ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NHẬP MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Automatic alignment between a SinoNom character-image and its Vietnamese translation text

GIẢNG VIÊN HƯỚNG DẪN
- Đinh Điền
- Nguyễn Hồng Bửu Long
- Lương An Vinh

SINH VIÊN THỰC HIỆN
- 22127220 - Nguyễn Anh Kiệt
- 22127275 - Trần Anh Minh
- 22127280 - Đoàn Đặng Phương Nam

Thành phố Hồ Chí Minh, 2024

Mục lục

1) Lời nói đầu

Trong thời đại công nghệ số, việc bảo tồn và nghiên cứu các ngôn ngữ cổ trở nên cấp thiết hơn bao giờ hết, đặc biệt là với di sản văn hóa ngôn ngữ của dân tộc. Chữ Hán - Nôm, từng là phương tiện truyền tải tri thức và văn hóa quan trọng trong lịch sử Việt Nam, đang dần mai một do sự thay thế của hệ thống chữ Quốc ngữ. Chính vì vậy, việc áp dụng các phương pháp công nghệ hiện đại để nghiên cứu và gắn kết ngôn ngữ này với tiếng Việt hiện đại là một nhiệm vụ vừa mang tính khoa học, vừa mang ý nghĩa bảo tồn văn hóa.

Đề tài **"Tự động gióng hàng giữa ảnh ký tự SinoNom và văn bản dịch tiếng Việt"** hướng đến việc khai thác và ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và nhận dạng ký tự quang học (OCR) để xây dựng hệ thống tự động hóa quá trình gióng hàng giữa chữ Hán - Nôm và tiếng Việt. Bằng cách kết hợp các công cụ trích xuất dữ liệu, phân tích, và đánh giá, nhóm nghiên cứu không chỉ mong muốn tạo ra một giải pháp hỗ trợ nghiên cứu ngôn ngữ hiệu quả mà còn góp phần lưu giữ giá trị văn hóa lâu đời của dân tộc. Hy vọng rằng, kết quả của nghiên cứu này sẽ không chỉ mang lại ý nghĩa về mặt công nghệ, mà còn tạo nền tảng cho những bước tiến sâu rộng hơn trong lĩnh vực bảo tồn ngôn ngữ cổ ở Việt Nam.

Lời cuối cùng, nhóm xin gửi lời cảm ơn đến thầy Đinh Điền cũng như các thầy cô/anh chị khác trong nhóm học tập vì đã cung cấp những bộ dữ liệu cũng như giúp chúng em có được định hướng đúng đắn trong quá trình học tập và thực hiện đồ án.

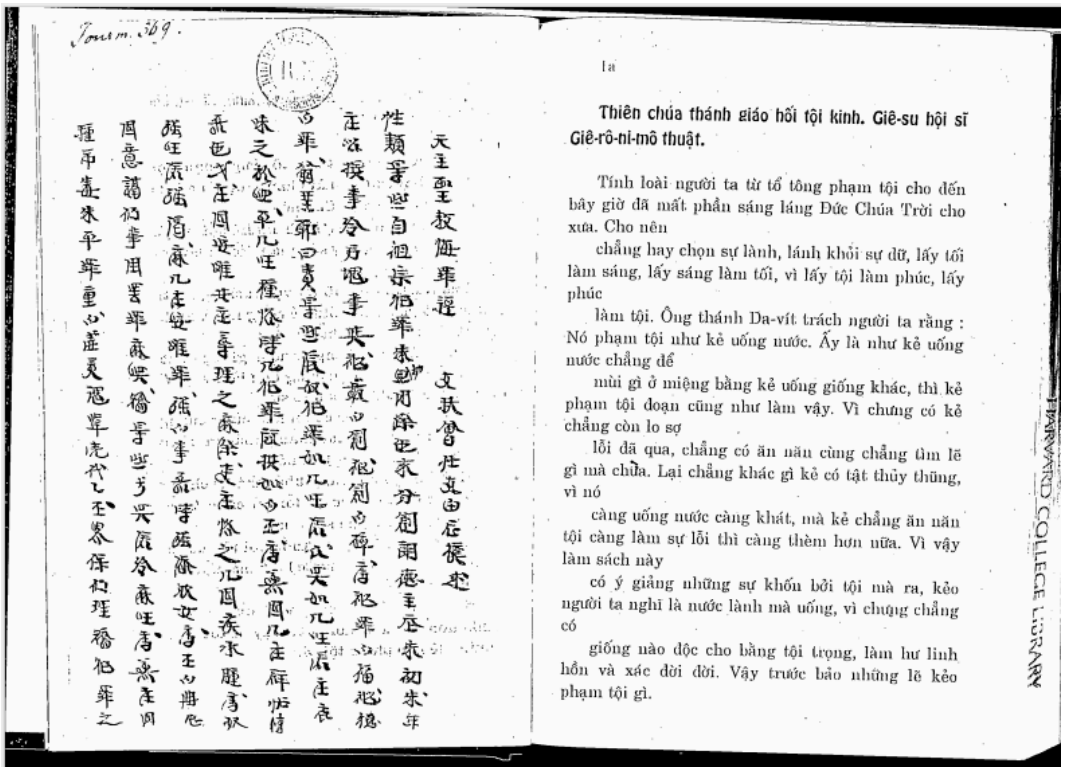
2) Dữ liệu đồ án

2.1) Khái quát chung

Để phục vụ cho việc triển khai đề tài, nhóm cần có những bộ dữ liệu song ngữ **Hán Nôm - Quốc Ngữ** để tiến hành xử lý và gióng hàng tự động. Nhờ có sự giúp đỡ và hỗ trợ tận tình của các thầy, nhóm đã thành công thu thập và gom về được bản PDF của **bốn bộ dữ liệu** đúng với mong muốn tìm kiếm của nhóm, ba trong số đó đến từ các quyển sách chữ Nôm là **CÁC THÁNH TRUYỆN** và **THIÊN CHÚA THÁNH MẪU** của nhà truyền giáo Girolamo Maiorica, bộ dữ liệu còn lại đến từ quyền sách **THIÊN CHÚA THÁNH GIÁO HỐI TỘI KINH** của cha Jeronymo MAYORICA SJ.

Điểm chung của cả bốn bộ dữ liệu trên nằm ở việc **chúng đều chứa các ảnh văn bản Hán Nôm, đi kèm với đó là bản dịch Quốc Ngữ cho từng câu Hán Nôm ở ngay trang sau của mỗi ảnh**. Tính cấu trúc này mang một ý nghĩa rất quan trọng vì nó giúp nhóm dễ dàng và thuận tiện hơn khi xử lý dữ liệu, góp phần hỗ trợ cho quy trình gióng hàng tự động của nhóm. Ngoài ra, các văn bản Hán Nôm trong mỗi quyền sách đều là các **văn bản viết tay**, điều này sẽ tạo ra khó khăn nhưng cũng là cơ hội để tăng thêm tính chính xác và hiệu quả cho việc OCR cũng như việc gióng hàng tự động

Riêng đối với sách THIÊN CHÚA THÁNH GIÁO HỐI TỘI KINH, **các bản dịch Quốc Ngữ cũng được thể hiện dưới dạng ảnh**, điều này cũng đã mang đến cho nhóm những thử thách trong việc xử lý một cách tự động và đồng bộ giữa các bộ dữ liệu.

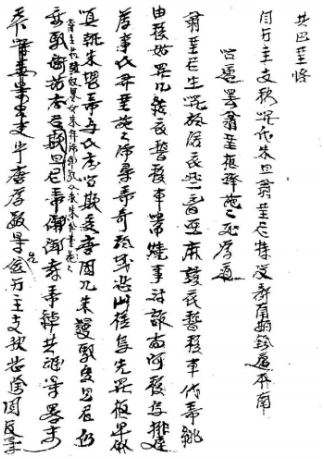


Hình 1: Một cặp ảnh văn bản Hán Nôm - Quốc ngữ trong sách THIÊN CHÚA THÁNH GIÁO HỐI TỘI KINH

2.2) Những khó khăn

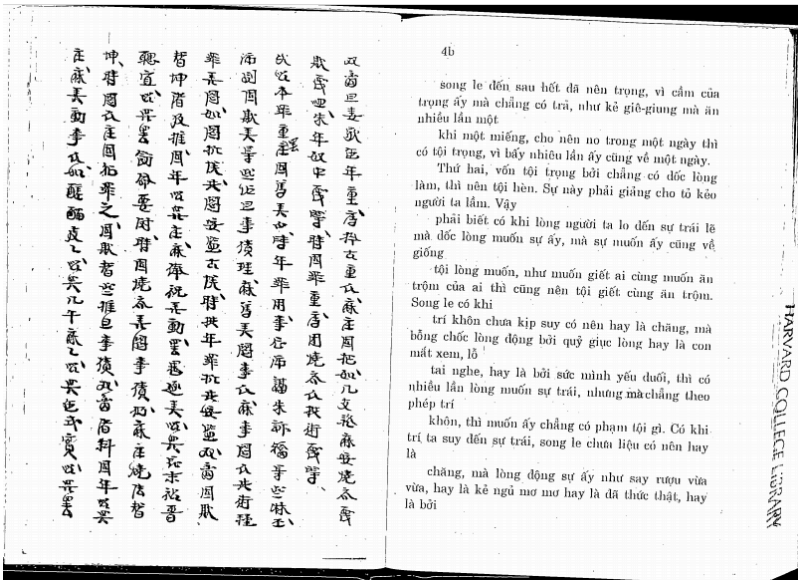
Như đã đề cập ở trên, việc **các bộ dữ liệu** đều ở dạng **văn bản viết tay** tạo ra rất nhiều khó khăn không chỉ cho việc gióng hàng tự động mà còn cho cả việc nhận diện chữ viết trong văn bản. Một số khó khăn nổi bật từ các bộ dữ liệu được nhóm tổng hợp lại như sau:

- ***Gạch xoá, ghi đè chữ***



Hình 2: Trang Hán Nôm có nhiều chỗ gạch xoá, ghi đè

- ***Trang sách được chụp không thẳng***



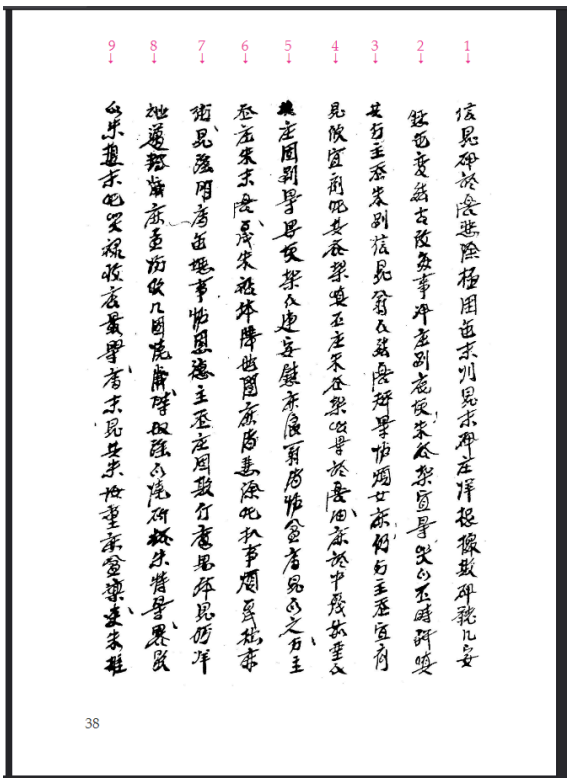
Hình 3: Trang sách được chụp không thẳng

- ***Dùng số thay chữ viết trong bản dịch Quốc ngữ***

3. Từ Đức Chúa Giê-su ra đời cho đến ông Thánh này được 308 năm.

Hình 4: Viết số 308 thay vì “ba trăm linh tám”

- ***Ảnh bị nhoè, mờ chữ hoặc bị lem mực***



Hình 5: Lem mực, ảnh bị mờ

- ***OCR thiếu box***

Hình 6: OCR thiếu bounding box

Ngoài ra còn có nhiều những khó khăn tiểu tiết khác khiến việc tiền xử lý các bộ dữ liệu trở thành một vấn đề thách thức trước khi thực hiện đến các bước gióng hàng tự động về sau.

2.3) Hướng xử lý khó khăn

Trong suốt quá trình làm, nhóm đã sử dụng *API Kim Hán Nôm* do thầy Đinh Điền cung cấp để có thể trích xuất ra được các từ Nôm từ hình ảnh trong các ngữ liệu mà thầy đã cung cấp. Tuy nhiên, công cụ trên đôi lúc cũng sẽ cho ra các kết quả sai sót tương tự những khó khăn trong tiền xử lý dữ liệu được trình bày ở trên, vậy nên đa phần trong số đó được nhóm xử lý một cách thủ công vì chúng xảy ra với số lượng nhỏ, như kẻ thêm ô bằng các công cụ như PPOCRLabel thuộc hệ sinh thái của PaddleOCR.

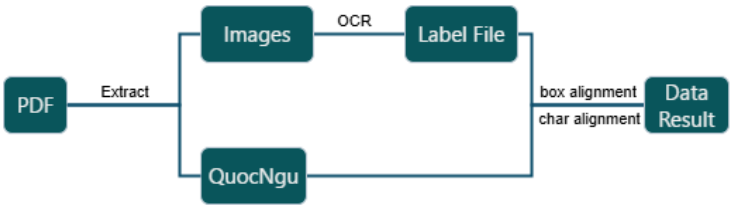
Đối với các khó khăn liên quan đến chữ quốc ngữ, nhóm đề xuất việc sử dụng thư viện *regex* của Python trong các trường hợp liên quan đến số, đến các tên nước ngoài được phiên âm nhưng chưa được ngăn cách. Một số khó khăn khác như ảnh bị nghiêng, tận dụng việc Google Vision API có thể trả về box của từng từ một, ta sẽ tiến hành gắn các từ trên cùng một dòng với nhau. Phương pháp tương tự với việc gắn các sắp xếp các bounding box theo cột (được đề cập ở mục 3.3.3) nhưng được chỉnh sửa lại để có thể xử lý theo dòng cho phù hợp. Vì phần lớn các ảnh đều không nghiêng đủ nhiều để khiến từ cuối của một hàng và từ đầu của hàng kế được coi là cùng hàng nhau, nên phương pháp này tương đối hiệu quả trong việc xử lý vấn đề này.

3) Quy trình thực hiện

3.1) Quy trình chuẩn bị và xử lý dữ liệu

Nhắc lại quy trình thực hiện giai đoạn giữa kỳ được thiết kế nhằm chuẩn bị và xử lý dữ liệu cho bài toán gióng hàng giữa chữ Hán-Nôm và Quốc Ngữ. Cụ thể các bước bao gồm:

- Chuẩn bị đầu vào:
  - Ngữ liệu đầu vào là một file PDF chứa hình ảnh các câu chữ Hán-Nôm và các câu dịch nghĩa Quốc Ngữ tương ứng. Trong trường hợp ngữ liệu Quốc Ngữ được lưu dưới dạng hình ảnh, cần tiến hành OCR để trích xuất nội dung văn bản.
  - Sau khi thu thập, các hình ảnh được tiền xử lý (phương pháp khử nhiều cùng các phương pháp khác) nhằm cải thiện, tối ưu hoá đầu vào cho quá trình OCR.
- Trích xuất và xử lý ngữ liệu
  - Các hình ảnh sau khi đã được xử lý sẽ sẽ được OCR để thu được văn bản SinoNom (thông qua API *Kim Hán Nôm*).
  - Các cặp câu SinoNom và Quốc Ngữ sẽ được xử lý gióng hàng thông qua thuật toán MED, kết hợp cùng bộ các bộ từ điển được cung cấp sẵn để so sánh để đánh giá độ chính xác của ngữ liệu.

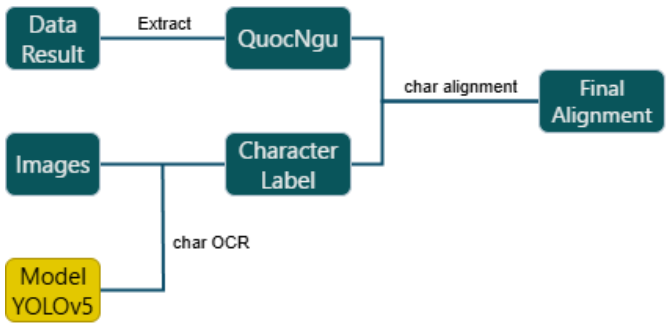


Hình 7: Quy trình chuẩn bị và xử lý dữ liệu

3.2) Quy trình gióng hàng các cặp chữ

Từ dữ liệu đã thu thập và xử lý ở giai đoạn giữa kỳ, quá trình gióng hàng sẽ được tiếp tục tiến hành ở cấp độ từ.

- Các hình ảnh chứa các chữ SinoNom sẽ được phân tích thông qua mô hình *YOLOv5*, nhằm xác định được các chỉ số vị trí của từng ký tự trong ảnh.
- Các ký tự này sẽ được tiếp tục được xử lý để tiến hành so sánh, gióng hàng và kiểm tra độ chính xác của từng cặp chữ.



Hình 8: Quy trình gióng hàng các cặp chữ

3.3) Một số điểm nhấn nổi bật

3.3.1) Xử lý bộ từ điển

Về các bộ từ điển sử dụng cho sản phẩm đồ án của nhóm, nhóm đã tận dụng lại hai bộ từ điển ***QuocNgu_SinoNom_Dic*** cho việc tra cứu các chữ Hán Nôm tương ứng với một chữ Quốc Ngữ cho trước và ***SinoNom_similar_Dic*** cho việc tra cứu các chữ Hán Nôm tương đồng với một chữ Hán Nôm cho trước. Xuyên suốt quá trình làm việc và tận dụng các bộ từ điển, nhóm đã nhận thấy một số đặc điểm sau:

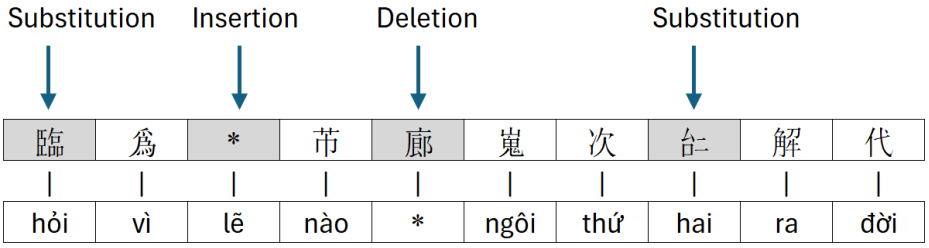
- **Một chữ Hán Nôm sẽ luôn tương đồng với chính nó**, do đó, cần thêm nó vào danh sách các chữ cái tương đồng với chữ Hán Nôm đang xét.
- Nếu A là một chữ Hán Nôm có 2 chữ Hán Nôm tương đồng là B và C, khi đó, **B cũng là chữ Hán Nôm tương đồng với C**.

Ngoài ra, trong quá trình xử lý, nhóm cũng đã lưu ý để tránh các trường hợp **"tương đồng bắc cầu"**, tức chữ Hán Nôm A tương đồng với chữ Hán Nôm B, chữ Hán Nôm B tương đồng với chữ Hán Nôm C, nhưng A và C có thể không tương đồng nhau.

3.3.2) Thuật toán MED

Thuật toán MED (Minimum Edit Distance) là một kỹ thuật trong xử lý ngôn ngữ tự nhiên, được dùng để đo độ khác biệt giữa hai chuỗi ký tự. Nó tính số lượng thao tác chỉnh sửa tối thiểu cần thực hiện để biến đổi chuỗi này thành chuỗi kia. Các thao tác bao gồm:

- ***Thay thế (substitution)***: Thay thế một ký tự trong chuỗi bằng một ký tự khác.
- ***Chèn (Insertion)***: Thêm một ký tự vào chuỗi.
- ***Xoá (Deletion)***: Loại bỏ một ký tự khỏi chuỗi.



Hình 9: So sánh 2 chuỗi theo MED

Thuật toán thường được triển khai dưới dạng quy hoạch động, sử dụng ma trận để lưu kết quả các phép tính con. Công thức tổng quát như sau:

$$D[i][j] = \min \begin{cases} D[i-1][j] + 1 & (\text{xoá}) \\ D[i][j-1] + 1 & (\text{chèn}) \\ D[i-1][j-1] + c & (\text{thay thế}) \end{cases} \quad (1)$$

Trong đó:

$c = 0$ nếu hai ký tự tương ứng giống nhau.

$c = 2$ nếu hai ký tự khác nhau.

Ở đây, chúng ta mong muốn là giảm tỉ lệ "xóa" và "chèn" cũng như tăng tỉ lệ của "thay thế". Vậy nên trong trường hợp cả ba giá trị trả về bằng nhau, chúng ta sẽ ưu tiên chọn "thay thế" để tiếp tục quá trình backtracking.

Tuy nhiên, khi xử lý văn bản giữa SinoNom và Quốc Ngữ, việc xác định toán tử *"bằng nhau"* không thể được thực hiện một cách thông thường. Vấn đề này sẽ được đề cập cụ thể hơn ở phần **Char alignment**.

Sau khi có bảng khoảng cách, tiến hành *backtrack* để truy ngược lại quá trình thay đổi. Điều này giúp xác định các thao tác đã được thực hiện (thêm, xóa, thay thế) để chuỗi này thành chuỗi kia. Kết quả này cung cấp thông tin cụ thể về sự khác biệt giữa hai chuỗi. (Song, kết quả từng lần chạy có thể khác nhau bởi rằng sẽ có những tổ hợp phép biến đổi khác nhau có cùng số thao tác).

代	9	10	9	10	9	8	7	6	7	6
解	8	9	8	9	8	7	6	5	6	7
𠊎	7	8	7	8	7	6	5	4	5	6
次	6	7	6	7	6	5	4	5	6	7
嵬	5	6	5	6	5	4	5	6	7	8
廊	4	5	4	5	4	5	6	7	8	9
芾	3	4	3	4	3	4	5	6	7	8
爲	2	3	2	3	4	5	6	7	8	9
臨	1	2	3	4	5	6	7	8	7	8
	0	1	2	3	4	5	6	7	8	9
		hỏi	vì	lẽ	nào	ngôi	thứ	hai	ra	đời

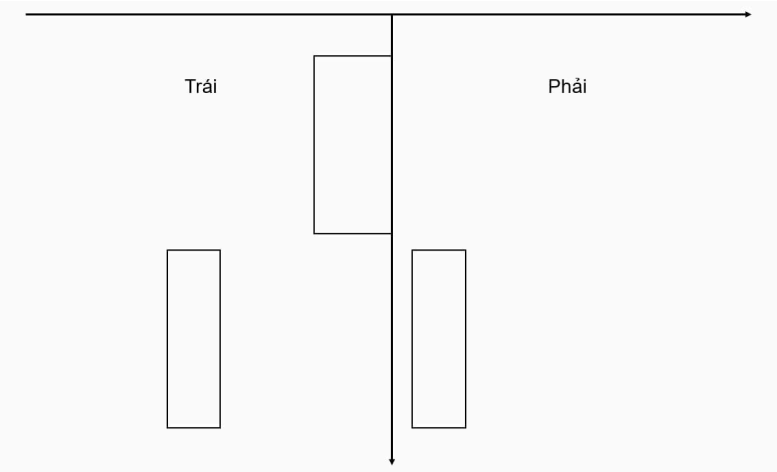
Hình 10: Quá trình backtrack

3.3.3) Sắp xếp các bounding box

Các bounding box có được sau khi thực hiện OCR có thể sẽ nằm ở các vị trí lộn xộn, do vậy, việc sắp xếp chúng theo đúng thứ tự chúng ta muốn là một nhu cầu cần thiết. Trong phạm vi của các bộ dữ liệu đã đề cập, nhóm mong muốn vị trí của các bounding box sẽ được sắp xếp theo thứ tự **từ phải sang trái**, **từ trên xuống dưới**. Dựa vào đó, nhóm đã đưa ra ý tưởng cho thuật toán sắp xếp như sau:

- ***Bước 1***: Xác định tất cả các bounding box **ngoài cùng bên phải** chưa xét

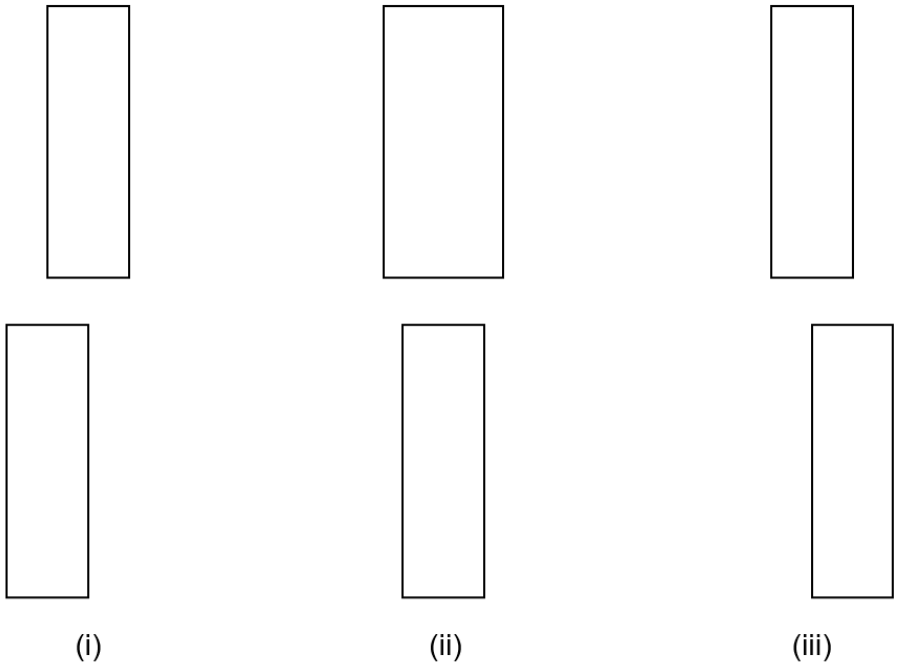
  Ở bước này, ta sẽ lấy ô bounding box ngoài cùng bên phải đầu tiên, gọi là box X, định nghĩa **"bên phải"** của nhóm được minh hoạ như sau:



Hình 11: Định nghĩa một box nằm "bên phải"
một bounding box cho trước

Sau đó, định nghĩa Y là tập các box **ngoài cùng bên phải**, ban đầu, Y chỉ gồm box X, sau đó, với từng box khác với các box trong Y, ta tiến hành kiểm tra xem box đó có **"nằm cùng cột"** với bất cứ box nào trong Y không, nếu có, ta thêm box đó vào tập Y.

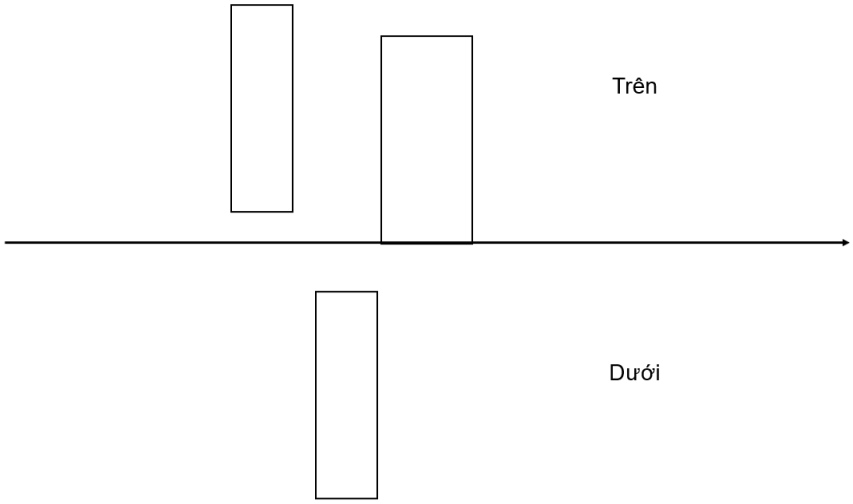
Hai box được gọi là **"nằm cùng cột"** nếu sự tương quan vị trí của chúng thuộc về một trong ba trường hợp sau đây:



Hình 12: Định nghĩa hai box "nằm cùng cột"

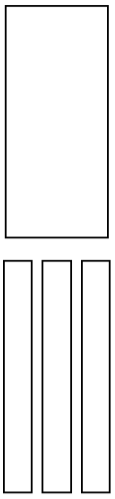
- ***Bước 2***: Sắp xếp các box ngoài cùng bên phải theo thứ tự **từ trên xuống dưới**

  Trước hết, nhóm định nghĩa khái niệm **"trên"** và **"dưới"** qua hình sau:



Hình 13: Định nghĩa hai box “nằm cùng cột”

Trong quá trình sắp xếp, nhóm cũng lưu ý đến trường hợp các box có thể **“nằm trên cùng 1 hàng”**, minh hoạ bằng hình sau:



Hình 14: Các box cùng “nằm trên một hàng”

Khi đó, việc sắp xếp các box trên sẽ được thực hiện tuân theo thứ tự **từ phải sang trái**

- ***Bước 3***: Đánh dấu các ô bên phải ngoài cùng là đã xét, nếu đã xét tất cả các box, thuật toán dừng lại, ngược lại, quay về **Bước 1**

3.3.4) Box alignment

Box alignment là quá trình liên kết các bounding box của văn bán Hán-Nôm với các câu Quốc Ngữ tương ứng, được thực hiện thông qua việc đánh giá chiều cao của từng box so với độ dài của câu Quốc Ngữ, và thông qua thuật toán MED. Điều này sẽ đảm bảo tính chính xác trong việc sắp xếp và liên kết nội dung, xử lý được các trường hợp xử lý thừa hay thiếu bounding box.

3.3.4.1) Phép ước lượng

Bằng cách tính tổng chiều cao của tất cả các bounding box và chia cho tổng số từ trong các câu Quốc Ngữ, ta có thể xấp xỉ được chiều cao trung bình của một ký tự Hán-Nôm. Từ đó có thể lấy chiều cao của box chia cho chiều cao trung bình của ký tự để xác định được số chữ trong box, cung cấp cơ sở để đối chiếu nội dung.

Song, phép xấp xỉ này có thể bị ảnh hưởng bởi sự chênh lệch kích thước chữ viết trong các bounding box khác nhau. Để khắc phục sai lệch kích thước này, sẽ cần **đặt một sai số ở mức 3 chữ (ký tự)**. Điều này sẽ giúp giải quyết được trường hợp khi câu Quốc Ngữ được dịch có nhiều chữ hơn so với văn bản gốc (Hán-Nôm) hoặc ngược lại.

3.3.4.2) Sử dụng thuật toán MED

Thuật toán MED được sử dụng để xử lý các sai sót trong quá trình liên kết bounding box và câu Quốc Ngữ. MED sẽ giúp điều chỉnh cho các trường hợp:

- Thiếu hoặc thừa bounding box do OCR.
- Sai số trong việc nhận dạng số từ trong bounding box.

3.3.5) Char alignment

3.3.5.1) Định nghĩa toán tử *"bằng nhau"*

Khi so sánh ký tự Hán-Nôm và từ Quốc Ngữ, ta không thể sử dụng toán tử bằng thông thường. Ví dụ, không thể đơn giản kết luận rằng "hello" và "xin chào" là khác nhau chỉ dựa vào cách thể hiện văn bản. Để giải quyết vấn đề này, cần truy xuất vào bộ từ điển được cung cấp để xác định xem hai từ có thể được xem là "bằng nhau" hay không.

Quá trình thực hiện cụ thể như sau:

- **Giả định đầu vào**:
  - Ký tự Hán-Nôm đầu vào: A.
  - Từ Quốc Ngữ cần so sánh: B.
  - Bộ từ điển các ký tự Hán-Nôm gần giống nhau: S_1.
  - Bộ từ điển dịch nghĩa từ Quốc Ngữ sang Hán-Nôm: S_2.
- **Các bước thực hiện**:
  - Dò từ B trong S_2 để thu được tập hợp các ký tự Hán-Nôm tương ứng về nghĩa, gọi tập hợp này là s_2.
  - Kiểm tra nếu $A \in s_2$: Nếu có, kết luận rằng A và B là giống nhau.
  - Dò ký tự A trong S_1 để thu được tập hợp top-20 các ký tự Hán-Nôm gần giống nhất với A, gọi tập hợp này là s_1.
  - Tìm phần giao giữa s_1 và s_2:
    - Nếu $s_1 \cap s_2 \neq \emptyset$, kết luận rằng A và B là tương tự nhau (vẫn có sự khác biệt nhỏ).

– Nếu $s_1 \cap s_2 = \emptyset$, kết luận rằng A và B hoàn toàn không giống nhau.

Quy trình này đảm bảo rằng việc xác định "bằng nhau" giữa ký tự Hán-Nôm và từ Quốc Ngữ không chỉ dựa trên cách thể hiện văn bản mà còn dựa trên ngữ nghĩa và mức độ tương đồng của các ký tự.

3.3.5.2) So sánh giữa câu Quốc Ngữ và câu Hán-Nôm

Quá trình so sánh bắt đầu bằng việc sử dụng câu Hán-Nôm thu được từ quá trình OCR và câu Quốc Ngữ tương ứng đã được xác định qua bước **Box Alignment**. Thuật toán MED được áp dụng để tính toán bảng khoảng cách tối thiểu cần thiết để biến đổi chuỗi Hán-Nôm thành chuỗi Quốc Ngữ.

Sau khi đã tiến hành *backtrack* dựa trên bảng khoảng cách, thu được các phép biến đổi đã được thực hiện để biến chuỗi Hán-Nôm thành Quốc Ngữ, ta cần tiếp tục đánh giá và thể hiện tính chính xác của cặp câu. Các trường hợp cần được cân nhắc bao gồm:

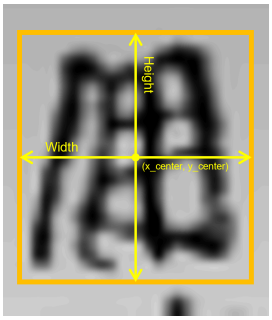
- Từ Hán Nôm tự OCR thiếu
- Từ Hán Nôm được OCR dư
- Từ Hán Nôm bị OCR sai
- Từ Hán Nôm OCR tương đồng với từ OCR đúng
- Từ Hán Nôm được OCR đúng

Các từ hán Nôm thuộc vào ba trường hợp đầu tiên thì việc OCR ra từ đó được xem là sai và từ đó được thể hiện bằng màu đỏ, còn với 2 trường hợp còn lại, việc OCR sẽ được xem là đúng và được thể hiện bằng màu đen.

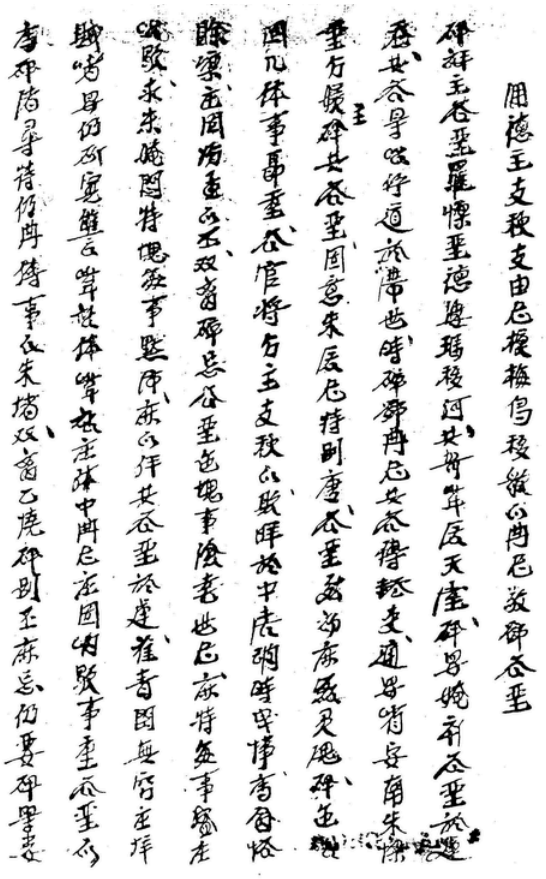
3.3.6) Mô hình YOLOv5

Mô hình YOLOv5 là một mô hình phát hiện đối tượng nổi tiếng, thuộc họ mô hình *"You Only Look Once"*. Trong quy phạm bài toán gióng hàng này, mô hình YOLO được sử dụng để phát hiện các ký tự Hán-Nôm trên hình ảnh (đã được trích xuất ra lúc OCR). Kết quả của quá trình này là dữ liệu bounding box bao quát từng ký tự được phát hiện trong ảnh, được cấu trúc dưới dạng: `center_x, center_y, width, height`, với:

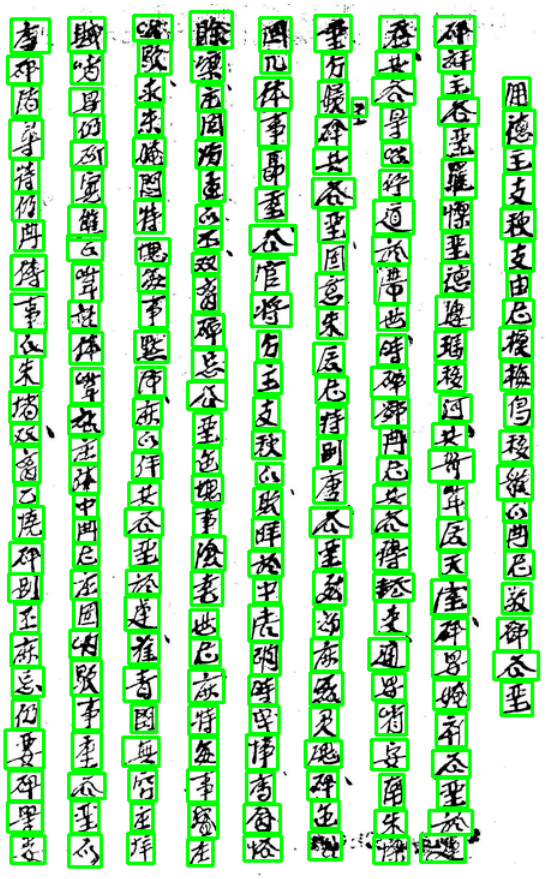
- `center_x`, `center_y` là toạ độ tâm của bounding box.
- `width`, `height` là chiều dài và chiều cao của bounding box.



Hình 15: Bounding box của ký tự



Hình 16: Văn bán Hán-Nôm gốc



Hình 17: Văn bản Hán-Nôm sau khi được OCR bởi Yolov5

Sau khi xác định được các *bounding box* của các ký tự Hán-Nôm trên trang, quá trình tiếp theo tiến hành gióng hàng các ký tự. Cụ thể, các *bounding box* sẽ được sắp xếp và căn chỉnh (*box alignment*) để tạo ra kết quả cuối cùng, thể hiện mối quan hệ gióng hàng giữa từng ký tự Hán-Nôm và từ Quốc Ngữ tương ứng.

3.4) Tóm tắt

Quy trình tổng thể bắt đầu với đầu vào là một file PDF chứa ngữ liệu SinoNom và Quốc Ngữ. Dữ liệu này sẽ được xử lý qua các bước gióng hàng câu và ký tự, kết hợp phân tích và đánh giá chất lượng. Kết quả cuối cùng là các cặp ký tự Hán-Nôm – Quốc Ngữ đã được xử lý, gióng hàng và so sánh, tạo cơ sở cho các nghiên cứu tiếp theo.

4) Kết quả đạt được

4.1) Độ đo đánh giá

Để đánh giá hiệu quả của quy trình dóng hàng tự động, nhóm đã sử dụng độ đo đánh giá là **tỷ lệ ảnh ký tự Hán Nôm có bản dịch Quốc Ngữ tương ứng** *(tức ký tự Hán Nôm được OCR đúng)*. Độ đo này sẽ phản ánh được khả năng dóng hàng tự động khi sử dụng quy trình dóng hàng tự động của nhóm, đồng thời cũng có thể giúp nhóm định hướng nhóm trong việc cải thiện hiệu quả của quy trình.

4.2) Kết quả thử nghiệm

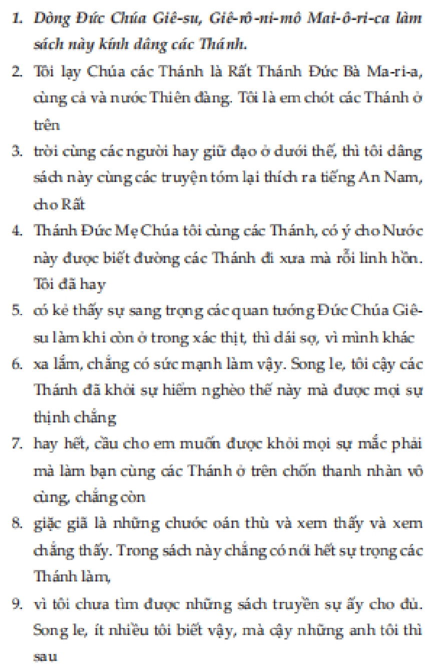
Sau khi đã thực hiện các cài đặt bên cạnh những thử nghiệm và quan sát, sử dụng độ đo đánh giá đã đề cập, kết quả tổng thể thu được trên cả bốn bộ dữ liệu là ***46.71%***, trong đó, kết quả cụ thể thu được trên từng bộ dữ liệu như sau:

Bộ dữ liệu	Kết quả đánh giá (%)
CÁC THÁNH TRUYỆN - THÁNG 1	40.123
CÁC THÁNH TRUYỆN - THÁNG 5 + THIÊN CHÚA THÁNH MẪU	49.659
THIÊN CHÚA THÁNH GIÁO HỐI TỘI KINH	56.126

Về **quy trình chuẩn bị và xử lý dữ liệu**, trên từng bộ dữ liệu Hán Nôm, kết quả của nhóm được thể hiện qua các hình ảnh sau (*vì một bộ dữ liệu có nhiều trang văn bản, nhóm chỉ chọn* ***tràng đầu tiên làm trang đại diện***, *kèm theo kết quả tương ứng của quy trình trên trang đó*)

- ***CÁC THÁNH TRUYỆN - THÁNG 1***

Hình 18: Văn bản Hán Nôm



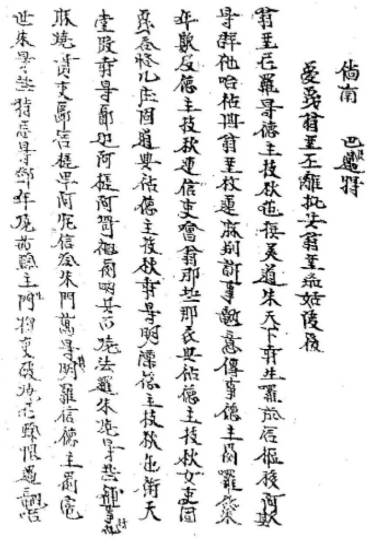
1. *Dòng Đức Chúa Giê-su, Giê-rô-ni-mô Mai-ô-ri-ca làm sách này kính dâng các Thánh.*
2. Tôi lạy Chúa các Thánh là Rất Thánh Đức Bà Ma-ri-a, cùng cả và nước Thiên đàng. Tôi là em chót các Thánh ở trên
3. trời cùng các người hay giữ đạo ở dưới thế, thì tôi dâng sách này cùng các truyện tóm lại thích ra tiếng An Nam, cho Rất
4. Thánh Đức Mẹ Chúa tôi cùng các Thánh, có ý cho Nước này được biết đường các Thánh đi xưa mà rỗi linh hồn. Tôi đã hay
5. có kẻ thấy sự sang trọng các quan tướng Đức Chúa Giê-su làm khi còn ở trong xác thịt, thì dái sợ, vì mình khác
6. xa lắm, chẳng có sức mạnh làm vậy. Song le, tôi cậy các Thánh đã khỏi sự hiểm nghèo thế này mà được mọi sự thịnh chẳng
7. hay hết, cầu cho em muốn được khỏi mọi sự mắc phải mà làm bạn cùng các Thánh ở trên chốn thanh nhàn vô cùng, chẳng còn
8. giặc giã là những chước oán thù và xem thấy và xem chẳng thấy. Trong sách này chẳng có nói hết sự trọng các Thánh làm,
9. vì tôi chưa tìm được những sách truyền sự ấy cho đủ. Song le, ít nhiều tôi biết vậy, mà cậy những anh tôi thì sau

Hình 19: Bản dịch Quốc Ngữ

SinoNom OCR	Chữ Quốc Ngữ
湘禬主支殺退由尼□𢷮閤殺縱□𠇕尼茲綿烝垂	Dòng Đức Chúa Giê-su, Giê-rô-ni-mô Mai-ô-ri-ca làm sách này kính dâng các Thánh.
外紆主恪望□惚酌慈遊鳴移悶吶術牢泛乏産鄧𦋦𡛔□尺酌於	Tôi lạy Chúa các Thánh là Rất Thánh Đức Bà Ma-ri-a, cùng cả và nước Thiên đàng. Tôi là em chót các Thánh ở trên
𤵶沙冬𡬼默㝐過於帯世渚碎舜□尼汝怒𠰳松变𨒺□術妥閑朱派	trời cùng các người hay giữ đạo ở dưới thế, thì tôi dâng sách này cùng các truyện tóm lại thích ra tiếng An Nam, cho Rất
□□擬斧拱忍望固慈來退尼渚別𥪝烝嗟忿断麻訟災𥘷纀包𪓇	Thánh Đức Mẹ Chúa tôi cùng các Thánh, có ý cho Nước này được biết đường các Thánh đi xưa mà rỗi linh hồn. Tôi đã hay
紅心碎孕昂蠡□宦將□尘退秋□欺群於平店□辞㛪慱薺翁□	có kẻ thấy sự sang trọng các quan tướng Đức Chúa Giê-su làm khi còn ở trong xác thịt, thì dái sợ, vì mình khác
𠰉深庄悶海湿𠆳企退𧼋彈𠆳分望𢀭魂事險嗟𠊛尼麻時□孕蛋花	xa lắm, chẳng có sức mạnh làm vậy. Song le, tôi cậy các Thánh đã khỏi sự hiểm nghèo thế này mà được mọi sự thịnh chẳng
𢧚埮□涞𡛔悶㝐魂㖡享默濕麻𠆳群𠀧怒塗於達准術閑𡜐恪庄濕	hay hết, cầu cho em muốn được khỏi mọi sự mắc phải mà làm bạn cùng các Thánh ở trên chốn thanh nhàn vô cùng, chẳng còn
𢴑□界仍𢧚逐□勾𨁪𥘷碎滲𦋦庄体中𠮩爲産固氿𢩽卒座忍筆勾	giặc giã là những chước oán thù và xem thấy và xem chẳng thấy. Trong sách này chẳng có nói hết sự trọng các Thánh làm,
爲爲𩵜簒密仍□涞孕□朱𩵜及包乙脱鄉別全麻急仙□碑署𡛤	vì tôi chưa tìm được những sách truyền sự ấy cho đủ. Song le, ít nhiều tôi biết vậy, mà cậy những anh tôi thì sau

Hình 20: Kết quả quy trình tiền xử lý dữ liệu

- ***CÁC THÁNH TRUYỆN - THÁNG 5***



Hình 21: Văn bản Hán Nôm

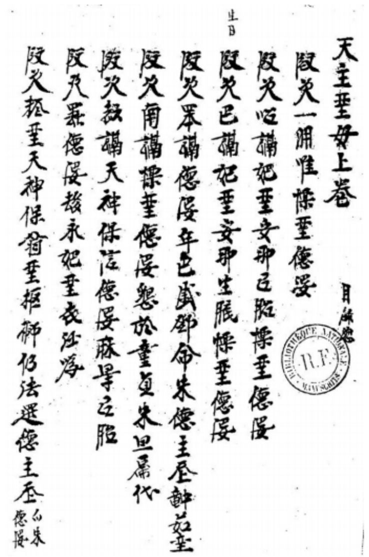
1. Tháng Năm: Ba mươi mốt ngày MỒNG MỘT.
2. Ông Thánh Phi-li-phê cùng ông Thánh Gia-cô-bê hậu.
3. Ông Thánh này là người Đức Chúa Giê-su đã chọn mở đạo cho thiên hạ, thì sinh ra ở nước Giudia. Khi
4. người còn trẻ hay xem sách ông Thánh Mai-sen, mà biết tỏ sự trong ấy truyện sự Đức Chúa Blời ra đời. Cho
5. nên khi gặp Đức Chúa Giê-su liền tin, lại gọi ông Na-tan-ay đến xem Đức Chúa Giê-su nữa. Lại có
6. một lần khác, kẻ chẳng có đạo đến xem Đức Chúa Giê-su, thì người mừng lắm. Đức Chúa Giê-su đã về thiên
7. đàng đoạn, thì người sang bên A-si-a cả, lấy lời nói cùng làm nhiều phép lạ cho nhiều người ta bỏ sự dối
8. mà theo thật. Lại sang nước Sitia đem tin lành cho muôn vàn người được mừng là tin Đức Chúa Blời xuống thế
9. cho người ta được cậy. Người dựng nên nhiều nhà thờ Chúa sinh muôn vật, lại phá nhiều nơi thờ quấy quá. Trong hai

Hình 22: Bản dịch Quốc Ngữ

SinoNom OCR	Chữ Quốc Ngữ
尚言也遣㖇	Tháng Năm: Ba mươi mốt ngày
夢箋翁至丕䵷执□舘至茹姑□後	MỒNG MỘT. Ông Thánh Phi-li-phê cùng ông Thánh Gia-cô-bê hậu.
𦞐垂尼羅骨德主技秋包梅灵道朱天下軒生𨤮執泛樞移倒欺	Ông Thánh này là người Đức Chúa Giê-su đã chọn mở đạo cho thiên hạ, thì sinh ra ở nước Giudia. Khi
㝵群𨀈哈枯□篭産枚蓮麻别謝事𨑮意僊事德主𡖡羅冬	người còn trẻ hay xem sách ông Thánh Mai-sen, mà biết tỏ sự trong ấy truyện sự Đức Chúa Blời ra đời. Cho
牢欺退德主後秋連信吏噲翁那些那冬典祜德主技状女吏圖	nên khi gặp Đức Chúa Giê-su liền tin, lại gọi ông Na-tan-ay đến xem Đức Chúa Giê-su nữa. Lại có
蔑丞終几□固道典□德主𨁥秋事骨㘈檪𥘀主拱秋包衕天	một lần khác, kẻ chẳng có đạo đến xem Đức Chúa Giê-su, thì người mừng lắm. Đức Chúa Giê-su đã về thiên
童□事㝵節迎紅庄倒䠷𥘀𡖡吶妥忍堯法遙朱□㝵些絰七	đàng đoạn, thì người sang bên A-si-a cả, lấy lời nói cùng làm nhiều phép lạ cho nhiều người ta bỏ sự dối
麻饒黄吏壽唁𧿆罕閟叽信派朱門萬骨明舘信德主𡖡嵬	mà theo thật. Lại sang nước Sitia đem tin lành cho muôn vàn người được mừng là tin Đức Chúa Blời xuống thế
世東景些時忌㝵蜂年㛑茄懿主吶將□□冲忌蜍恨遙𨤮	cho người ta được cậy. Người dựng nên nhiều nhà thờ Chúa sinh muôn vật, lại phá nhiều nơi thờ quấy quá. Trong hai

Hình 23: Kết quả quy trình tiền xử lý dữ liệu

- ***THIÊN CHÚA THÁNH MẪU***



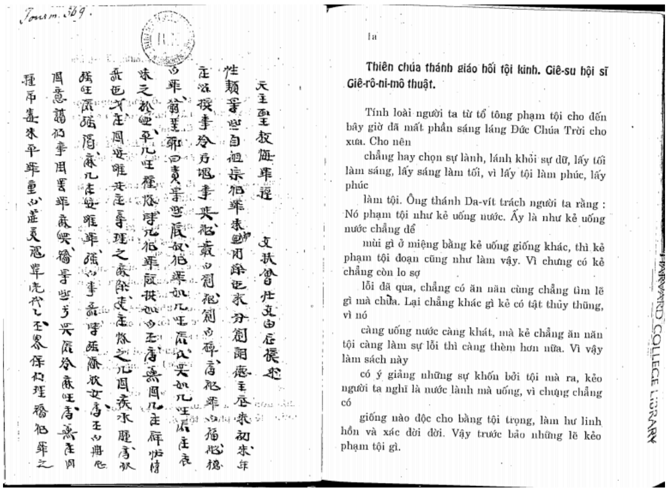
Hình 24: Văn bản Hán Nôm

Hình 25: Bản dịch Quốc Ngữ

天主聖母上卷目綠之	THIÊN CHÚA THÁNH MẪU. Thượng Quyển. MỤC LỤC TỔNG.
段次一洇憚操靈德□	Đoạn thứ nhất: Dòng dõi Rất Thánh Đức Bà.
段次□䋥妑雯乂命包□操雯怒邊	Đoạn thứ hai: Giảng bà Thánh An-na chịu thai Rất Thánh Đức Bà.
段次𠯕䋥姤雪姿𢁉堂復悌窒㥛□	Đoạn thứ ba: Giảng bà Thánh An-na sinh đẻ Rất Thánh Đức Bà.
段次罘䋥𥘀侈乎色歳𣼽命求𢹣主全𧈬茹些	Đoạn thứ bốn: Giảng Đức Bà nên ba tuổi dâng mình cho Đức Chúa Trời trong nhà thánh.
段次南講操嘆怨涹㦖於管𢟘朱退爲人	Đoạn thứ năm: Giảng Rất Thánh Đức Bà khấn ở đồng trinh cho đến trọn đời.
段次紹嬋永𢶒保園德□麻㝵尋醒	Đoạn thứ sáu: Giảng Thiên Thần bảo tin Đức Bà mà Người chịu thai.
段□羅𥘀侈𧼋永妑望眾沙𩿠	Đoạn thứ bảy: Đức Bà đi viếng bà Thánh I-sa-ve.
段次澄棄天碎保箱雲皐□仍渃霳㗂笠仝□朱德昇	Đoạn thứ tám: Thánh Thiên Thần bảo ông Thánh Giu-se những phép lạ Đức Chúa Trời làm cho Đức Bà.

Hình 26: Kết quả quy trình tiền xử lý dữ liệu

- ***THIÊN CHÚA THÁNH GIÁO HỐI TỘI KINH***



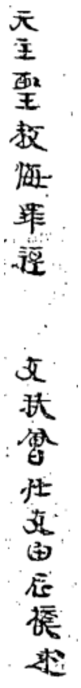
Hình 27: Văn bản Hán Nôm và bản dịch Quốc ngữ tương ứng

SinoNom OCR	Chữ Quốc Ngữ
天至聖-王 叔洒罪經又扶會尹元申忌揆追	Thiên chúa thánh giáo hối tội kinh Giê su hội sĩ Giê rô ni mô thuật
性類差-啞自 桓蒜油舘求 𥚄闭鈽尼求分創口𣘈主丕朱初朱年	Tính loài người ta từ tổ tông phạm tội cho đến bây giờ đã mất phần sáng láng Đức Chúa Trời cho xưa Cho nên
庄紅 拱事冷另塊事 唭𥚄嫩⺈包 𥚄創⺈碎爲 𥚄罪⺈絶 𥚄-桃	chẳng hay chọn sự lành lánh khỏi sự dữ lấy tối làm sáng lấy sáng làm tối vì lấy tội làm phúc lấy
⺈車鑰 悪鄧紅素骨 -啾浪叙 肔澤紅几丐 潙衣口佃几 旺寇 庄枚	làm tội Ông thánh Da vít trách người ta rằng Nó phạm tội như kẻ uống nước Ấy là như kẻ uống nước chẳng để
味 爻於 刨平几 旺口沁肆口 𣳔罪 𣜿拱 𣛢⺈死香烝 𡆫几庄碎 悀核	mùi gì ở miệng bằng kẻ uống giống khác thì kẻ phạm tội đoạn cũng như làm vậy Vì chưng có kẻ chẳng còn lo sợ
口包戈庄悶 娑 灘决泛辱理之麻徐吏庄 烙泠口固疾永 䐉爲双	lỗi đã qua chẳng có ăn năn cùng chẳng tìm lẽ gì mà chữa Lại chẳng khác gì kẻ có tật thủy thũng vì nó
濡征 灖𦧲阑麻几悴綫 灘罕 濡冲事口肆濡派欣 爻爲丕⺈典尼	càng uống nước càng khát mà kẻ chẳng ăn năn tội càng làm sự lỗi thì càng thèm hơn nữa Vì vậy làm sách này
悶意-𨔈禍事 淵罷罪麻魂 𩩖骨 -啞 㝵口底於麻旺爲-烝産回	có ý giảng những sự khốn bởi tội mà ra kẻo người ta nghĩ là nước lành mà uống vì chưng chẳng
口桴-審朱平罪重⺈ 庄灵口弹光-代丿 丕東 鐌悶埋橋犯罪之	giống nào độc cho bằng tội trọng làm hư linh hồn và xác đời đời Vậy trước bảo những lẽ kéo phạm tội gì

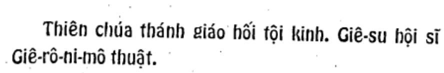
Hình 28: Kết quả quy trình tiền xử lý dữ liệu

Về ***quy trình gióng hàng các cặp chữ***, kết quả của nhóm được thể hiện qua các hình ảnh sau (*nhóm sẽ chỉ chọn ra **một cột Hán Nôm bất kỳ của trang Hán Nôm đầu tiên làm đại diện**, kèm theo bản dịch Quốc ngữ tương ứng của câu đó*)

- ***CÁC THÁNH TRUYỆN - THÁNG 1***



Hình 29: Văn bản Hán Nôm



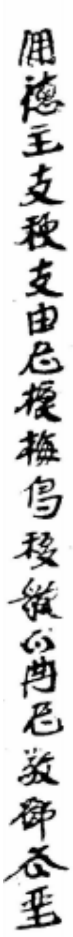
Thiên chúa thánh giáo hối tội kinh. Giê-su hội sĩ
Giê-rô-ni-mô thuật.

Hình 30: Bản dịch Quốc Ngữ

SinoNom OCR	Chữ Quốc Ngữ
天	thiên
至	chúa
聖	thánh
王	
叔	giáo
洒	hối
罪	tội
經	kinh
又	giê
扶	su
會	hội
尹	sĩ
元	giê
申	rô
忌	ni
揆	mô
追	thuật

Hình 31: Kết quả quy trình gióng hàng tự động

- ***CÁC THÁNH TRUYỆN - THÁNG 5***



Hình 32: Văn bản Hán Nôm

Dòng Đức Chúa Giê-su, Giê-rô-ni-mô Mai-ô-ri-ca làm sách này kính dâng các Thánh.

Hình 33: Bản dịch Quốc Ngữ

洞	dòng
德	đức
主	chúa
支	giê
殺	su
退	giê
由	rô
尼	ni
□	
瓏	
圖	mô
	mai
殺	ô
	ri
縱	ca
□	làm
册	sách
尼	này
茲	kính
	dâng
綿	các
柔	
垂	thánh

Hình 34: Kết quả quy trình gióng hàng tự động

- ***THIÊN CHÚA THÁNH MẪU***

Hình 35: Văn bản Hán Nôm

Đoạn thứ nhất: Dòng dõi Rất Thánh Đức Bà.

Hình 36: Bản dịch Quốc Ngữ

段	đoạn
次	thứ
一	nhất
洞	dòng
憚	dõi
操	rất
靈	thánh
德	đức
□	bà

Hình 37: Kết quả quy trình gióng hàng tự động

- ***THIÊN CHÚA THÁNH GIÁO HỐI TỘI KINH***

Hình 38: Văn bản Hán Nôm

Tháng Năm: Ba mươi mốt ngày

Hình 39: Bàn dịch Quốc Ngữ

尚	tháng
言	năm
也	ba
遣	mươi
	mốt
喝	ngày

Hình 40: Kết quả quy trình gióng hàng tự động

5) Hạn chế và định hướng cải tiến

5.1) Hạn chế

Mặc dù đã đạt được những kết quả nhất định, bài toán gióng hàng vẫn tồn tại một số hạn chế đáng lưu ý, ảnh hưởng đến hiệu quả và độ chính xác của mô hình:

5.1.1) Ngữ liệu đầu vào không nhất quán

- *Đa dạng cấu trúc PDF*: Mỗi file PDF có cấu trúc khác nhau, ví dụ như định dạng văn bản, cách sắp xếp câu hoặc độ phân giải hình ảnh. Điều này làm tăng độ phức tạp khi xây dựng quy trình tiền xử lý phù hợp cho từng loại tài liệu.
- *Chất lượng hình ảnh thấp*: Các hình ảnh được quét thường gặp phải tình trạng như:
  - Chữ bị mờ, nhòe, hoặc mất nét do quá trình scan.
  - Sự xuất hiện của các yếu tố không mong muốn như nhiễu, vết lem mực, hoặc các phần tử dư thừa (chẳng hạn như dấu chấm, vạch kẻ).
- *Đặc thù chữ viết tay*: Với ngữ liệu viết tay, sự đa dạng trong phong cách chữ viết (nét đậm, nét nhạt, độ nghiêng) gây ra khó khăn trong nhận dạng và xử lý.

5.1.2) Hạn chế của công cụ OCR

- Độ chính xác thấp với tiếng Việt cổ: Do thiếu dữ liệu huấn luyện và đặc thù của ngôn ngữ Việt cổ, các công cụ OCR hiện tại thường gặp khó khăn trong việc:
  - Nhận dạng chính xác các ký tự SinoNom, đặc biệt là các ký tự ít phổ biến hoặc có nét phức tạp.
  - Xử lý các dấu đặc trưng của tiếng Việt, dẫn đến lỗi nhận dạng ở các ký tự có dấu (dấu thanh hoặc dấu phụ).
- Tác động của ngoại cảnh: Các yếu tố như độ sáng, độ tương phản, hoặc chất lượng giấy gốc có thể làm giảm hiệu quả của OCR, dẫn đến kết quả đầu ra không nhất quán.

5.1.3) Hạn chế trong dữ liệu từ điển

- *Thiếu ký tự đặc thù*: Bộ từ điển tham chiếu hiện tại chưa bao quát được toàn bộ các ký tự SinoNom, đặc biệt là các ký tự cổ ít được sử dụng.
- *Không có ngữ cảnh*: Việc so sánh dựa trên từ điển thường bỏ qua ngữ cảnh của câu, dẫn đến các lỗi gióng hàng khi ký tự có nhiều nghĩa hoặc cách đọc khác nhau.
- *Khả năng xử lý ký tự không hợp lệ*: Các ký tự bị lỗi (mất nét, viết đè) không được xử lý hiệu quả, dẫn đến sai lệch trong kết quả gióng hàng.

5.1.4) Khó khăn trong gióng hàng từng ký tự

- *Độ chính xác của vị trí ký tự*: Mặc dù mô hình YOLOv5 cung cấp vị trí các ký tự trong hình ảnh, sai lệch nhỏ trong tọa độ hoặc kích thước vùng chọn có thể làm giảm độ chính xác khi gióng hàng.
- *Phân biệt ký tự tương đồng*: Một số ký tự có hình dạng tương tự nhau (ví dụ: “日” và “曰” trong Hán-Nôm) gây khó khăn trong việc phân biệt, đặc biệt là khi chất lượng hình ảnh thấp.
- *Xử lý chữ chồng chéo*: Trong nhiều trường hợp, các ký tự bị viết đè hoặc chèn lẫn nhau, gây khó khăn lớn cho cả việc nhận dạng và gióng hàng.

5.1.5) Hạn chế về tài nguyên và thời gian

- *Yêu cầu xử lý thủ công*: Một số bước trong quy trình (như tiền xử lý dữ liệu hoặc kiểm tra kết quả gióng hàng) vẫn đòi hỏi sự can thiệp của con người, làm tăng chi phí và thời gian thực hiện.
- *Hiệu suất mô hình*: Khi xử lý lượng lớn dữ liệu hoặc các file PDF có độ phức tạp cao, thời gian xử lý của hệ thống tăng đáng kể, ảnh hưởng đến tính khả thi trong triển khai thực tế.

5.2) Định hướng cải thiện

5.2.1) Xây dựng và mở rộng bộ dữ liệu phong phú hơn

- *Tăng cường dữ liệu ký tự SinoNom*:
  - Thu thập thêm ngữ liệu từ các tài liệu Hán-Nôm, đặc biệt là các tài liệu ít phổ biến hoặc có tính chất lịch sử cao.
  - Kết hợp với các chuyên gia ngôn ngữ để xây dựng danh mục đầy đủ hơn về các ký tự SinoNom, bao gồm cả những ký tự hiếm hoặc ít xuất hiện.
- *Xử lý dữ liệu lỗi*:
  - Ghi nhận và bổ sung các trường hợp ký tự bị lỗi (chữ lem, mất nét, viết đè) để cải thiện khả năng nhận dạng của mô hình OCR.
  - Áp dụng các kỹ thuật data augmentation (tăng cường dữ liệu) như làm mờ, chỉnh độ sáng, hoặc thêm nhiễu để tạo ra các mẫu dữ liệu đa dạng hơn.

5.2.2) Nâng cấp thuật toán OCR

- *Lựa chọn mô hình OCR khác hiệu quả hơn*:

- Sử dụng các mô hình hỗ trợ tiếng Việt tốt hơn để đảm bảo nó hiểu rõ hơn về các trúc câu từ và dấu câu.
- *Tối ưu hoá cho tiếng việt cổ*:
  - Phát triển và tối ưu thêm việc xử lý các ký tự phức tạp hoặc không phổ biến.

5.2.3) Tích hợp thêm công cụ nhận dạng ngữ cảnh

- *Tích hợp ngữ cảnh vào quá trình gióng hàng*:
  - Sử dụng các mô hình ngôn ngữ (Language Models) như GPT hoặc BERT để hiểu và xử lý ngữ cảnh của câu, từ đó giảm thiểu sai sót khi gióng hàng các ký tự đa nghĩa.
- *Xây dựng từ điển ngữ cảnh*:
  - Phát triển một bộ từ điển ngữ cảnh chứa thông tin về cách sử dụng các ký tự SinoNom trong ngữ pháp tiếng Việt, giúp cải thiện độ chính xác khi xử lý các ký tự.

5.2.4) Tăng cường khả năng kiểm định chất lượng

- *So sánh kết quả gióng hàng*:
  - Tích hợp thêm các bước kiểm định chất lượng, so sánh kết quả gióng hàng với các tài liệu tham chiếu hoặc ngữ liệu có độ chính xác cao.
- *Phản hồi từ chuyên gia*:
  - Nhận phản hồi từ các chuyên gia Hán-Nôm để đánh giá và cải thiện kết quả gióng hàng, đặc biệt trong các trường hợp ký tự hiếm hoặc phức tạp.

6) Nguồn tham khảo

[1] N. H. Bửu Long, *“Slide bài giảng lý thuyết nhập môn NLP”*.

[2] *“Hướng dẫn bài tập thực hành NLP”*.

[3] *YOLOv5 PyTorch Format*