NGHIÊN CỨU XÂY DỰNG BỘ DỮ LIỆU PHỤ ĐỀ VIDEO HỌC LIỆU NGÀNH CÔNG NGHỆ THÔNG TIN

Trần Thị Thu Phương, Nguyễn Quốc Tuấn, Lê Thị Hằng
Trường Đại học Thủ đô Hà Nội

Tóm tắt: *Nghiên cứu này tập trung vào việc xây dựng một bộ dữ liệu phụ đề video chuyên ngành CNTT nhằm nâng cao khả năng tiếp cận học liệu và hỗ trợ phát triển các ứng dụng NLP trong giáo dục. Nghiên cứu đề xuất một quy trình thu thập và xử lý dữ liệu bài bản, từ lựa chọn nguồn, trích xuất, làm sạch, chuẩn hóa đến kiểm định chất lượng dữ liệu. Kết quả là một bộ dữ liệu có tính học thuật cao, làm nền tảng cho các nghiên cứu và ứng dụng trong lĩnh vực giáo dục CNTT*

Từ khóa: *phụ đề học liệu; xử lý ngôn ngữ tự nhiên; trí tuệ nhân tạo trong giáo dục; học liệu số ngành CNTT.*

Nhận bài ngày 20.04.2025; gửi phản biện, chỉnh sửa, duyệt đăng ngày 30.05.2025
Liên hệ tác giả: Nguyễn Quốc Tuấn; email: nqtuan@daihocthudo.edu.vn

1. ĐẶT VẤN ĐỀ

Trong bối cảnh thế giới đang bước vào cuộc cách mạng công nghiệp lần thứ tư, dữ liệu đã trở thành nền tảng cho đổi mới sáng tạo trong nhiều lĩnh vực, đặc biệt là giáo dục và Công nghệ Thông tin (CNTT). Sự bùng nổ của dữ liệu trên toàn cầu, với ước tính đạt 175 zettabyte vào năm 2025 [1], đặt ra yêu cầu cấp thiết về việc tổ chức và khai thác dữ liệu hiệu quả trong giáo dục đại học. Trong đó, các tài liệu học liệu kỹ thuật số, như video giảng dạy và khóa học trực tuyến, ngày càng đóng vai trò quan trọng trong quá trình đào tạo.

Tuy nhiên, việc thiếu các bộ dữ liệu phụ đề học thuật chất lượng cao và mang tính chuyên ngành đang là một thách thức lớn, đặc biệt đối với các video giáo dục trong lĩnh vực CNTT. Nghiên cứu của World Economic Forum (2020) chỉ ra rằng, mặc dù các kỹ năng CNTT như AI, học máy và lập trình là những năng lực quan trọng trong tương lai, nhiều người học vẫn gặp rào cản về ngôn ngữ và khả năng tiếp cận học liệu số [2]. Điều này càng cho thấy sự cần thiết của các giải pháp hỗ trợ như phụ đề, dịch thuật và chatbot học tập để tăng tính cá nhân hóa và khả năng tiếp cận nội dung số, một xu hướng được 74% các cơ sở giáo dục đại học quan tâm (EDUCAUSE, 2022).[3]

Tại Việt Nam, nhu cầu xây dựng cơ sở dữ liệu học liệu số chuyên ngành CNTT, bao gồm cả phụ đề video, đang tăng nhanh cùng với quá trình số hóa của các trường đại học. Tuy nhiên, việc đảm bảo chất lượng của các dữ liệu này vẫn còn là một vấn đề nan giải do thiếu các quy trình chuẩn hóa và bộ tiêu chí toàn diện, gây khó khăn cho việc ứng dụng AI và các công nghệ Xử lý ngôn ngữ tự nhiên (NLP) trong giáo dục.

Để giải quyết những thách thức trên, nghiên cứu này đặt mục tiêu xây dựng một bộ dữ liệu phụ đề video chuyên ngành CNTT, dựa trên các tiêu chuẩn học thuật rõ ràng và có tính ứng dụng thực tiễn cao trong giáo dục đại học. Nghiên cứu không chỉ tập trung vào việc thu thập và xử lý dữ liệu mà còn hướng đến việc đánh giá khả năng ứng dụng của các mô hình tóm tắt, dịch thuật và trích xuất thông tin, những công nghệ có tiềm năng định hình tương lai của giáo dục số. Hướng tiếp cận này phù hợp với xu thế phát triển của các nền tảng học trực tuyến hàng đầu như Coursera, edX và Udacity, nơi mà việc tích hợp phụ đề đa ngôn ngữ và

các công cụ hỗ trợ truy cập ngày càng được chú trọng để nâng cao trải nghiệm học tập. Các nghiên cứu trong lĩnh vực giáo dục trực tuyến đã chứng minh rằng, phụ đề chính xác và hỗ trợ ngôn ngữ không chỉ cải thiện khả năng tiếp thu kiến thức mà còn tăng cường tương tác và tỷ lệ hoàn thành khóa học, đặc biệt đối với người học không sử dụng ngôn ngữ bản địa hoặc có nhu cầu hỗ trợ đặc biệt.

Tóm lại, việc xây dựng dữ liệu phụ đề học liệu CNTT có vai trò quan trọng trong việc nâng cao chất lượng đào tạo và tạo điều kiện thuận lợi cho việc tích hợp AI vào giáo dục một cách toàn diện và bền vững.

2. NỘI DUNG

2.1. Tổng quan về xử lý phụ đề và chuẩn hóa dữ liệu video

2.1.1. Khái niệm và vai trò của phụ đề

Phụ đề là phần văn bản hiển thị đồng bộ với nội dung âm thanh của video, thường thể hiện lời thoại, nội dung thuyết minh hoặc mô tả âm thanh. Trong bối cảnh giáo dục số, đặc biệt là các video học liệu chuyên ngành, phụ đề không chỉ hỗ trợ người học tiếp cận tốt hơn nội dung bài giảng mà còn đóng vai trò quan trọng trong việc:[4]

- Hỗ trợ người học có nhu cầu đặc biệt (ví dụ: người khiếm thính),
- Tăng khả năng tiếp thu nội dung trong môi trường đa nhiệm (xem video không bật âm thanh),
- Tạo nền tảng cho các ứng dụng học máy như tóm tắt nội dung, phân tích dữ liệu ngôn ngữ, dịch tự động, tìm kiếm nội dung theo văn bản...

2.1.2. Các loại phụ đề và định dạng kỹ thuật

Phụ đề có thể được chia thành:

- Phụ đề cứng (open captions): gắn liền với video, không thể tắt hoặc chỉnh sửa.
- Phụ đề mềm (closed captions): tồn tại dưới dạng file riêng biệt, có thể bật/tắt hoặc xử lý lại.

Các định dạng phổ biến trong phụ đề học liệu gồm:

- SRT (SubRip Subtitle): đơn giản, dễ sử dụng, phù hợp cho nghiên cứu NLP.
- VTT (WebVTT): thường dùng cho nền tảng web, hỗ trợ thêm siêu dữ liệu.
- ASS/SSA: hỗ trợ định dạng phức tạp, chủ yếu dùng trong phụ đề phim.

Việc chọn định dạng phụ đề cần phù hợp với mục tiêu xử lý dữ liệu, đảm bảo khả năng đồng bộ, dễ trích xuất và tích hợp với các mô hình AI.

2.1.3. Quy trình xử lý và tạo phụ đề

- Quá trình tạo phụ đề có thể thực hiện theo ba cách:
- Thủ công: nghe – gõ lại lời thoại – căn thời gian. Tuy chính xác nhưng mất thời gian.
- Bán tự động: sử dụng hệ thống nhận dạng giọng nói (như Whisper hoặc Google ASR) để tạo bản nháp, sau đó con người hiệu đính và căn chỉnh lại thời gian. Đây là hướng tiếp cận được sử dụng trong nghiên cứu này để cân bằng giữa độ chính xác và hiệu suất.
- Tự động hoàn toàn: sử dụng các hệ thống AI tạo và căn chỉnh phụ đề, nhưng độ chính xác còn phụ thuộc nhiều vào âm thanh, ngữ cảnh và từ vựng chuyên ngành.

2.1.4. Chuẩn hóa dữ liệu

Để phục vụ mục tiêu nghiên cứu và ứng dụng AI, việc chuẩn hóa dữ liệu là bước then chốt. Các tiêu chí chuẩn hóa bao gồm:

- Định dạng thống nhất: sử dụng định dạng phụ đề (SRT/VTT), định dạng video (MP4).
- Cấu trúc metadata rõ ràng: mỗi mẫu dữ liệu cần gắn thông tin về tiêu đề, lĩnh vực, giảng viên, thời lượng, ngôn ngữ, nguồn gốc...

- Từ vựng chuyên ngành được xử lý nhất quán: chuẩn hóa thuật ngữ, xử lý viết tắt, đồng bộ giữa các ngôn ngữ.
- Thời gian hiển thị phụ đề chính xác: đảm bảo sai số không vượt quá ±0,5 giây để tối ưu cho cả người học và các hệ thống tự động

2.1.5. Kết quả kỳ vọng và tiềm năng ứng dụng của bộ dữ liệu phụ đề

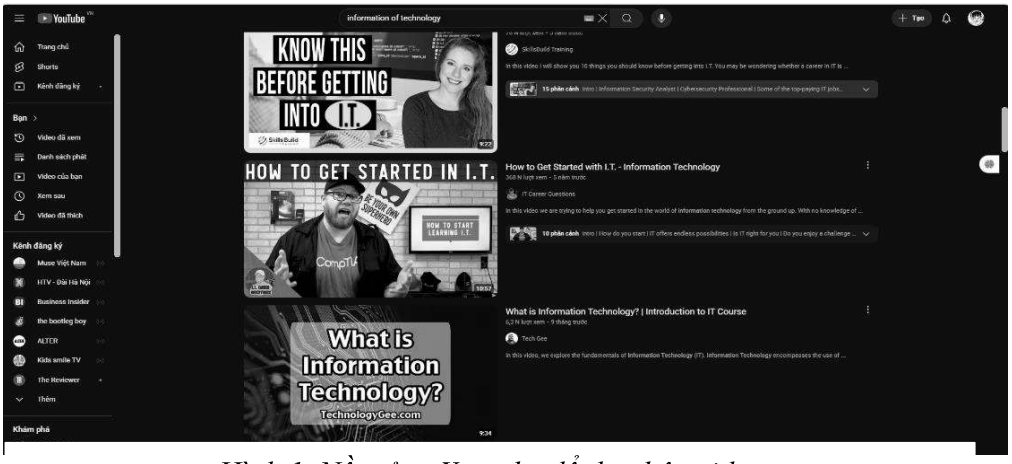
Kết quả từ việc thu thập và xây dựng bộ dữ liệu phụ đề video chuyên ngành CNTT dự kiến sẽ đạt được nhiều thành tựu quan trọng. Nghiên cứu dự kiến tạo ra bộ dữ liệu phụ đề gồm 15k mẫu nội dung giáo dục CNTT đa dạng. Dựa trên bộ dữ liệu này, ít nhất ba mô hình NLP chuyên biệt sẽ được phát triển, bao gồm mô hình tóm tắt, mô hình dịch thuật và mô hình trích xuất thông tin. Kết quả nghiên cứu cũng sẽ góp phần nâng cao khả năng tiếp cận học liệu CNTT cho ba nhóm đối tượng chính: người khiếm thính, người học không bản địa và người có khó khăn học tập.

Ngoài ra, nghiên cứu sẽ cung cấp các công cụ và thuật toán hỗ trợ cho việc xây dựng hệ thống trợ lý học tập thông minh, hệ thống đề xuất học tập cá nhân hóa và công cụ đánh giá tự động. Bộ dữ liệu và các công cụ xử lý liên quan sẽ được công bố để đóng góp cho cộng đồng nghiên cứu, tạo nền tảng cho các nghiên cứu tiếp theo trong lĩnh vực NLP và AI ứng dụng trong giáo dục CNTT. Việc xây dựng bộ dữ liệu phụ đề video chuyên ngành CNTT này không chỉ tạo ra nguồn tài nguyên học tập có giá trị, mà còn mở ra nhiều cơ hội mới cho việc ứng dụng công nghệ xử lý ngôn ngữ tự nhiên trong giáo dục và đào tạo.

2.2. Nguồn thu thập phụ đề video

2.2.1. Phương pháp lựa chọn nền tảng

Sau khi phân tích các nền tảng chia sẻ video phổ biến, YouTube được xác định là nguồn chính để thu thập phụ đề video chuyên ngành công nghệ thông tin (CNTT). Việc lựa chọn này dựa trên một số tiêu chí quan trọng. Trước hết, YouTube cung cấp tính đa dạng nội dung, bao gồm nhiều tài liệu giáo dục về CNTT từ cơ bản đến nâng cao. Khả năng tiếp cận cũng là yếu tố quan trọng khi nền tảng này cung cấp API cho phép trích xuất phụ đề một cách tự động và hiệu quả. Ngoài ra, YouTube hỗ trợ nội dung bằng nhiều ngôn ngữ, trong đó có tiếng Anh và tiếng Việt, đáp ứng đúng yêu cầu của dự án. Độ tin cậy của nguồn cũng được đảm bảo nhờ sự hiện diện của nhiều kênh giáo dục uy tín, với lượng người đăng ký lớn và các đánh giá tích cực từ người xem.



Hình 1. Nền tảng Youtube để thu thập video

2.2.2. Đánh giá và lựa chọn kênh

Độ chính xác của nội dung được đánh giá dựa trên sự so sánh với các nguồn học thuật và ý kiến từ chuyên gia trong lĩnh vực. Chuyên môn của người tạo nội dung được xem xét