Multimodal Text Processing System: An Integrated Approach

Heet Manish Kanani, Alp Yalcinkaya, Purva Agarwal,
Utkarsh Gogna, Yi Zu

September 2024

1 Project Overview

This project is aimed at building an all-in-one system that brings together several important NLP tasks such as text summarization, speech-to-text translation, and text classification based on transformer models such as BERT, T5, Whisper, and pre-trained LLM model for multimodal data processing. This approach is expected to be more effective in not only presenting the data in different ways but also in creating poetry, lyrics, automatic transcription, summarization, and many other applications.

Key features of the system include:

- **Speech-to-Text Translation:** The project demonstrates the practical lingual capacity of the Whisper technology as a tool for decoding language automatically and effectively in multilingual communication.
- **Text Summarization:** Developing advanced transformer architectures such as BART that prove their worth with superior results in many text abstract topics like summarization.[1].
- **Text Classification:** The LSTM-CRF model has demonstrated state-of-the-art performance in Named Entity Recognition (NER) without handcrafted features [2].

2 Related Work

Each of these fields has made a tremendous amount of progress individually:

- **Speech-to-Text:** Whisper and Google's ASR are examples of the best working systems for such tasks. Whisper's multilingual fast learning has been easily shown, and zero-shot data has been directly given without fine-tuning [3].
- **Text Summarization:** BART, a sequence-to-sequence model, is a promising technology because it improves the quality of summaries for abstractive summarization tasks. BART excels in summarizing highly abstractive datasets such as XSum, with gains of up to six points in the ROUGE score [1].
- **Text Classification:** The LSTM-CRF model offers better performance than RNN-CRF for most classification tasks in the Russian corpus, even when using a smaller language model [2].

Few projects have tried incorporating these tasks into one unified system.

3 Proposed Methods

We propose to develop a multimodal text processing system that integrates speech-to-text translation, text summarization, and text classification into a seamless workflow. The system will utilize Whisper technology for accurate transcription of multilingual audio inputs, allowing for recognition across diverse accents and varying noise levels. We will explore pre-trained models to enhance transcription quality and may incorporate the Google Speech-to-Text API for real-time capabilities in commercial settings.

Once the speech is transcribed into text, the pre-trained LLM model will be employed to generate concise summaries, distilling the main points from the transcriptions into more digestible formats. This

process will involve fine-tuning a pre-trained LLM model on relevant datasets to optimize its performance for the specific contexts we encounter.

Following summarization, we will implement the LSTM-CRF model for text classification and named entity recognition (NER). This model excels at identifying entities such as names, dates, and locations within the summarized text, leveraging contextual relationships captured by LSTM while ensuring accurate sequence labeling through the CRF layer.

The overall workflow will function as follows: audio input will be transcribed into text using Whisper, which will then be summarized using a pre-trained LLM model, followed by classification and entity extraction with the LSTM-CRF model. Each component will be evaluated for effectiveness through performance metrics such as Word Error Rate (WER) for transcription, ROUGE scores for summarization, and precision, recall, and F1 scores for classification. This integrated approach aims to create a robust system that streamlines the conversion of speech into actionable insights, suitable for applications such as customer support, content generation, and data analysis.

4 Evaluation Metrics

Each component of the system will be evaluated using the following metrics:

- **Speech-to-Text (S2T):** The performance of the Whisper model will be measured using the Word Error Rate (WER) [3], defined as:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Words Spoken}}$$

  WER measures the percentage of incorrect words in the transcriptions compared to the actual spoken words, ensuring the accuracy of speech-to-text translation.

- **Text Summarization (TS):** The performance of the pre-trained LLM model summarization model will be evaluated using ROUGE scores [4]. ROUGE-N (n-gram overlap) and ROUGE-L (longest common subsequence) are key metrics:
  - **ROUGE-N**: Measures the overlap of n-grams (e.g., unigram, bigram) between the generated summary and the reference summary.
  - **ROUGE-L**: Measures the longest common subsequence (LCS) between the generated summary and the reference, providing insight into the fluency and coverage of the summary.

- **Text Classification (TC):** The classification model's performance will be assessed using:
  - **Accuracy**: The ratio of correct predictions to the total predictions.
  - **Precision**: The ratio of true positive predictions to the total number of positive predictions.
  - **Recall**: The ratio of true positive predictions to the total number of actual positive instances.
  - **F1 Score**: The harmonic mean of Precision and Recall, providing a balanced view of performance [2].

5 Timeline

- Weeks 1-2: We will begin by integrating and testing one model (either the speech-to-text model using Whisper, the text summarization model using BART, or the classification model using LSTM-CRF). This initial implementation will help us assess the feasibility and performance of our approach.

- Weeks 3-4: If the first model implementation is successful, we will proceed to implement the second model. This could involve either fine-tuning and validating the text summarization model or the classification model, depending on which was implemented first.

- Weeks 5-6: Following the successful integration of the second model, we will focus on implementing the third model. This phase will include fine-tuning and testing to ensure all components work well together.

- Weeks 7-8: Finally, we will develop and test the complete system with multimodal input, ensuring seamless integration between the speech-to-text, summarization, and classification processes.

References

[1] Lewis, Mike and Liu, Yinhan and Goyal, Naman and Ghazvininejad, Marjan and Mohamed, Abdelrahman and Levy, Omer and Stoyanov, Ves and Zettlemoyer, Luke, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, arXiv preprint arXiv:1910.13461, 2019.

[2] Lample, Guillaume and Ballesteros, Miguel and Subramanian, Sandeep and Kawakami, Kazuya and Dyer, Chris, *Neural Architectures for Named Entity Recognition*, arXiv preprint arXiv:1603.01360, 2016.

[3] Radford, Alec and others, *Robust Speech Recognition via Large-Scale Weak Supervision*, arXiv preprint arXiv:1910.13461, 2019.

[4] Lin, Chin-Yew, *ROUGE: A package for automatic evaluation of summaries*, In Text Summarization Branches Out, pages 74–81, 2004.