Project Report
BioASQ Task-13b - Biomedical Semantic QA
Information Retrieval and Web Search

University of Mannheim

Onur Can Memiş (2040827), Abdullah Al Momin (1869371),
Marc Popescu-Pfeiffer (1638093), Enes Saban Tanrikulu(2055430),
Basar Temur (1909058)

November 27, 2024

Contents

1 Introduction

Biomedical question answering (QA) is a complex task requiring accurate information retrieval and precise answer generation from large, domain-specific datasets. In this project, we participated in BioASQ Task-13b, leveraging the PubMed dataset to explore and evaluate independent models for document retrieval and QA. For retrieval, we used Vector Space Models (TF-IDF with raw counts and logarithmic smoothing) and the probabilistic BM25 model, ranking documents based on relevance. For QA, we employed Llama 3.1, a large language model, and `deepset/roberta-base-squad2`, a fine-tuned span-based QA model. Each model was developed and assessed independently, providing valuable insights into their performance in biomedical QA.

This report outlines our approach's methodology, implementation, and evaluation for BioASQ Task-13b, providing insights into its performance and limitations. The following sections detail our system design, dataset preprocessing, model training, and results.

2 Dataset Description

For the BioASQ Task-13b, the PubMed dataset was utilized as the primary source for question-answering. It provides rich, structured metadata such as titles, questions, answers, contexts, relevant documents, publication dates, and answer begin and end indices. These attributes enable targeted retrieval and semantic understanding of biomedical queries.

For our project, we specifically focused on curated subsets of PubMed data relevant to the BioASQ challenge. This involved pre-processing abstracts and associated metadata to align with the question-answering requirements of Task-13b. The dataset's breadth and quality presented both opportunities and challenges in system development, necessitating efficient retrieval strategies and sophisticated NLP techniques to handle the complexity of biomedical language.

The following sections will detail our preprocessing steps, the challenges encountered with the dataset, and how we addressed them to develop an effective question-answering system.

3 Methodology

Our approach for BioASQ Task-13b aimed to evaluate the effectiveness of different methodologies in biomedical question answering. Each component was implemented and assessed independently, focusing on specific tasks such as relevant document retrieval and answer generation. The following models and techniques were used:

3.1 Relevant Document Retrieval

To identify documents from the PubMed dataset that were most relevant to a query, we implemented and evaluated three independent retrieval models:

1. **Vector Space Models (VSM):**

   - **TF-IDF Model with Raw Counts:**
     This model ranked documents by computing Term Frequency-Inverse Document Frequency (TF-IDF) scores using raw term frequencies. It served as a baseline for retrieval performance.
   - **TF-IDF Model with Logarithmic Smoothing:**
     Logarithmic scaling was applied to term frequencies in this variation, reducing the influence of excessively frequent terms and enhancing the balance between term importance and query relevance.

2. **Probabilistic Model:**

   - **BM25:**
     The BM25 model was implemented to rank documents probabilistically, accounting for term frequency, inverse document frequency, and document length. Its ability to adapt to varying document lengths in the PubMed dataset made it particularly effective for biomedical literature.

3.2 Question Answering (QA)

For generating precise answers to biomedical questions, we evaluated two advanced natural language processing models independently:

1. **Large Language Model (LLM):**

   - **Llama 3.1 (`llama3.1:8b`):**
     This 8-billion-parameter model was utilized to understand and generate comprehensive answers directly from biomedical texts. The Llama 3.1 model was evaluated for its ability to generate contextually relevant answers in response to domain-specific queries.

2. **Fine-Tuned QA Model:**

   - `deepset/roberta-base-squad2`**:**
     A pre-trained RoBERTa model fine-tuned on the SQuAD2.0 dataset was employed to extract span-based answers from relevant documents. It focused on identifying concise and precise responses within text segments.

3.3 Approach

Each model was run independently on the provided questions from BioASQ Task-13b, and their respective outputs were collected and analyzed. The results from retrieval models were evaluated based on their ability to fetch relevant documents, while the QA models were assessed for the quality and accuracy of their answers.

This modular approach allowed us to analyze the individual strengths and weaknesses of each model. The findings will provide insights into how these models can be further refined or integrated in future iterations for improved performance in biomedical question-answering tasks.

4 Implementation

The implementation of our system for BioASQ Task-13b focused on evaluating individual models for document retrieval and question answering. Each component was implemented and assessed independently without connecting the outputs of the retrieval models to the QA models. Below, we outline the implementation details.

4.1 Relevant Document Retrieval

1. **Data Preprocessing**
    - Titles, abstracts, and metadata from the PubMed dataset were extracted for document indexing.
    - Standard preprocessing steps such as lowercasing, punctuation removal, tokenization, and stopword filtering were applied.
    - Documents were vectorized for computational efficiency.
2. **Implementation of Retrieval Models**
    - **TF-IDF with Raw Counts:**
        - Implemented using scikit-learn's `TfidfVectorizer`.
        - Queries and documents were vectorized, and cosine similarity was computed to rank documents by relevance.
    - **TF-IDF with Logarithmic Smoothing:**
        - Logarithmic scaling was applied to term frequencies during TF-IDF computation.
        - Document-query relevance was evaluated using cosine similarity, emphasizing balanced term importance.
    - **BM25:**
        - The `rank_bm25` library was used for probabilistic document ranking.

    - Model parameters k_1 and b were tuned to optimize retrieval for biomedical text.

4.2 Question Answering (QA)

Implementation of QA Models

- **Llama 3.1 (`llama3.1:8b`):**
    - Implemented using the Ollama API.
    - Question and related context were provided directly to the model.
    - The model generated detailed answers based on the context and its pre-trained knowledge.
- **`deepset/roberta-base-squad2`:**
    - Implemented using Hugging Face's Transformers library.
    - The model was pre-trained on the SQuAD2.0 dataset and fine-tuned on BioASQ questions.
    - It provided concise, span-based answers, drawing from a static subset of the PubMed data rather than dynamically retrieved documents.

5 Evaluation

The evaluation of our BioASQ Task-13b system focused on independently assessing the performance of each model in the document retrieval and question-answering phases. Below, we outline the evaluation criteria and results for each model.

5.1 Relevant Document Retrieval

- **Vector Space Models (VSM):**
    - **Evaluation Metric:**
      The TF-IDF models (with raw counts and logarithmic smoothing) were evaluated based on cosine similarity, ranking documents by their relevance scores for a given query.
    - **Results:**
      The logarithmic smoothing model showed improved balance in rankings for queries involving common biomedical terms compared to the raw count model.
- **BM25:**
    - **Evaluation Metric:**
      The BM25 model ranked documents based on a probability score, reflecting the likelihood of document relevance to a given query.

    - **Results:**
      BM25 demonstrated robust performance, especially for queries where document length normalization played a critical role. Its parameter tuning improved retrieval effectiveness across varied query types.

5.2 Question Answering (QA)

1. **Llama 3.1** (`llama3.1:8b`)**:**

    - **Evaluation Metric:**
      The outputs of the Llama 3.1 model were evaluated manually for correctness and relevance.
    - **Results:**
        - The model provided detailed and contextually appropriate answers for 64 out of 85 queries.
        - Some answers included irrelevant information, reflecting the model's reliance on its pre-trained knowledge rather than document-specific inputs.

2. `deepset/roberta-base-squad2`**:**

    - **Evaluation Metrics:**
        - **Exact Match (EM):** Measures the percentage of answers that exactly match the ground truth.
        - **F1 Score:** Evaluates the overlap between the predicted and ground truth answers by considering precision and recall.
    - **Results:**
        - **Average Exact Match (EM):** 0.817
        - **Average F1 Score:** 0.856

      These results, achieved on the validation dataset, indicate that the model effectively identified correct answer spans for most queries, balancing precision and recall well.

5.3 Summary of Evaluation

- Retrieval models effectively ranked relevant documents, with BM25 outperforming the TF-IDF variants in scenarios involving lengthy or complex queries.

- Llama 3.1 demonstrated versatility in generating detailed answers, though it required manual evaluation due to the lack of input from retrieved documents.

- `deepset/roberta-base-squad2` excelled in span-based QA tasks, achieving high accuracy and consistency as reflected in the EM and F1 scores.

The independent evaluation of these models provided valuable insights into their performance and highlighted areas for future integration and improvement in biomedical question-answering systems.

6 Challenges

The development and evaluation of our system for BioASQ Task-13b presented several challenges, stemming from the complexity of the task, the size of the dataset, and the computational demands of advanced NLP models. Below, we outline the key challenges encountered:

1. **Dataset Size and Preprocessing**
   - The PubMed dataset, exceeding $100GB$, posed significant challenges in terms of storage, management, and processing.
   - Efficiently indexing and preprocessing such a large corpus required substantial memory and computational resources. Standard text normalization tasks such as tokenization, stopword removal, and vectorization became time-intensive due to the dataset's scale.
2. **Computational Complexity for Fine-Tuning**
   - Fine-tuning large language models such as Llama 3.1 or domain-specific models like `deepset/roberta-base-squad2` is computationally expensive.
   - Limited access to high-performance GPUs constrained the fine-tuning process, impacting the ability to fully adapt models to the biomedical domain.
3. **Time and Resource Constraints**
   - The iterative process of model training, evaluation, and debugging was time-consuming, further compounded by limited computational resources.
   - The size and complexity of the task required significant time for data preparation, model execution, and result analysis.

Addressing these challenges would require greater computational resources, improved dataset-handling strategies, and potential workflow integration. Despite these obstacles, our modular approach allowed for a comprehensive evaluation of each model's capabilities.

7 Related Works

Biomedical question answering has been an active area of research, with prior works employing various approaches, including traditional information retrieval

Model/Approach	Document Retrieval Metric	QA Metric	Results
PubMedBERT + BM25 [2]	Mean Average Precision (MAP)	F1 Score	MAP: 0.404, F1: 0.753
BioBERT + BM25 [4]	Mean Reciprocal Rank (MRR)	Exact Match (EM), F1	MRR: 0.416, EM: 0.692, F1: 0.771
Llama 2 (7B) [5]	Manual (QA relevance scoring)	Manual	High relevance for general biomedical questions
BioASQ Official Baseline [3]	Mean Average Precision (MAP)	F1 Score	MAP: 0.308, F1: 0.622
RoBERTa + BM25 [1]	Normalized Discounted Cumulative Gain (NDCG)	Exact Match (EM), F1	NDCG: 0.372, EM: 0.716, F1: 0.782

Table 1: Top 5 results from related papers

methods, fine-tuned machine learning models, and large language models. Below, we compare some notable solutions to BioASQ tasks and their reported performances.

7.1 Comparison and Observations

1. **Traditional Retrieval Models with BERT Variants:**
   - PubMedBERT and BioBERT models have shown notable improvements in biomedical tasks, particularly when combined with BM25 for document retrieval.
   - BioBERT, fine-tuned on biomedical literature, demonstrates higher EM and F1 scores compared to the official baseline.
2. **Large Language Models (LLMs):**
   - Models like Llama 2, while evaluated manually, excel in generating contextually relevant answers, highlighting their potential in open-domain biomedical QA.
3. **Baseline Systems:**
   - The official BioASQ baselines provide a reference point for evaluating newer models. Our `roberta-base-squad2` outperforms the baseline in both EM and F1 metrics, showcasing the effectiveness of fine-tuned span-based QA models.
4. **RoBERTa Variants:**
   - RoBERTa, particularly when paired with retrieval models like BM25, demonstrates competitive performance, striking a balance between retrieval and QA tasks.

Our project extends this body of work by independently evaluating multiple retrieval and QA models on the BioASQ Task-13b dataset. The insights gained can inform future efforts in designing integrated workflows and optimizing model performance.

8 Conclusion

In this project, we independently evaluated multiple models for document retrieval and question answering in the context of BioASQ Task-13b. Retrieval models like TF-IDF and BM25 effectively ranked documents based on relevance, while QA models such as Llama 3.1 and `deepset/roberta-base-squad2` demonstrated strong performance in generating accurate answers, with the latter achieving high Exact Match and F1 scores. Although the components were not integrated into a unified workflow, this modular approach provided valuable insights into the individual capabilities of each model. These findings highlight the potential for future systems to combine these models for enhanced performance in biomedical question answering.

References

[1] somosnlp-hackathon-2022/roberta-base-biomedical-es-squad2-es · Hugging Face, January 2024.

[2] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, September 2021. arXiv:2007.15779.

[3] Anastasia Krithara, Anastasios Nentidis, Eirini Vandorou, Georgios Katsimpras, Yannis Almirantis, Magda Arnal, Adomas Bunevicius, Eulalia Farre-Maduell, Maya Kassiss, Vasileios Konstantakos, Sherri Matis-Mitchell, Dimitris Polychronopoulos, Jesus Rodriguez-Pascual, Eleftherios G Samaras, Martina Samiotaki, Despina Sanoudou, Aspasia Vozi, and Georgios Paliouras. BioASQ Synergy: a dialogue between question-answering systems and biomedical experts for promoting COVID-19 research. *Journal of the American Medical Informatics Association*, 31(11):2689–2698, November 2024.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining, October 2019. arXiv:1901.08746.

[5] Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. BioInstruct: Instruction Tuning of Large Language Models for Biomedical Natural Language Processing, June 2024. arXiv:2310.19975 [cs].