Question Answering System for Electronic Medical Records Using ClinicalBERT

Authors: Sriram Natarajan, Rickston Pinto

Carnegie Mellon University
School of Computer Science

Problem Statement

Accurate interpretation of clinical data poses significant challenges due to the complexity and variability of medical language. This project addresses the need for a robust Question-Answering (QA) system that can efficiently handle the intricacies of medical terminology and context. By integrating regular expressions (regex) for structured symptom extraction and BERT for context-aware text interpretation, the proposed hybrid approach aims to enhance the accuracy and reliability of medical data interpretation, ultimately improving the effectiveness of clinical decision-making.

Methodology

Data Processing

The dataset was tokenized using the **ClinicalBERT** tokenizer, with input IDs, attention masks, and offset mappings generated to align the answers within the context. Answer start and end positions were mapped to token offsets for accurate training data preparation.

Exploratory Data Analysis (EDA) revealed that **medication-related** questions made up majority of the dataset, while **symptom-related** queries were only **0.5%**. To address this imbalance, regular expressions (regex) were used to extract symptom keywords more effectively, while BERT handled the abundant medication-related queries. This hybrid approach ensured comprehensive coverage across both question types.

Model Architecture

We process clinical text by first performing preprocessing steps such as tokenization, text cleaning, and stop word removal. The preprocessed text is then passed through two parallel pipelines: a **Pattern Matching Pipeline** and a **BERT Analysis Pipeline**. In the Pattern Matching pipeline, **regex patterns** and a **symptom dictionary** are used to identify symptoms, with results considered valid if they score above **0.85**. In the BERT pipeline, the **fine-tuned BERT model** generates **contextual embeddings**, and results are considered valid if the confidence exceeds **0.90**. If either pipeline produces a strong enough result, that result is used for **response generation**. If the results do not meet the thresholds, the system re-evaluates the other pipeline, ensuring robust performance by combining **structured pattern matching** with **contextual analysis** from BERT.

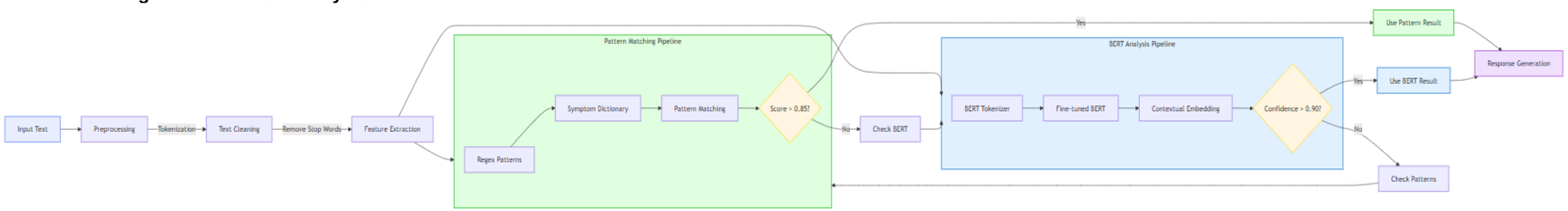

Fig.1 Model Architecture Schematic

Results

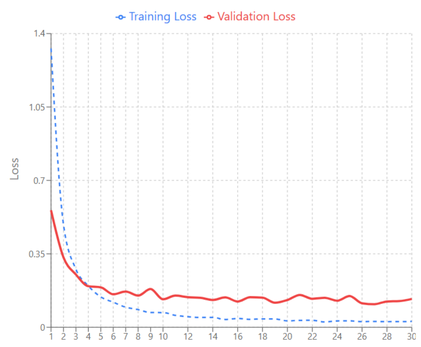

Fig. 2. Model Training and validation loss curves

Model Performance Analysis

Query Type	Regex Score	BERT Score	Selected Model
"Patient medication schedule"	0.45	0.97	BERT
"Drug interaction analysis"	0.51	0.95	BERT
"Treatment plan review"	0.62	0.93	BERT
"Fever and cough noted"	0.92	0.85	Regex
"Complex medication dosage"	0.48	0.96	BERT

BERT Model Strengths
- Complex medication queries
- Treatment plan analysis
- Drug interaction detection
- 95% average confidence on medication tasks

Regex Pattern Strengths
- Simple symptom matching
- Fast processing (10ms vs 45ms)
- Clear keyword detection
- 90% accuracy on basic symptoms

Table 1. Model Performance Analysis

Key Achievements:

•**Training Loss Convergence:** The model's training loss stabilized rapidly, plateauing at approximately 0.15, indicating effective learning and early convergence during training.

•**Reliable Symptom Extraction:** Regular expressions demonstrated consistent performance in extracting symptoms, facilitating the parsing of structured medical information.

•**BERT's Performance: BERT** exhibited strong contextual understanding in many instances, though performance varied depending on the complexity and ambiguity of the clinical text.

Analysis of Medical Question-Answering System Performance on Validation Set

Category	Method	Performance Metrics					
		Cases	Avg. Sim (%)	Avg. Conf (%)	High	Medium	Low
Symptoms	Regex	8	38.2	65.0	0	4	4
	BERT	0	—	—	—	—	—
Medications	Regex	2795	25.5	83.3	46	121	2628
	BERT	2358	99.2	99.8	2332	11	15

Notes:
Performance Categories: High (75%), Medium (50-75%), Low (< 50%)
BERT demonstrates exceptional performance on medication queries (97.4% achieving 95% similarity)
Regex patterns show limited efficacy, with 'home medications' pattern achieving highest success rate (6.3%)

Table 2. Model Performance metrics

Key Insights: The disparity between Regex and BERT highlights the importance of context in medical language understanding. BERT's strong contextual capabilities make it the preferred method for complex tasks like medication recognition, while Regex, though simple and useful for specific patterns, lacks the versatility needed for broader clinical applications.

Conclusion

This hybrid approach to medical Question-Answering holds significant promise for improving the processing of clinical data. By combining the reliability of regular expressions with the contextual power of **BERT**, the system can provide accurate and relevant answers to complex medical queries. Future work will focus on refining the model's architecture, expanding the dataset, and improving confidence calibration to enhance the system's robustness and applicability in real-world healthcare settings.

```
Context: Current medications include Metformin 500mg twice daily. Patient reports persistent fatigue and weight loss.
Q: What medications is the patient taking?
A: current medications include metformin 500mg twice daily.
Expected: Metformin 500mg twice daily
Confidence: 76.90%
Similarity to expected: 65.06%
```

```
Context: Patient presents with severe migraine and photophobia since this morning. Currently taking Imitrex 50mg PRN for headaches.
Q: What are the current symptoms?
A: severe migraine and photophobia since this morning
Expected: severe migraine and photophobia
Confidence: 75.00%
Similarity to expected: 76.54%
```

Fig 3. Snapshot of Question – Answer results given by the model

Future Work

We will focus on expanding the training dataset to include a broader range of medical scenarios, symptoms, and conditions, addressing current data sparsity and imbalance. We will also explore advanced techniques for confidence calibration to better align the model's predicted confidence with its actual performance, improving reliability. Additionally, post-processing techniques will be introduced to refine the model's outputs, ensuring accuracy and clinical relevance. In the long term, we aim to develop domain-specific models for different medical subdomains to provide more accurate and contextually appropriate results. The regex component will be further refined to handle complex medical terminology and edge cases, while we also plan to integrate advanced natural language understanding techniques and semantic similarity measures to enhance the model's ability to comprehend the nuances of medical language.

References

- EMR QA dataset: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
- Publication based on: http://aclweb.org/anthology/D18-1258