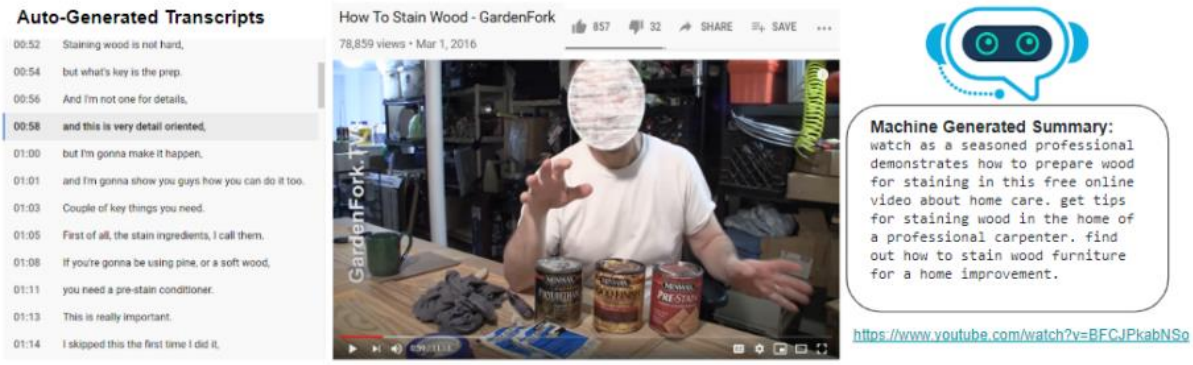
BERTSum – ỨNG DỤNG CỦA BERT TRONG TÓM TẮT CHỈ DẪN

Đoàn Phạm Thảo Như, Lý Gia Bảo, Nguyễn Linh Chi, Trần Thanh Phương
Trường Đại học Kinh tế Tài chính TP.HCM, nhudpt20@uef.edu.vn
Trường Đại học Kinh tế Tài chính TP.HCM, baolg220@uef.edu.vn
Trường Đại học Kinh tế Tài chính TP.HCM, chinl220@uef.edu.vn
Trường Đại học Kinh tế Tài chính TP.HCM, phuongtt20@uef.edu.vn



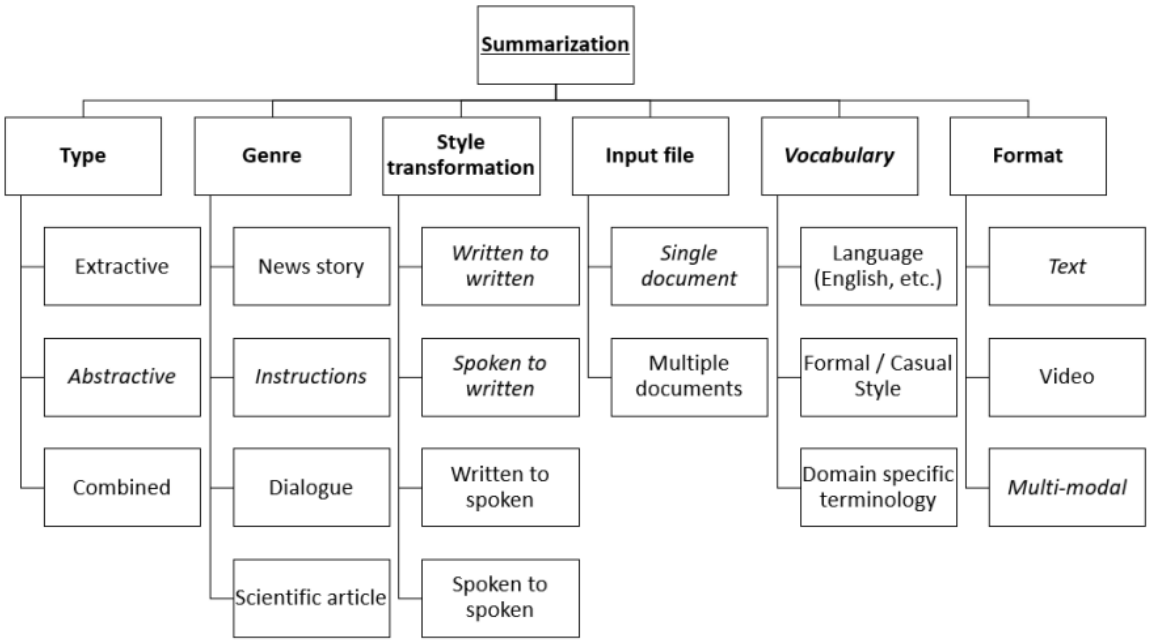
Hình 1: Ảnh chụp màn hình video How2 trên Youtube với bản ghi và bản tóm tắt do mô hình tạo ra.

Tóm tắt: *lời nói* ***trừu tượng*** *là một vấn đề đầy thách thức vì tính tự phát của nhịp điệu, tính bất ổn, và các vấn đề khác thường không có trong văn bản viết. Trong nghiên cứu của chúng tôi, mô hình BERTSum được sử dụng lần đầu tiên trong ngôn ngữ nói. Chúng tôi tạo ra những bản tóm tắt trừu tượng của các video hướng dẫn kể chuyện trên một loạt chủ đề, bao gồm làm vườn, nấu ăn, cài đặt phần mềm và thể thao. Chúng tôi sử dụng học chuyển giao và tiền huấn luyện mô hình trên một số tập dữ liệu đa ngành nghề, bao gồm cả tiếng Anh viết và nói, nhằm mở rộng từ vựng. Để sửa các lỗi phân đoạn câu và dấu chấm câu trong kết quả đầu ra của hệ thống nhận dạng giọng nói tự động (ASR), chúng tôi cũng tiền xử lý các bản ghi âm. Chúng tôi phân tích kết quả từ các tập dữ liệu How2 và WikiHow bằng cách sử dụng các phương pháp đánh giá ROUGE và Content-F1. Chúng tôi sử dụng người đánh giá là con người để chấm điểm một số bản tóm tắt được chọn ngẫu nhiên từ bộ dữ liệu được tổng hợp từ YouTube và HowTo100M. Kết quả đạt được một mức độ trôi chảy và tính hữu ích của văn bản dựa trên đánh giá ngang ngửa với các bản tóm tắt được viết bởi con người. Khi áp dụng trên các bài viết WikiHow với nhiều chủ đề và phong cách khác nhau, mô hình của chúng tôi vượt trội hơn so với các mô hình tốt nhất hiện có, đồng thời không có sự giảm hiệu suất trên tập dữ liệu tiêu chuẩn CNN/DailyMail. Phương pháp này có tiềm năng lớn để tăng khả năng khám phá và tiếp cận với tài liệu trực tuyến do tính tổng quát mạnh mẽ qua các phong cách và chủ đề. Chúng tôi tưởng tượng việc tích hợp chức năng này vào các trợ lý ảo thông minh, cho phép chúng tóm tắt nhanh chóng cả nội dung nói và viết của hướng dẫn.*

Từ khóa: *BERT, Xử lý ngôn ngữ tự nhiên, Tóm tắt văn bản, Truy xuất thông tin, Mô hình ngôn ngữ, Trợ lý ảo, Trừu tượng, Mạng thần kinh, Video hướng dẫn tường thuật.*

1. GIỚI THIỆU

Nghiên cứu của chúng tôi được thúc đẩy bởi nhu cầu cung cấp sự tiếp cận cho lượng lớn tài liệu trên internet do người dùng tạo ra. Nhiệm vụ của chúng tôi là tăng cường các công cụ tóm tắt tự động nhằm hỗ trợ người dùng trong việc tiếp thu thông tin. Các nhà sản xuất nội dung trực tuyến thường sử dụng nhiều ngôn ngữ thông tục, các cụm từ phụ và thuật ngữ ngành. Do đó, việc tóm tắt văn bản không chỉ đơn giản là rút ra thông tin chính từ nguồn mà còn là biến đổi nó thành một cái gì



Hình 2: *Phân loại các phương pháp và loại tóm tắt.*

đó có tính liên kết và tổ chức hơn. Bài nghiên cứu này tập trung vào tóm tắt trích xuất và trừu tượng của các hướng dẫn nói và viết đã được kể. Các câu quan trọng nhất trong tài liệu được xác định bằng cách trích xuất, một vấn đề phân loại đơn giản cũng xác định xem một câu có thuộc vào tóm tắt hay không. Ngược lại, tóm tắt trừu tượng đòi hỏi khả năng tạo ra từ và cụm từ mới không có trong văn bản gốc. Vấn đề về mạch lạc, dễ hiểu là quan trọng với các mô hình ngôn ngữ được sử dụng để tóm tắt văn bản hội thoại. Đây là lần đầu tiên việc tóm tắt ngôn ngữ nói từ đầu vào ASR (chuyển đổi giọng nói thành văn bản) sử dụng một mô hình dựa trên BERT. Mục tiêu của chúng tôi là tạo ra một công cụ phổ biến có thể được áp dụng cho nhiều bài viết và video How2. Nếu vấn đề này có thể được giải quyết, mô hình tóm tắt có thể được mở rộng sang các ứng dụng khác trong lĩnh vực này, như tóm tắt cuộc trò chuyện trong các hệ thống trò chuyện giữa con người và robot.

Những phần còn lại của bài nghiên cứu này:

- Đánh giá về các kỹ thuật tóm tắt hiện đại.
- Mô tả về tập dữ liệu văn bản, cuộc hội thoại và tóm tắt được sử dụng cho quá trình huấn luyện.
- Sử dụng các mô hình tóm tắt văn bản dựa trên BERT và tinh chỉnh các tập lệnh được tạo tự động từ các video hướng dẫn.
- Đề xuất cải tiến phương pháp đánh giá các số liệu được sử dụng bởi các nghiên cứu trước đây.
- Phân tích kết quả thử nghiệm và so sánh với kết quả chuẩn

2. NGHIÊN CỨU TRƯỚC ĐÂY

Hình 2 hiển thị phân loại các loại và phương pháp tóm tắt. Trước năm 2014, trọng tâm chính của việc tóm tắt là nỗ lực trích xuất các dòng từ các tài liệu đơn lẻ bằng cách sử dụng các mô hình thống kê và mạng lưới thần kinh [23] [17]. Công trình của Sutskever và cộng sự [22] [2] về các mô hình sequence-to-sequence đã mở ra con đường mới cho các mạng thần kinh trong xử lý ngôn ngữ tự nhiên. Trong ngành công nghiệp, LSTM (một phần RNN) nổi lên như một chiến lược hàng đầu trong năm 2014 và 2015, tạo ra những kết quả vượt trội. Thành công đã đạt được với những sửa đổi kiến trúc này trong các ứng dụng bao gồm nhận dạng giọng nói, dịch máy, phân tích cú pháp và chú thích hình ảnh. Các kết quả này đã mở đường cho tóm tắt trừu tượng, bắt đầu vượt trội so với tóm tắt trích xuất về hiệu

***Bảng 1*:** Các bộ dữ liệu training và testing

Total Training Dataset Size	**535,527**
CNN/DailyMail	90,266 and 196,961
WikiHow Text	180,110
How2 Videos	68,190
Total Testing Dataset Size	**5,195 videos**
YouTube (DIY Videos and How-to Videos)	1,809
HowTo100M	3,386

suất. Vấn đề vector có độ dài cố định đã được giải quyết trong một bài báo của Vaswani et al. [25] vào năm 2017, cho phép mạng neural tập trung vào thông tin quan trọng để thực hiện các nhiệm vụ dự đoán. Các kỹ thuật chú ý dựa trên Transformer trở nên phổ biến hơn khi thực hiện các nhiệm vụ như dịch và tóm tắt. Các mô hình kết hợp LSTM và các biến thể mạng neural sâu lớn hiện nay là những mô hình tốt nhất cho tóm tắt video trừu tượng. Công việc gần đây về tóm tắt đa phương tiện bao gồm các phương thức âm thanh và hình ảnh vào các mô hình ngôn ngữ ngoài các đầu vào văn bản để cung cấp tóm tắt thông tin video. Tuy nhiên, vẫn khó để tạo ra các tóm tắt thú vị từ các văn bản trò chuyện bằng cách sử dụng bản ghi âm hoặc sự kết hợp của các phương thức. Số lượng các bộ dữ liệu thử nghiệm có sẵn cho loại nghiên cứu này bị hạn chế do thiếu dữ liệu được chú thích bởi con người [18] [10]. Các câu chuyện tin tức có cấu trúc là nền tảng của hầu hết các công việc trong lĩnh vực tóm tắt tài liệu. Việc sắp xếp dữ liệu với thời gian, chủ đề và phong cách đã được định nghĩa là tâm điểm của việc tóm tắt video [4]. Các phương pháp xử lý ngôn ngữ tự nhiên cũng đã được sử dụng truyền thống để phân tách và ghép các khung hình video quan trọng để tạo tóm tắt video [5]. Trên hết, việc chuyển đổi ngôn ngữ nói thành văn bản viết có thể gặp khó khăn do sự thay đổi phong cách thường xuyên và không nhất quán. Sử dụng sự kết hợp của các video hướng dẫn được thuật lại, các văn bản và các tin tức với các phong cách, chiều dài và chất văn học khác nhau, chúng tôi tiếp cận việc tóm tắt video trong nghiên cứu này bằng cách mở rộng các mô hình tóm tắt văn bản từng tài liệu đơn lẻ hiệu quả nhất [19] để tóm tắt các video.

Bảng 2: Các thống kê bổ sung

YouTube Min/Max Length	4/1,940 words
YouTube Avg Length	259 words
HowTo100M Sample Min/Max Length	5/6,587 words
HowTo100M Sample Avg Length	859 words

3. PHƯƠNG PHÁP LUẬN

3.1 Thu thập dữ liệu

Chúng tôi đưa ra giả thuyết rằng việc đào tạo trên các bộ dữ liệu lớn hơn sẽ cải thiện khả năng mô hình của chúng tôi để tạo ra các bản tóm tắt mạch lạc trên các văn bản khác nhau. Các kích thước tập dữ liệu văn bản và video khác nhau được hiển thị trong Bảng 1. Các bản tóm tắt bằng văn bản được bao gồm cho mỗi tập dữ liệu huấn luyện. Ngôn ngữ và độ dài của dữ liệu nằm trong phạm vi từ thông tục đến trang trọng và từ câu đơn đến đoạn văn ngắn.

- **Dữ liệu CNN/DailyMail** [7]: CNN và DailyMail chứa một bộ sưu tập các câu chuyện tin tức và tóm tắt, với độ dài trung bình của bài báo là 119 từ và độ dài tóm tắt là 83 từ. Các bài viết từ năm 2007 đến 2015 được tập hợp lại.
- **Dữ liệu Wikihow** [9]: một cơ sở dữ liệu văn bản khá lớn với hơn 200.000 bản tóm tắt của các tài liệu riêng lẻ. Wikihow là tập hợp các văn bản hướng dẫn "Cách thực hiện" gần đây được lấy từ wikihow.com, bao gồm mọi thứ từ "Cách chơi Uno" đến "Cách đối phó với lo lắng về vi-rút corona". Các bài viết này có kích thước và chủ đề khác nhau, nhưng

tất cả chúng đều được viết với tâm trí người dùng. Tóm tắt của bài báo được tạo thành từ các câu đầu tiên của mỗi đoạn văn.

- **Dữ liệu How2** [20]: Phần tổng hợp YouTube này có các video (8.000 video - khoảng 2.000 giờ) dài trung bình 90 giây và 291 từ trong bản dịch. Nó bao gồm các bản tóm tắt do con người viết mà chủ sở hữu video được hướng dẫn viết bản tóm tắt để tối đa hóa lượng khán giả. Tóm tắt có độ dài từ hai đến ba câu với độ dài trung bình là 33 từ.

Bất chấp sự phát triển của các bộ dữ liệu hướng dẫn như Wikihow và How2, những tiến bộ trong việc tóm tắt đã bị hạn chế bởi sự sẵn có của các bản tóm tắt và bảng điểm có chú thích của con người. Những bộ dữ liệu như vậy rất khó lấy và tốn kém để tạo ra, thường dẫn đến việc sử dụng lặp đi lặp lại dữ liệu có cấu trúc cao và có nhiệm vụ đơn lẻ. Như đã thấy với các mẫu trong tập dữ liệu How2, chỉ những video có độ dài nhất định và tóm tắt có cấu trúc mới được sử dụng để đào tạo và thử nghiệm. Để mở rộng phạm vi nghiên cứu của mình, chúng tôi đã bổ sung các bộ dữ liệu tóm tắt được gắn nhãn hiện có bằng các tập lệnh video hướng dẫn được tạo tự động và các mô tả do con người quản lý.

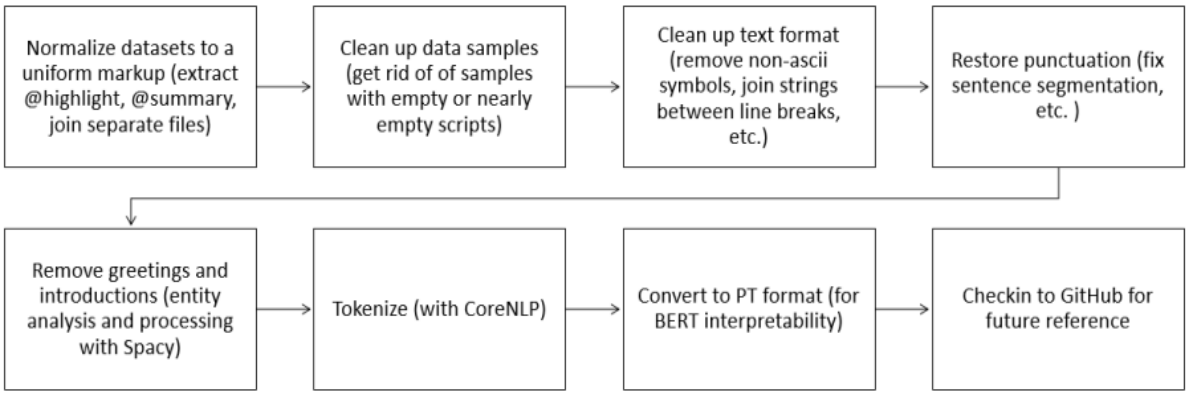
Chúng tôi giới thiệu bộ dữ liệu mới thu được từ việc kết hợp một số danh sách phát YouTube 'Hướng dẫn thực hiện' và 'Tự làm' cùng với các mẫu từ Bộ dữ liệu HowTo100Million đã xuất bản [16]. Để kiểm tra tính hợp lý của việc sử dụng mô hình này trong thực tế, chúng tôi đã chọn các video trên các văn bản hội thoại khác nhau không có tóm tắt tương ứng hoặc chú thích của con người. Bộ dữ liệu 'Hướng dẫn' [24] và 'Tự làm' [6] được chọn là danh sách phát hướng dẫn bao gồm các chủ đề khác nhau từ việc tải xuống phần mềm chính thống đến cải tiến nhà cửa. Danh sách phát 'Hướng dẫn' sử dụng phần thuyết minh bằng máy trong video để hỗ trợ hướng dẫn trong khi danh sách phát 'Tự làm' có các video có người thuyết trình là con người. Bộ dữ liệu HowTo100Million là một bộ dữ liệu quy mô lớn gồm hơn 100 triệu video clip được lấy từ các video hướng dẫn được tường thuật trên 140 danh mục. Bộ dữ liệu của chúng tôi kết hợp một mẫu trên tất cả các danh mục và sử dụng các chú thích ngôn ngữ tự nhiên từ các tường thuật được sao chép tự động do YouTube cung cấp.

3.2 Tiền xử lý dữ liệu

Do tính đa dạng và phức tạp của dữ liệu đầu vào, chúng tôi đã xây dựng một quy trình tiền xử lý để sắp xếp dữ liệu theo một định dạng chung. Chúng tôi đã quan sát thấy các vấn đề thiếu dấu chấm câu, từ ngữ không chính xác và phần giới thiệu không liên quan đã ảnh hưởng đến việc đào tạo người mẫu. Với những thách thức này, mô hình của chúng tôi đã hiểu sai ranh giới của phân đoạn văn bản và tạo ra các bản tóm tắt chất lượng kém. Trong những trường hợp đặc biệt, mô hình không thể tạo ra bất kỳ bản tóm tắt nào. Để duy trì sự trôi chảy và mạch lạc trong các bản tóm tắt do con người viết, chúng tôi đã làm sạch và khôi phục cấu trúc câu như trong Hình 3. Chúng tôi đã áp dụng tính năng phát hiện thực thể từ thư viện phần mềm nguồn mở để xử lý ngôn ngữ tự nhiên nâng cao có tên là spacy [8] và nltk: Bộ công cụ ngôn ngữ tự nhiên để xử lý ngôn ngữ tự nhiên tượng trưng và thống kê [15] để loại bỏ phần giới thiệu và ẩn danh các đầu vào của mô hình tóm tắt của chúng tôi. Chúng tôi tách các câu và mã hóa bằng bộ công cụ Stanford Core NLP trên tất cả các bộ dữ liệu và xử lý trước dữ liệu theo cùng một phương pháp [21].

3.3 Mô hình:

Chúng tôi đã sử dụng các mô hình BertSum được đề xuất trong [14] cho nghiên cứu của chúng tôi. Điều này bao gồm cả các mô hình tóm tắt theo phương pháp trích xuất và tạo ra, trong đó sử dụng một bộ mã hóa cấp độ tài liệu dựa trên Bert. Kiến trúc transformer áp dụng một bộ mã hóa BERT đã được tiền huấn luyện kết hợp với một bộ giải mã Transformer được khởi tạo ngẫu nhiên. Nó sử dụng hai tốc độ học khác nhau: tốc độ học thấp cho bộ mã hóa và một tốc độ học riêng biệt cao hơn cho bộ giải mã để nâng cao khả năng học. Chúng tôi đã sử dụng một máy Linux 4-GPU và khởi tạo một mô hình cơ bản bằng cách huấn luyện một



Hình 3: Một quy trình tiền xử lý dữ liệu cho việc tóm tắt văn bản

mô hình trích xuất trên 5.000 mẫu video từ bộ dữ liệu How2. Ban đầu, chúng tôi áp dụng BERT base uncased với 10.000 bước và điều chỉnh mô hình tóm tắt và lớp BERT để chọn kích thước epoch cho phù hợp nhất. Sau đó, chúng tôi tiếp tục huấn luyện mô hình tóm tắt sáng tạo trên How2 và WikiHow một cách riêng lẻ. Phiên bản tốt nhất của mô hình tóm tắt sáng tạo đã được huấn luyện trên tập dữ liệu tổng hợp gồm CNN/DailyMail, Wikihow và How2 với tổng cộng 535.527 ví dụ và 210.000 bước. Chúng tôi sử dụng kích thước lô huấn luyện là 50 và chạy mô hình trong 20 epoch. Bằng cách kiểm soát thứ tự của các tập dữ liệu mà chúng tôi huấn luyện mô hình, chúng tôi đã cải thiện khả năng thông suốt của các bản tóm tắt. Như đã nêu trong các nghiên cứu trước đây, mô hình ban đầu chứa hơn 180 triệu tham số và sử dụng hai bộ tối ưu hóa Adam với β_1=0,9 và β_2=0,999 cho bộ mã hóa và bộ giải mã lần lượt. Bộ mã hóa sử dụng tốc độ học là 0,002 và bộ giải mã có tốc độ học là 0,2 để đảm bảo bộ mã hóa được huấn luyện với độ dốc chính xác hơn trong khi bộ giải mã trở nên ổn định hơn. Kết quả của các thí nghiệm được thảo luận trong Mục 4. Chúng tôi giả thuyết rằng thứ tự huấn luyện là quan trọng đối với mô hình giống như con người học. Ý tưởng áp dụng học trình độ trong xử lý ngôn ngữ tự nhiên đã trở thành một chủ đề nổi bật trong nghiên cứu [1] [26]. Chúng tôi bắt đầu huấn luyện trên các mẫu có cấu trúc cao trước khi chuyển sang cấu trúc ngôn ngữ phức tạp hơn nhưng dễ dự đoán. Chỉ sau khi huấn luyện các kịch bản văn bản chúng tôi tiếp tục huấn luyện với các kịch bản video, trong đó mô hình phải đối mặt với thách thức về luồng không cố định và ngôn ngữ trò chuyện.

3.4 Đánh giá kết quả

Kết quả đánh giá bằng ROUGE, thước đo tiêu chuẩn cho tóm tắt trừu tượng [11]. Mặc dù chúng tôi mong đợi mối tương quan giữa tóm tắt tốt và điểm ROUGE cao, nhưng chúng tôi đã quan sát các ví dụ về tóm tắt kém với điểm số cao, chẳng hạn như trong Hình 10 và tóm tắt tốt với điểm ROUGE thấp. Ví dụ minh họa về lý do tại sao chỉ số ROUGE là không đủ được trình bày trong Phụ lục, Hình 10. Ngoài ra, chúng tôi đã thêm tính năng chấm điểm Nội dung F1, một chỉ số do Đại học Carnegie Mellon [3] đề xuất để tập trung vào mức độ liên quan của nội dung. Tương tự như ROUGE, Nội dung F1 cho điểm tóm tắt bằng điểm f có trọng số và hình phạt cho thứ tự từ không chính xác. Nó cũng giảm giá các từ dừng và buzz thường xuyên xuất hiện trong miền Hướng dẫn, chẳng hạn như "học cách thực hiện từ các chuyên gia trong video trực tuyến miễn phí này". Để chấm điểm các đoạn văn không có phần tóm tắt bằng văn bản, chúng tôi đã khảo sát các giám khảo con người bằng một khung đánh giá bằng Python, Google Biểu mẫu và bảng tính Excel. Các bản tóm tắt trong các cuộc khảo sát được lấy mẫu ngẫu nhiên từ bộ dữ liệu của chúng tôi để tránh sai lệch. Để tránh thông tin bất đối xứng giữa bản tóm tắt do con người tạo ra và do máy tạo, chúng tôi đã xóa văn bản viết hoa. Chúng tôi đã hỏi hai loại câu hỏi: Một câu hỏi kiểm tra Turing để người tham gia phân biệt AI với các

mô tả do con người tạo ra. Thứ hai liên quan đến việc lựa chọn xếp hạng chất lượng cho tóm tắt. Dưới đây là các định nghĩa về tiêu chí rõ ràng:

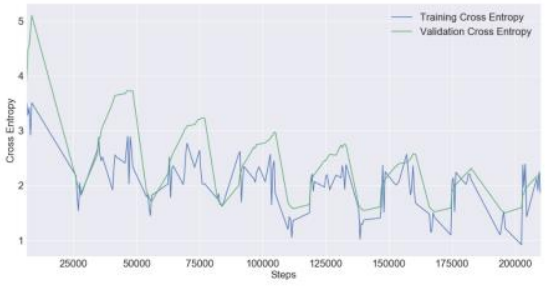
- Sự trôi chảy: Liệu văn bản có nhịp điệu tự nhiên không?
- Sự hữu dụng: Có đủ thông tin để người dùng quyết định liệu họ có muốn dành thời gian để xem video không?
- Súc tích: Văn bản có gọn gàng hay có sự lặp lại không cần thiết?
- Tính ổn định: Có các câu không liên quan, không rõ nghĩa hoặc mơ hồ, gây nhầm lẫn hoặc mâu thuẫn trong văn bản không?
- Độ thực tế: Có điều gì có vẻ hư cấu và kỳ quặc trong cách kết hợp từ và không "bình thường" không?

Các tùy chọn cho việc đánh giá bản tóm tắt như sau: 1: Tệ 2: Dưới mức trung bình 3: Trung bình 4: Tốt 5: Xuất sắc.

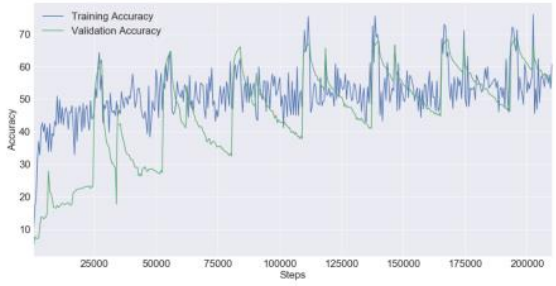
4. QUÁ TRÌNH HUẤN LUYỆN VÀ KẾT QUẢ

4.1 Quá trình huấn luyện:

Mô hình BertSum là mô hình có hiệu suất tốt nhất trên tập dữ liệu CNN/DailyMail, mang lại kết quả đạt trạng thái tối ưu (Dòng 6) 3. Mô hình BertSum hỗ trợ cả kỹ thuật tóm tắt trích dẫn và tóm tắt trừu tượng. Kết quả cơ sở của chúng tôi được thu được bằng cách áp dụng mô hình trích dẫn BertSum được tiền huấn luyện trên CNN/DailyMail vào các video How2. Tuy nhiên, mô hình đã cho điểm rất thấp cho kịch bản của chúng tôi. Các bản tóm tắt được tạo ra từ mô hình không mạch lạc, lặp đi lặp lại và thiếu thông tin. Mặc dù hiệu suất kém, mô hình hoạt động tốt hơn trong phần phụ đề sức khỏe trong các video



Hình 4: Entropy chéo: Huấn luyện và đánh giá



Hình 5: Độ chính xác: Huấn luyện và đánh giá

How2. Chúng tôi giải thích điều này như một triệu chứng của sự tập trung cao trong các bản tin tin tức được tạo ra bởi CNN/DailyMail. Chúng tôi nhận ra rằng tóm tắt trích dẫn không phải là mô hình mạnh nhất cho mục tiêu của chúng tôi: hầu hết các video trên YouTube được trình bày theo phong cách trò chuyện tự nhiên, trong khi các bản tóm tắt có tính trang trọng cao hơn. Chúng tôi đã chuyển sang tóm tắt trừu tượng để cải thiện hiệu suất.

Mô hình trừu tượng sử dụng một kiến trúc mã hóa-giải mã, kết hợp cùng bộ mã hóa BERT được tiền huấn luyện và bộ giải mã Transformer được khởi tạo ngẫu nhiên. Nó sử dụng một kỹ thuật đặc biệt, trong đó phần mã hóa gần như được giữ nguyên với một tốc độ học thấp và tạo ra một tốc độ học riêng cho bộ giải mã để tăng khả năng học của nó. Để tạo ra một mô hình trừu tượng có tính tổng quát, chúng tôi đã tiến hành huấn luyện ban đầu trên một nguồn văn bản tin tức lớn. Điều này giúp mô hình của chúng tôi hiểu được các văn bản có cấu trúc. Sau đó, chúng tôi đã giới thiệu Wikihow, giúp mô hình tiếp cận với lĩnh vực "Làm thế nào". Cuối cùng, chúng tôi đã huấn luyện và xác thực trên tập dữ liệu How2, thu hẹp phạm vi của mô hình vào một định dạng có cấu trúc được lựa chọn một cách chọn lọc. Ngoài việc huấn luyện theo thứ tự, chúng tôi đã thử nghiệm việc huấn luyện mô hình bằng cách sử dụng các tập hợp ngẫu nhiên của các mẫu đồng nhất. Chúng tôi phát hiện rằng huấn luyện mô hình bằng cách sử dụng một tập hợp được sắp xếp của các mẫu đã có hiệu suất tốt hơn so với các tập hợp ngẫu nhiên.

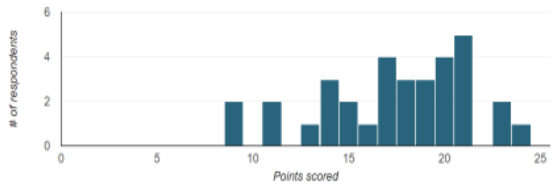
Biểu đồ entropy chéo trong Hình 4 cho thấy rằng mô hình không bị quá khớp hoặc thiếu

khớp với dữ liệu huấn luyện. Sự phù hợp tốt được biểu thị bằng sự hội tụ của các đường huấn luyện và xác thực. Hình 5 cho thấy chỉ số độ chính xác của mô hình trên các tập huấn luyện và xác thực. Mô hình được xác thực bằng cách sử dụng tập dữ liệu How2 so với tập dữ liệu huấn luyện. Mô hình cải thiện như dự kiến với nhiều bước hơn.

4.2 Kết quả và đánh giá

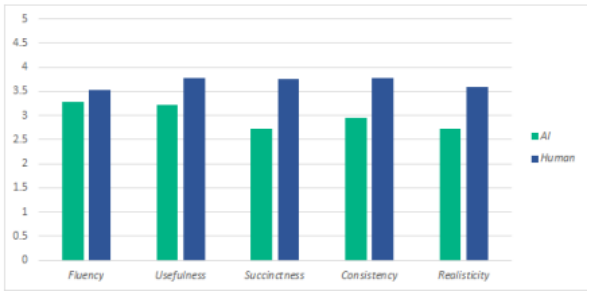
Mô hình BertSum được huấn luyện trên tập dữ liệu CNN/DailyMail [14] đã mang lại những điểm số đẹp đẽ khi áp dụng vào các mẫu từ các tập dữ liệu đó. Tuy nhiên, khi thử nghiệm trên tập dữ liệu Test của How2, mô hình đã cho hiệu suất kém và thiếu tính tổng quát trong việc tóm tắt (xem hàng 1 trong Bảng 3). Khi xem xét dữ liệu, chúng tôi nhận thấy rằng mô hình thường chọn một hoặc hai câu đầu tiên cho việc tóm tắt. Chúng tôi giả thiết rằng việc loại bỏ phần giới thiệu khỏi văn bản sẽ giúp cải thiện các điểm ROUGE. Mô hình của chúng tôi đã cải thiện được một vài điểm ROUGE sau khi áp dụng phương pháp tiền xử lý được mô tả trong Phần 3.2 ở trên. Một cải tiến khác đến từ việc thêm tính năng loại bỏ các từ trùng lặp ở đầu ra của mô hình, vì chúng tôi quan sát thấy điều này xảy ra trên các từ hiếm gặp đối với mô hình. Tuy nhiên, chúng tôi vẫn chưa đạt được điểm số cao hơn 22.5 ROUGE-1 F1 và 20 ROUGE-L F1 (điểm số ban đầu đạt được từ việc huấn luyện chỉ trên tập dữ liệu CNN/DailyMail và kiểm tra trên dữ liệu How2). Xem xét điểm số và văn bản của các tóm tắt riêng lẻ, chúng tôi nhận thấy mô hình hoạt động tốt hơn trên một số chủ đề như y tế, trong khi điểm số thấp hơn trên các chủ đề khác như thể thao.

Sự khác biệt trong phong cách đàm thoại giữa kịch bản video và các bài tin tức (mà mô hình đã được huấn luyện trước đó) ảnh hưởng đến chất lượng đầu ra của mô hình. Trong việc áp dụng ban đầu của mô hình tóm tắt trích xuất được huấn luyện trên tập dữ liệu CNN/DailyMail, các lỗi phong cách được hiển thị một cách đặc biệt. Mô hình coi những câu giới thiệu ban đầu là quan trọng trong việc tạo ra các tóm tắt (hiện tượng này được [15] gọi là N-lead [15], trong đó N là số lượng câu đầu tiên quan trọng). Mô hình của chúng tôi đã tạo ra các tóm tắt ngắn, đơn giản như "hi!"



Hình 4: Điểm của các giám khảo trong thử thách phân biệt các bản tóm tắt được tạo bởi máy học từ các chú thích video thực tế trên YouTube

Hình 5: Phân phối tỷ lệ của báo động nhầm (FP) và bỏ sót (FN) trong từng câu hỏi



Hình 6: Đánh giá chất lượng các bản tóm tắt được tạo ra (AI so sánh với người)

và "hello, this is <họ và tên>".

Việc huấn luyện lại mô hình BertSum trừu tượng trên tập dữ liệu How2 đã mang lại một kết quả thú vị và ngoài mong đợi - mô hình hội tụ vào trạng thái tạo ra cùng một tóm tắt vô nghĩa với những từ khoá nhấn mạnh chung cho hầu hết các video, bất kể lĩnh vực: "học cách làm cách cách một trong những video miễn phí này về làm cách cách và chuyên gia chuyên gia trong cách cách độc đáo và chuyên gia và chuyên gia. để sử dụng và chuyên nghiệp. Chuyên gia độc đáo này dành cho chuyên gia."

Experiment					
Model	Pretraining Data	Test Set	Rouge-1	Rouge-L	Content-F1
1. BertSum, BertSum with pre+post processing	CNN/DM	How2	18.08 to 22.47	18.01 to 20.07	26.0
2. BertSum with random training	How2, 1/50 Sampled-WikiHow, CNN/DM	How2	24.4	21.45	18.7
3. BertSum with random training and postprocessing	How2, 1/50 Sampled-WikiHow, CNN/DM	How2	26.32	22.47	32.9
4. BertSum with ordered training	How2, WikiHow, CNN/DM	How2	48.26	44.02	36.4
5. BertSum	WikiHow	WikiHow	35.91	34.82	29.8
6. BertSum [14]	CNN/DM	CNN/DM	43.23	39.63	Out of Scope
7. Multi-modal Model [18]	How2	How2	59.3	59.2	48.9
8. MatchSum (BERT-base) [27]	WikiHow	WikiHow	31.85	29.58	Not Available
9. Lead 3 for WikiHow [9]	Not Applicable	CNN/DM	40.34	36.57	Not Available
10. Lead 3 for CNN/DM [9]	Not Applicable	WikiHow	26.00	24.25	Not Available
11. Lead 3 for How2 [9]	Not Applicable	How2	23.66	20.69	16.2

Bảng 3: So sánh các kết quả

Trong chuỗi các thí nghiệm tiếp theo của chúng tôi, chúng tôi đã sử dụng tập dữ liệu mở rộng để huấn luyện. Mặc dù sự khác biệt trong các điểm ROUGE của kết quả từ BertSum Model 1 (xem Bảng 3) không quá đáng kể so với BertSum Model 2 và 3, chất lượng của các tóm tắt từ quan điểm của các thẩm định viên có sự khác biệt về mặt chất lượng. Kết quả tốt nhất của chúng tôi trên các video How2 (xem thí nghiệm 4 trong Bảng 3) đạt được bằng cách tận dụng toàn bộ tập dữ liệu đã được gán nhãn (CNN/DM, WikiHow và các video How2) với cấu hình bảo tồn thứ tự. Các điểm ROUGE tốt nhất mà chúng tôi đạt được cho tóm tắt video có thể so sánh với kết quả tốt nhất trong tài liệu tin tức [14] (xem hàng 9 trong Bảng 3).

Cuối cùng, chúng tôi đã vượt qua kết quả tốt nhất hiện tại trên tập dữ liệu WikiHow. Điểm số Rouge-L hiện tại cho tập dữ liệu WikiHow là 26.54 ở hàng 8. Mô hình của chúng tôi sử dụng mô hình tóm tắt trừu tượng BERT để đạt được điểm Rouge-L là 36.8 ở hàng 5, vượt qua điểm chuẩn hiện tại 10.26 điểm. So với mô hình Pointer Generator + Coverage, cải tiến trên Rouge-1 và Rouge-L là khoảng 10 điểm mỗi mục. Chúng tôi đạt được cùng kết quả khi thử nghiệm trên tập dữ liệu WikiHow bằng cách sử dụng BertSum với việc huấn luyện có thứ tự trên toàn bộ tập dữ liệu How2, WikiHow và CNN/DailyMail.

Với kết quả ban đầu của chúng tôi, chúng tôi đã đạt được các mô tả video trôi chảy và dễ hiểu, mang lại một ý tưởng rõ ràng về nội dung. Điểm số của chúng tôi không vượt qua các điểm số từ các nhà nghiên cứu khác [20] mặc dù chúng tôi đã sử dụng BERT. despite employing BERT. Tuy nhiên, các tóm tắt của chúng tôi dường như có nội dung trôi chảy hơn và hữu ích hơn đối với người dùng xem các tóm tắt trong How-To. Một số ví dụ được đưa ra trong [Phụ lục **C**].

Tóm tắt trừu tượng đã giúp giảm thiểu ảnh hưởng của lỗi chuyển đổi từ giọng nói sang văn bản, được quan sát trong một số bản ghi lại, đặc biệt là các phụ đề tự động được tạo ra trong tập dữ liệu bổ sung như một phần của dự án này (các bản ghi lại trong video How2 đã được xem xét và sửa lỗi một cách thủ công, do đó lỗi chính tả ở đó ít xuất hiện hơn). Ví dụ, trong một mẫu của tập dữ liệu kiểm tra, phụ đề đã gây nhầm lẫn từ "how you get a text from a YouTube video" thành "how you get attacks from a YouTube video". Vì thông thường có nhiều sự trùng lặp trong các giải thích, mô hình vẫn có thể tìm ra ngữ cảnh đủ để tạo ra một tóm tắt có ý nghĩa. Chúng tôi không quan sát thấy các tình huống mà các tóm tắt không khớp với chủ đề của video do ảnh hưởng từ các lỗi chính tả thường xảy ra trong các kịch bản được tạo ra bằng ASR mà không có sự giám sát của con người, nhưng đảm bảo đúng ranh giới giữa các câu bằng cách sử dụng Spacy để sửa các lỗi dấu câu trong giai đoạn tiền xử lý đã tạo ra một sự khác biệt rất lớn.

Dựa trên những quan sát này, chúng tôi quyết định rằng mô hình đã tạo ra những kết quả mạnh mẽ tương đương với các mô tả do con người viết. Để phân tích sự khác biệt về chất lượng của tóm tắt, chúng tôi sử dụng sự giúp đỡ của các chuyên gia để đánh giá các đặc điểm hội thoại giữa các tóm tắt của chúng tôi và các mô tả mà người dùng cung cấp cho video của họ trên YouTube. Chúng tôi đã tuyển dụng một nhóm đa dạng gồm hơn 30 tình nguyện viên để đánh giá một tập hợp gồm 25 tóm tắt video được tạo ra bởi mô hình của chúng tôi và các mô tả video từ tập dữ liệu hội thoại. Chúng tôi tạo ra hai loại câu hỏi: loại đầu tiên, một phiên bản của thử thách Turing nổi tiếng, là một thử thách để phân biệt AI và các mô tả được tạo ra bởi con người và sử dụng framework được mô tả trong mục 3.4. Người tham gia được thông báo sẽ có khả năng như nhau rằng một số, tất cả hoặc không có tóm tắt nào được tạo ra bởi máy trong nhiệm vụ phân loại này. Câu hỏi thứ hai thu thập một phân phối các điểm đánh giá liên quan đến chất lượng hội thoại. Kết quả tổng hợp cho cả hai đánh giá được trình bày trong các hình từ 6-8. Chúng tôi quan sát thấy không có điểm hoàn hảo nào trong câu trả lời của thử thách Turing. Kết quả bao gồm nhiều kết quả bị báo động nhầm (False Positive) và bỏ sót (False Negative) [Phụ lục **D**].

Chất lượng đầu ra kiểm tra của chúng ta tương đương với các bản tóm tắt trên YouTube. "Văn bản thực tế" là cơ hội phát triển chính vì mô hình trừu tượng có xu hướng tạo ra những câu không liên kết và không logic mặc dù vẫn đúng ngữ pháp. Các tác giả con người thường mắc phải lỗi sử dụng ngôn ngữ. Ưu điểm của việc sử dụng mô hình tóm tắt trừu tượng là cho phép chúng ta giảm nhẹ một số vấn đề về ngữ pháp của tác giả video.

5 KẾT LUẬN

Công trình nghiên cứu của chúng tôi giải quyết nhiều vấn đề mà chúng tôi đã xác định trong quá trình nỗ lực tổng quát hóa mô hình BertSum để tóm tắt các kịch bản video hướng dẫn trong suốt quá trình training.

- Chúng tôi đã khám phá cách các sự kết hợp khác nhau giữa dữ liệu training và các tham số sẽ ảnh hưởng đến hiệu suất training của mô hình tóm tắt trừu tượng BertSum như thế nào.
- Chúng tôi đã tạo ra các bước tiền xử lý mới cho các kịch bản phụ đề tự động được tạo ra trước quá trình tóm tắt.
- Chúng tôi đã tổng quát hóa mô hình tóm tắt trừu tượng BertSum cho các kịch bản video hướng dẫn tự động với chất lượng gần bằng các mô tả được chọn ngẫu nhiên do người dùng YouTube tạo ra.
- Chúng tôi đã thiết kế và triển khai một framework mới cho review không thiên vị, từ đó tạo ra các điểm số có tính hiện thực hóa và khách quan hơn, bổ sung cho ROUGE, BLEU và Content F1. Tất cả các điểm được liệt kê ở trên đều có sẵn trong kho lưu trữ của chúng tôi để phục vụ cho các mục đích nghiên cứu trong tương lai.

Nhìn chung, kết quả mà chúng tôi đạt được cho đến nay từ các video hướng dẫn nghiệp dư khiến chúng tôi tin rằng mình đã tạo ra một mô hình huấn luyện có khả năng tạo ra các bản tóm tắt từ các kịch bản ASR (chuyển đổi giọng nói thành văn bản) với chất lượng cạnh tranh so với các mô tả được lựa chọn bởi con người trên YouTube, dù chỉ có số lượng hạn chế các tập dữ liệu tóm tắt đã được gán nhãn có sẵn. Mặc dù bài báo này tập trung vào các mô hình BERTSum và mang lại kết quả tốt nhất trong nhiệm vụ tóm tắt, BERTSum chỉ là một biến thể của BERT và chỉ hoạt động tốt trên một số tập dữ liệu mà chúng tôi đã đề cập trong bài báo này. Điều này có nghĩa là mô hình BERTSum cần phát triển sâu hơn để có khả năng tổng quát hóa và áp dụng vào nhiều công việc hơn trong Xử lý Ngôn ngữ Tự nhiên (NLP). Trong tương lai, chúng tôi có thể tiếp tục làm việc với BERT nhưng trên các tập dữ liệu tiếng Việt để không chỉ mang lại các video thú vị từ nước ngoài vào Việt Nam, mà còn ngược lại. Hơn nữa, chúng tôi có thể học và xem xét các thuật toán mới nhất và xử lý tốt nhiệm vụ tóm tắt như Pegasus [28], T5 [29], ProphetNet [30] để cải thiện mô hình trong việc tóm tắt hướng dẫn.

TRI ÂN

Trước hết chúng tôi xin gửi lời cảm ơn đến giảng viên Nguyễn Sơn Lâm, người đã đưa ra

những hướng dẫn và góp ý đầy giá trị trong suốt quá trình thực hiện nghiên cứu. Chúng tôi cũng xin tri ân đến những khảo sát viên đã dành thời gian cho phần nghiên cứu đánh giá của chúng tôi.

THAM KHẢO

[1] Yoshua Bengio, J'me Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML (ACM International Conference Proceeding Series)*, Andrea Pohoreckyj Danyluk, L'on Bottou, and Michael L. Littman (Eds.), Vol. 382. ACM, 41–48. http://dblp.uni-trier.de/db/conf/icml/icml2009.html #BengioLCW09

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. CoRR abs/1406.1078 (2014). arXiv:1406.1078 http://arxiv.org/abs/1406.1078

[3] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: *Language Specific Translation Evaluation for Any Target Language*. In Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.

[4] B. Erol, D. . Lee, and J. Hull. 2003. *Multimodal summarization of meeting recordings*. In 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), Vol. 3. III–25.

[5] Berna Erol, Dar-Shyang Lee, and Jonathan J. Hull. 2003. *Multimodal summarization of meeting recordings*. In Proceedings of the 2003 IEEE International Conference on Multimedia and Expo, ICME 2003, 6-9 July 2003, Baltimore, MD, USA. IEEE Computer Society, 25–28. https://doi.org/10.1109/ICME.2003.1221239.

[6] GardenFork. 2018. *DIY How-to Videos*. Retrieved July 15, 2020 from https://www.youtube.com/playlist?list=PL05C1F99A68D37472.

[7] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. *Teaching Machines to Read and Comprehend*. In Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 1693–1701. http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf

[8] Matthew Honnibal and Ines Montani. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. (2017). To appear.

[9] Mahnaz Koupaee and William Yang Wang. 2018. *WikiHow: A Large Scale Text Summarization Dataset*. CoRR abs/1810.09305 (2018). arXiv:1810.09305 http://arxiv.org/abs/1810.09305.

[10] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. *Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video*. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, 1092–1102. https://doi.org/10.18653/v1/D17-1114.

[11] Chin-Yew Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013.

[12] Chin-Yew Lin and Eduard Hovy. 2002. *Manual and Automatic Evaluation of Summaries*. In Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4 (Phildadelphia, Pennsylvania) (ASR 02). Association for Computational Linguistics, USA. https://doi.org/10.3115/1118162.1118168.

[13] Chunyi Liu, Peng Wang, Xu Jiang, Zang Li, and Jieping Ye. 2019. *Automatic Dialogue Summary Generation for Customer Service*. 1957–1965. https://doi.org/10.1145/3292500.3330683.

[14] Yang Liu and Mirella Lapata. 2019. *Text Summarization with Pretrained Encoders*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3730–3740. https://doi.org/10.18653/v1/D19-1387.

[15] Edward Loper and Steven Bird. 2002. *NLTK: The Natural Language Toolkit*. CoRR cs.CL/0205028 (2002). http://dblp.unitrier.de/db/journals/corr/corr0205.html#cs-CL-0205028.

[16] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips.* CoRR abs/1906.03327 (2019). arXiv:1906.03327 http://arxiv.org/abs/1906.03327 .

[17] Ani Nenkova. 2005. *Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference*. 1436–1441.

[18] Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. *Multimodal*

Abstractive Summarization for How2 Videos. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 6587–6596. https://doi.org/10.18653/v1/P19-1659.

[19] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. *A Neural Attention Model for Abstractive Sentence Summarization*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, 379–389. https://doi.org/10.18653/v1/D15-1044.

[20] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. *How2: A Large-scale Dataset for Multi-modal Language Understanding*. CoRR abs/1811.00347 (2018). arXiv:1811.00347 http://arxiv.org/abs/1811.00347

[21] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. *Get To The Point: Summarization with Pointer-Generator Networks.* CoRR abs/1704.04368 (2017). arXiv:1704.04368 http://arxiv.org/abs/1704.04368.

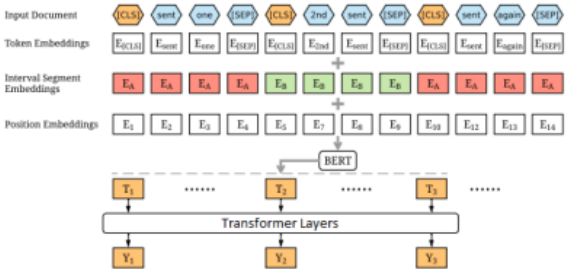
[22] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. *Sequence to Sequence Learning with Neural Networks.* CoRR abs/1409.3215 (2014). arXiv:1409.3215 http://arxiv.org/abs/1409.3215.

[23] Krysta M. Svore, Lucy Vanderwende, and Christopher J. C. Burges. 2007. *Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources*. In EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, Jason Eisner (Ed.). ACL, 448–457. https://www.aclweb.org/anthology/D07-1047/

[24] How to Videos. 2020. *How-to Videos*. Retrieved July 20, 2020 from https://www.youtube.com/channel/UC_qTn8RzUXBP5VJ0q2jROGQ

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need*. CoRR abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762.

[26] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. *Curriculum Learning for Natural Language Understanding*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 6095–6104. https://www.aclweb.org/anthology/2020.acl-main.542



Hình 97: Mô hình BERTSum

[27] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and X. Huang. 2020. *Extractive Summarization as Text Matching*. ArXiv abs/2004.08795 (2020)

[28] Zhang, J., Zhao, Y., Saleh, M., Liu, P. J., & Chang, S. F. (2020). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) https://arxiv.org/abs/1912.08777

[29] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2020). *"BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension."* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871-7880 https://arxiv.org/abs/1910.13461

[30] Yang, L., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R. R., & Le, Q. V. (2019). *ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training*. Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL), 2019, 2292-2305. https://arxiv.org/pdf/2001.04063v3.pdf

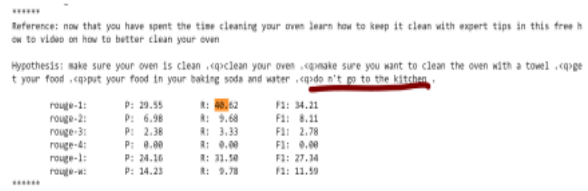
A CHI TIẾT MÔ HÌNH

Tóm tắt trích xuất (extractive summarization) thông thường là một công việc phân loại nhị phân có gắn nhãn cho biết liệu các câu có nên được cho vào trong bản tóm tắt hay không. Ngược lại, tóm tắt trừu tượng (abstractive summarization) yêu cầu khả năng tạo ra ngôn ngữ để tạo ra các bản tóm tắt chứa các từ và cụm từ mới không có trong văn bản nguồn.

Hình 9 thể hiện mô hình BertSum. Nó sử dụng một bộ mã hóa mới dựa trên BERT có thể mã hóa một tài liệu và thu được biểu diễn của các câu. CLS token được thêm vào mỗi câu thay vì chỉ có 1 CLS token trong mô hình

BERT gốc. Mô hình trừu tượng sử dụng một kiến trúc mã hóa giải mã, kết hợp bộ mã hóa BERT được tiền huấn luyện cùng với một bộ giải mã Transformer được khởi tạo ngẫu nhiên. Mô hình sử dụng một kỹ thuật đặc biệt, trong đó phần mã hóa được giữ gần như không thay đổi với một tỷ lệ học thấp rất thấp và một tỷ lệ học riêng biệt được sử dụng cho bộ giải mã để giúp nó học tốt hơn.

B HÌNH MINH HỌA LÝ DO CHỈ SỐ ROUGE BỊ HẠN CHẾ



Hình 80: Một ví dụ mà chỉ số ROUGE gây hiểu lầm

C VÍ DỤ SO SÁNH ĐẦU RA MÔ HÌNH VỚI ĐIỂM CHUẨN VÀ CÁC TÓM TẮT THAM KHẢO

Các ví dụ dưới đây đã được chọn để minh họa một số khía cạnh của vấn đề. Đầu tiên, chúng tôi chia sẻ URL của video để người đọc có thể xem nội dung gốc. Thứ hai, chúng tôi chia sẻ kết quả cuối cùng của quá trình tóm tắt trừu tượng với phiên bản mô hình tốt nhất hiện tại của chúng tôi (Summary Abs). Để so sánh, chúng tôi cung cấp các bản tóm tắt từ các video Điểm chuẩn của How2 hiện tại đang bỏ qua mô hình của chúng tôi về mặt điểm số, nhưng, như có thể thấy trong các ví dụ này, không phải về mức độ trôi chảy và hữu ích. Tài liệu tham khảo đại diện cho mô tả video thực tế trên YouTube do các tác giả tuyển chọn. Ngược lại, chúng tôi hiển thị Summary Ext - kết quả của tóm tắt trích xuất, điều này giải thích tại sao tóm tắt trừu tượng lại phù hợp hơn cho mục đích này, vì chúng tôi đang cố gắng thực hiện chuyển đổi phong cách từ văn bản nói cho văn bản nguồn sang văn bản tóm tắt. Vì BertSum không viết hoa, nên tất cả các văn bản bên dưới đã được chuyển đổi thành chữ thường để thống nhất.

- Video 1: https://www.youtube.com/watch?v=F_4UZ3bGMP8.
- Summary Abs 1: growing rudbeckia requires full hot sun and good drainage. grow rudbeckia with tips from a gardening specialist in this free video on plant and flower care. care for rudbeckia with gardening tips from an experienced gardener.
- Benchmark 1: growing black - eyed - susan is easy with these tips, get expert gardening tips in this free gardening video. Reference 1: growing rudbeckia plants requires a good deal of hot sun and plenty of good drainage for water. start a rudbeckia plant in the winter or anytime of year with advice from a gardening specialist in this free video on plant and flower care.
- Summary Ext 1: make sure that your plants are in your garden. get your plants. don’t go to the flowers. go to your garden’s soil. put them in your plants in the water. take care of your flowers.
- Video 2: https://www.youtube.com/watch?v=LbsGHj2Akao
- Summary 2: camouflage thick arms by wearing sleeves that are not close to the arms and that have a line that goes all the way to the waist. avoid wearing jackets and jackets with tips from an image consultant in this free video on fashion. learn how to dress for fashion modeling.
- Benchmark 2: hide thick arms and arms by wearing clothes that hold the arms in the top of the arm. avoid damaging the arm and avoid damaging the arms with tips from an image consultant in this free video on fashion.
- Reference 2: hide thick arms by wearing clothes sleeves that almost reach the waist to camouflage the area .conceal the thickness at the top of the arms with tips from an image consultant in this free video on fashion.

- Summary Ext 2: make sure that you have a look at the top of the top. if you want to wear the right arm. go to the shoulder. wear a long-term shirts. keep your arm in your shoulders. don’t go out.

D VÍ DỤ CHO VIỆC BÁO ĐỘNG NHẦM VÀ BỎ SÓT TỪ KẾT QUẢ KHẢO SÁT

False Negative (FN): Các khảo sát viên cho rằng những mẫu tóm tắt được viết bởi máy trong khi chúng được viết bởi con người. False Positive (FP): Các khảo sát viên cho rằng những mẫu tóm tắt được viết bởi con người trong khi chúng được viết bởi máy.

Ví dụ FN:

- "permanently fix flat atv tires with tireject ??. dry rot, bead leaks, nails, sidewall punctures are no issue. these 30yr old atv tires permanently sealed and back into service in under 5 min. they sealed right up and held air for the first time in a long time. this liquid rubber and kevlar are a permanent repair and will protect from future punctures."
- "how to repair a bicycle tire : how to remove the tube from bicycle tires. by using handy tire levers, expert cyclist shows how to remove the tube from bicycle tires, when changing a flat tire, in this free bicycle repair video."

Ví dụ FP:

- "learn about the parts of a microscope with expert tips and advice on refurbishing antiques from a professional photographer in this free video clip on home astronomy and buildings. learn more about how to use a light microscope with a demonstration from a science teacher."
- "watch as a seasoned professional demonstrates how to use a deep fat fryer in this free online video about home pool care. get professional tips and advice from an expert on how to organize your kitchen appliance and kitchen appliance for special occasions.