THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/TAWZ4LSmgRo*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/binh120702/CS2205.CH183/blob/main/X%C3%82Y%20D%E1%BB%B0NG%20%C4%90%E1%BB%92%20TH%E1%BB%8A%20TRI%20TH%E1%BB%A8C%20T%E1%BB%AA%20%20V%C4%82N%20B%E1%BA%A2N%20PH%C3%81P%20L%C3%9D%20V%E1%BB%9AI%20NH%E1%BA%ACN%20%20DI%E1%BB%86N%20TH%E1%BB%B0C%20TH%E1%BB%82%20D%E1%BB%B0A%20TR%C3%8AN%20LLM.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Phan Doãn Thái Bình
- MSSV: 240101003

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 6
- Số câu hỏi QT của cả nhóm: 6
- Link Github: https://github.com/binh120702/CS2205.CH183

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

XÂY DỰNG ĐỒ THỊ TRI THỨC TỪ VĂN BẢN PHÁP LÝ VỚI NHẬN DIỆN THỰC THỂ DỰA TRÊN LLM

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

BUILDING KNOWLEDGE GRAPHS FROM LEGAL DOCUMENTS WITH LLM-BASED ENTITY RECOGNITION

TÓM TẮT *(Tối đa 400 từ)*

Trong lĩnh vực pháp lý, khả năng đưa ra quyết định kịp thời và chính xác phụ thuộc nhiều vào việc khai thác và xử lý các nguồn dữ liệu văn bản như bản án, điều luật, hợp đồng và tài liệu quy định. Nhận diện thực thể pháp lý (Legal NER) đóng vai trò quan trọng trong việc trích xuất và tổ chức thông tin từ các tài liệu này, hỗ trợ xây dựng đồ thị tri thức phục vụ truy xuất và phân tích dữ liệu. Nghiên cứu này hướng đến việc phát triển một hệ thống nhận diện thực thể sử dụng mô hình ngôn ngữ lớn (LLM), kết hợp với bộ luật và quy tắc hậu xử lý để cải thiện độ chính xác. Hệ thống sẽ giúp xây dựng đồ thị tri thức (Knowledge Graph) pháp lý, cung cấp cấu trúc và quan hệ giữa các thực thể như tòa án, luật, điều khoản, nguyên đơn và bị đơn. Dữ liệu đầu vào bao gồm các văn bản pháp lý tiếng Việt, và kết quả đầu ra là một đồ thị tri thức có thể được sử dụng để hỗ trợ tìm kiếm thông tin pháp lý và phân tích tri thức.

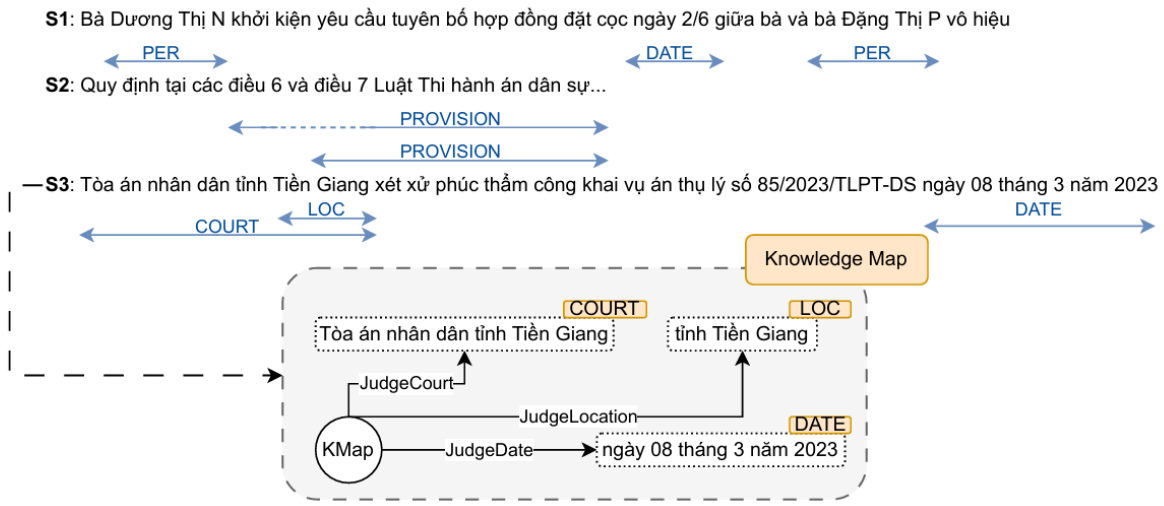
GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong lĩnh vực pháp lý, các văn bản như luật, điều khoản, hợp đồng, và án lệ chứa đựng khối lượng thông tin khổng lồ và có vai trò quan trọng trong việc xác định quyền lợi, trách nhiệm và quy trình xử lý tranh chấp. Tuy nhiên, các tài liệu này thường có ngôn ngữ phức tạp, sử dụng nhiều thuật ngữ chuyên ngành và chứa đựng các quan hệ chồng chéo giữa các thực thể pháp lý. Điều này đặt ra thách thức lớn đối

với việc truy xuất và tổ chức thông tin từ các nguồn dữ liệu này.

Nhận diện thực thể tên (NER) là một trong những nhiệm vụ quan trọng trong xử lý ngôn ngữ tự nhiên, giúp xác định và phân loại các thực thể như cá nhân, tổ chức, địa điểm, điều khoản pháp lý, và các bên liên quan trong vụ án. Đặc biệt, trong lĩnh vực pháp lý, hệ thống NER cần có khả năng nhận diện các thực thể có cấu trúc đa tầng, chẳng hạn như một điều khoản thuộc một điều luật cụ thể.

Nghiên cứu này đề xuất ứng dụng mô hình ngôn ngữ lớn (LLM) để nhận diện thực thể trong các tài liệu pháp lý tiếng Việt, kết hợp với các kỹ thuật xử lý hậu kỳ nhằm cải thiện độ chính xác và tính nhất quán của kết quả trích xuất. Bên cạnh đó, các thực thể được nhận diện sẽ được tổ chức thành một đồ thị tri thức, giúp biểu diễn rõ ràng các mối quan hệ giữa chúng. Cách tiếp cận này không chỉ hỗ trợ truy xuất thông tin hiệu quả hơn mà còn mở ra nhiều ứng dụng tiềm năng như hỗ trợ ra quyết định, tư vấn pháp lý tự động, và phân tích tri thức từ dữ liệu pháp lý.



Hình 1. Một số ví dụ về thực thể đa tầng trong văn bản pháp lý và ứng dụng.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

- Xây dựng một hệ thống nhận diện thực thể pháp lý sử dụng mô hình LLM để trích xuất các thực thể quan trọng trong văn bản pháp lý.
- Tổ chức thông tin thành đồ thị tri thức, liên kết các thực thể pháp lý thông qua các quan hệ có ý nghĩa.

- Đánh giá hệ thống thông qua bộ dữ liệu pháp lý tiếng Việt nhằm đảm bảo độ chính xác và tính khả thi của mô hình.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Thu thập và tiền xử lý dữ liệu
   - Sử dụng bộ dữ liệu văn bản pháp lý tiếng Việt, bao gồm các bản án, luật, điều khoản. Nguồn dữ liệu uy tín: https://thuvienphapluat.vn/ .
   - Tiền xử lý văn bản: chuẩn hóa định dạng, loại bỏ dữ liệu nhiễu, tách câu và đoạn văn.
2. Nhận diện thực thể bằng mô hình LLM
   - Xây dựng một danh sách thực thể, dự kiến bao gồm:
     - **Thực thể chung:** PERSON, ORG, LOC, DATE.
     - **Thực thể pháp lý:** STATUTE, PROVISION, PETITIONER, RESPONDENT, COURT.
   - Áp dụng kỹ thuật nhận diện thực thể sử dụng LLM, kết hợp với prompt engineering để tối ưu kết quả.
   - Sử dụng bộ quy tắc xử lý hậu kỳ nhằm giảm lỗi nhận diện và cải thiện độ chính xác.
3. Xây dựng đồ thị tri thức
   - Thiết kế ontology pháp lý dựa trên các thực thể đã nhận diện.Trích xuất và ánh xạ quan hệ giữa các thực thể để xây dựng đồ thị tri thức.
   - Sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên để xác định quan hệ như "ban hành bởi", "có hiệu lực từ", "liên quan đến", hoặc những mối quan hệ dựa trên khoảng cách các thực thể trong một câu (proximity-based).
4. Đánh giá hệ thống
   - Xây dựng và sử dụng bộ dữ liệu pháp lý tiếng Việt để đánh giá độ chính xác của hệ thống.
   - So sánh với các phương pháp nhận diện thực thể truyền thống nhằm đánh giá hiệu suất.

- Thống kê các mối quan hệ của đồ thị tri thức nhằm đưa ra cái nhìn tổng quan về quá trình trích xuất mối quan hệ.

KẾT QUẢ MONG ĐỢI

- Hệ thống có khả năng nhận diện chính xác các thực thể pháp lý với độ chính xác cao, đặc biệt trong việc phân biệt và xử lý các thực thể có cấu trúc lồng nhau trong văn bản pháp lý.
- Đồ thị tri thức pháp lý được xây dựng sẽ không chỉ hỗ trợ tìm kiếm thông tin mà còn phục vụ các ứng dụng phân tích tri thức chuyên sâu như lập luận pháp lý tự động, truy xuất quan hệ giữa các thực thể và hỗ trợ soạn thảo tài liệu pháp lý.
- Bộ dữ liệu gán nhãn thực thể pháp lý được cung cấp sẽ bao gồm nhiều loại thực thể với ngữ cảnh phong phú, giúp cải thiện độ chính xác của các mô hình NLP pháp lý và hỗ trợ các nghiên cứu mở rộng về khai phá tri thức trong lĩnh vực pháp luật.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Erwin Filtz, Sabrina Kirrane, and Axel Polleres. The linked legal data landscape: linking legal data across different countries. Artif. Intell. Law 29, 4 (Dec. 2021): 485–539.

[2]. Hu Zhang, Jiayu Guo, Yujie Wang, Zhen Zhang, and Hansen Zhao. Judicial nested named entity recognition method with MRC framework. International Journal of Cognitive Computing in Engineering 4 (2023): 118–126.

[3]. Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. SemEval-2023 Task 6: LegalEval - Understanding Legal Texts. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023): 2362–2374.

[4]. H. Keshavarz *et al*., "Named Entity Recognition in Long Documents: An End-to-end Case Study in the Legal Domain," *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, 2022: 2024-2033.

[5]. Hung Q. Ngo, Hien D. Nguyen, and Nhien-An Le-Khac. Ontology Knowledge Map Approach Towards Building Linked Data for Vietnamese Legal Applications. Vietnam Journal of Computer Science 11, 02 (2024): 323–342.