NÂNG CAO TÓM TẮT TRÍCH XUẤT: CẢI TIẾN VÀ ĐỔI MỚI TRONG BERTSUM

Trần Hữu Lộc - 22520796 **Trần Tuấn Khoa - 22520692**

Trường Đại học Công nghệ thông tin,

Mục tiêu

Chúng tôi thử nghiệm cải tiến và đổi mới BERTSUM bằng việc:

- Tích hợp các mô hình transformer tiên tiến như: RoBERTa và Transformer-XL.
- Sử dụng kỹ thuật mã hóa dựa trên đồ thị: GCNs

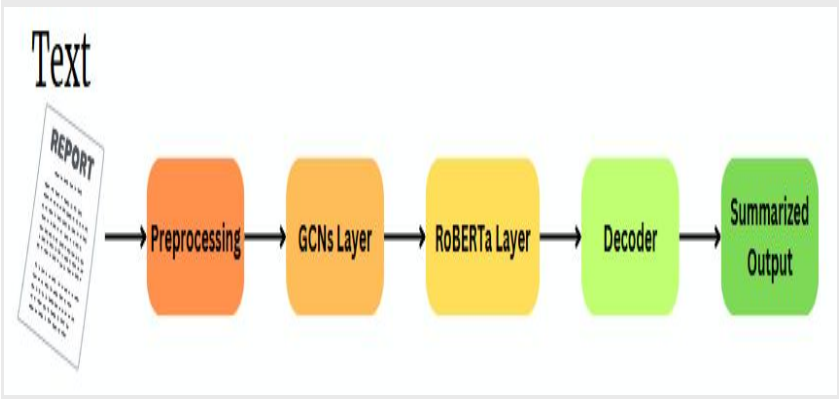
Chúng tôi còn thu thập và tạo ra một bộ dữ liệu mới cho vấn đề tóm tắt trích xuất văn bản.

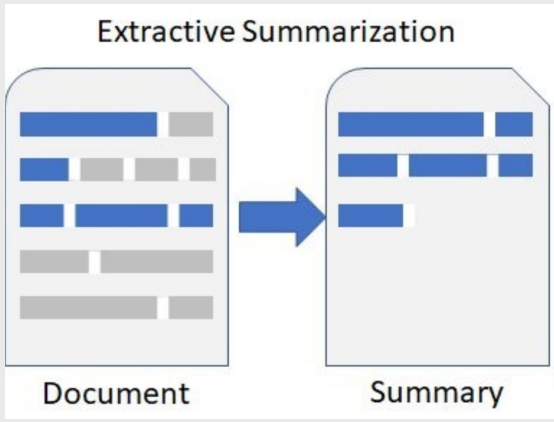
Lí do chọn đề tài?

Tóm tắt trích xuất đóng vai trò quan trọng trong việc rút ngắn thông tin văn bản dài thành các bản tóm tắt ngắn gọn nhưng vẫn giữ lại được thông tin quan trọng. BERTSUM, sử dụng BERT, là 1 mô hình mạnh mẽ trong việc tạo ra các bản tóm tắt bằng cách tóm tắt trích xuất nhưng lại gặp khó khăn trong việc xử lý các phụ thuộc xa và quan hệ giữa các câu trong văn bản. Vì thế, chúng tôi thử nghiệm cải tiến và đổi mới BERTSUM.
Ngoài ra, chúng tôi muốn tạo thêm bộ dữ liệu mới đủ tin cậy để phát triển vấn đề tóm tắt trích xuất văn bản.

Tổng quan

Cải tiến và đổi mới BERTSUM → Chương trình minh họa → Bộ dữ liệu mới






```
Original Paragraph:

"Artificial intelligence (AI) is revolutionizing healthcare. It helps doctors diagnose diseases more accurately and create personalized treatment plans. AI also speeds up drug discovery, making it faster and more cost-effective."

Extractive Summary:

"Artificial intelligence (AI) is revolutionizing healthcare. It helps doctors diagnose diseases more accurately and create personalized treatment plans."
```

Mô tả

1. Nội dung

- Tìm hiểu câu trả lời cho câu hỏi: **"Liệu việc áp dụng tích hợp mô hình transformer tiên tiến như RoBERTa và Transformer-XL và sử dụng kỹ thuật mã hóa dựa trên đồ thị có nâng cao được hiệu quả cho bài toán tóm tắt trích xuất văn bản so với các phương pháp truyền thống và các mô hình tóm tắt trích xuất khác hay không?"**
- Tích hợp các mô hình transformer tiên tiến
- Sử dụng kỹ thuật mã hóa dựa trên đồ thị
- Đánh giá và so sánh hiệu quả
- Tạo bộ dữ liệu mới

3. Kết quả dự kiến

- BERTSUM có hiệu quả tốt hơn sau khi áp dụng các cải tiến và đổi mới
- Chương trình minh họa
- Một bộ dữ liệu mới có độ tin cậy cao

2. Phương pháp

- **Thu thập và tiền xử lý dữ liệu:** Sử dụng các bộ dữ liệu phù hợp như CNN/DailyMail, XSUM để huấn luyện và đánh giá BERTSUM, bao gồm việc tiền xử lý để chuẩn bị dữ liệu cho quá trình huấn luyện và kiểm thử. Ngoài ra, chúng tôi còn tự tạo ra 1 bộ dữ liệu mới đóng góp cho vấn đề tóm tắt trích xuất.
- **Huấn luyện và fine-tuning mô hình:**
  - Biểu diễn mối quan hệ ngữ nghĩa giữa các câu dưới dạng đồ thị.
  - Sử dụng GCNs để mã hóa đồ thị ngữ nghĩa này, giúp nắm bắt các mối quan hệ phức tạp và phụ thuộc xa giữa các câu.
  - Huấn luyện các mô hình trên các bộ dữ liệu lớn như CNN/DailyMail và XSUM.
  - Fine-tuning các mô hình transformer như RoBERTa và Transformer-XL để tối ưu hóa hiệu suất của BERTSUM trong việc tóm tắt trích xuất.
- **Đánh giá và so sánh kết quả:** Đánh giá khách quan hiệu quả của BERTSUM sau khi áp dụng các cải tiến và đổi mới
- **Phân tích kết quả và đề xuất:** Phân tích các kết quả thu được từ nghiên cứu và đề xuất các hướng phát triển tiếp theo để cải thiện khả năng tóm tắt trích xuất của BERTSUM trong các ứng dụng thực tế.

NII **Trần Hữu Lộc– Trường Đại học Công nghệ thông tin, ĐHQG-HCM**
TEL : 0815750047 Email : trhuuloc04@gmail.com