SPEECH TO TEXT AND NAMED ENTITY RECOGNITION MODELS DOCUMENTATION

Sections:

1. Project Description
2. Researches
3. Pipeline
4. Model and Dataset
5. Challenges and Solutions
6. Results
7. Further Plans

Project Description

This project focuses on developing and fine-tuning Speech-to-Text (STT) and Named Entity Recognition (NER) models for Uzbek. The STT model is based on Whisper v2 and has been fine-tuned using LoRA on audio samples to enhance transcription accuracy for the target language. The NER model has been built by fine-tuning Multilingual RoBERTa to recognize and classify named entities in text.

Researches

Some researches was applied before starting projects. Several existence projects related to STT (uzbekvoice.ai, aisha.group) was compared for the performance and quality. According to NER, many pre-trained models (xlm-RoBERTa, mBERT) was studied to determine how model perform on Uzbek language.

Pipeline

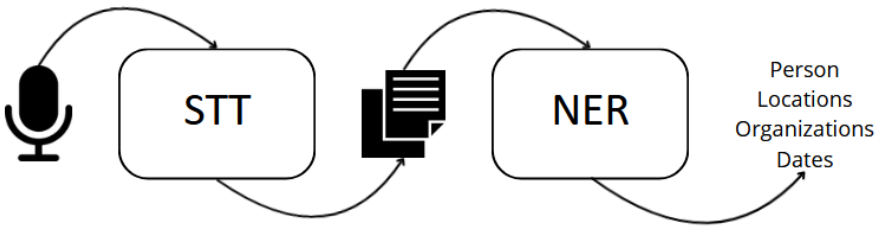
This project integrates Speech-to-Text and Named Entity Recognition into a seamless pipeline for automatic transcription and entity extraction. The pipeline follows these steps:

Speech-to-Text (STT):

Audio input is processed using a fine-tuned STT model, converting speech into text with high accuracy.

Named Entity Recognition (NER):

The transcribed text is then passed to NER model, which identifies and classifies entities such as names, locations, organizations, and dates.



Models and Dataset

For Named Entity Recognition (NER), we utilize XLM-RoBERTa-Large with LoRA (Low-Rank Adaptation) to efficiently fine-tune the model on a custom Uzbek NER dataset hosted on Hugging Face. XLM-RoBERTa, a multilingual transformer-based model, offers robust language representation across multiple languages, including Uzbek. Using LoRA reduces computational cost while maintaining performance, making the model more adaptable for resource-constrained environments.

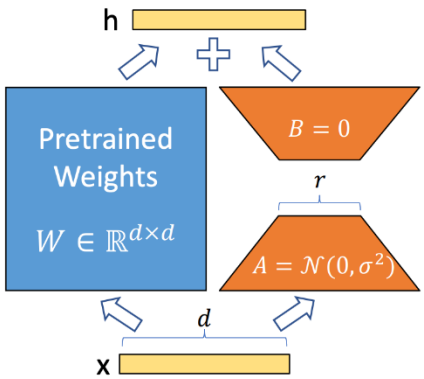
For Speech-to-Text (STT), we used Whisper-Large-v2 with LoRA, leveraging the Common Voice 13 Uzbek dataset. Whisper, developed by OpenAI, is a state-of-the-art automatic speech recognition (ASR) model that performs well across various languages and accents. Fine-tuning with LoRA allows domain adaptation while maintaining Whisper's efficiency in transcribing Uzbek speech accurately.

Problems

One of the key challenges in training and fine-tuning large NLP models for Uzbek is the lack of computational resources. Training XLM-RoBERTa-Large and Whisper-Large-v2 requires substantial GPU, CPU, and RAM, which can be a limiting factor, especially when working with large datasets like Common Voice 13 for STT and custom NER datasets. High memory consumption and long training times make it difficult to fine-tune these models on standard hardware, necessitating efficient optimization techniques.

Solution

To overcome these resource constraints, we leverage LoRA (Low-Rank Adaptation), a parameter-efficient fine-tuning method. LoRA significantly reduces memory usage by freezing the pretrained model weights and injecting trainable low-rank matrices into specific layers. This allows fine-tuning large models like XLM-RoBERTa-Large and Whisper-Large-v2 without the need for extensive GPU resources, making the process more efficient and accessible.



Result

- **STT Model: WER (Word Error Rate) = 47.4**
  Normalized WER = 33.1

- **NER Model: Accuracy = 93.5 %**
  Precision = 76.5 %
  Recall = 73.1 %
  F1-score = 74.8 %

Further Plans

For NER, we plan to further improve entity recognition for Uzbek by:

- Expanding the training dataset with more domain-specific data (e.g., legal, medical, or news texts).
- Exploring adapter-based fine-tuning to enhance model efficiency.
- Deploying the fine-tuned NER model as an API for broader usability in real-world applications.

For STT, future improvements include:

- Enhancing transcription accuracy by fine-tuning Whisper-Large-v2 on additional dialects and accents.
- Implementing post-processing techniques (e.g., spell-checking and punctuation restoration) to refine outputs.
- Deploying the STT model in a real-time speech recognition system for Uzbek language applications.

Resources for Learning

- Huggingface NLP Course
- Huggingface Audio Course
- Deeplearning.AI : Natural Language Processing Specialization
- StatQuest with Josh Starmer