THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo): *https://www.youtube.com/watch?v=vlIDONTg-7c*
- Link slides: *https://github.com/thuynguyen2003/CS519.011-PPLKCKH/blob/main/Slide.pdf*

• Họ và Tên: Nguyễn Thị Thùy • MSSV: 21521514 	• Lớp: CS519.O11 • Tự đánh giá (điểm tổng kết môn): 9.5/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 11 • Link Github: https://github.com/thuynguyen2003/CS519.011-PPLKCKH • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: Tất cả các công việc của đồ án đều được thực hiện bởi thành viên duy nhất là sinh viên Nguyễn Thị Thùy

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TRÍCH XUẤT THÔNG TIN SO SÁNH TRONG BÌNH LUẬN TIẾNG VIỆT DỰA TRÊN MÔ HÌNH TẠO SINH VĂN BẢN

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

COMPARATIVE OPINION QUINTUPLE EXTRACTION IN VIETNAMESE REVIEW BASED ON THE TEXT GENERATION MODELS

TÓM TẮT

Trong bối cảnh thị trường kinh doanh trực tuyến ngày càng phát triển, nhu cầu tìm hiểu về sản phẩm thông qua nhận xét của các người dùng khác ngày càng tăng, cộng đồng mua bán trực tuyến thường xuyên tạo ra những đánh giá chi tiết và so sánh, đặt ra một thách thức lớn về việc tổng hợp thông tin từ nhiều nguồn. Điều này dẫn đến một vấn đề: Làm thế nào để người tiêu dùng có thể so sánh và tổng hợp thông tin một cách thuận tiện, nhanh chóng trong việc lựa chọn được sản phẩm phù hợp?

Đề tài tập trung vào nghiên cứu và áp dụng các mô hình tạo sinh văn bản, kết hợp với kỹ thuật Xử lý ngôn ngữ Tự nhiên, để giải quyết bài toán trích xuất nhóm ý kiến so sánh (COQE) trong tiếng Việt. Mục tiêu là tổng hợp và triển khai các phương pháp và mô hình hiệu quả để giải quyết bài toán này, mang lại lợi ích cho người tiêu dùng, doanh nghiệp và cộng đồng trực tuyến. Nghiên cứu có tiềm năng thúc đẩy phát triển trong lĩnh vực thương mại điện tử và đánh giá sản phẩm, dịch vụ tại Việt Nam.

Bài toán COQE có nhiệm vụ xác định câu so sánh và trích xuất ra các yếu tố so sánh có trong câu. Để giải quyết bài toán này, đề tài tập trung vào nghiên cứu và triển khai các mô hình tạo sinh văn bản dựa trên kiến trúc transformer và đánh giá trên bộ dữ liệu tiếng Việt được thu thập từ cuộc thi "ComOM - Comparative Opinion Mining from Vietnamese Product Reviews".

GIỚI THIỆU

Trong thời đại kinh doanh trực tuyến ngày càng phát triển và đa dạng, người tiêu dùng

đang đối diện với một lượng khổng lồ các sản phẩm và dịch vụ. Sự bùng nổ của các nền tảng kinh doanh trực tuyến đã tạo ra một cảnh ngạc nhiên về sự đa dạng và phong phú trong thị trường. Điều này đặt ra câu hỏi cốt lõi: Làm thế nào để người tiêu dùng có thể tìm kiếm, so sánh và tổng hợp thông tin một cách thuận tiện, nhanh chóng để đưa ra quyết định mua sắm thông thái và hợp lý?

Hiện nay, trong các phân khúc thường sẽ xuất hiện nhiều sản phẩm hoặc dịch vụ của các nhãn hàng, công ty với nhau. Điều này dẫn đến việc người dùng có xu hướng tìm kiếm sự so sánh giữa các sản phẩm, dịch vụ để đưa ra được sự quyết định phù hợp với nhu cầu của mình. Bên cạnh đó, hiện nay cũng có nhiều cộng đồng trực tuyến trong mỗi lĩnh vực sản phẩm thường tạo ra bài đánh giá chi tiết và so sánh. Điều này đặt ra một thách thức lớn về việc tổng hợp thông tin từ nhiều nguồn để đưa ra quyết định mua sắm thông minh. Chúng ta cần một cách hiệu quả để so sánh và tìm kiếm thông tin về sự khác biệt giữa các sản phẩm và phiên bản khác nhau, và việc tự động hóa quá trình này trở nên cực kỳ quan trọng.

Đề tài này nghiên cứu, áp dụng các mô hình tạo sinh văn bản cùng với các kỹ thuật trong lĩnh vự Xử lý ngôn ngữ Tự nhiên để giải quyết những thách thức này, mang lại lợi ích to lớn cho người tiêu dùng, doanh nghiệp và cộng đồng trực tuyến. Nghiên cứu trong khóa luận này có tiềm năng thúc đẩy phát triển trong lĩnh vực thương mại điện tử và đánh giá sản phẩm, dịch vụ tại Việt Nam.

Xác định bài toán:

Bài toán trích xuất nhóm ý kiến so sánh (Comparative Opinion Quintuple Extraction [1] - COQE) nhận vào một câu nhận xét, nếu câu đó mang ý so sánh thì sẽ trả một tập hợp 5 phần tử gọi là “quintuple” bao gồm:

- Subject (s): Chủ thể của sự so sánh.
- Object (o): Khách thể được so sánh với chủ thể.
- Comparative Aspect (ca): Từ hoặc cụm từ về tính năng hoặc thuộc tính của chủ thể và khách thể đang được so sánh.
- Comparative Opinion (co): Từ hoặc cụm từ so sánh thể hiện ý kiến so sánh.

- Comparative Preference (cp): Mức độ so sánh.

- Đầu vào: Một câu nhận xét sản phẩm
- Đầu ra: Danh sách các quintuple trong câu nếu có.

$$S = \{..., (s, o, ca, co, cp), \ldots\}$$

MỤC TIÊU

- Nghiên cứu các bài báo, báo cáo khoa học, luận văn, ... liên quan đến bài toán trích xuất nhóm ý kiến so sánh (COQE) và sự phát triển của các mô hình tạo sinh văn bản cho tiếng Việt. Từ đó tổng hợp được các phương pháp và mô hình hiệu quả để giải quyết bài toán này.
- Triển khai và tinh chỉnh các mô hình tạo sinh văn bản cùng với phương pháp đã tìm được để giải quyết bài toán COQE tiếng Việt.
- Đánh giá hiệu quả của các mô hình này trên các tập dữ liệu tiếng Việt.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung:

- Nghiên cứu các bài báo, báo cáo khoa học, luận văn, ... liên quan đến bài toán khai thác, phân tích, trích xuất ý kiến so sánh và sự phát triển của các mô hình tạo sinh văn bản cho tiếng Việt.
- Tìm hiểu về các bài toán tiền đề của COQE như: Comparative Sentence Identification – CSI [2], Comparative Element Extraction – CEE [2], Comparative Preference Classification – CPC [3].
- Nghiên cứu các mô hình tạo sinh văn bản dựa trên kiến trúc transformer [4] như: T5, viT5, PhoGPT, GPT2 ….
- Đọc và tổng hợp phương pháp giải quyết bài toán COQE sử dụng các mô hình tạo sinh văn bản như UniCOQE [5], MvP [6], Multitask Finetuning [7].
- Thu thập và chuẩn bị dữ liệu từ cuộc thi ComOM để sử dụng trong quá trình đánh giá hiệu quả bằng các thang đo như độ chính xác, Macro F1, Micro F1 …

Phương pháp:

- Tiến hành đánh giá và phân tích các công trình nghiên cứu liên quan về khai

thác, phân tích và trích xuất ý kiến so sánh.

- Tìm hiểu về cấu trúc transformer và sự phát triển của các mô hình tạo sinh văn bản có thể sử dụng cho tiếng Việt.
- Nghiên cứu về các bài toán tiền đề để hiểu rõ bối cảnh của bài toán COQE.
- Tìm hiểu các phương pháp giải quyết bài toán COQE dựa trên các mô hình tạo sinh văn bản.
- Triển khai và tinh chỉnh các mô hình tạo sinh văn bản dựa trên kiến trúc Transformer, kết hợp với phương pháp giải quyết bài toán COQE.
- Sử dụng bộ dữ liệu tiếng Việt từ cuộc thi ComOM để huấn luyện và đánh giá các mô hình và phương pháp đã tìm hiểu.

KẾT QUẢ MONG ĐỢI

- Nâng cao hiệu quả hiệu suất trên bộ dữ liệu ComOM dựa trên các mô hình tạo sinh văn bản.
- Báo cáo về các phương pháp và mô hình ngôn ngữ đã được sử dụng trong việc trong bài toán này, cùng với các kết quả thực nghiệm, so sánh, đánh giá giữa các phương pháp và đề xuất được một số hướng nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

[1] Ziheng Liu, Rui Xia, and Jianfei Yu. 2021. Comparative Opinion Quintuple Extraction from Product Reviews. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3955–3965.

[2] Jindal, N., & Liu, B. (2006). Mining Comparative Sentences and Relations. AAAI Conference on Artificial Intelligence.

[3] Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing Comparative Sentences. In Proceedings of the 6th Workshop on Argument Mining, pages 136–145, Florence, Italy. Association for Computational Linguistics.

[4] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information

Processing Systems.
[5] Zinong Yang, Feng Xu, Jianfei Yu, and Rui Xia. 2023. UniCOQE: Unified Comparative Opinion Quintuple Extraction As A Set. In Findings of the Association for Computational Linguistics: ACL 2023, pages 12229–12240.
[6] Gou, Z., Guo, Q., & Yang, Y. (2023). MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction. Annual Meeting of the Association for Computational Linguistics.
[7] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M., Shen, S., Yong, Z., Schoelkopf, H., Tang, X., Radev, D.R., Aji, A., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023). Crosslingual Generalization through Multitask Finetuning. Annual Meeting of the Association for Computational Linguistics.