آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی
Language Processing and Digital Humanities

Drug Question Answering

Final Project - NLP Course - Dr. Asgari

Sara Azarnoush
MohammadReza Daviran
Nona Ghazizadeh
Sina Abdous
Hadis Ahmadian
Mahsa Yazdani

August 2023

Outline

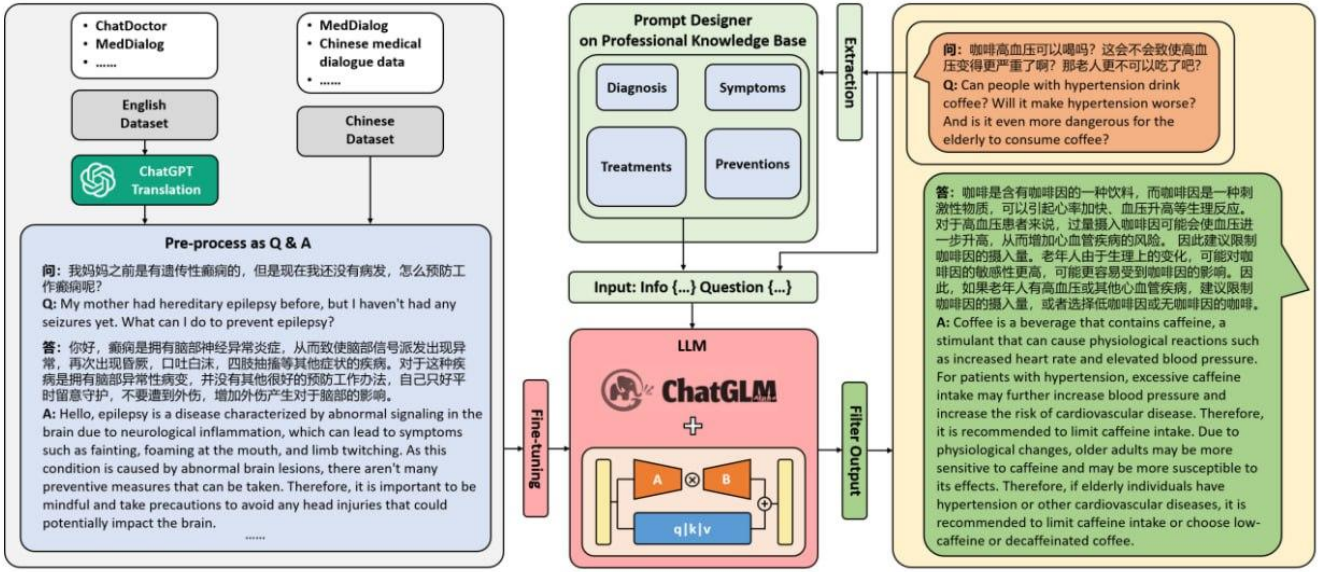
- Introduction
- Datasets
  - PubMed
  - PubMed QA
  - Translation
- Models
  - Translation
  - Embedding
  - Elastic Search
  - Summarization
    - Bert
    - LLM
- Web Based Demo
- Results

Introduction

The Drug Question Answering System presented in this work aims to provide comprehensive answers to queries in both English and Persian languages within the medical domain. The system utilizes a combination of natural language processing techniques, embedding generation, translation, information retrieval, and summarization to achieve its goal.

Related Works

DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task

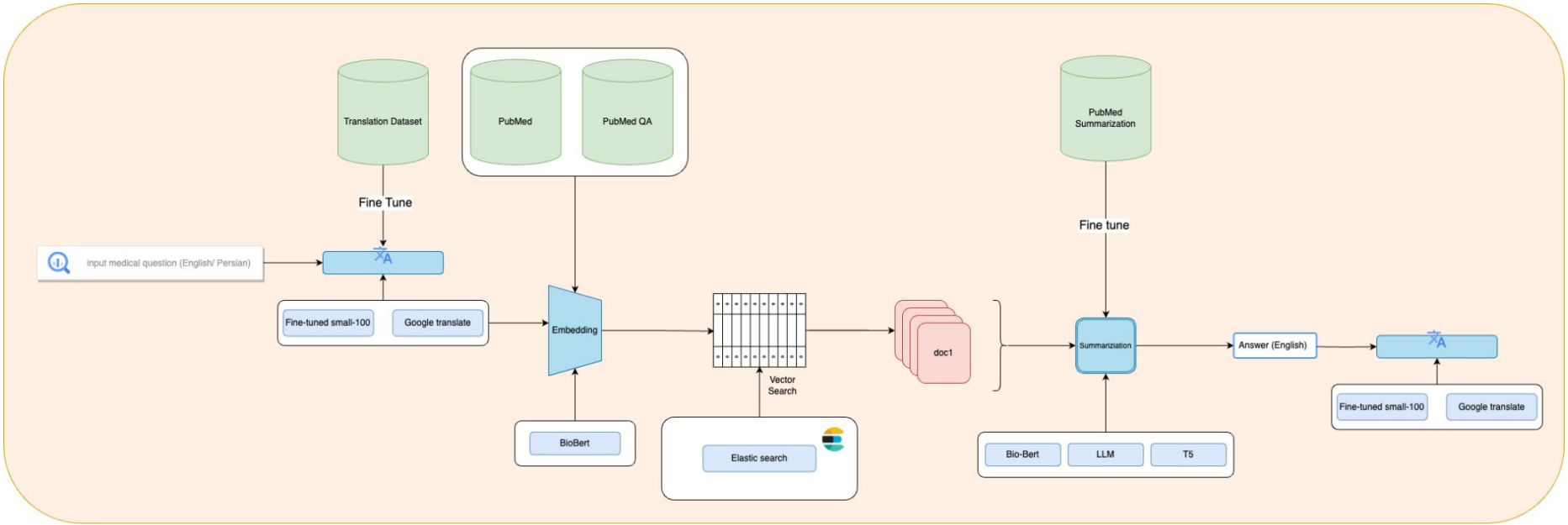


Xiong, Honglin et al. “DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task.” *ArXiv* abs/2304.01097 (2023): n. pag.

Approach

- Create datasets for medical papers and medical question answering
- Create datasets for medical translation
- Finetune translation model
- Create embedding for our data and query
- Use Elastic-Search for vector searching
- Fine-tune summarization model
- Use LLMs for summarization

Pipeline


Translation Dataset
PubMed
PubMed QA
PubMed Summarization
Fine Tune
Fine tune
input medical question (English/ Persian)
Fine-tuned small-100
Google translate
Embedding
Vector Search
doc1
Summarization
Answer (English)
Fine-tuned small-100
Google translate
BioBert
Elastic search
Bio-Bert
LLM
T5

Datasets

Pubmed Dataset

- Free resource for biomedical and life sciences literature search and retrieval.
- Contains over 35 million citations and abstracts of biomedical literature.
- Does not provide full-text journal articles (Links to full text available from other source)
- Exceeds 100GB in size
  - Impractical to download the entire dataset into Google Colab
  - Download focused on abstracts of articles published in 2023
- Initial attempt used E-utilities API for data retrieval.
  - Encountered limitations in data retrieval process.
- Shifted to using EDirect for data retrieval.
  - Utilized EDirect on a Unix system.
- Retrieved data saved into a text file.
- Processed text file into a structured CSV file.

PubMed QA Dataset

- Innovative biomedical question answering (QA) dataset
- Answer research questions with yes/no/maybe
- Utilize corresponding abstracts for answering
- 1k expert-annotated QA instances
- 61.2k unlabeled QA instances
- 211.3k artificially generated QA instances
- Question: Derived from research article titles or content
- Context: Abstract content without conclusion
- Long Answer: Abstract conclusion, likely answers the research question
- Yes/No/Maybe Answer: Summarizes the conclusion's stance

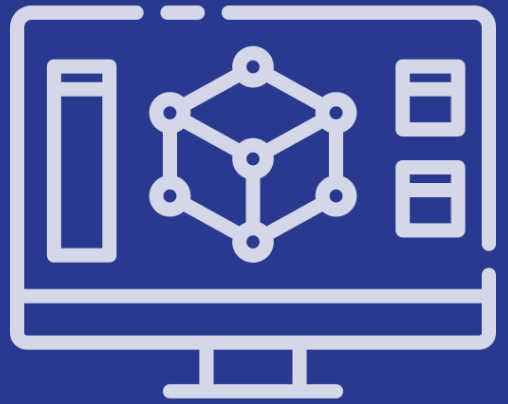
PubMed Summarization Dataset

- Comprises around 133,000 PubMed articles and their abstracts
- Abstracts inherently function as "summaries" of articles
- utilized to fine-tune base models for summarization tasks
- Summarization tasks involve input of biomedical text (article) and output of summary (abstract)
- Used subset of dataset employed due to resource constraints
- Used to fine-tuned T5, BioBart, Falcon-7B models

Translation Dataset

- Translate medical terminology for diseases and drugs between English and Persian
- Scarcity of existing datasets in medical translation domain
- Generate a dataset for fine-tuning translation models
- Curate disease names in English and Persian
- Three synthetic medical sentences in both English and Persian for each disease
- Content covers symptoms, treatments, and pharmaceutical interventions
- Contextually relevant and coherent content ensured by ChatGPT API
- High-quality dataset suitable for training and evaluation

Models



Translation

- English to Persian translation for medical context
- Employed "SMaLL-100" model architecture
    - Designed for low-resource languages and complex translations
- Enhance model's proficiency in translating disease-related text
- Model's refined ability to translate between English and Persian especially in the domain of medical conditions
- Improving language understanding and translation skills in medical context
- Enabling two-way translation for medical content
- Evaluation based on accurate translation of medical terminology and disease descriptions (BLEU)

Translation Evaluation

English to Farsi

BLUE-1	0.60297
BLUE-2	0.49301
BLUE-3	0.43094
BLUE-4	0.39818

Farsi to English

BLUE-1	0.78158
BLUE-2	0.71473
BLUE-3	0.66086
BLUE-4	0.60995

Translation Comparison

Question	
Google Translate	Medical fa_en
نشانه های حمله قلبی چییست	
بر اساس حقایق پزشکی ارائه شده، علائم حمله قلبی شامل آنژین صدری (درد یا ناراحتی قفسه سینه)، ایسکمی بخش ST (که با تغییرات در بخش ST نوار قلب مشخص می شود)، افت فشار خون (فشار خون پایین) و علائم مربوط به فشار خون پایین است. مناسبت ها.	علائم حملات قلبی شامل آنژین (درد سینه یا ناراحتی)، تنگی نفس، درد یا ناراحتی در بدن بالا (مانند دست ها، پشت، گردن، فک یا معده) است، تهوع، خستگی سر و عطسه سرما.
درمان سرطان ریه چیست	
بر اساس حقایق پزشکی ارائه شده، درمان سرطان ریه شامل درمان سیتواستاتیک، درمان تسکینی و رادیوتراپی است. نشان داده شده است که درمان سیتواستاتیک باعث بهبود بقا در بیماران مبتلا به کارسینوم سلول غیر کوچک در مراحل IV یا IIIB با درگیری پلور یا N2-N3 و با وضعیت عملکرد ۲ یا کمتر است. درمان تسکینی برای بیماران مبتلا به سرطان کیسه صفرا غیرقابل برداشت استفاده می شود و هدف آن بهبود علائم و کیفیت زندگی است. رادیوتراپی به عنوان درمان کمکی یا تسکین دهنده برای متاستاز ریه در کارسینوم نازوفارنکس استفاده می شود و نشان داده شده است که بقای کلی را بهبود می بخشد. انتخاب گزینه های رادیوتراپی برای سرطان پروستات به ترجیح بیمار بستگی دارد و برخی از بیماران به دلیل نگرانی در مورد سمیت، دوز کمتری از پرتودرمانی را انتخاب می کنند.	بر پایه اطلاعات پزشکی ارائه شده، درمان سرطان ریه شامل شیمی درمانی، درمان تشعشعی و درمان آرامشی است. شیمی درمانی مزایای درمان پیشرفته سرطان ریه غیر کوچک را نشان داده است، به ویژه در بیماران با سرطان سلول های غیر کوچک در مراحل چهارم یا IIIB با شرکت سلول یا ن۲-ن۲ و با وضعیت عملکرد ۲ یا کمتر است. درمان سیتواستاتیک، نوعی درمان، پیش بینی مستقل از بقاء است. گزینه های درمان آرامشی برای سرطان ریه شامل درمان تشعشعی با دوز مختلف، مانند ۲۵ گی در بخش های ۵ یا ۵۰ گی در بخش های ۲۰ است. انتخاب درمان به وضعیت کلی بیمار بستگی دارد، هیچ تفاوتی در بقاء بین دو گزینه مشاهده نشده است.
علایم سرطان ریه چیست	
علائم سرطان ریه ذکر شده در حقایق پزشکی ارائه شده، تنگی نفس، درد و دیسفاژی است.	علائم سرطان ریه ذکر شده در حقایق پزشکی ارائه شده، تنگی نفس، درد و دیسفاژی است.

Embedding

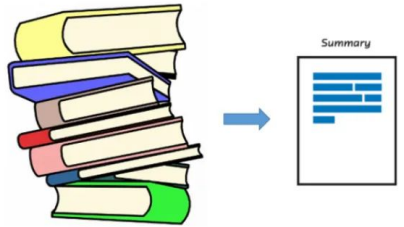
- Read and process context from datasets
- Embed data using BioBert model
- Embed context and translated input question
- BioBert is a specialized variant of BERT (Bidirectional Encoder Representations from Transformers)
- Embeddings are numerical representations capturing semantic meaning and contextual information
- Enable comparison and measurement of similarity between text components

Elastic Search

- Searching among embeddings
- Locate specific embeddings efficiently
- Create a database containing collected embeddings
    - Utilize Elastic Search for database creation
    - A common tool for vector searching due to its efficient indexing process
- Search query embedding within the saved embeddings database
- Elastic Search facilitates this search.
- Elastic Search's indexing algorithms are optimized for vector searching

Summarization - BioBert / T5

- Due to resource constraints, GPT (large model) couldn't be fine-tuned for biomedical QA domain
- Experimented with smaller, fine-tuned language models for summarization tasks
- Choose BioBart and T5 large models for fine-tuning
- Using a dataset of 1000 pubmed articles and their abstracts
  (Pubmed summarization dataset)
- Fine-tuned T5 model demonstrated acceptable and comparable summarization performance to GPT
- GPT still displayed superior performance compared to fine-tuned T5 and BioBart models



Summarization - LLM

- Attempted to enhance model performance using LoRA and Falcon-7b
    - Limited by resource constraints.
- Relied on a dataset of 3000 PubMed articles with abstracts.
- Overcame limitations with GPT-3.5-turbo model API for text summarization.
- Addressed binary questions using summarization outputs
- Utilized Bio-Bart and T5 models alongside GPT model
- Meticulously compared and contrasted summarized outputs
    - GPT-based approach consistently demonstrated superior accuracy

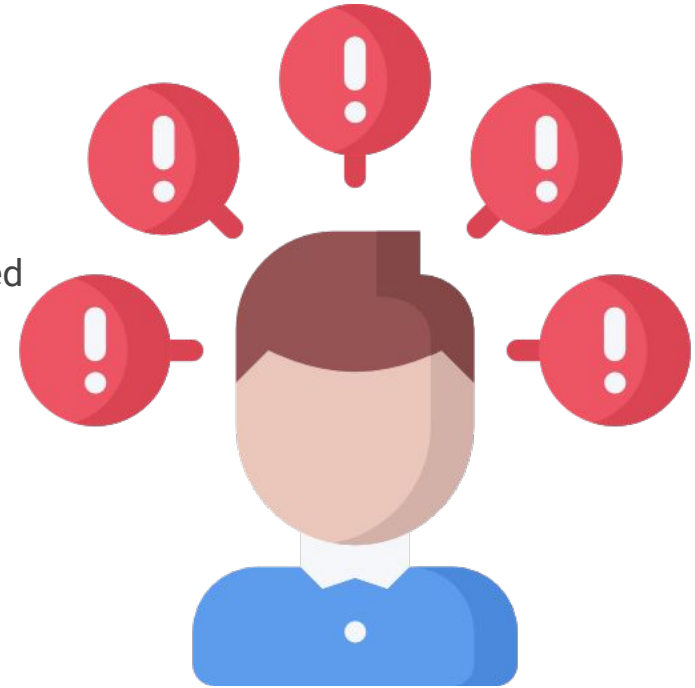
Summarization Evaluation

Comparison of different summarization components with 10 samples

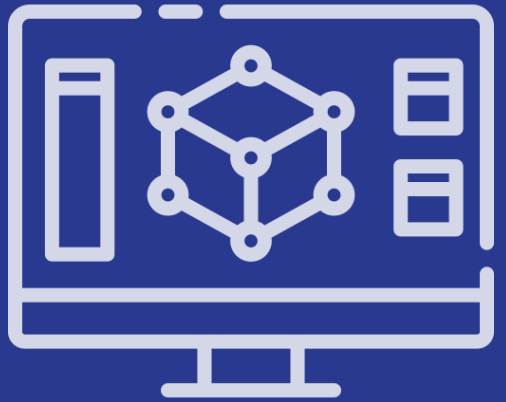
Model	Accuracy
GPT	0.7
BioBert	0.4
T5	0.6

Proposed Model Problems

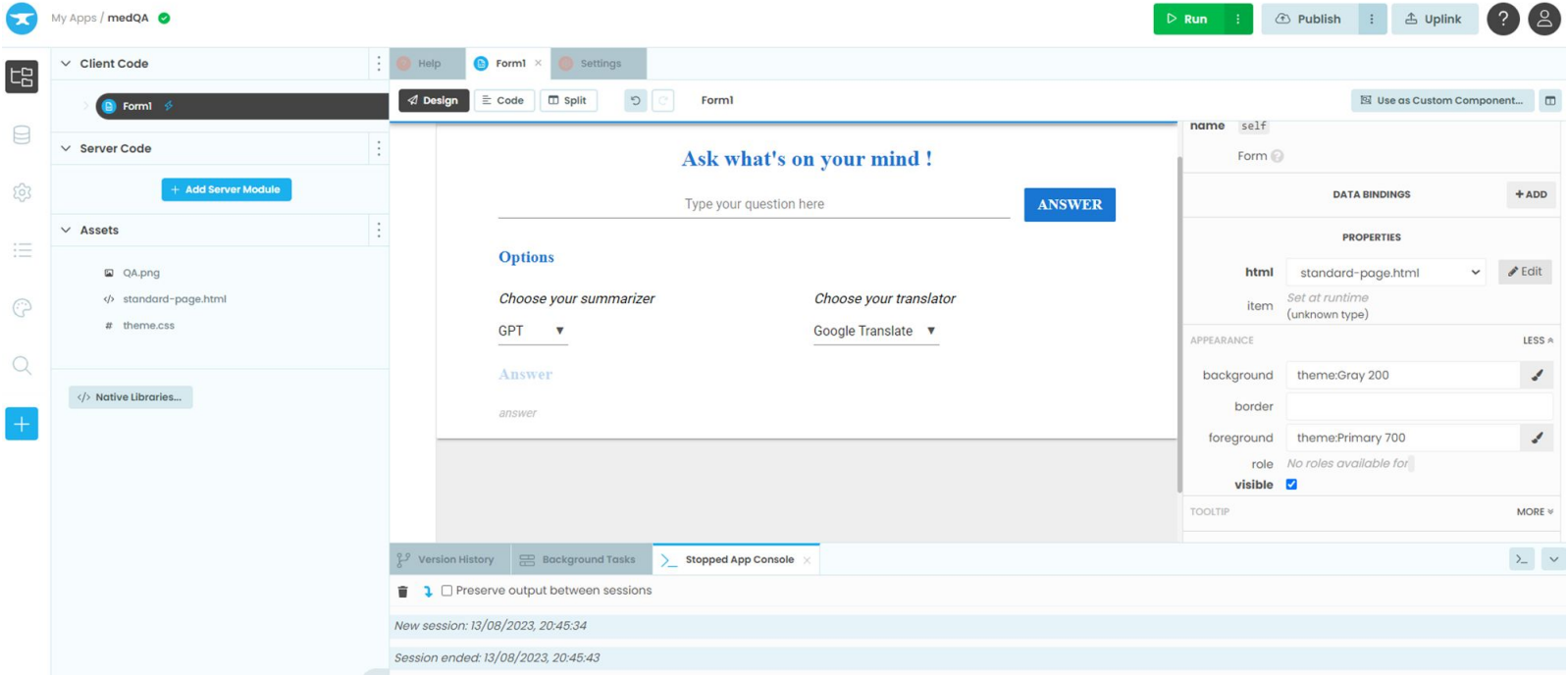
- Resource Problems
  - Generating datasets
  - Training Process
  - Cannot Make Architecture More Complicated
  - Storage



Web Based Demo



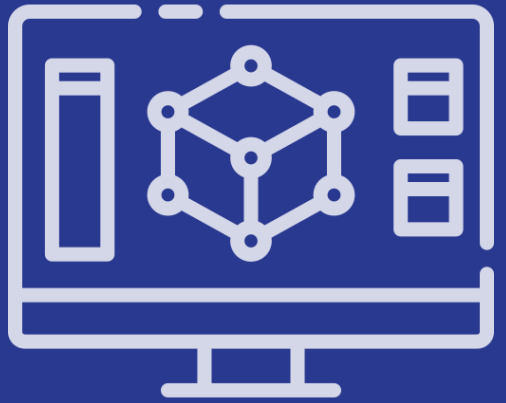
Web-Based Demo



Web-Based Demo (English)

Web-Based Demo (Farsi)

Evaluation



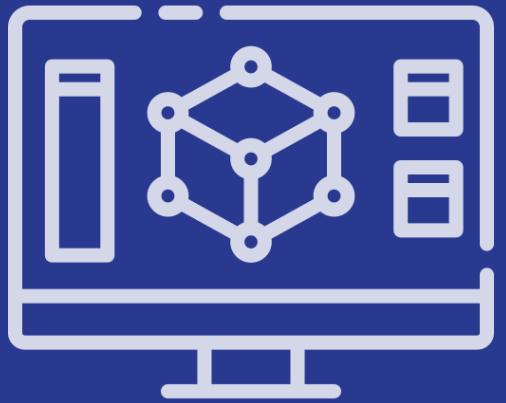
Results

Comparison of overall performance with different summarization components with 50 samples on PubMed QA

yes/no/maybe questions

Summarization Model	Pipeline Accuracy
GPT	0.61
BioBert	0.28
T5	0.39

Conclusion
&
Future Work



Conclusion & Future Works

- Work on Architecture of Models: better LLMs
- Data: Farsi data, process
- Train our model on more data
- Improve Web Based Demo of Models
- bias , security, validity

References

- Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages
- Evidence Extraction to Validate Medical Claims in Fake News Detection
- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Any Questions?

Thank you