Xây Dựng Đồ Thị Tri Thức Từ Văn Bản Pháp Lý Với Nhận Diện Thực Thể Dựa Trên LLM

Phan Doãn Thái Bình [1]

[1] Trường Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh

What ?

Nghiên cứu giới thiệu một phương pháp để trích xuất và tổ chức thông tin từ văn bản pháp lý, bao gồm:

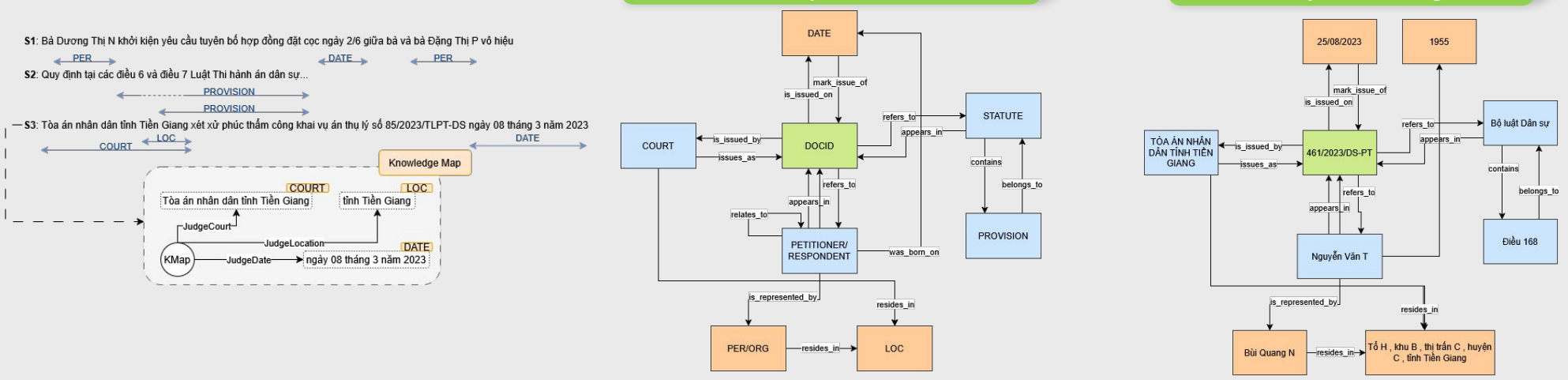
- Sử dụng mô hình ngôn ngữ lớn (LLM) để nhận diện thực thể pháp lý.
- Xây dựng đồ thị tri thức giúp tổ chức thông tin và quan hệ giữa các thực thể.
- Hỗ trợ truy xuất thông tin và phân tích tri thức từ dữ liệu pháp lý.

Why ?

- Văn bản pháp lý chứa thông tin quan trọng về quyền lợi, nghĩa vụ và quy trình pháp lý. Việc truy xuất thông tin trong các tài liệu này rất khó khăn do ngôn ngữ chuyên ngành và cấu trúc phức tạp.Các nghiên cứu trước đây chủ yếu tập trung vào NER thông thường, chưa khai thác tốt quan hệ giữa các thực thể trong bối cảnh pháp lý.
- **Giải pháp của chúng tôi:** Kết hợp LLM , xây dựng đồ thị tri thức để tổ chức thông tin pháp lý có hệ thống.

Overview

Nhận diện thực thể pháp lý → Trích xuất quan hệ giữa các thực thể. → Ứng dụng đồ thị tri thức vào truy xuất thông tin



Description

1. Nhận diện thực thể pháp lý

- Crawl dữ liệu từ nguồn uy tín như: thuvienphapluat, sau đó tiền xử lí văn bản.
- Sử dụng LLM để xác định các thực thể như **Tòa án, Điều luật, Bị đơn, Nguyên đơn…**
- Áp dụng kỹ thuật **in-context learning** để tinh chỉnh mô hình.
- Hậu xử lý để loại bỏ nhiễu, thực thể không chính xác.

```
Given the following set of labels: {set_of_labels}

Based on the provided set of labels,
identify named entities (which may be nested) in
the following passage:

    {legal_document}

Question: Which entities belong to the {target_label} class
in the passage above?
Answer in a single line, listing the entities separated by
"___", do not add any spare information.
```

Figure 1. Cấu trúc prompt để nhận diện thực thể pháp lý.

	PER	ORG	LOC	DATE	PETITIONER	RESPONDENT	COURT	STATUTE	PROVISION	AMT	DOC_ID
PER	represents for	works for	resides in	was born on	represents for	represents for				pays to	
ORG	employs as		is located in	was founded on	represents for	represents for					issues as
LOC	is address for	is host to	belongs_to		is address for	is address for	is host to				
DATE	marks birth of	marks founding of			marks birth of	marks birth of	marks hearing at	marks effect of	marks effect of		mark issue of
PETITIONER	is represented by	is represented by	resides in	was born on	relates_to	disputes with	files with			claims for	appears in
RESPONDENT	is represented by	is represented by	resides in	was born on	is disputed by	relates_to	appears before	relates to	relates to	owes to	appears in
COURT			is located in	holds session on	receives case from	issues summons to		applies in	applies in	orders for	issues as
STATUTE				came into effect on		relates to	is enforced by	relates_to	contains within		appears in
PROVISION				came into effect on		relates to	is enforced by	belongs to	relates_to	specifies for	appears in
AMT	is paid by				is claimed by	is owed by			is specified by		appears in
DOC_ID		is issued by		is issued on	refers to	refers to	is issued by	refers to	refers to	states about	relates to

Figure 2. Định nghĩa một số quan hệ khả dĩ giữa các khái niệm trong đồ thị tri thức.

2. Trích xuất quan hệ giữa các thực thể.

- Định nghĩa các quan hệ khả dĩ trong văn bản pháp lý.
- Sử dụng các luật để xác định quan hệ như:
  - "Ban hành bởi" giữa **STATUTE** và **COURT**
  - "Liên quan đến" giữa **PETITIONER** và **RESPONDENT**
- Sử dụng thuật toán **proximity-based** để phát hiện quan hệ dựa trên vị trí trong văn bản.

3. Ứng dụng đồ thị tri thức vào truy xuất thông tin

- Xây dựng **ontology pháp lý** phản ánh quan hệ giữa các thực thể.
- Biểu diễn dữ liệu dưới dạng **đồ thị tri thức** để hỗ trợ truy vấn và phân tích.
- Cấu trúc đồ thị tri thức:
  - **Nút (Nodes):** Đại diện cho thực thể pháp lý.
  - **Cạnh (Edges):** Biểu diễn quan hệ giữa các thực thể..

NII

Phan Doãn Thái Bình – Trường Đại học Công nghệ Thông tin
TEL : ******* Email : binhpdt.19@grad.uit.edu.vn**