


Họ và tên	NGUYỄN NHƯ THANH MSHV: CH2001015
Ảnh	
Số buổi vắng	1
Bonus	8
Tên đề tài (VN)	HỆ THỐNG HỎI - ĐÁP TỰ ĐỘNG HỖ TRỢ TÌM HIỂU KIẾN THỨC VỀ LUẬT ĐẤT ĐAI TẠI VIỆT NAM
Tên đề tài (EN)	
Giới thiệu	<ul style="list-style-type: none"> <i>Bài toán/vấn đề mà đề tài muốn giải quyết</i> Bài toán hỏi - đáp tự động cho một lĩnh vực cụ thể (closed-domain question answering). Lĩnh vực cụ thể mà đề tài tập trung giải quyết là về luật đất đai tại Việt Nam. <i>Lý do chọn đề tài, khả năng ứng dụng thực tế:</i> Luật về đất đai là một trong những bộ luật rất quan trọng vì nó gắn liền với nhiều tình huống trong đời sống thực tế của đại đa số người dân đang sinh sống tại Việt Nam. Hiện nay có rất nhiều điều luật, văn bản

pháp luật về đất đai được ban hành, nên để người học có thể tìm hiểu và nắm rõ là một việc cần nhiều thời gian và công sức.

Từ lý do trên, cùng với sự phát triển mạnh mẽ của Khoa Học Máy Tính nói chung và lĩnh vực Xử Lý Ngôn Ngữ Tự Nhiên (NLP) nói riêng, tác giả nhận thấy có thể áp dụng một số kỹ thuật trong NLP để hỗ trợ cho việc tìm hiểu về luật của người học, cụ thể ở đây là tìm hiểu về luật về đất đai của Việt Nam. Người học có thể đặt câu hỏi theo cách gần gũi với cách hỏi tự nhiên, hệ thống sẽ cho câu trả lời là điều luật/văn bản chứa thông tin trả lời cho câu hỏi đầu vào.

● **Phạm vi & đối tượng của đề tài**

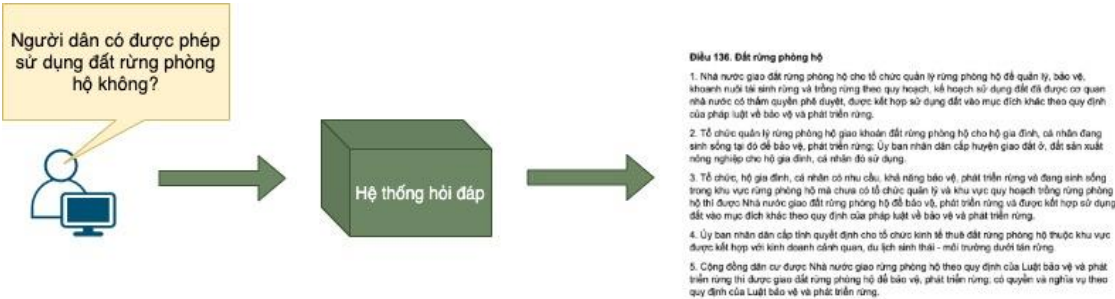
○ *Phạm vi đề tài:*

■ **Hỏi - đáp tiếng Việt tự động** hỗ trợ tìm hiểu về **luật đất đai** hiện hành (thời điểm hiện tại là năm 2021).

○ *Đối tượng:* dành cho các sinh viên ngành luật, đang học ở các trường đại học/cao đẳng luật, hoặc các cá nhân muốn tìm hiểu về luật đất đai.

● **Mô tả input và output:**

- Input: Câu hỏi tiếng Việt có liên quan đến luật đất đai Việt Nam.
- Output: Văn bản/điều luật chứa thông tin trả lời cho câu hỏi.



Hình 1. Mô hình tổng quan bài toán

Mục tiêu

- **Mục tiêu 1:** Xây dựng được một bộ dữ liệu về luật đất đai hiện hành (năm 2021) của Việt Nam.

	<ul style="list-style-type: none"> ● Mục tiêu 2: Xây dựng được một hệ thống hỏi & đáp tự động về luật đất đai, có độ chính xác cao ($> 85\%$), thời gian trả lời chấp nhận được (từ 1-3 giây), và cho phép người dùng đặt câu hỏi gần gũi với cách hỏi tự nhiên. ● Mục tiêu 3: Phổ biến tới các sinh viên đang học về luật tại TPHCM và lấy ý kiến đánh giá.
Nội dung và phương pháp thực hiện	<ul style="list-style-type: none"> ● Nội dung 1: Tìm hiểu về cấu trúc & thành phần của bộ dữ liệu dành cho bài toán hỏi - đáp trong NLP. <i>Phương pháp thực hiện:</i> <ul style="list-style-type: none"> ○ Tìm hiểu cấu trúc & thành phần của bộ dữ liệu cho bài toán hỏi - đáp tự động, tham khảo một số bộ dữ liệu phổ biến như: SQuAD 2.0 [1]. ○ Tìm hiểu cấu trúc & thành phần của bộ dữ liệu tiếng Việt [2][3], ưu tiên chọn những bài báo làm về luật (Legal question answering dataset). ○ Viết chú thích, hướng dẫn về cấu trúc bộ dữ liệu, và các bước thêm, cập nhật dữ liệu. <p>Kết quả đạt được dự kiến: nắm được cấu trúc & thành phần cần có của một bộ dữ liệu dành cho bài toán hỏi - đáp tự động.</p> ● Nội dung 2: Tìm hiểu về bài toán hỏi & đáp tự động trong NLP. <i>Phương pháp thực hiện:</i> <ul style="list-style-type: none"> ○ Tìm hiểu kiến trúc tổng quan của bài toán hỏi - đáp. ○ Tìm hiểu các phương pháp state of the art cho bài toán hỏi - đáp: <ul style="list-style-type: none"> ■ Hiện nay, các kỹ thuật deep learning được áp dụng cho bài toán hỏi - đáp trong NLP đạt được những kết quả cao, vì vậy tác giả sẽ tập trung tìm hiểu về kỹ thuật transformer [4]

và những mô hình deep learning ví dụ như: BERT [5], ALBERT [6],....

- Đọc các bài báo đề xuất phương pháp giải quyết bài toán hỏi - đáp về kiến thức luật, ưu tiên những bài xử lý dữ liệu tiếng Việt → chọn ra một số phương pháp/mô hình phù hợp nhất với đề tài.

Kết quả đạt được dự kiến: Hiểu về kiến trúc bài toán, nắm được những phương pháp state of the art phù hợp với đề tài → làm nền tảng để xây dựng hệ thống hỏi & đáp tự động.

- **Nội dung 3:** Xây dựng bộ dữ liệu về luật đất đai hiện hành tại Việt Nam và thực hiện thí nghiệm trên tập dữ liệu nhỏ..

Phương pháp thực hiện:

- Dữ liệu sẽ được thu thập & chọn lọc bởi các chuyên gia về luật cộng tác cùng tác giả như: một số luật sư, giảng viên luật. Các chuyên gia thu thập sẽ dựa vào tài liệu hướng dẫn ở **Nội dung 1** để thu thập và nhập dữ liệu.
- Kiểm tra và rà soát lại bộ dữ liệu.
- Chọn tập dữ liệu con có kích thước nhỏ từ bộ dữ liệu đã xây dựng để tiến hành cài đặt với các phương pháp đã chọn ở **Nội dung 2** → bước này sẽ giúp xác nhận lại xem bộ dữ liệu có chuẩn hay không, và giúp chọn ra phương pháp/mô hình giải quyết phù hợp.

Kết quả đạt được dự kiến: Xây dựng được một bộ dữ liệu về luật đất đai (hoàn thành **Mục tiêu 1**). Và chọn ra được một số phương pháp/ mô hình phù hợp với đề tài.

- **Nội dung 4:** Xây dựng hệ thống hỏi & đáp tự động bằng tiếng Việt hỗ trợ tìm hiểu về luật đất đai tại Việt Nam.

Phương pháp thực hiện:

	<ul style="list-style-type: none"> ○ Cài đặt thử nghiệm các phương pháp/ mô hình đã chọn ở Nội dung 3 với bộ dữ liệu đã xây dựng. ○ So sánh, đánh giá → chọn ra phương pháp phù hợp với yêu cầu của đề tài. <p>Kết quả đạt được dự kiến: Xây dựng được một hệ thống hỏi & đáp tự động tiếng Việt về luật đất đai tại Việt Nam (hoàn thành Mục tiêu 2).</p> <ul style="list-style-type: none"> ● Nội dung 5: Phổ biến ứng dụng đến các sinh viên luật - đối tượng hướng đến của đề tài, nhằm mục đích hỗ trợ sinh viên học về luật. <p>Phương pháp thực hiện:</p> <ul style="list-style-type: none"> ○ Xây dựng & triển khai (deploy) ứng dụng web. ○ Liên hệ với các chuyên gia đã giúp đỡ xây dựng bộ dữ liệu để mời dùng thử và giới thiệu tới các sinh viên. ○ Khảo sát và lấy ý kiến từ các giảng viên, sinh viên đã dùng thử. <p>Kết quả đạt được dự kiến: Xây dựng được một ứng dụng & triển khai cho người dùng sử dụng (hoàn thành Mục tiêu 3)</p>
Kết quả dự kiến	<ul style="list-style-type: none"> ● 01 bộ dữ liệu về luật đất đai hiện hành của Việt Nam. ● Mã nguồn (source code) liên quan đến hệ thống & các thí nghiệm đã thực hiện. ● Ứng dụng cho phép người dùng đặt các câu hỏi liên quan tới luật đất đai của Việt Nam. ● Báo cáo chi tiết về phương pháp cài đặt, quy trình xây dựng bộ dữ liệu. ● Báo cáo khảo sát ý kiến người dùng.
Tài liệu tham khảo	<p>[1] P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July</p>

15-20, 2018, Volume 2: Short Papers, 2018, pp. 784–789. doi: 10.18653/v1/P18-2124.

[2] P. M. Kien, H.-T. Nguyen, N. X. Bach, V. Tran, M. L. Nguyen, and T. M. Phuong, “Answering Legal Questions by Learning Neural Attentive Text Representation,” in Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, 2020, pp. 988–998. doi: 10.18653/v1/2020.coling-main.86.

[3] K. V. Nguyen, V. Nguyen, A. H.-T. Nguyen, and N. Nguyen, “A Vietnamese Dataset for Evaluating Machine Reading Comprehension,” in Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, 2020, pp. 2595–2605. doi: 10.18653/v1/2020.coling-main.233.

[4] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. doi: 10.18653/v1/n19-1423.

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” 2020. [Online]. Available: <https://openreview.net/forum?id=H1eA7AEtvS>