MAI 475– Large Language Model

IV MSAIM

01-07-2024

Regular lab Question

Lab Exercise 4:

Implement a Comparative Study of Foundation, Indic, and International Language Models in Voice-Based Document QA Systems

Objective:

Implement and analyze the performance of three transformer-based language models:

1. **A Foundation Model for English** (e.g., FLAN-T5, GPT, or BERT)
2. **An Indic Language Model** (e.g., IndicBERT, MuRIL, or KannadaBERT)
3. **A Non-English International Language Model** (e.g., CamemBERT for French, BETO for Spanish, or GermEval for German)

You will build a system that:

- Accepts **PDF documents as input**
- Takes **questions via voice input**
- Extracts **relevant answers from the document**
- Returns **voice output in the respective language**

Tasks:

i. Model Selection and Description

- Choose one model each from:
  - Foundation: FLAN-T5, GPT-3.5, or BERT
  - Indic: IndicBERT (or language-specific model like MalayalamBERT, KannadaBERT)
  - International (non-English): BETO (Spanish), CamemBERT (French), or GePpeTto (Italian)
- Explain the background, training corpus, and primary use-cases of each.

ii. Implementation Task

- Accept a **PDF document** in each language (English, one Indic language, and one non-English international language).
- Process the document and **embed it for question answering** using the selected model.
- Use **voice input** for questions (speech-to-text).
- Generate answers and **convert back to speech** (text-to-speech).
- Use the **same prompt/question in all three languages**. (You can use Machine Translation and check the quality of completion/answers)

iii. Dataset and Preprocessing (Optional)

- Use existing documents or create synthetic PDF content in English, Indic (e.g., Kannada), and International (e.g., French).
- Detail your preprocessing: PDF parsing, tokenization, text normalization, truncation/padding, prompt design, etc.

iv. Evaluation and Results

Outputs:

- Provide at least **3 examples of question-answer interactions** per language.
- Add **screenshots or transcripts** of the voice inputs and outputs.
- Describe any **differences in language structure, extraction quality, and response accuracy**.

v. Evaluation Metrics:

- **ROUGE / BLEU** scores (if comparing generated text to reference)
- **Human Evaluation:**
    - Coherence / Relevance / Fluency /Voice output clarity and pronunciation accuracy

vi. Prepare a comparison table as follows (flexible aspects):

Aspect	Foundation Model (English)	Indic Model (e.g., Kannada)	International Model (e.g., French)
Language Fluency	✔ / ✘	✔ / ✘	✔ / ✘
Named Entity Handling	✔ / ✘	✔ / ✘	✔ / ✘
Accuracy of Extraction	High / Medium / Low	High / Medium / Low	High / Medium / Low
Speech-to-Text Accuracy	✔ / ✘	✔ / ✘	✔ / ✘
Text-to-Speech Quality	Good / Moderate / Poor	Good / Moderate / Poor	Good / Moderate / Poor

NB: No marks will be credited for the pipeline implementation.

Program Evaluation Rubrics

Model Selection and Implementation	**6 Marks**
Timely Submission	**2 Marks**
Viva	**2 Marks**

General Instructions

- The file you have to save with your name, last 3 digits of register number and program number “Anto_501_Lab1”.
- The implemented code you have to download and upload in the Google Class room in the given scheduled time.