Development of a Biomedical Question Answering System Based on Transformer Models*

Notebook for the VerbaNex AI Lab at CLEF 2025

Lila López[1,*,†], Juan C. Martinez-Santos[2] and Edwin Puertas[3]

[1]Universidad Tecnologica de Bolivar, School of Engineering, Cartagena de Indias 130010, Colombia.

Abstract

Recent advances in artificial intelligence have enabled the automation of complex tasks in the biomedical domain, such as automatic question-answering. Within this framework, the BioASQ international challenge encourages the development of systems capable of understanding natural language questions and generating accurate answers based on the scientific literature. This work aims to design a system that classifies the types of questions and produces suitable responses accordingly. We implemented a modular pipeline with six main stages: (1) question type classification, (2) linguistic preprocessing, (3) Dynamic Routing and specialized model, (4) hyperparameters, (5) context retrieval, and (6) performance evaluation, including predicted type, execution time, and per-instance metrics. The system demonstrated strong performance in both the classification and answer generation tasks. In addition, a detailed analysis of each question helped identify specific errors and areas for improvement, depending on the question category.

Keywords

Biomedical, natural language processing, automatic response generation, classification, evaluation

1. Introduction

In recent years, the rapid growth of the biomedical literature has led to significant advances. However, it has also posed major challenges in the retrieval and synthesis of scientific knowledge. The medical community requires intelligent systems. These systems must enable efficient access to relevant, accurate and up-to-date information. Some authors claim that natural language questions should be the way to do it [1].

The BioASQ Challenge has been held annually since 2013. Its goal is to evaluate the ability of automated systems to answer complex biomedical questions [2]. They used natural language processing (NLP), information retrieval, and machine learning technologies [3]. These initiatives have incorporated Transformer-based models such as BigBird, BART Large CNN, and LongT5 to process long-form texts effectively [4].

Retrieving biomedical information from abstracts is a growing challenge. This challenge is due to information overload and the increasing volume of scientific literature. Biomedical experts have reported difficulties in finding precise information within full-length documents [5].

However, in QA systems that respond to user-formulated questions, some works have reported limited performance. For example, a study using the SQuAD data set achieved a precision of 69.69 % and an average correct response rate of 69.93 % [1].

The proposed system integrates advanced natural language processing techniques, allowing the identification of the type of question posed and the generation of coherent and relevant answers [6]. Furthermore, the authors implemented an automatic evaluation methodology to analyze the

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

lillopez@utb.edu.co (L. López); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

0009-0008-7763-9933 (L. López); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

system's performance in terms of accuracy, coverage, and relevance of the responses [7]. This proposal contributes to the advancement of intelligent solutions for information retrieval in the biomedical field.

2. Materials and Methods

Now, we present the techniques used in each methodological phase. The general architecture of the system is also described [8]. We developed this design in the context of the BioASQ challenge.

2.1. Data

The dataset used originates from Task B of the BioASQ challenge. This task assesses the ability of NLP models to answer biomedical questions formulated in natural language [9]. The corpus contains approximately 5,389 questions categorized into four types: yes/no, factoid, list, and summary [10]. The distribution is as follows: 1,600 factoid questions (30%), 1,459 yes/no (27%), 1,283 summary (24%), and 1,047 list (19%) This diversity enables evaluation of model performance on both closed and open question answering tasks, as illustrated in Table 1.

Table 1
Class distribution in question type

Question type	Total	Percentages
Factoid	1600	30%
Yes/no	1459	27%
Summary	1283	24%
List	1047	19%

2.2. Methodology

The proposed biomedical question-answer system follows a modular architecture composed of six main stages[11], as illustrated in Figure 1.

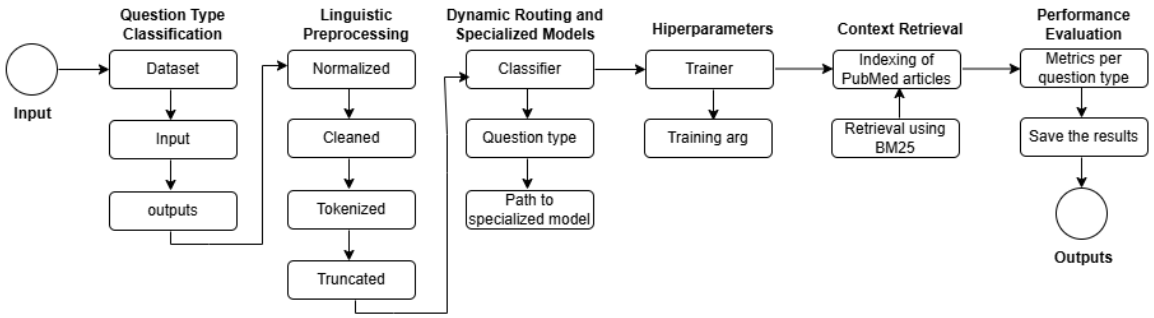

Figure 1: System pipeline.

2.3. Question Type Classification

Each instance in the dataset was processed using a supervised classification model. The goal was to predict its category among four possible types [7], as shown in Table 2. We used a Transformer-based model for this task [12], specifically DistilBERT. The model was trained on a labeled set of biomedical questions [13]. The model's output was the identification of the appropriate category for each question.

Table 2
Dataset version

Dataset	Total size	Question types	Format
BioASQ 13b	5389	Factoid, yes/no, list and summary	JSON(with fields)

2.4. Linguistic Preprocessing

First, we identified the type of question. Next, the text undergoes preprocessing. This process includes normalization, which converts the text to lowercase, removes duplicate spaces, and cleans special characters [14]. We also performed tokenization and truncation. These operations ensure compatibility with downstream language models.

2.5. Dynamic Routing and Specialized Models

Once the question type was classified, it was dynamically routed to a specialized answering model based on its category, as shown in Table 3. This phase employed pre-trained models adapted to specific tasks. We used generative models for summary-type questions. We applied extractive models to factoid and list questions. For yes/no questions, classification models were used [15].

Table 3
Specialized models by question type

Base model	Task	Tokenization	Specialized models
Distilbert	Multiclass classification (4)	AutoTokenizer	bart-large-mnli, and biobert

2.6. Hyperparameters

We trained the models using optimized hyperparameters to achieve the best performance. Key parameters included base model, number of epochs, batch size, maximum sequence length, and tokenization technique. These are summarized in Table 4.

Table 4
hyperparameters used for training the models

Base model	Number of epochs	batch size	maximum length	tokenization
Distilbert	4	8	512	Padding=True, truncation=True

In addition, we documented the hardware environment used during the experiments. It includes the operating system, virtual environment, RAM, CPU, and the estimated training time of the classifier [16]. Details are also shown in Table 5.

2.7. Context Retrieval

For questions requiring external information (factoids, lists, summaries), we retrieved relevant context. We performed the retrieval using a BM25-based technique. We applied this method over an index built from biomedical articles in PubMed. The retrieved context serves as input to generate more accurate answers.

Table 5
Hardware used in experiments

OS	Virtual environment	RAM	CPU	Estimated training time
MacOS(M1)	Python 3.9.13	8GB	MPS or no acceleration	2 hours and 44 minutes

2.8. Performance Evaluation

In summary, we applied question-type-specific evaluation metrics in accordance with the BioASQ guidelines. For yes/no questions, we used accuracy and F1 score. We evaluated factoid questions using strict and lenient accuracy. We assessed list questions with precision, recall, and F1 score. Finally, we evaluated summary questions using ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L).

3. Proposed System Architecture

The proposed architecture for the biomedical question-answering system [17] was designed based on the methodology shown in Figure 1. We defined hyperparameters, hardware configuration, and dataset version. Semantic retrieval and context expansion techniques were also integrated [18]. We employed this complete setup in the experiments conducted, as illustrated in Figure 2.

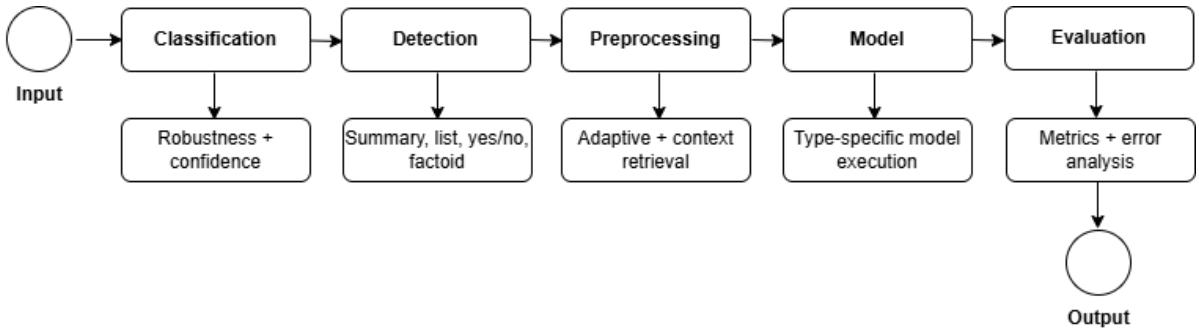

Figure 2: Architecture.

4. Results and Discussion

During the evaluation phase of the proposed system, we processed a total of 50 biomedical questions from the test set of the BioASQ Task B challenge [19]. These questions were pre-labeled according to the four types defined by the task: yes/no, factoid, list, and summary [20]. The system automatically classified the question type, retrieved the relevant context, applied specialized models—including fine-tuned versions of DistilBERT, BioBERT, and BERT-SQuAD2 per type—and generated an answer [21], which we evaluated using type-specific metrics.

4.1. Analysis of Results

The results from the experiments conducted for the BioASQ challenge show optimal performance for factoid and list questions. We achieved strict accuracy and F1 scores of 1.0 in several cases. In contrast, although yes/no questions reached 100% overall accuracy, their F1 score was 0.0 [22]. It indicates issues related to class imbalance between "yes" and "no" answers [4]. These results are detailed in Table 6.

The results are also compared with those obtained by BioASQ participants in previous years [23]. We can observe similar performance patterns, [24]. These findings are presented in Tables 7 and 8.

Table 6
Results by Yes/No, Factoid and List tasks for exact answers

System	Yes/no				Factoid			List		
	Acc.	F1_Yes	F1_No	Mac_F1	Str_Acc	Len_Acc.	MRR	Prec.	Recall	F-Meas
NLP-UTB4	0.7308	0.8444	-	0.4222	0.0455	0.0455	0.0455	0.0526	0.0132	0.0211
BioASQ-Baseline	0.3462	0.3704	0.3200	0.3452	0.1818	0.2727	0.2197	0.2243	0.3177	0.2439
edo	0.2692	-	0.4242	0.2121	0.0909	0.2273	0.0455	0.0526	0.0132	0.0211

Table 7
Results by summary tasks for ideal answers.

System	R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
NLP-UTB4	0.0277	0.0301	0.0322	0.0337
BioASQ-Baseline	-	-	-	-
edo	0.2797	0.2347	0.2816	0.2306

Table 8
Results of BioASQ Task Synergy: NLM runs for the passage retrieval task [24].

System	Mean precision	Recall	F-Measure
NLM-1	0.3927	0.1798	0.2153
NLM-2	0.4157	**0.2584**	**0.2712**

4.2. Discussion of Results

In this context, the results highlight the effectiveness of the proposed pipeline in both question-type classification and answer generation, particularly for factoid and list questions, where we accurately extracted specific information. The combination of supervised classification techniques and pre-trained language models enabled the generation of coherent and contextually relevant responses for yes/no and summary questions.

However, the low F1 scores observed for yes/no questions reveal a limitation of the model in accurately distinguishing between binary responses. We can attribute this issue to class imbalance during training, as well as the fact that we do not use models explicitly fine-tuned for the biomedical domain. Similarly, the variability in ROUGE scores for summary questions indicates that the quality of the generated responses is highly dependent on the content and relevance of the retrieved context [25].

5. Conclusion

The development of the biomedical question-answering system proposed in this work demonstrates its feasibility. We implemented a modular and specialized pipeline. It integrates multiple machine learning models and information retrieval techniques. The system automatically classifies the type of question. Then, it routes each instance to the appropriate model based on its category. Finally, it evaluates the responses using type-specific metrics. In this way, the solution addresses both closed and open question types.

In summary, the developed system represents a significant contribution to the field. It advances artificial intelligence tools applied to biomedical knowledge retrieval and understanding. This approach

opens opportunities for future applications in clinical and research settings.

6. Future Work

We proposed a Reinforcement Learning with Human Feedback (RLHF) approach to enhance response selection. The initial policy will rely on pre-trained generative models, such as BART or T5, which will produce multiple candidate responses.

We proposed addressing poor performance in yes/no questions by using specialized and fine-tuned models. Train these models on balanced and enriched datasets. Use corpora such as BioASQ yes/no and PubMedYesNo. Explore binary classification models like BioBERT or RoBERTa-bio.

For summary-type questions, the goal is to enhance the coverage and fidelity of generated responses. We propose fine-tuned encoder-decoder models such as BART or BioPEGASUS. Biomedical multi-reference summary datasets, such as PubMedQA summaries or MEDIQA, will be used.

CRediT authorship contribution statement

Lila López: Methodology, data curation, system, writing, original draft. **Juan C. Martinez-Santos**: Review and evaluation. **Edwin Puertas**: Review, formal analysis, and validation.

Declaration on Generative AI

The authors used generative AI tools, specifically ChatGPT and Grammarly, to support the writing process. These tools were employed for grammar and spelling checks, as well as for paraphrasing and rewording parts of the text. The authors take full responsibility for the content of the paper.

Acknowledgments

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [1], affiliated with the UTB, for their contributions to this project.

References

[1] M. V. Sadhuram, A. Soni, Natural language processing based new approach to design factoid question answering system, in: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 2020, pp. 276–281.

[2] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, E. Gaussier, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 138. doi:10.1186/s12859-015-0564-6.

[3] R. Devendra Kumar, K. Srihari, C. Arvind, W. Viriyasitavat, Biomedical event extraction on input text corpora using combination technique based capsule network, Sādhanā 47 (2022) 198.

[4] I. Naveed, M. Wasim, Ideal answer generation for biomedical questions using abstractive summarization, in: 2023 25th International Multitopic Conference (INMIC), IEEE, 2023, pp. 1–6.

[5] W. Yoon, Y. So, J. Lee, J. Kang, Pre-trained language model for biomedical question answering, in: Proceedings of the BioNLP Workshop 2020, Association for Computational Linguistics, 2020, pp. 79–85. doi:10.18653/v1/2020.bionlp-1.10.

[1] https://github.com/VerbaNexAI

[6] D. Weissenborn, M. Schroeder, G. Tsatsaronis, Answering complex questions with open-domain reading comprehension systems, arXiv preprint arXiv:1906.01071 (2019). URL: https://arxiv.org/abs/1906.01071.
[7] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: CEUR Workshop Proceedings, volume 2959, 2021. URL: http://ceur-ws.org/Vol-2959.
[8] A. Mutawa, S. Sruthi, A comparative evaluation of transformers and deep learning models for arabic meter classification, Applied Sciences 15 (2025) 4941.
[9] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 1–28.
[10] J. Mey, International conference on computational linguistics, STUF-Language Typology and Universals 18 (1965) 589–592.
[11] V. Sharmila, S. Kannadhasan, A. R. Kannan, P. Sivakumar, V. Vennila, Challenges in Information, Communication and Computing Technology: Proceedings of the 2nd International Conference on Challenges in Information, Communication, and Computing Technology (ICCICCT 2024), April 26th & 27th, 2024, Namakkal, Tamil Nadu, India, CRC Press, 2024.
[12] M. Lutfillayev, O. Narkulov, Artificial intelligence, blockchain, computing and security: Volume 2, 2023, 2, DOI: https://doi. org/10.1201/9781032684994-115 (????) 712–718.
[13] A. M. Striuk, Embracing emerging technologies: Insights from the 6th workshop for young scientists in computer science & software engineering, CEUR Workshop Proceedings, 2024.
[14] E. Martinez, J. Cuadrado, J. C. M. Santos, E. Puertas, Verbanex ai at clef exist 2024: detection of online sexism using transformer models and profiling techniques, environments 5 (2024) 7.
[15] E. H. Yossy, D. Suhartono, A. Trisetyarso, W. Budiharto, Question classification of university admission using named-entity recognition (ner), in: 2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), IEEE, 2023, pp. 20–25.
[16] H. Dong, V. Suárez-Paniagua, W. Whiteley, H. Wu, Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation, Journal of biomedical informatics 116 (2021) 103728.
[17] Y. Yan, B.-W. Zhang, X.-F. Li, Z. Liu, List-wise learning to rank biomedical question-answer pairs with deep ranking recursive autoencoders, PloS one 15 (2020) e0242061.
[18] A. B. Barlybayev, A. S. Mukanova, Advancements in geospatial question-answering systems: A case study on the implementation in the kazakh language, in: 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), IEEE, 2024, pp. 1710–1715.
[19] M. Sarrouti, S. O. El Alaoui, Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions, Artificial intelligence in medicine 102 (2020) 101767.
[20] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of bioasq tasks 12b and synergy12 in clef2024, Working Notes of CLEF 2024 (2024).
[21] K. Khelil, G. Besbes, H. Baazaoui-Zghal, Semantic question answering: Deep learning and nosql solution for the medical domain, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 6486–6493.
[22] Y. Du, Q. Li, L. Wang, Y. He, Biomedical-domain pre-trained language model for extractive summarization, Knowledge-Based Systems 199 (2020) 105964.
[23] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, et al., Bioasq at clef2025: The thirteenth edition of the large-scale biomedical semantic indexing and question answering challenge, in: European Conference on Information Retrieval, Springer, 2025, pp. 407–415.
[24] M. Sarrouti, D. Gupta, A. B. Abacha, D. Demner-Fushman, Nlm at bioasq synergy 2021: Deep learning-based methods for biomedical semantic question answering about covid-19., in: CLEF

(Working Notes), 2021, pp. 335–350.
[25] M.-T. C. Evans, M. Latifi, M. Ahsan, J. Haider, Leveraging semantic text analysis to improve the performance of transformer-based relation extraction, Information 15 (2024) 91.