better readability among the methods. YAKE+DBPEDIA sits between GoldSack and the reference in both readability and sentence length (Flesch 53.14; 17.14 words). BART produces the longest sentences (20.46 words), while GoldSack has the lowest lexical diversity (TTR 0.5459). The reference shows the highest lexical diversity (TTR 0.6182) and an average word length of 5.18 characters. This suggests that GoldSack produces more readable but simpler outputs, with YAKE+DBPEDIA close behind. We can also see that the reference summaries retain richer vocabulary and slightly more complex phrasing.

6. Conclusion

In this work, we present an open-resource approach to biomedical lay summarization. Based on the reported performance, our approach is competitive, and information extraction can be performed through mechanisms beyond named entities, such as YAKE!. To support these claims, we presented experiments on both the training and test sets and discussed the strengths and weaknesses of our method. Futhermore, a scientific paper covering this research was submitted to an international peer-reviewed workshop. The work underwent an independent evaluation process by three external reviewers who had no prior contact with this project, resulting in its acceptance and publication. This external validation highlights the relevance of our contributions to the biomedical natural language processing community. The published article, is acessible at:

Veloso, J. P., & Amorim, E. (2026). *An open-resource knowledge augmentation for biomedical lay summarization*. In *Proceedings of the 6th Clinical Natural Language Processing Workshop (CL4Health) at the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026)*. European Language Resources Association (ELRA).

However, not everything worked smoothly, and this work has several limitations. One limitation concerns computational resource constraints, which prevented us from testing a wider range of model parameters, including YAKE! settings. Time constraints introduced an additional limitation: we evaluated only one LLM for classification, even though many other models could now be tested. Because this was a zero-shot approach with a simple prompt, the results were not optimal, and a few-shot approach could improve them. Another limitation is that the evaluation was conducted by only one person; broader human evaluation would have strengthened the conclusions. Our approach also struggles with abstracts that contain many highly similar, interconnected technical entities (e.g., protein families with multiple related members). In graph-based representations, the structure benefits most documents but can increase confusion in some cases by creating very dense concept clusters that lead to repetitive generation. Finally, a standard BERT-based classifier trained on combined general and biomedical data could improve classification and might be more efficient than the in-context learning approach based on LLMs.

Nevertheless, the proposed approach remains competitive. Future work can improve results by exploring adaptive graph pruning so that edge density is reduced when concepts exceed a similarity threshold in localized regions of the graph. Additional improvements may come from fewer experimental constraints, better computational resources, more robust classification strategies, broader parameter exploration, and additional datasets.

Appendix A

This Appendix contains more detailed information about the results of the experiments mentioned on Section 5.3:

Table 1: Number of keywords that were considered relevant to biomedicine/biology by human evaluation that did not have DBpedia entries related to biomedicine/biology

File	Num. Keywords
Test	212
Val	191
Train	179

Table 2: Confusion matrix components (TP, FP, FN, and TN) for biomedical keyword classification across dataset splits

Dataset	TP	FP	FN	TN
Validation	176	61	5	28
Train	183	40	4	33
Test	181	46	5	28

Table 3: Performance metrics of the LLM on biomedical keyword classification

Dataset	Precision	Recall	F1-score	Accuracy
Validation	0.743	0.972	0.842	0.756
Train	0.821	0.979	0.893	0.831
Test	0.797	0.973	0.877	0.804

Table 4: Validation results using XGBoost classifier using TF-IDF for downsampled train/test datasets (eLife articles + AG News + 20 Newsgroups)

Split Type	Set	Precision	Recall	F1-Score	Accuracy
Fixed Split	Train	0.6608	0.9994	0.7956	0.7432
	Test	0.6329	0.9976	0.7744	0.7094
Random Split	Train	0.6555	0.9994	0.7917	0.7371
	Test	0.6267	1.0000	0.7705	0.7022

Table 5: Validation results using XGBoost classifier using TF-IDF for non-downsampled train/test datasets (eLife articles + AG News + 20 Newsgroups)

Split Type	Set	Precision	Recall	F1-Score	Accuracy
Fixed Split	Train	0.9649	0.0666	0.1246	0.9870
	Test	0.8077	0.0508	0.0957	0.9866
Random Split	Train	0.9457	0.0527	0.0998	0.9868
	Test	1.0000	0.0460	0.0880	0.9867

Appendix B

This Appendix contains more detailed information about the results of the experiments mentioned on Section 5.4.2 and 5.5:

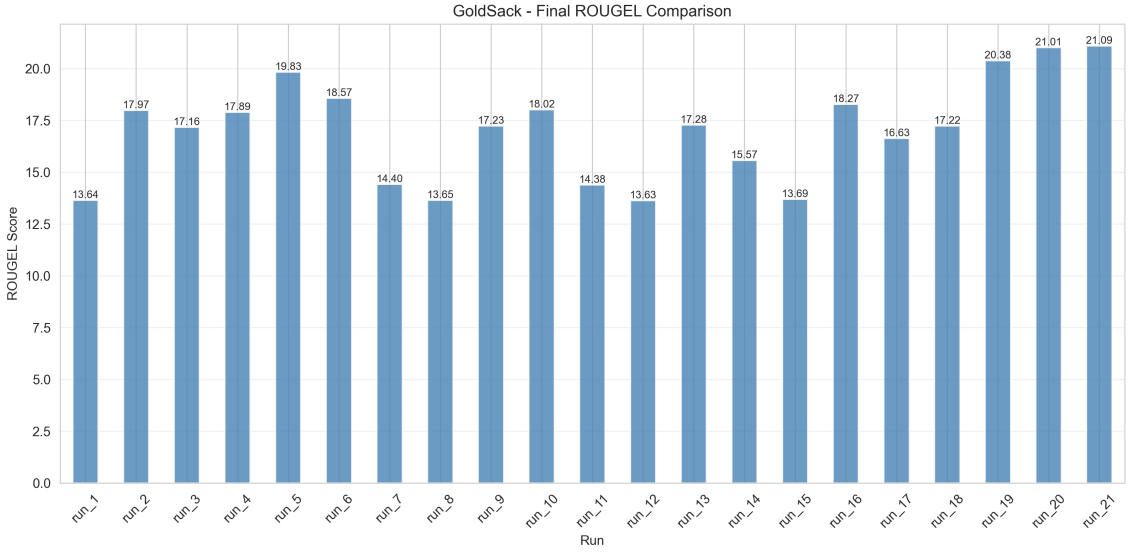

Figure 1: GoldSack - All RUNS ROUGEL VALUES

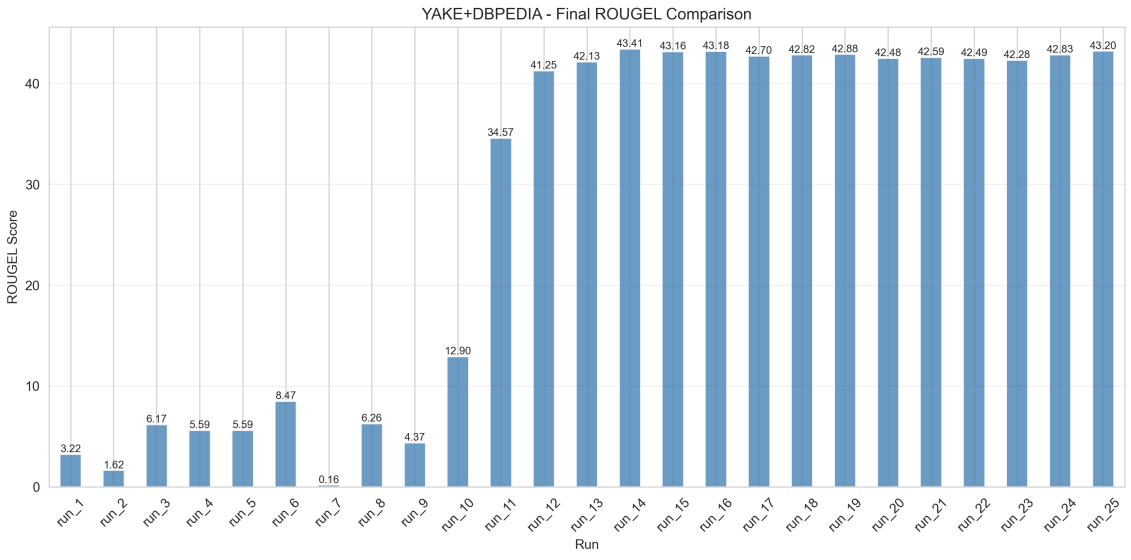

Figure 2: YAKE+DBPEDIA - All RUNS ROUGEL VALUES

Table 6: Comparative Analysis: Example 102 (GoldSack Success)

Example 102: Preexisting Memory CD4 T Cells (GoldSack BLEU: 17.51, YAKE BLEU: 5.90)
Reference: Immune cells called CD4 T cells help the body build immunity to infections caused by bacteria and viruses, or after vaccination. Receptor proteins on the outside of the cells recognize pathogens, foreign molecules called antigens, or vaccine antigens. Vaccine antigens are usually inactivated bacteria or viruses, or fragments of these pathogens. After recognizing an antigen, CD4 T cells develop into memory CD4 T cells ready to defend against future infections with the pathogen. People who have never been exposed to a pathogen, or have never been vaccinated against it, may nevertheless have preexisting memory cells ready to defend against it. This happens because CD4 T cells can recognize multiple targets, which enables the immune system to be ready to defend against both new and familiar pathogens. Elias, Meysman, Bartholomeus et al. wanted to find out whether having preexisting memory CD4 T cells confers an advantage for vaccine-induced immunity. Thirty-four people who were never exposed to hepatitis B or vaccinated against it participated in the study. These individuals provided blood samples before vaccination, with 2 doses of the hepatitis B vaccine, and at 3 time points afterward. Using next generation immune sequencing and machine learning techniques, Elias et al. analyzed ...
GoldSack: un cells called T4 T cells are to body to a by diseases. by viruses and viruses. and by a. Theactive proteins on CD surface of the T help the and and particles that peptigens, and pept-igens. Theines moleculesigens activate produced injected the when, viruses, but they of these viruses. The the the antigen CD4 T cells divide into a T4 T cells. to respond against future infections. the sameogen. However with have been been exposed to a pathogen can such who never had vaccinated against it, may have have memoryxisting memory CD that to defend against future. The is because the4 T cells are recognize and different, including can them immune system to mount more to defend against a the infections old infections. However et Farjer et andolomewu et al. have to find out whether individuals axisting memory CD4 T cells thaters an advantage in people-el immunity. The healthythree healthy had had vaccinated exposed to hepatitis B were hepatitis against it had in a study. The individuals had a samples that and and and the doses of the hepatitis B vaccine administered and after the and points after. The high- sequencing sequencing, machine learning„ Elias, al. found the
YAKE: une cells called T4 T cells recognize the body to a. many. by viruses, viruses. such by vaccination. Whencept proteins on the surface of the T help the, and particles, peptigens, and pept-igens. Whenines-igens activate recognized recognized the„ viruses, but are of viruses viruses. However the the antigen the4 T cells specialize antibodies antibodies T4 T cells, to respond against any infections. the sameogen. However who have been been vaccinated to a vaccineogen are but even no been vaccinated, one, are not have memoryxisting memory CD that to fight themselves any. However suggests because the4 T cells are recognize specific different, including allows them immune system to mount prepared to fight against any infections and old infections. However et Becjman et Scholomi et al. have to find out if memory memoryxisting memory T4 T cells helpsers an advantage to the protectionim immunity. Healthy healthyfive healthy were had all vaccinated to a B were any against it were in a experiments. The individuals were the samples from and, after the doses of the vaccine B vaccine, and after the days points after. The high- sequencing sequencing, machine learning„ Elias, al. found the

Table 7: Comparative Analysis: Example 238 (YAKE Success)

Example 238: The effect of hybridization on transposable element accumulation in an undomesticated fungal species (GoldSack BLEU: 5.51, YAKE BLEU: 12.53)
Reference: Hybrids arise when two populations of organisms that are related but genetically different mate and produce offspring. Hybridization has long been regarded as one of the many ways species evolve. Studying the changes in the genome that result from this process can provide insights into evolutionary history and predict the outcome of mixing between genetically different populations. In fact, the inability of two organisms to mate and produce viable and fertile hybrids has been used as a way to define species. It has been speculated that the infertility of many hybrids is due to short sequences of DNA in the genome called transposable elements. These elements are sequences of DNA that, when active, can move to a different position in the genome, causing mutations. It is thought that the process of hybridization may be activating transposable elements leading to the infertility often observed in hybrids. The activation of transposable elements in hybrids has been studied in animals and plants, and usually, the hybrids studied were either generated in the laboratory or found in the wild. Fungal species, such as the yeast Saccharomyces paradoxus, have hundreds of wild strains, including many hybrids, and can also be crossed in the laboratory to produce new hybrids, allowing a combined approach to studying the activation of transposable elements. Hén ...
GoldSack: brid are when two species of organisms cross are genetically to have distinct from together then offspring. Theization can been been recognized as an of the most ways that evolve. Howeverying hybrid effects that hybrid genomes of result from hybrid process is reveal insights into how processes. the how fate of future populations species similar species. However particular, hybrid hybrid of a populations to reproduce and produce offspring offspring reprodu offspring has been linked to a model to study the. Hybrid is been suggested that certain genetic of hybrid hybrid could a to mutations- of DNA called their DNA that transposable elements (These sequences can often of DNA that can when dupl, can duplicate around other new location within the genome. causing mutations that Hybrid has thought that hybrid presence of hybridization could trigger responsible theseposable elements, to the accumulation of seen in hybrids. However hypothesis of transposable elements is hybrids is been suggested in many, plants, but it has it effects are have hybrids hybrids by the wild or natural to natural wild. Howeverungi hybrids, however as yeast yeast Saccharomyces cereus, are been of thousands- that but hybrids hybrids. that hybrid be hybrid hybrid with the laboratory. produce hybrids strains. but the fungus species to study hybrid effects of transposable elements. However
YAKE: brid are when two closely of a that are genetically to have distinct from together reproduce offspring that Theization can been been thought as a of the major ways that can. Howeverying hybrid evolutionary that hybrid genomes that result from hybrid hybrid has reveal insights into how processes. how how future of hybrid up two similar species. However particular, hybrid accumulation of hybrid closely to copy and produce offspring offspring distinct offspring has been linked to a powerful to study the. Mut is been suggested that hybrid genetic of hybrid hybrid may one to a sections of DNA called the genome called transposable elements, Trans sections can present of DNA that can when they, can duplicate around other different position in the genome. and the that Trans was thought that hybrid increased of hybridization could have one transposable elements and to the accumulation of in in hybrid. However yeast of transposons elements is hybrid is been suggested in many, plants, but has has it effects accumulate have genetically sterile by a laboratory or natural to natural natural. Howeverungi hybrid that such as the yeast Saccharomyces paradoxus, are been of trans hybrid that but a hybrids that that are be be studied with the laboratory. create hybrid species. but the more hybrid to studying hybrid evolution of transposable elements. However énault

Table 8: Comparative Analysis: Example 26 (GoldSack ≈YAKE)

Example 26: Greatwall promotes cell transformation by hyperactivating AKT in human malignancies (GoldSack BLEU: 17.61, YAKE BLEU: 17.12)
Reference: In order to form a tumour, cancer cells have to overcome the controls that normally prevent cells from dividing too often. These controls include an enzyme called PP2A, which inhibits cell division by regulating the activity of other proteins. In a normal cell, PP2A may be temporarily deactivated from time to time to enable the cell to divide. However, PP2A is permanently deactivated in many cancer cells, which allows these cells to divide many times in quick succession. A protein called Greatwall is involved in deactivating PP2A in healthy cells, but it was not clear whether increases in Greatwall activity can promote the formation of tumours. Here, Vera et al. use a variety of cell biology techniques to address this question. The experiments show that increasing the amount of Greatwall in human and mouse cells can promote some of the transformations needed for these cells to become cancerous. Also, Greatwall increases the growth of tumours in mice. These effects are caused by the over-activation of a protein called AKT, which is already known to promote the formation of many cancers. Vera et al. show that Greatwall regulates AKT activity using a different pathway to how it influences PP2A activity. Further experiments revealed that many human tum ...
GoldSack: humans to divide a newour, cells cells need to divide the normal of keep prevent them from dividing. much. One controls are the enzyme called PP2A, which is the division. phosph the activity of a proteins. PP cancer tum cell, PP2A is be inactive inactiveactivated to the to time, allow the cell to divide properly However, in2A is often deactivated in cancer cancers cells. and is the cells to divide more times faster a succession. This protein called Greatwall (involved in theactivating PP2A. cancer cells. but it is not clear how this in thewall levels could also cancer growth of tumours. To, Cheng et al. used a technique of genetic models techniques to investigate this question. The experiments show that Great the levels of Greatwall activity cells cells mouse cells increases promote tum tum the cancer that for tum cells to divide tumous. This, thewall is the activity of tumours by mice, The findings are independent by the phosphproductionexpression of a protein called AKT, which is involved involved to be tum formation of tum tum. However et al. then that thewall activates theT by by a new pathway than the it normally other2A activity. This experiments show that the cancers cancersours
YAKE: humans to produce a newour, cells cells need to grow many barriers of normally exist them from dividing. quickly. One controls involve the enzyme called AK2A, which is other division. binding the activity of other proteins. However cancer cancer cell, PP2A is not in inactivated when being to time. prevent cells cell to divide correctly However, in2A is often activatedactivated in cancer cancers cells. and is them cells to grow more times more a succession. Many protein called AKwall- involved in controllingactivating PP2A. cancer cells and but it is not clear whether it in itswall levels cause also cancer formation of tumours. To, V- al. studied genetic range of cell models techniques to investigate this question. The experiments show that Great the levels of awall protein cells cells other cells increases promote the cancers these cell that to tum cancers to transform cancerous. However, increasingwall increases the ability of tumours by human and Further findings were not by increasing additionproductionexpression of a protein called AKT, which is normally involved to be tum formation of tum tum. Further et al. also that AKwall increases AKT by by a combination pathway to those other normally other2A.. This experiments show that AK cancers cancersours

Table 9: Comparative Analysis: Example 54 (Both Low)

Example 54: Kinesin-4 KIF21B limits microtubule growth to allow rapid centrosome polarization in T cells (GoldSack BLEU: 6.18, YAKE BLEU: 7.33)
The immune system is composed of many types of cells that can recognize foreign molecules and pathogens so they can eliminate them. When cells in the body become infected with a pathogen, they can process the pathogen's proteins and present them on their own surface. Specialized immune cells can then recognize infected cells and interact with them, forming an 'immunological synapse'. These synapses play an important role in immune response: they activate the immune system and allow it to kill harmful cells. To form an immunological synapse, an immune cell must reorganize its internal contents, including an aster-shaped scaffold made of tiny protein tubes called microtubules. The center of this scaffold moves towards the immunological synapse as it forms. This re-orientation of the microtubules towards the immunological synapse is known as 'polarization' and it happens very rapidly, but it is not yet clear how it works. One molecule involved in the polarization process is called KIF21B, a protein that can walk along microtubules, building up at the ends and affecting their growth. Whether KIF21B makes microtubules grow more quickly, or more slowly, is a matter of debate, and the ...
GoldSack: cyt system is made of a different of cells, are be and molecules and respond. that can attack them.When a detect the immune are infected, a foreignogen, they must use the infectionogen by□s DNA to release them to the surface surface. Thisized proteins cells called recognize recognize the cells and send with them to and a immun□immunological synapse'. This interactionsapses are a important role in the responses. when allow the immune system by help the to recognize the pathogens. However do this immunological synapse, a immune cell forms firstize its cyt structure, which the organiskshaped compartmentold called of proteins fil fil called microtubules. The micro of the reorganold is toward the contactological synapse, the reorgan, However process-organation is the microtubule is the synological synapse is driven as centpolarization'. it is in quickly. but little is not clear clear how micro happens. One of called in this reorgan of is a KIF21B. which k that is cause on thetubules. but up a the end of causing how length. However orIF21B is atubules grow or or or or if slowly, is not question of debate. but it answer
YAKE: cyt system is made of many different of cells, are communicate and objects, respond. they can defend them. For a encounter the immune encounter infected, a foreignogen, they activate produce the foreignogen by□s DNA and release them to the own.. Toized structures cells called recognize recognize the or and respond with them. for a immun□immunemunological synapse'. When eventsapses are an important role in the responses. when help the immune system□ help the to recognize the pathogens. When achieve the immunological synapse, the immune cell moves moveize its structure structure, including the arrayisksshaped structureold made of fil fil fil called microtubules. The micro of the structureold is toward the exposedological synapse, the moves. However movementorganorganation of the centtubule is the antigenological synapse requires driven as centcentolarization'. involves takes quickly quickly. and little is not clear clear how the works. Previous k involved in reorgan reorgan of is a KIF21B, which k that is be along microtubules and and them forces the end of pulling the behavior. Previous KIF21B is atubules grow or or or or restricts slowly, is not topic of debate. and

- **Green:** Accurate preservation matching reference
- **Blue:** Acceptable rephrasing (different wording, correct meaning)
- **Yellow:** Grammatical/syntax errors (spacing, word repetition, wrong articles)
- **Orange:** Entity fragmentation (incomplete entity names, split tokens, corrupted biological terms)
- **Red:** Semantic loss or critical meaning errors

References

[1] L. Bornmann and R. Mutz, “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references,” *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 11, p. 2215–2222, Nov. 2015. [Online]. Available: https://doi.org/10.1002/asi.23329 [Cited on page 1.]

[2] N. Fraser, L. Brierley, G. Dey, J. K. Polka, M. Pálfy, F. Nanni, and J. A. Coates, “The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape,” *PLOS Biology*, vol. 19, no. 4, pp. 1–28, Apr. 2021, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pbio.3000959 [Cited on page 1.]

[3] E. Landhuis, “Scientific literature: Information overload,” *Nature*, vol. 535, no. 7612, pp. 457–458, Jul. 2016. [Online]. Available: https://doi.org/10.1038/nj7612-457a [Cited on page 1.]

[4] Q. Xie, Z. Luo, B. Wang, and S. Ananiadou, “A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models,” Jul. 2023, arXiv:2304.08763 [cs]. [Online]. Available: http://arxiv.org/abs/2304.08763 [Cited on page 1.]

[5] A. Chaves, C. Kesiku, and B. Garcia-Zapirain, “Automatic Text Summarization of Biomedical Text Data: A Systematic Review,” *Information*, vol. 13, no. 8, p. 393, Aug. 2022, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-2489/13/8/393

[6] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, “A systematic review of automatic text summarization for biomedical literature and EHRs,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 28, no. 10, pp. 2287–2297, Aug. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8449627/ [Cited on page 1.]

[7] T. Goldsack, C. Scarton, M. Shardlow, and C. Lin, “Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles,” 2024. [Online]. Available: https://arxiv.org/abs/2408.08566 [Cited on page 1.]

[8] J. Zhou, C. Ye, P. Xu, and H. Xin, “Team YXZ at BioLaySumm: Adapting Large Language Models for Biomedical Lay Summarization,” in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, and J. Tsujii, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 818–825. [Online]. Available: https://aclanthology.org/2024.bionlp-1.76/ [Cited on pages 1 and 2.]

[9] Z. You, S. Radhakrishna, S. Ming, and H. Kilicoglu, “UIUC_bionlp at BioLaySumm: An Extract-then-Summarize Approach Augmented with Wikipedia Knowledge for Biomedical Lay Summarization,” in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, and J. Tsujii, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 132–143. [Online]. Available: https://aclanthology.org/2024.bionlp-1.11/ [Cited on pages 1 and 2.]

[10] “Infodemic.” [Online]. Available: https://www.who.int/publications/e/item/the-covid-19-infodemic [Cited on page 2.]

[11] “Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation.” [Online]. Available: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation [Cited on page 2.]

[12] T. Goldsack, Z. Zhang, C. Lin, and C. Scarton, “Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10 589–10 604. [Online]. Available: https://aclanthology.org/2022.emnlp-main.724/ [Cited on pages 4, 17, 21, and 30.]

[13] U. Kamath, K. Keenan, G. Somers, and S. Sorenson, *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Cham: Springer Nature Switzerland, 2024. [Online]. Available: https://link.springer.com/10.1007/978-3-031-65647-7 [Cited on page 5.]

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 2019,

arXiv:1810.04805 [cs]. [Online]. Available: http://arxiv.org/abs/1810.04805 [Cited on pages 5 and 26.]

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, arXiv:1907.11692 [cs]. [Online]. Available: http://arxiv.org/abs/1907.11692 [Cited on page 5.]

[16] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024. [Cited on page 6.]

[17] 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," Oct. 2019, arXiv:1910.13461 [cs]. [Online]. Available: http://arxiv.org/abs/1910.13461 [Cited on pages 6 and 8.]

[18] 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," Sep. 2023, arXiv:1910.10683 [cs]. [Online]. Available: http://arxiv.org/abs/1910.10683 [Cited on page 6.]

[19] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025. [Cited on page 7.]

[20] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," in *ACM Transactions on Computing for Healthcare (HEALTH)*, 2021. [Online]. Available: https://arxiv.org [Cited on page 7.]

[21] "Automatic summarization," Jul. 2024, page Version ID: 1236297853. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Automatic_summarization&oldid=1236297853 [Cited on page 8.]

[22] "Techniques for automatic summarization of documents using language models | AWS Machine Learning Blog." [Online]. Available: https://aws.amazon.com/pt/blogs/machine-learning/techniques-for-automatic-summarization-of-documents-using-language-models/ [Cited on page 8.]

[23] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in *International conference on machine learning*. PMLR, 2020, pp. 11 328–11 339. [Cited on page 8.]

[24] Amazon Web Services, “Amazon Titan Text models - Amazon Bedrock User Guide,” https://docs.aws.amazon.com/bedrock/latest/userguide/titan-text-models.html, 2024, accessed: 2024-05-23. [Cited on page 8.]

[25] Anthropic, “Text processing - Claude Documentation,” https://docs.anthropic.com/claude/docs/text-processing, 2024, accessed: 2024-05-23. [Cited on page 8.]

[26] OpenAI, “Summarizing long documents - OpenAI Cookbook,” https://cookbook.openai.com/examples/summarizing_long_documents, 2023, accessed: 2024-05-23. [Cited on page 8.]

[27] V. Tretyak and D. Stepanov, “Combination of abstractive and extractive approaches for summarization of long scientific texts,” *arXiv preprint arXiv:2006.05354*, 2020. [Cited on page 9.]

[28] D. Gunasekaran, A. Rana, C. Rodriguez, S. Hlaing, L. Nguyen, and Y. Li, “Techniques for automatic summarization of documents using language models,” AWS Machine Learning Blog, December 2023. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/techniques-for-automatic-summarization-of-documents-using-language-models/ [Cited on page 9.]

[29] “Search.” [Online]. Available: https://www.dbpedia.org/resources/lookup/ [Cited on page 11.]

[30] “Unified Medical Language System (UMLS),” publisher: U.S. National Library of Medicine. [Online]. Available: https://www.nlm.nih.gov/research/umls/index.html [Cited on page 11.]

[31] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, “A Text Feature Based Automatic Keyword Extraction Method for Single Documents,” in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham: Springer International Publishing, 2018, pp. 684–691. [Cited on page 13.]

[32] M. Grootendorst, “Keybert: Minimal keyword extraction with bert.” 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265 [Cited on page 13.]

[33] R. Xu and P. i. B. I. Stanford University, “Automatic biomedical information extraction from free text,” 2010. [Online]. Available: https://purl.stanford.edu/xn198gf0515 [Cited on pages 15 and 16.]

[34] Z. Luo, Q. Xie, and S. Ananiadou, “Readability Controllable Biomedical Document Summarization,” May 2023, arXiv:2210.04705 [cs]. [Online]. Available: http://arxiv.org/abs/2210.04705 [Cited on page 15.]

[35] C.-Y. Lin, “ROUGE: a Package for Automatic Evaluation of Summaries,” Jul. 2004, pp. 74–81. [Online]. Available: https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/ [Cited on pages 15 and 34.]

[36] Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, “Pre-trained language models with domain knowledge for biomedical extractive summarization,” *Knowledge-Based Systems*, vol. 252, p. 109460, Sep. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122007328 [Cited on page 15.]

[37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, arXiv:1901.08746 [cs]. [Online]. Available: http://arxiv.org/abs/1901.08746 [Cited on page 15.]

[38] Q. Xie, P. Tiwari, and S. Ananiadou, “Knowledge-Enhanced Graph Topic Transformer for Explainable Biomedical Text Summarization,” *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1836–1847, Apr. 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10227516 [Cited on page 16.]

[39] “Introduction to MeSH.” [Online]. Available: https://www.nlm.nih.gov/mesh/introduction.html [Cited on page 16.]

[40] “Overview of SNOMED CT.” [Online]. Available: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html [Cited on page 17.]

[41] S. Ming, Y. Guo, and H. Kilicoglu, “Towards Knowledge-Guided Biomedical Lay Summarization using Large Language Models,” in *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, S. Ananiadou, D. Demner-Fushman, D. Gupta, and P. Thompson, Eds. Albuquerque, New

Mexico: Association for Computational Linguistics, May 2025, pp. 285–297. [Online]. Available: https://aclanthology.org/2025.cl4health-1.24/ [Cited on page 17.]

[42] T. M. G. Pakull, H. Damm, A. Idrissi-Yaghir, H. Schäfer, P. A. Horn, and C. M. Friedrich, "WisPerMed at BioLaySumm: Adapting Autoregressive Large Language Models for Lay Summarization of Scientific Articles," Sep. 2024, arXiv:2405.11950 [cs]. [Online]. Available: http://arxiv.org/abs/2405.11950 [Cited on page 18.]

[43] Y. Guo, W. Qiu, Y. Wang, and T. Cohen, "Automated Lay Language Summarization of Biomedical Scientific Reviews," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 160–168, May 2021, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16089 [Cited on page 18.]

[44] "BioLaySumm." [Online]. Available: https://biolaysumm.org/2023/ [Cited on page 19.]

[45] "BioLaySumm." [Online]. Available: https://biolaysumm.org/2024/ [Cited on page 19.]

[46] TfidfVectorizer. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html [Cited on page 25.]

[47] XGBoost documentation — xgboost 3.1.1 documentation. [Online]. Available: https://xgboost.readthedocs.io/en/stable/ [Cited on page 26.]

[48] "AG's corpus of news articles." [Online]. Available: http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html [Cited on page 28.]

[49] K. Lang, "NewsWeeder: Learning to Filter Netnews (To appear in ML 95)," Jan. 2000. [Cited on page 28.]

[50] T. Goldsack, Z. Zhang, C. Lin, and C. Scarton, "Making science simple: Corpora for the lay summarisation of scientific literature," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10 589–10 604. [Online]. Available: https://aclanthology.org/2022.emnlp-main.724 [Cited on page 28.]

[51] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015. [Cited on page 29.]

[52] K. Lang, “Newsweeder: Learning to filter netnews,” in *Machine learning proceedings 1995*. Elsevier, 1995, pp. 331–339. [Cited on page 29.]

[53] eLife Features Team. (2017) Plain-language summaries: How to write an elife digest. Accessed: 2025-08-11. [Online]. Available: https://elifesciences.org/inside-elife/85518309/plain-language-summaries-how-to-write-an-elife-digest [Cited on pages 29 and 30.]

[54] S. R. King, E. Pewsey, and S. Shailes, “Plain-language summaries of research: An inside guide to elife digests,” *eLife*, vol. 6, p. e25410, 2017. [Cited on page 30.]

[55] T. Goldsack, “TGoldsack1/Corpora_for_lay_summarisation,” May 2025, original-date: 2022-10-10T10:25:01Z. [Online]. Available: https://github.com/TGoldsack1/Corpora_for_Lay_Summarisation [Cited on page 30.]

[56] Deucalion – rede nacional de computação avançada. [Online]. Available: https://rnca.fccn.pt/deucalion/ [Cited on page 33.]

[57] “quickumls: QuickUMLS is a tool for fast, unsupervised biomedical concept extraction from medical text.” [Online]. Available: https://github.com/Georgetown-IR-Lab/QuickUMLS [Cited on page 34.]

[58] “Semantic Network,” in *UMLS® Reference Manual [Internet]*. National Library of Medicine (US), Aug. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9679/ [Cited on page 34.]

[59] “IBM watsonx as a Service,” Oct. 2025. [Online]. Available: https://www.ibm.com/docs/en/watsonx/saas?topic=metrics-bleu [Cited on page 36.]