CHUYỂN GIỌNG NÓI THÀNH VĂN BẢN TIẾNG VIỆT ĐA PHƯƠNG NGỮ DỰA TRÊN HỌC CÓ GIÁM SÁT VỚI KIẾN TRÚC TRANSFORMER

Hoàng Tiến Dũng - 240101006

Tóm tắt

- Lớp: **CS2205.CH183**
- Link Github của nhóm: https://github.com/htdung167/CS2205.CH183
- Link YouTube video: https://www.youtube.com/watch?v=LXraMvyJpAU
- Ảnh + Họ và Tên của các thành viên:

  Hoàng Tiến Dũng - 240101006

Giới thiệu

- Chuyển giọng nói thành văn bản (Speech-to-Text, STT) đã có **nhiều ứng dụng trong nhiều lĩnh vực quan trọng:** trợ lý ảo, tổng hợp phụ đề tự động, dịch thuật thời gian thực và hỗ trợ tiếp cận cho người khuyết tật.
- Các mô hình STT cho tiếng Anh đã đạt kết quả chính xác cao.
- Mô hình STT cho tiếng Việt đã có nhiều tiến bộ, tuy nhiên vẫn gặp thách thức trong việc **xử lý sự đa dạng các phương ngữ địa phương**. Chưa có nhiều mô hình và bộ dữ liệu đa phương ngữ.

  → Xây dựng bộ dữ liệu và mô hình cho Chuyển đổi giọng nói thành văn bản tiếng Việt đa phương ngữ

❖ **Đầu vào:** Một đoạn âm thanh tối đa 30 giây có giọng nói của người.

❖ **Đầu ra:** Một đoạn chuỗi ký tự tương ứng với giọng nói trong âm thanh

Mục tiêu

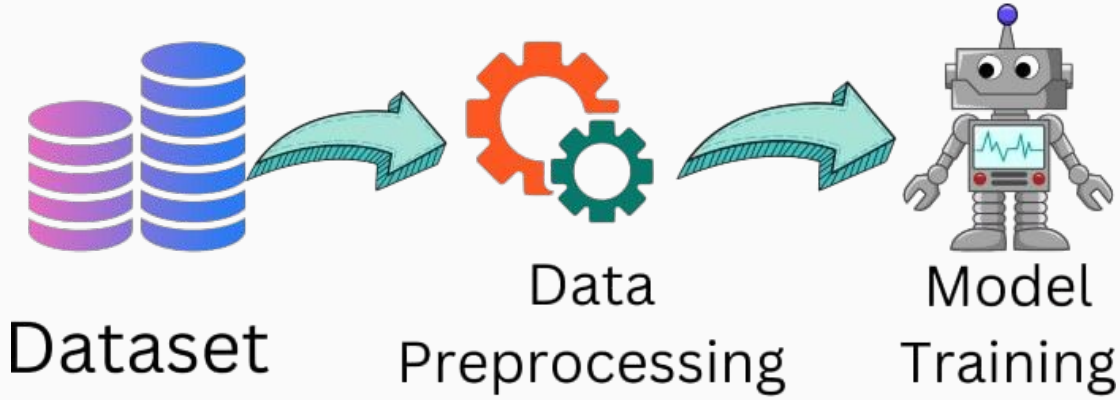
- **Mục tiêu 1:** Xây dựng bộ dữ liệu BanhMi Multi-Dialect Dataset **hơn 200 giờ ghi âm của ba phương ngữ**: **phương ngữ Bắc (Bắc Bộ), phương ngữ Trung (Bắc Trung Bộ), phương ngữ Nam (Nam Trung Bộ và Nam Bộ)** ở dạng **"văn viết"**.
- **Mục tiêu 2:** Huấn luyện **mô hình BanhMiWhisper** - mô hình chuyển giọng nói sang văn bản đa phương ngữ cho Tiếng Việt. Đảm bảo mô hình cải thiện chỉ số Word Error Rate (WER) và Character Error Rate (CER) trên tất cả các bộ đánh giá tiếng Việt nói chung, tiếng Việt đa phương ngữ nói riêng.
- **Mục tiêu 3:** Áp dụng mô hình vào một ứng dụng cụ thể: **Chatbot**. Xây dựng ứng dụng để chạy thử nghiệm mô hình có giao diện trực quan, thân thiện với người dùng.

Nội dung và Phương pháp

Nội dung 1: Thực hiện thu thập dữ liệu âm thanh tiếng Việt ở dạng thô.

Phương pháp thực hiện:

- Sử dụng các công cụ tự động (Selenium, BeautifulSoup) để thu thập dữ liệu âm thanh từ Youtube, Tiktok, Facebook, các trang sách nói, kể chuyện và các kênh truyền hình địa phương từ nhiều tỉnh thành.
- Áp dụng các kỹ thuật tiền xử lý đơn giản, có quy luật để **tiền xử lý bước đầu** như rút trích âm thanh từ video, xác định có tồn tại giọng nói trong video.



Nội dung và Phương pháp

Nội dung 2: Xây dựng quy trình tự động hóa tạo bộ dữ liệu đa phương ngữ tiếng Việt BanhMi Multi-Dialect Dataset.

Phương pháp thực hiện:

- Phát triển một pipeline tích hợp các module xử lý dữ liệu tự động:
  - Phân đoạn vùng có giọng nói
  - Rút trích văn bản với độ chính xác cao bằng cách kết hợp nhiều mô hình STT sẵn có
  - Lọc nhiễu và chú thích thời gian, mỗi phân đoạn không quá 30 giây
  - Đảm bảo đủ và cân bằng ba phương ngữ: phương ngữ Bắc (Bắc Bộ), phương ngữ Trung (Bắc Trung Bộ), phương ngữ Nam (Nam Trung Bộ và Nam Bộ)
- Kiểm tra ngẫu nhiên tập con để đảm bảo chất lượng. Chia bộ dữ liệu thành 3 phần: training, validation, testing với tỷ lệ hợp lý.

Nội dung và Phương pháp

Nội dung 3: Huấn luyện và đánh giá mô hình BanhMiWhisper.

Phương pháp thực hiện:

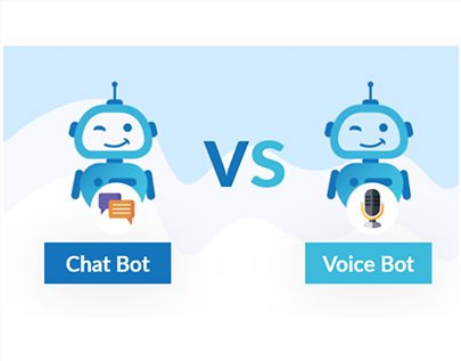
- Sử dụng bộ dữ liệu BanhMi Multi-Dialect đã xây dựng để huấn luyện mô hình BanhMiWhisper từ các mô hình Whisper (5 phiên bản), PhoWhisper (5 phiên bản) và Whisper Turbo.
- Thí nghiệm với các chiến lược:
    - Thay đổi kiến trúc mô hình
    - Kết hợp huấn luyện trộn lẫn cùng với các bộ dữ liệu công khai tiếng Việt hiện nay
    - Các cách phân bổ độ dài âm thanh khác nhau
- So sánh kết quả mô hình BanhMiWhisper với các mô hình STT hiện có để đánh giá ưu nhược điểm. Các chỉ số đánh giá gồm WER và CER trên các bộ kiểm thử của các bộ dữ liệu tiếng Việt hiện nay (Vivos, Common Voice Vi, VLSP 2020, ViMD) và tập kiểm thử của BanhMi Multi-Dialect.

Nội dung và Phương pháp

Nội dung 4: Xây dựng ứng dụng Chatbot có sử dụng BanhMiWhisper như một phần của hệ thống.

Phương pháp thực hiện:

- Phát triển kiến trúc hệ thống cho Chatbot, trong đó BanhMiWhisper đảm nhận vai trò STT để xử lý truy vấn người dùng với Python là ngôn ngữ lập trình chính.
- Triển khai BanhMiWhisper thành API. Xây dựng giao diện người dùng thân thiện, dễ sử dụng với thư viện Streamlit, tích hợp liền mạch giữa STT và các chức năng xử lý ngôn ngữ tự nhiên cho Chatbot.



Kết quả dự kiến

- **BanhMi Multi-Dialect Dataset:** Bộ dữ liệu chuyển đổi giọng nói thành văn bản đa phương ngữ cho tiếng Việt với hơn 200 giờ âm thanh có chỉ số WER dưới 1%.
- **BanhMiWhisper**: Mô hình chuyển đổi giọng nói sang văn bản đa phương ngữ cho tiếng Việt với nhiều phiên bản kích thước mô hình khác nhau, WER dưới 3% và CER dưới 1% trên các bộ đánh giá tiếng Việt nói chung, tiếng Việt đa phương ngữ nói riêng và BanhMi Multi-Dialect Dataset. Đạt State-of-the-Art cho bài toán chuyển đổi giọng nói sang văn bản đa phương ngữ tiếng Việt.
- **Ứng dụng Chatbot:** Ứng dụng có áp dụng mô hình BanhMiWhisper cho chuyển đổi giọng nói sang văn bản, văn bản đó được đưa vào Chatbot trả phản hồi cho người dùng. Với giao diện trực quan, thân thiện với người dùng và dễ sử dụng.

Tài liệu tham khảo

[1]. Radford, Alec and Kim, Jong Wook and Xu, Tao and Brockman, Greg and McLeavey, Christine and Sutskever, Ilya: Robust speech recognition via large-scale weak supervision. ICML 2023: 1-27

[2] Thanh-Thien Le and Linh The Nguyen and Dat Quoc Nguyen: PhoWhisper: Automatic Speech Recognition for Vietnamese. ICLR 2024 Tiny Papers track: 1-3

[3] Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, Kiet Van Nguyen: Multi-Dialect Vietnamese: Task, Dataset, Baseline Models and Challenges. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: 7476-7498