ĐỀ CƯƠNG ĐỀ TÀI LUẬN VĂN THẠC SĨ

TÊN ĐỀ TÀI: NGHIÊN CỨU THỰC NGHIỆM CÁC MÔ HÌNH BERT CHO TÁC VỤ TÓM TẮT ĐA VĂN BẢN TRÊN TIẾNG VIỆT

TÊN ĐỀ TÀI (tiếng Anh): EMPIRICAL STUDY OF TEXT SUMMARIZATION ON MULTI-DOCUMENTS IN VIETNAMESE USING BERT

Thời gian thực hiện: 6 tháng

Học viên thực hiện:

Tô Quốc Huy

Giới thiệu đề tài:

- Tóm tắt văn bản là việc cô đọng thông tin từ một hoặc nhiều đoạn văn bản thành một đoạn văn bản ngắn hơn. Tuy giảm thiểu số lượng câu chữ nhưng vẫn phải đảm bảo các yếu tố như thông tin và ý nghĩa về mặt nội dung. Các ứng dụng của tóm tắt văn bản tự động bao gồm: phân loại văn bản lớn, Question Answering, tóm tắt văn bản pháp lý, tóm tắt tin tức, tạo tiêu đề tự động.
- Việc tự động hóa công việc tóm tắt đang ngày càng phổ biến và độ hiểu quả cải thiện dần theo thời gian**.** Trên Tiếng Việt, các mô hình tóm tắt tự động đã được đề xuất như TSGVi [1], CFVi [2] và một số mô hình khác dựa trên thuật toán TextRank.
- Đề tài nghiên cứu độ hiệu quả của mô hình pre-train BERT, một mô hình hiện đại đã được thực nghiệm và chứng minh độ hiệu quả trên các tác vụ khác trong lĩnh vực xử lý ngôn ngữ tự nhiên. Tuy đã được thực nghiệm trên Tiếng Anh [3] nhưng chưa được thực nghiệm và đánh giá trên Tiếng Việt.

Mục tiêu đề tài:

- Nghiên cứu các mô hình BERT, các kỹ thuật có liên quan cho bài tóm tắt đa văn bản trên Tiếng Việt

- Chạy thực nghiệm để kiểm chứng độ chính xác và đánh giá hiệu suất của các mô hình.
- Cải thiện độ chính xác của mô hình, chọn ra mô hình tốt nhất cho bài toán tóm tắt đa văn bản.

Nội dung nghiên cứu:

- Tạo tự động các đoạn tóm tắt sử dụng các mô hình BERT đa ngôn ngữ và đơn ngôn ngữ kết hợp với thuật toán K-Means clustering trên bộ dữ liệu VietnameseMDS.
- Đánh giá và so sánh độ hiệu quả trên độ đo ROUGE giữa các mô hình BERT và với các mô hình đã được đề xuất trước đó.
- Tối ưu mô hình dựa trên các kết quả phân tích để cho kết quả tốt nhất.

Phương pháp thực hiện:

- Nghiên cứu phương pháp thực nghiệm:
    - Thực nghiệm trên các mô hình pre-trained BERT đa ngôn ngữ và đơn ngôn ngữ, điều chỉnh tham số để tìm ra mô hình phù hợp với bài toán đặt ra, như:
        - BERT-multilingual
        - XML-Roberta
        - DistilBERT-multilingual
        - PhoBERT
        - ViBERT4News
    - Đồng thời áp dụng thuật toán như K-Means Clustering để trích xuất các câu có liên quan với chủ đề.
    - Các phương pháp được thực nghiệm trên bộ dữ liệu VietnameseMDS, bao gồm 200 cụm văn bản chuyên dùng cho tóm tắt đa văn bản.
    - Tối ưu các giá trị tham số của mô hình dựa trên kết quả phân tích từ kết quả độ do đánh giá trên từng mô hình.

Kết quả mong đợi:

- Đề xuất được mô hình pre-train BERT tốt nhất cho tác vụ tóm tắt dã văn bản trên Tiếng Việt

Tài liệu tham khảo:

[1] T. A. Nguyen Hoang, H. K. Nguyen and Q. V. Tran, "An Efficient Vietnamese Text Summarization Approach Based on Graph Model," *2010 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, 2010, pp. 1-6, doi: 10.1109/RIVF.2010.5633162.

[2] Ung, V.-G., Luong, A.-V., Tran, N.-T., & Nghiem, M.-Q. (2015). Combination of Features for Vietnamese News Multi-document Summarization. *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE).* doi:10.1109/kse.2015.71

[3] Miller, Derek. "Leveraging BERT for extractive text summarization on lectures." *arXiv preprint arXiv:1906.04165* (2019).

Kế hoạch thực hiện:

Chúng tôi tiến hành thực hiện nghiên cứu đề tài đọc hiểu văn bản tự động trong vòng 6 tháng và kế hoạch thực hiện chi tiết trong bảng 1.

Bảng 1 - Kế hoạch thực hiện đề tài trong 6 tháng

Công việc	T1	T2	T3	T4	T5	T6
Tìm hiểu tổng quan						
Nghiên cứu phương pháp						
Báo cáo						