

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.011

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM



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

Nguyễn Thị Thùy - 21521514

Tóm tắt

- Lớp: CS519.011
- Link Github của nhóm:
<https://github.com/thuynghuyen2003/CS519.011-PPLKCKH>
- Link YouTube video:
<https://www.youtube.com/watch?v=vIIDONTq-7c>



Nguyễn Thị Thùy - 21521514

Giới thiệu

- Đặt 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.

Giới thiệu

- **Xác định bài toán COQE:**

Bài toán trích xuất 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): Khía cạnh so sánh
- + Comparative Opinion (co): Từ hoặc cụm từ so sánh thể hiện sự 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.

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

Mục tiêu

- Nghiên cứu, tổng hợp các phương pháp và mô hình hiệu quả để giải quyết bài toán trích xuất ý kiến so sánh (COQE).
- 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 trên tiếng Việt.
- Đánh giá hiệu quả của các mô hình này trên tập dữ liệu tiếng Việt.

Nội dung và Phương pháp

Nội dung:

- Tìm hiểu các nghiên cứu 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 CSI, CEE [2], CPC [3]
- Nghiên cứu các mô hình tạo sinh văn bản dựa trên kiến trúc transformer [4] như: T5, viT5, PhoGPT, GPT2 ... và các phương pháp giải quyết bài toán COQE 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 ...

Nội dung và Phương pháp

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 các mô hình tạo sinh văn bản có thể sử dụng cho tiếng Việt cùng với các phương pháp giải quyết bài toán COQE dựa trên các mô hình tạo sinh văn bản.
- Nghiên cứu về các bài toán tiền đề để hiểu rõ bối cảnh của bài toán COQE.
- Triển khai, tinh chỉnh các mô hình và phương pháp đã tìm được và đánh giá trên bộ dữ liệu tiếng Việt

Kết quả dự kiến

- Tìm ra giải pháp nâng cao hiệu suất trên bộ dữ liệu ComOM dựa trên các mô hình tạo sinh văn bản.
- Một bài báo cáo tổng kết đầy đủ kết quả thực hiện nghiên cứu đề tài:
 - Tổng quan về bài toán khai thác phân tích ý kiến so sánh.
 - Phân tích các phương pháp và mô hình tạo sinh văn bản đã sử dụng
 - Kết quả thực nghiệm, so sánh, đánh giá giữa các phương pháp
 - Đánh giá hiệu quả của mô hình trên các tập dữ liệu
 - Đề xuất các 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
- [2] Jindal, N., & Liu, B. (2006). Mining Comparative Sentences and Relations. AAAI Conference on Artificial Intelligence.
- [3] Alexander Panchenko, et al. 2019. Categorizing Comparative Sentences. In Proceedings of the 6th Workshop on Argument Mining, Florence, Italy. Association for Computational Linguistics.
- [4] Vaswani, et al. 2017. Attention is All you Need. Neural Information Processing Systems.
- [5] Zinong Yang, et al. 2023. UniCOQE: Unified Comparative Opinion Quintuple Extraction As A Set. In Findings of the Association for Computational Linguistics: ACL 2023
- [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, et al. (2023). Crosslingual Generalization through Multitask Finetuning. Annual Meeting of the Association for Computational Linguistics.