Vietnamese Text-to-Speech and Named Entity Recognition (NER) Report

Abstract

This report covers two tasks: the development of a Vietnamese Text-to-Speech (TTS) system and a Named Entity Recognition (NER) system for resume extraction. The TTS system aims to convert Vietnamese text into speech, while the NER system focuses on extracting names and emails from CVs in PDF format. Both tasks use machine learning models and NLP techniques such as SpaCy, Regex, and gTTS for TTS. The report outlines the methodology, challenges, and future improvements for both systems.

Task 1: Named Entity Recognition (NER) – Resume Extraction

Objective

The objective of this task is to extract names and emails from CVs in PDF format using NLP tools, specifically SpaCy and Regex.

Methodology

The following steps were taken to build the NER system:

1. PDF Text Extraction: Used PyMuPDF (fitz) to extract raw text from PDF documents.

2. Name Extraction: Applied SpaCy's `en_core_web_sm` model and filtered entities labeled as `PERSON`.

3. Email Extraction: Used Regular Expressions (Regex) to capture common email formats.

Example Output

Using the test file `cv_1.pdf`, the system extracted the following information:

- Name: Nguyen Thanh Xuan

- Email: xuanusm@gmail.com

Sample Output:

📄 File: cv_1.pdf
👤 Name: Nguyen Thanh Xuan
📧 Email: xuanusm@gmail.com

Challenges

The challenges encountered during the NER task include:

1. Inconsistent Layouts: Resumes with varied layouts may affect SpaCy's ability to accurately recognize names and entities.

2. Missing or Split Names: SpaCy may miss names or split them incorrectly if uncertain.

3. Non-standard Email Formats: Some emails may not match common formats captured by simple regex.

Future Improvements

The following improvements could be implemented:

1. Train a Custom NER Model for Vietnamese names using SpaCy or HuggingFace Transformers.

2. Improve Email Detection using multiple regex checks or layout heuristics.

3. Add PDF Layout Parsing to handle complex CV layouts, such as tables or headings.

Task 2: Text-to-Speech (TTS) for Vietnamese

Objective

The objective of this task is to develop a Vietnamese Text-to-Speech (TTS) system that converts text into speech using NLP tools and the gTTS library.

Methodology

The following steps were taken to build the TTS system:

1. Text Normalization: Handle numbers, abbreviations, and special characters.

2. Grapheme-to-Phoneme Conversion (G2P): Use a simplified rule-based approach to convert Vietnamese words into phonemes.

3. Speech Synthesis: Use the gTTS library to convert normalized text into speech and save it as an MP3 file.

Example Output

For the input text 'Xin chào mọi người.', the following steps were executed:

1. Text Normalization: 'Xin chào mọi người' (no change in this case).

2. Phoneme Conversion: 's i n c h a o m o i n g u o i'.

3. Speech Synthesis: The speech was synthesized and saved as `output.mp3`.

Sample Output:

Audio saved as output.mp3

Challenges

The challenges encountered during the TTS task include:

1. Phoneme Mapping: Vietnamese phonemes, especially tones, require a complex model for accurate conversion.

2. Limited Data for Vietnamese TTS: The gTTS library uses a generic voice model that may not fully capture Vietnamese speech nuances.

3. Pronunciation Variability: Regional accents in Vietnamese may lead to pronunciation differences.

Future Improvements

The following improvements could be implemented:

1. Custom Phoneme Model: Train a more sophisticated G2P model for Vietnamese.

2. Tonal Speech Synthesis: Train on a tonal dataset to better capture the six tones of Vietnamese.

3. Emotion Modeling: Add emotion detection to generate more natural speech.

4. Regional Accent Support: Train accent-specific models for better pronunciation handling.