Group 1

Medical AI QA System

Problem Statement

- Many people find it hard to get the health advice they need.
- Reasons include the high cost of seeing a doctor, medical information that's difficult to understand, and not being able to get advice quickly.
- This situation creates a gap in healthcare where people can't get the help they need when they need it.

Motivation

- We're inspired to use new NLP technology to close this gap.
- Our project is about making a Question/Answer system that can talk to people and give them health advice right away.
- By doing this, we want to make it easier for everyone to get the health information they need without the high cost or long wait.
- This system will also help doctors and nurses by answering common questions, allowing them to focus on more critical tasks.

Related Papers

1. The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications
2. Task-oriented Dialogue System for Automatic Diagnosis
3. MIE: A Medical Information Extractor towards Medical Dialogues
4. Building blocks of a task-oriented dialogue system in the healthcare domain
5. A Multi-Persona Chatbot for Hotline Counselor Training
6. PlugMed: Improving Specificity in Patient-Centered Medical Dialogue Generation using In-Context Learning

PlugMed: Improving Specificity in Patient-Centered Medical Dialogue Generation using In-Context Learning

- PlugMed is a medical dialogue system that enhances the specificity of large language model (LLM) responses through in-context learning. It consists of two main components:
- A Prompt Generation (PG) module that retrieves relevant example dialogues from global and local views to generate prompts for the LLM.
- A Response Ranking (RR) module that selects the best response from the LLM's outputs using a fine-tuned small language model.
- New automatic evaluation metrics, such as intent accuracy and high-frequency medical term accuracy, are introduced to assess specificity. Experiments show that PlugMed improves the specificity of LLM responses by generating more accurate intents and medical terminology compared to baselines, as confirmed by human evaluations.

A Multi-Persona Chatbot for Hotline Counselor Training

- The paper proposes "Crisisbot", a chatbot that simulates hotline visitors with different personas to train human counselors in a realistic, low-risk environment. The authors develop a counselor strategy annotation scheme and a multi-task training framework to enable Crisisbot to generate persona-relevant responses.
- The multi-task framework consists of a Prompt Generation Module that retrieves relevant example exchanges, and a Response Ranking Module that ranks response candidates generated by a large language model. Automatic evaluation shows increased diversity and persona-relevance compared to baselines.
- Human evaluation reveals a discrepancy between crowdworkers' and experienced counselors' preferences. Counselors emphasize the importance of response variety for effective training, despite the need for improved conversation flow.
- The key contribution is the multi-task framework leveraging counseling strategies to curate varied personas, evaluated with specificity metrics. The mixed results highlight the importance of involving target users during system development.

Building Blocks of a Task-Oriented Dialogue System in the Healthcare

- The paper introduces a framework for healthcare dialogue systems, focusing on tasks like triage and diagnosis.
- It highlights three key areas: privacy in data collection, dialogue management, and evaluation.
- For privacy, it suggests creating simulated dialogue data with expert knowledge and crowdsourcing.
- The system uses a Reinforcement Learning (RL) agent, trained with simulations and crowdsource data.
- Evaluations use success rates and user feedback to ensure the system is both effective and user-friendly.

Task-oriented Dialogue System for Automatic Diagnosis

- The paper presents a dialogue system aimed at automating medical diagnoses by conversing with patients to gather symptoms.
- It utilizes a task-oriented dialogue framework that improves diagnosis accuracy by collecting additional symptoms during interactions.
- The system's core includes Natural Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation (NLG), with a focus on DM employing reinforcement learning.
- Experiments demonstrate the system's effectiveness in symptom collection, showcasing superior performance over baseline models.

MIE: A Medical Information Extractor towards Medical Dialogues

- MIE (Medical Information Extractor) is a deep matching model designed specifically for extracting medical data from doctor-patient dialogues.
- Its architecture consists of four key components: Annotation, Encoder, Matching, and Aggregate Modules, each addressing distinct challenges in medical dialogue interactions.
- Annotation Module employs a sliding window approach for precise labeling of relevant information within the dialogue.
- Encoder Module utilizes Bi-LSTM with self-attention mechanisms to effectively encode dialogue turns and capture nuanced information.
- Evaluation results demonstrate promising outcomes, particularly with the MIE-multi model, showcasing its efficacy in accurately extracting medical information from doctor-patient interactions for Electronic Medical Records (EMRs) conversion.

The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications

- The paper examines task-oriented dialogue systems in healthcare, focusing on their construction, management, and evaluation processes.
- Two primary approaches to designing these systems are outlined: the pipeline approach, which involves distinct components handling different tasks, and the end-to-end approach, where a single model manages all tasks simultaneously.
- Dialogue management strategies are discussed, including rule-based, intent-based, hybrid, and corpus-based approaches, each offering distinct methods for decision-making and system functionality.
- Modality, referring to how users interact with the system (e.g., text, speech, graphical interfaces), is explored, highlighting its impact on interaction quality.
- Evaluation methods, encompassing human feedback and automated measurements, are investigated to assess system performance, offering subjective insights into user satisfaction and objective metrics such as task completion rates and response times.

Solution Requirements

1. Deploy NLP to efficiently handle patient queries
2. Enhance speed and accuracy of information retrieval
3. Categorize database into bins or vector repositories
4. Use classifier to determine appropriate repository based on user input
5. Create efficient prompt templates for smooth user interactions
6. Fine-tune LLM with Patient QA dataset to improve responses
7. Experiment to find best classifier and LLM models
8. Assess feature impact via ablation study
9. Develop web chat interface for patients, researchers, and doctors

Architecture

Our proposed solution is centered around leveraging the capabilities of Large Language Models (LLM), with a focus on employing Long Short-Term Memory (LSTM) network as our baseline architecture. This section outlines the models we plan to use and integrate into our system to enhance its performance in understanding and responding to medical queries

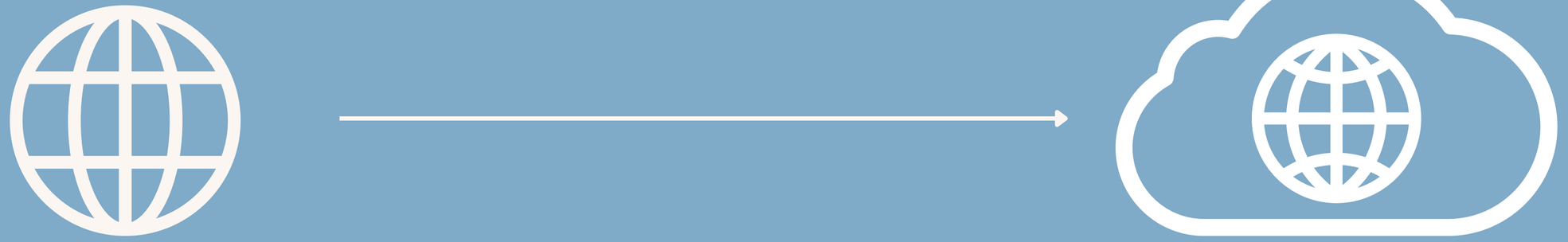
Experiment Design

We begin by combining the datasets from Hugging Face's "medical-qa-datasets" and Kaggle's "diagnoise-me" to form a comprehensive corpus for training and evaluation
The dataset shall be transformed into a vector database and subjected to a series of preprocessing steps to distill the necessary information, which will subsequently be furnished to our Natural Language Processing (NLP) model. The system's classifier is designed to partition the vector database, enabling the model to exclusively access information from the pertinent dataset. Furthermore, the Large Language Model (LLM) will undergo fine-tuning through an ablation study to achieve the desired outcomes. Comparative analysis will be conducted between the results obtained from the LLM and those derived from the baseline model (Long Short-Term Memory, LSTM), along with additional models, to ascertain their relative efficacy.

Evaluation Metrics

1. Intent and Entity Classification Metrics: Accuracy, Precision, Recall, and F1-score for measuring the correctness of intent classification.
2. Response Generation Metrics: BLEU, Perplexity, and Coherence Score for assessing the quality and relevance of generated responses.
3. Dialogue Quality Metrics: Task Completion Rate, Average Turns per Dialogue, and User Satisfaction Score for evaluating the effectiveness and user satisfaction of the conversation.
4. Human Evaluation Metrics: Intelligibility, Appropriateness, and Engagement for qualitative assessment of the chatbot's responses and user experience.

Website



Website

Flask

Hosted

Cloud Server

Task Contribution

Dataset preparation and cleaning	Model	Web Development
Myo Thiha Kaung Htet cho	All of us	Rakshya

Group 1

Thank you very much!