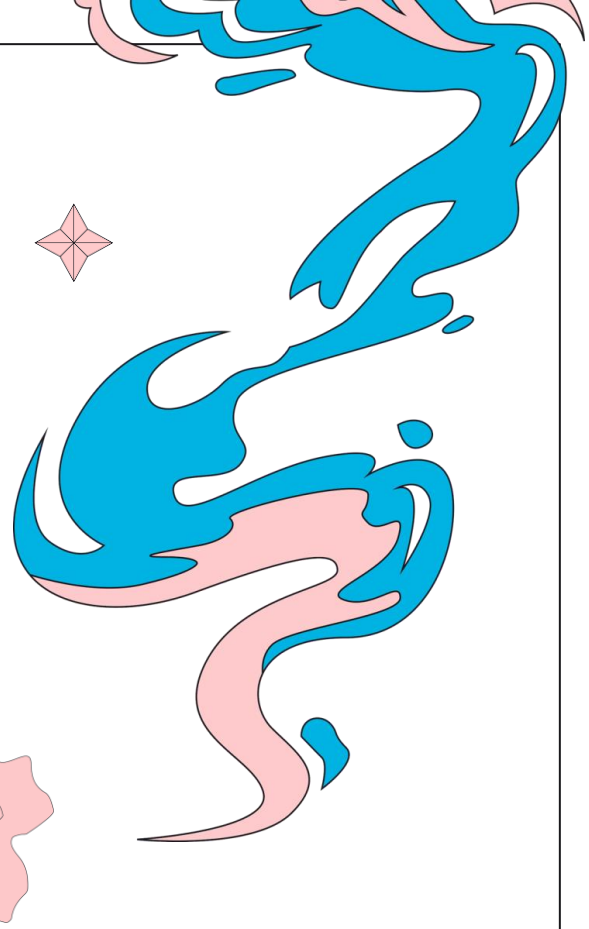
Nhận dạng ký tự quang học

Tài liệu cổ

Chữ Nôm

字喃

chữ Nôm



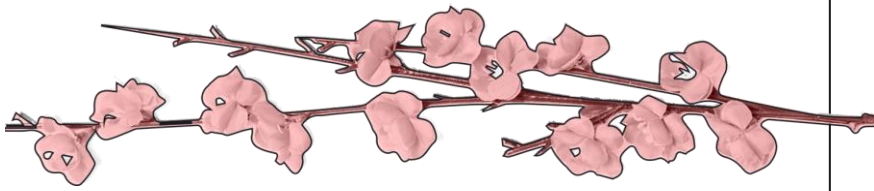
Thành viên

- 19127078 - Nguyễn Đỗ Thanh Trúc 
- 19127135 - Phạm Bảo Hân 

Giảng viên hướng dẫn

- Đinh Điền
- Lương An Vinh
- Lê Thị Thúy Hằng

Mục lục



01

Chữ Nôm

Nôm character

1. **Tổng quan:**
    - Chữ Nôm được hình thành và phát triển trong khoảng từ thế kỷ 10 - thế kỷ 19.
    - Hàn Thuyên được công nhận là người có công truyền bá chữ Nôm.
    - Không thể thống kê được có bao nhiêu chữ Nôm.
    - Có ít nhất 32,695 chữ Nôm theo *Hội Bảo tồn Di sản chữ Nôm*.

Hàn Thuyên

Nôm character

2. Đặc điểm

- **Mượn từ Hán Việt (Sino-Vietnamese)**
  - Chữ sơn nghĩa là ‘ngọn núi’ được viết 山 (phát âm: shān)
- **Sáng tạo bởi người Việt Nam**
  - Mượn 2 từ Hán: 1 để diễn tả âm thanh, 1 để diễn tả ý
    - Từ lửa nghĩa là ‘ngọn lửa’ được viết 焗 (gồm 2 thành phần: 火 (nghĩa: lửa), 呂 (phát âm: lǚ)
  - Mượn từ Hán để diễn tả âm thanh
    - Từ một nghĩa là ‘số một’ được viết 沒 (phát âm: mò)
  - Mượn 2 từ Hán để diễn tả ý nghĩa
    - Từ trời nghĩa là ‘bầu trời’ được viết 𡈽 (gồm 2 thành phần: 天 (nghĩa: thiên đường), 上 (nghĩa: phía trên)

Nôm character

3. Hệ thống chữ viết

- Ghi âm tiếng mẹ đẻ.
  => Ý thức tự chủ ngôn ngữ muốn có chữ viết riêng của dân tộc ta
- Lưu giữ lịch sử văn hóa và di sản vô giá của dân tộc Việt Nam. (văn học, triết học, lịch sử, luật pháp, y học, tôn giáo,…).
- Với sự ra đời của chữ Quốc Ngữ, chữ Nôm dần dần biến mất.
  => **Vấn đề cấp bách**: Bảo tồn di sản, văn hóa của dân tộc Việt Nam.

𤾓𢆥𥪞𡎝𠊛些
Trăm năm trong cõi người ta.
𡨸才𡨸命窖□恄饒
Chữ tài chữ mệnh khéo là ghét nhau.
𣦆戈没局滧桃
Trải qua một cuộc bể dâu,
仍調□□㐌□疸□
Những điều trông thấy đã đau đớn lòng.
遻之彼嗇斯豐
Lạ gì bỉ sắc tư phong,
歪撐涓貝䏧紅打桱
Trời xanh quen với má hồng đánh ghen.
稿蕎吝□𫅷畑
Cảo thơm lần giở trước đèn,
風情固録群傳史撐
Phong tình có lục còn truyền sử xanh.
浪䊷□靖朝明
Rằng: Năm Gia Tĩnh triều Minh,
罕方㛪㴉𠄩京凭鐄
Bốn phương phẳng lặng, hai kinh vững vàng.
固茹員外户王

Truyện Kiều

Nôm character

4. **Ý nghĩa:**
   - Ghi âm tiếng mẹ đẻ.
     => Ý thức tự chủ ngôn ngữ muốn có chữ viết riêng của dân tộc ta
   - Lưu giữ lịch sử văn hóa và di sản vô giá của dân tộc Việt Nam.
     (văn học, triết học, lịch sử, luật pháp, y học, tôn giáo,…).
   - Với sự ra đời của chữ Quốc Ngữ, chữ Nôm dần dần biến mất.
     => **Vấn đề cấp bách**: Bảo tồn di sản, văn hóa của dân tộc Việt Nam.

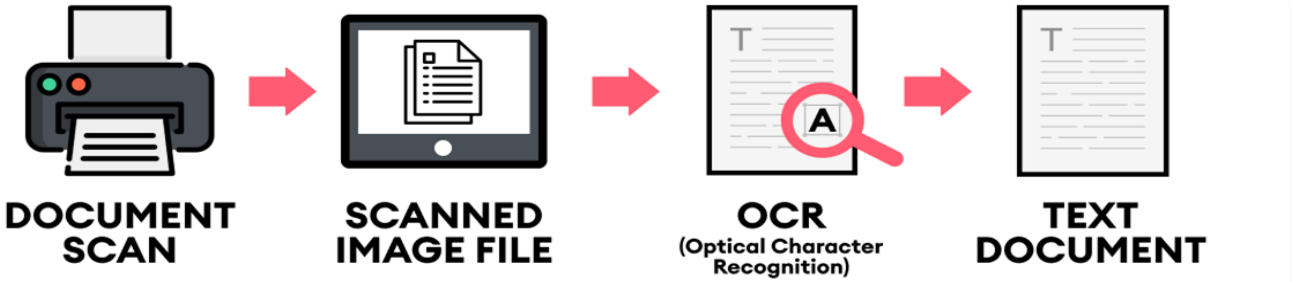
Bộ Đại Việt sử ký toàn thư
Kỷ Trưng Nữ Vương

02

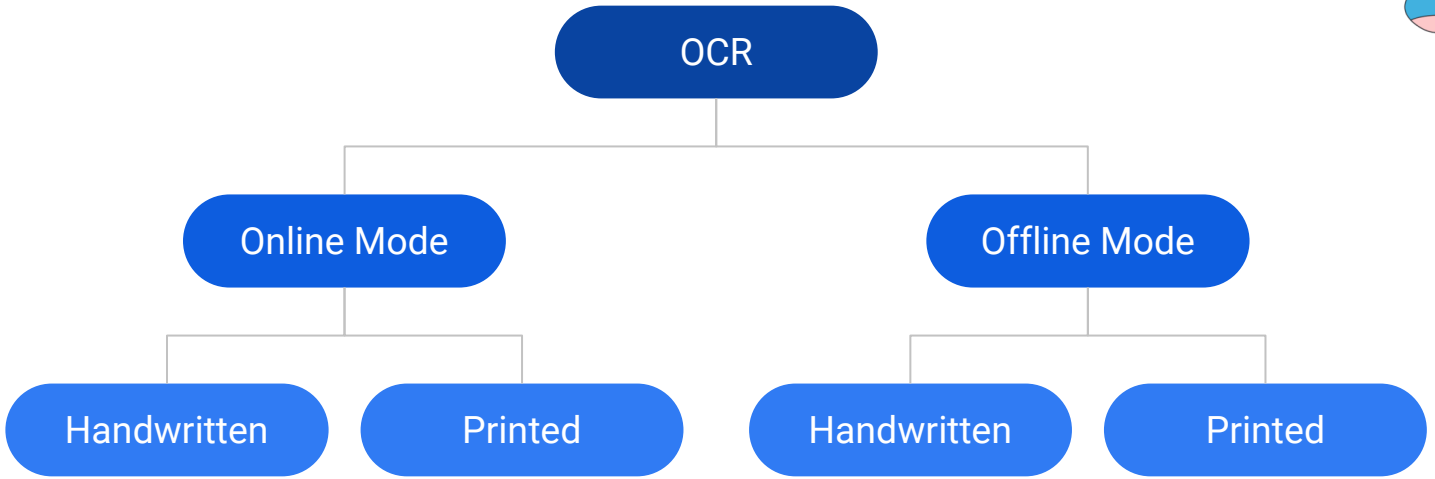
Nhận dạng ký tự quang học (OCR)

I.Định nghĩa

Nhận diện ký tự quang học là quá trình tự động nhận dạng và trích xuất các ký tự trong ảnh (ảnh văn bản viết tay, văn bản in hoặc quét), sau đó chuyển đổi thành định dạng kỹ thuật số có thể chỉnh sửa.



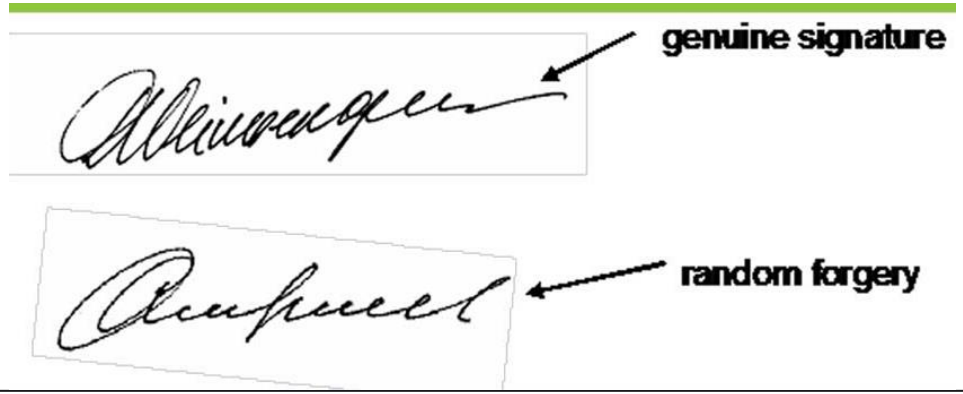
II. Phân loại

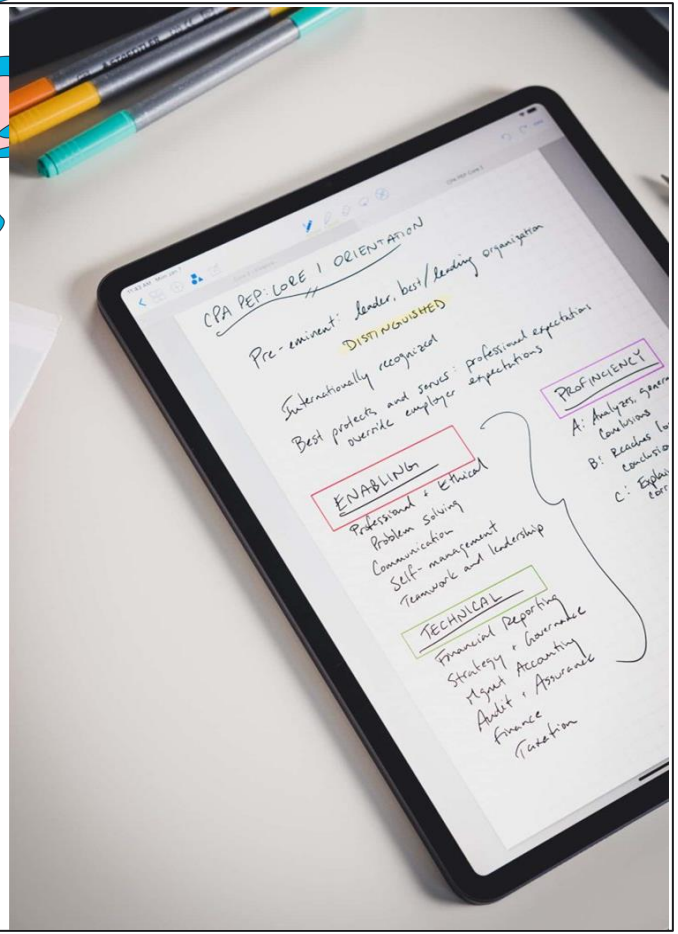

OCR
Online Mode
Offline Mode
Handwritten
Printed
Handwritten
Printed

III. Ứng dụng

Online OCR

- Bút cảm ứng dùng cho smart phone, tablet,...
- Xác minh chữ ký
- Xác thực danh tính khách hàng

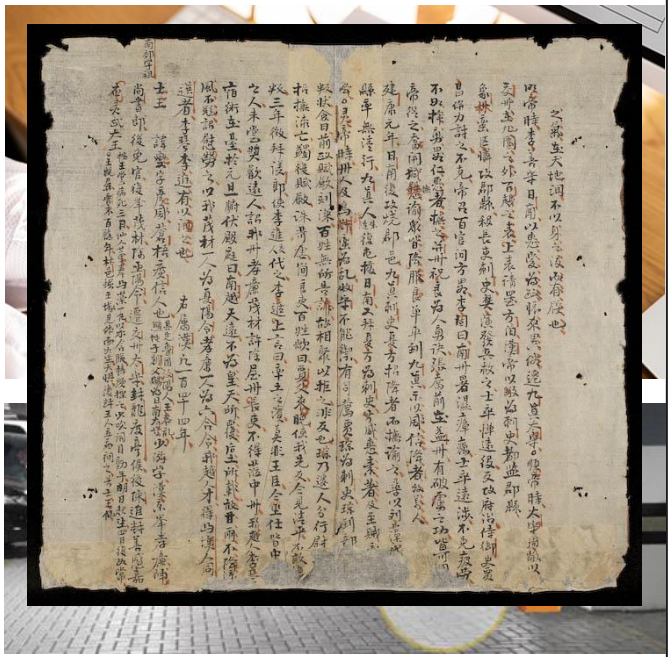




III. Ứng dụng

Offline OCR

- Chuyển đổi văn bản
- Nhận diện biển số xe
- **Số hóa các tài liệu cổ**: Những tài liệu cổ chứa đựng một lượng lớn thông tin và kiến thức lịch sử quan trọng. Việc số hóa các tài liệu này là phương pháp tốt nhất để bảo tồn chúng. Các bước chính của quá trình này bao gồm:
  - Scanning
  - OCR
  - Lưu trữ dữ liệu

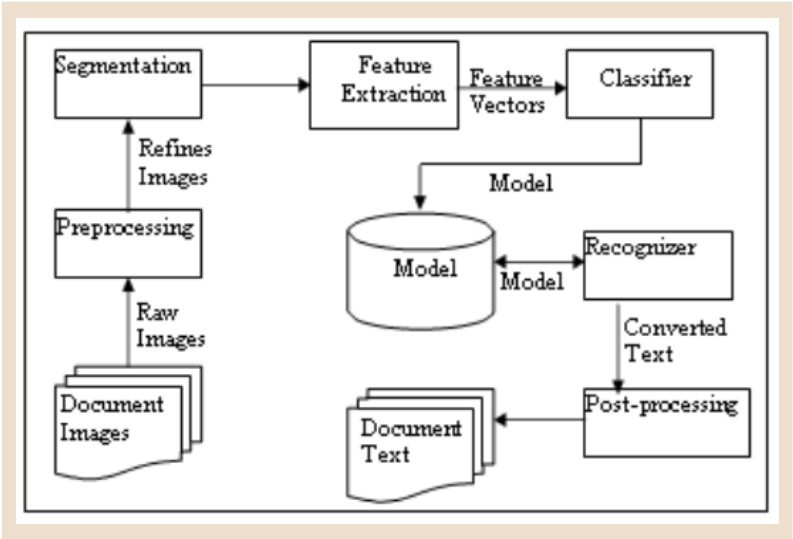


IV. Các bước thực hiện

3 Phân tách

2 Tiền xử lý

1 Thu nhận ảnh



Rút trích đặc trưng 4

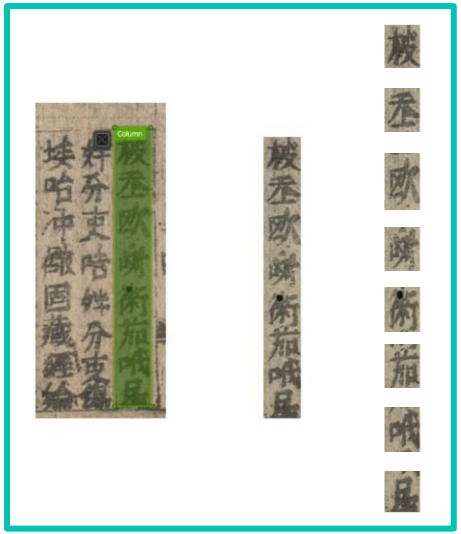
Phân loại ký tự 5

Hậu xử lý 6

IV. Các bước thực hiện

Thu nhận ảnh

Tiền xử lý



Phân tách

被 Bị
歪 trời
吹 thổi
喲 đuổi
術 về
茄 nhà
吪 dạy
民 dân

Rút trích đặc trưng
+
Phân loại ký tự

被 Bị
歪 trời
欧 xua
喲 đuổi
術 về
茄 nhà
吪 dạy
民 dân

Hậu xử lý

03

Sửa lỗi văn bản

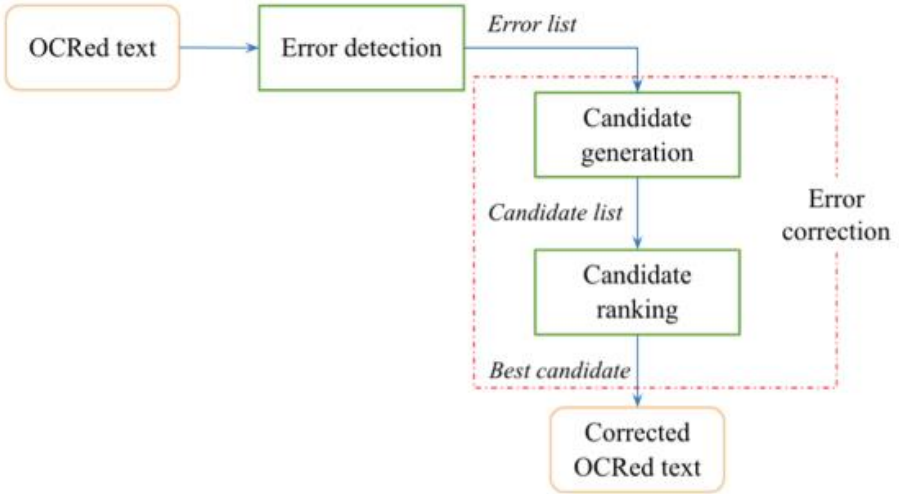
Sửa lỗi văn bản

Quy trình:

- Phát hiện lỗi: xác định các token lỗi từ văn bản đầu vào.
- Sửa lỗi: khắc phục các token bị sai (danh sách các từ ứng viên).

Yêu cầu:

- Chính xác, không sinh ra lỗi mới.
- Độ phức tạp về thời gian và bộ nhớ thấp.



Phân loại theo loại lỗi

- Lỗi chính tả (Spelling error): từ sai chính tả

VD: nghành (ngành), cọ sát (cọ xát)

Error Type	Example
Cognitive error:	I don't know the correct spelling of Levenstain distance.
Typographic:	THis sentence was typed in haser.
Typographic (OCR):	SUppLEMENTAhy lNFOhMATlON.
Typographic (Diacritic):	The authors of this article are Daniel Hladek, Matus Pleva and Jan Stas.

Note: The spelling errors are underlined.

- Lỗi ngữ pháp (Grammatical error): từ đúng chính tả nhưng sai về mặt ngữ nghĩa trong câu

Type	Example
IWC	自己有双聪明能干的手，什么都能做出来。 You have smart hands to do everything. (Tips: "hands" cannot be combined with "smart")
CM	绿色植物具有产生氧气。 Plants have (the ability) to produce oxygen. (Tips: Lack of object "the ability")
CR	我们已走了约十里左右的路程。 We had walked about 10 miles or so. (Tips: "about" and "or so" are redundant)
SC	交通事故发生的原因是开车看手机造成的。 Traffic accidents are caused by (because) looking at cell phones while driving. (Tips: the structure of "because" and "caused by" cannot appear together in one sentence)
IWO	我改正并认识了自己的错误。 I corrected and realized my fault. (Tips: realize the fault first and correct it later)
ILL	我们应该防止事故不发生。 We should prevent accidents from not occurring. (Tips: double negation causes illogical errors)
AM	刚一开门，看病的就进来了。 As the door opened, the doctor/patient came in. (Tips: there is an ambiguity about who comes in)

- Sai trật tự từ (IWO)
- Sai cách kết hợp từ (IWC)
- Thiếu thành phần (CM)
- Dư thành phần (CR)
- Nhầm lẫn cấu trúc (SC)
- Phi logic (Phi logic, ILL)
- Nhập nhằng (AM)

Phương pháp

Thủ công

- Độ chính xác cao
- Tốn kém về chi phí, thời gian
- Phụ thuộc vào trình độ của người review
- Cần người theo dõi, đánh giá hiệu suất công việc
- Không phù hợp với các văn bản có bản quyền

Phương pháp

- *Non-word errors: từ sai chính tả và không có trong từ điển (từ vô nghĩa)*
- *Real-word errors: từ được dùng sai ngữ cảnh nhưng có trong từ điển (từ có nghĩa)* Vd: *He is string (strong).*

Dựa trên luật

- Tốc độ xử lý nhanh.
- Sửa lỗi về non-word errors, một số real-word errors đơn giản.
- Thiếu sự uyển chuyển, đặc biệt đối với các tác phẩm văn học.
- Cần xây dựng bộ từ điển đủ lớn, các bộ luật đủ mạnh, văn phạm không nhập nhằng, ...
- Ví dụ:
  - Tra cứu từ điển dùng hash table, search tree
  - Thuật toán phân tích cú pháp Earley Parser

Dựa trên máy học:

- Tốc độ xử lý nhanh
- Sửa nhiều loại lỗi (real-word and non-word errors)
- Chi phí về việc huấn luyện mô hình (ngữ liệu huấn luyện, chi phí huấn luyện: thời gian, GPU, …)
- Phức tạp hơn các mô hình theo hướng dựa trên luật
- Học ngữ cảnh của văn bản => phù hợp cho nhiều dạng văn bản khác nhau
- Ví dụ:
  - Mô hình ngôn ngữ dạng học sâu: CNN, Transformer, …
  - Mô hình ngôn ngữ Ngram

04

Bài báo

TOWARDS NÔM HISTORICAL DOCUMENT OPTICAL CHARACTER RECOGNITION

Tran Thi Anh Thu, Le Pham Ngoc Yen, Tran Thai Son, Dinh Dien

Tổng quan

End-to-end model: Tesseract OCR, YOLOv5

Two-stage model:

- Mô hình nhận diện: YOLOv5
- Mô hình phân loại: CNN: AlexNet, ResNet

Table 5. Final top-1 and top-5 validation (mAP@0.5, %) on Nôm datasets of the three OCR baselines

Model	NomTrain	NomTest	
	K1871	K1902	LVT
End-to-end YOLO	0.0618	0.062	
YOLO-AlexNet	98.47 - 99.14	59.02 - 68.76	59.78 - 70.60
YOLO-ResNet101	98.91 - 99.32	71.80 - 80.54	72.96 - 82.36

Tổng quan

Mô hình ngôn ngữ

MacBERT: MLM as correction BERT

- MacBERT là mô hình dựa trên mô hình RoBERTa, nhưng được chỉnh sửa để dùng cho tiếng Trung.
- Thay vì sử dụng ký hiệu [MASK], MacBERT dùng các từ có nghĩa tương đương để thay thế. Mô hình sẽ tự động nhận diện chữ được mask, dự đoán và sửa đổi thành chữ ban đầu.

Original Sentence 使用语言模型来预测下一个词的概率。 we use a language model to predict the probability of the next word.

	Chinese	English
Original Masking	使 用 语 言 **[M]** 型 来 **[M]** 测 下 一 个 词 的 概 率 。	we use a language **[M]** to **[M]** **##di** **##ct** the **pro** **[M]** **##bility** of the next word .
+ WWM	使 用 语 言 **[M]** **[M]** 来 **[M]** **[M]** 下 一 个 词 的 概 率 。	we use a language **[M]** to **[M]** **[M]** **[M]** the **[M]** **[M]** **[M]** of the next word .
++ N-gram Masking	使 用 **[M]** **[M]** **[M]** **[M]** 来 **[M]** **[M]** 下 一 个 词 的 概 率 。	we use a **[M]** **[M]** to **[M]** **[M]** **[M]** the **[M]** **[M]** **[M]** **[M]** **[M]** next word .
+++ Mac Masking	使 用 语 法 建 模 来 预 见 下 一 个 词 的 几 率 。	we use a **text system** to **ca** **##lc** **##ulate** the **po** **##si** **##bility** of the next word .

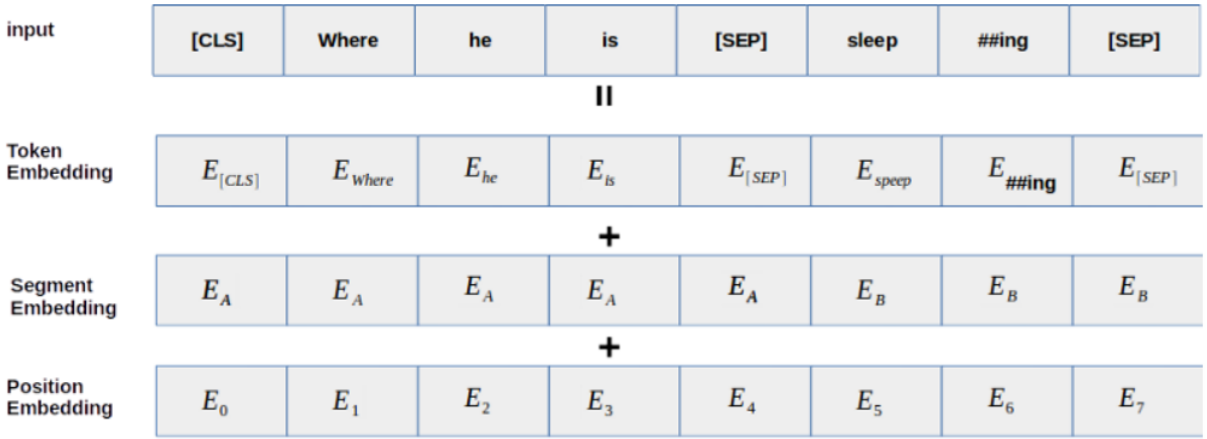
Phù hợp với bài toán Sửa lỗi văn bản cho chữ Nôm do cơ chế masking và sự tương đồng giữa chữ Nôm và Hán tự.

Tổng quan

Mô hình ngôn ngữ

MacBERT: MLM as correction BERT

- MacBERT sẽ hoạt động dựa trên các embeddings là Token, Segment, Position.



05

Quy trình

Dữ liệu

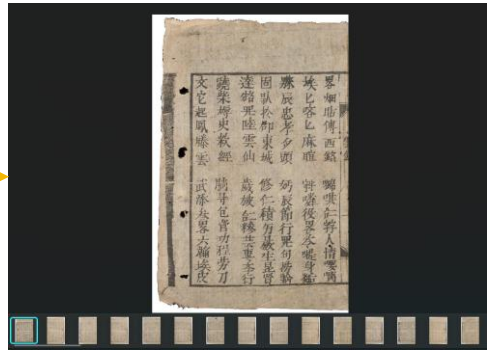
	OCR	Sửa lỗi văn bản
Chú thích	Có	Có
Đơn vị ngôn ngữ	Mẫu tự	Văn bản hoặc câu
Bình diện	Hình thái	Ngữ pháp Ngữ nghĩa
Tag set	Tất cả các mẫu tự chữ Nôm	Tất cả các mẫu tự chữ Nôm

Dữ liệu đầu vào

OCR

images

annotation.csv



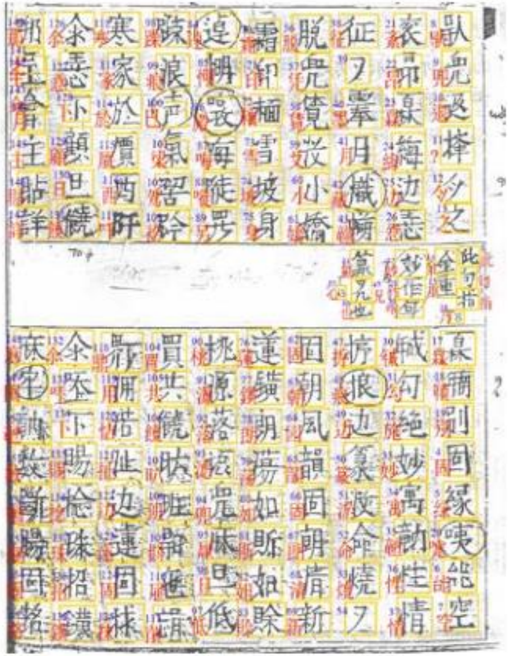
	A	B	C	D	E	F	G	H
1	label	x1	x2	y1	y2	source_im	img_w	img_h
2	7567	427	462	149	190	nlvnpf-005	493	760
3	7551	426	464	192	227	nlvnpf-005	493	760
4	4021	425	462	226	262	nlvnpf-005	493	760
5	50B3	426	463	261	302	nlvnpf-005	493	760
6	897F	426	464	304	339	nlvnpf-005	493	760
7	9298	426	464	338	374	nlvnpf-005	493	760
8	2110E	424	463	410	446	nlvnpf-005	493	760
9	552D	426	463	445	482	nlvnpf-005	493	760
10	20129	426	463	481	515	nlvnpf-005	493	760
11	21982	426	463	513	551	nlvnpf-005	493	760

Dữ liệu đầu ra

OCR

	A	B	C	D	E	F	G	H
1	label	x1	x2	y1	y2	source_im	img_w	img_h
2	7567	427	462	149	190	nlvnpf-005	493	760
3	7551	426	464	192	227	nlvnpf-005	493	760
4	4021	425	462	226	262	nlvnpf-005	493	760
5	50B3	426	463	261	302	nlvnpf-005	493	760
6	897F	426	464	304	339	nlvnpf-005	493	760

署 畑 貼 傳 西 銘
噃 唭 𧘇 䏧 人 情 嘤 噅
埃 匕 喀 匕 麻 暄
𤔤 嚹 役 署 苓 喂 身 𩵜
𪔩 辰 忠 孝 ▯ 頭
㚫 辰 節 行 ▯ 句 撈 躺
固 馭 扵 郡 東 城
修 仁 積 ▯ ▯ 生 昆 賢
達 㹸 ▯ 陸 雲 仙
歲 𢼂 𧘇 糝 芸 專 㝋 行
蹺 柴 熂 史 𣘃 經
腨 㝵 包 ☐ 功 程 劳 刀
文 它 起 鳳 滕 雲
武 添 叁 署 六 鞱 埃



(b) *Truyện Kiều* (1902) page 10

Dữ liệu

Sửa lỗi văn bản

Training

	A	B	C
1	original_text	correct_text	wrong_ids
2	黙藉擱撓欺𢚸墖	黙丕坦常欺𢚸墖	3 4 2
3	客賖神容餒迍㺞	客賖紅髮餒迍邅	3 7 4
4	撐箕㖭瀋層達	撐箕㖭瀋層達	
5	惢埃唅篇朱欄餒尼	爲埃𣛤孕朱铖餒尼	1 4 6 3
6	㰌長城龍□秦月	㰌長城龍□秦月	
7	媿汫泉須矘式蠢	媿甘泉睞矘式蠢	2 4
8	䰃吝鐮宝㩿秞	尯吝鐮宝㩿秞	1
9	姅𧊉傳檄定㐌識征	姅𧊉傳檄定𠰘出征	6 7
10	揞清平𠤩冪𧿨䮄	渃清平𠤩冪𧿨䮄	1

Inference

署 煩 貼 傳 西 銘
𠽋 嘆 極 㝵 人 情 㖼 𠶋
埃 匕 咯 匕 麻 䁮

署 畑 貼 傳 西 銘
𠽋 唭 𢀭 㝵 人 情 㖼 𠶋
埃 匕 咯 匕 麻 䁮

Note:

署 煩 貼 傳 西 銘
𠽋 嘆 極 㝵 人 情 㖼 𠶋
埃 匕 咯 匕 麻 䁮

Quy trình

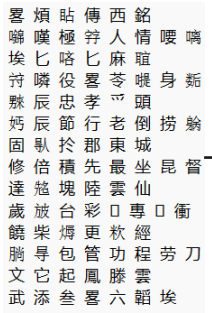
Input data → YOLOv5 → ResNet101 → top 5 ứng viên có score cao nhất cho mỗi token

+

my_annotations

label	conf	x1	y1	x2	y2	candidate	score_cls	source_img
7567	1.0	427	149	462	190	['7567' '5712' '5E	[9.9889982e-01 4.2262673	nlvnpf-0059-002.jpg
7551	1.0	426	192	464	227	['7551' '7169' '24	[0.72729665 0.22891313 0	nlvnpf-0059-002.jpg
4021	1.0	425	226	462	262	['4021' '23218' '2	[9.9990249e-01 3.9029237	nlvnpf-0059-002.jpg
50B3	1.0	426	261	463	302	['50B3' '2102F' '4	[9.9986184e-01 3.0877945	nlvnpf-0059-002.jpg
897F	1.0	426	304	464	339	['897F' '5339' '56	[9.9914682e-01 7.9658079	nlvnpf-0059-002.jpg
9298	1.0	426	338	464	374	['9298' '20E1C' '(	[9.9363434e-01 6.4138498	nlvnpf-0059-002.jpg
2110E	1.0	424	410	463	446	['2110E' '8832' '5	[0.6032809 0.03228584 0.	nlvnpf-0059-002.jpg
552D	1.0	426	445	463	482	['552D' '5922' '25	[0.05690898 0.04528534 0	nlvnpf-0059-002.jpg
20129	1.0	426	481	463	515	['20129' '5728' '2	[0.16578875 0.09892979 0	nlvnpf-0059-002.jpg
21982	1.0	426	513	463	551	['21982' '7279' '6	[9.9709940e-01 2.6703565	nlvnpf-0059-002.jpg

Quy trình



Huấn luyện trên tập dữ liệu chữ Nôm

MacBERT

Score của các ứng viên này trong mô hình MacBERT

score = w*score_cl + (1-w)*score_lm

Kết quả cuối cùng (Top 1 score)

Tập top 5 ứng viên có score cao nhất của mỗi token từ mô hình phân loại

Demo quy trình hậu xử lý

Câu đúng: Hôm nay tôi đi học.

Câu input (output OCR): Hôm nay tối đi học.

Tokenize: Hôm nay tối đi học.

VD: tokenize cho tiếng Anh (biến hình): he is sleep ##ing.

Demo quy trình hậu xử lý

OCR	Hôm: 0.7 Hôn: 0.2 Nộm: 0.1	nay: 1.0	tối: 0.4 tôi: 0.3 toi: 0.1	đi: 1.0	học: 1.0
OCR output	Hôm	nay	tối	đi	học
MacBERT	Hôm: 0.3 Hôn: 0.1 Nộm: 0.01	nay: 1.0	tôi: 0.5 tối: 0.3 toi: 0.1	đi: 1.0	học: 1.0
score = w*score_cl + (1-w)*score_lm (w = 0.5)	S(Hôm) = 0.5 S(Hôn) = 0.15 S(Nộm) = 0.055		S(tôi) = 0.4 S(tối) = 0.35 S(toi) = 0.1		
Output	Hôm	nay	tôi	đi	học

Khó khăn

OCR

Lý do khách quan:

- Chất lượng ảnh hoặc tài liệu không tốt. (nhiễu, mờ, xuống cấp,...)
- Trong văn bản viết tay, kể cả những chữ cùng loại thì vẫn có sự biến đổi trong cách viết.
- Người viết có thể viết các dòng không theo 1 đường thẳng cố định.

Khó khăn

OCR

Lý do chủ quan:

- Loại chữ gồm nhiều mẫu tự gây ra thách thức lớn cho quá trình phân loại: chữ Nôm, Hán tự, Hangul, Kanji ...

Ước tính có khoảng 50,000 ***Hán tự****, trong đó 3775 chữ được dùng thường xuyên*

vs

26 chữ cái được dùng trong ***Tiếng Anh***

- Cách viết phức tạp, có sự tương đồng nhau về hình dáng giữa các chữ
- Viết dọc ⇒ ranh giới giữa các chữ không rõ ràng

Khó khăn

Sửa lỗi văn bản

- Tìm kiếm mô hình phù hợp để ứng dụng vào việc sửa lỗi văn bản cho chữ Nôm
  - Mô hình về sửa lỗi chính tả (spelling errors) hay sửa lỗi ngữ pháp (grammatical errors)?
  - Huấn luyện mô hình từ đầu (train from scratch) hay sử dụng học chuyển giao (transfer learning)?

- Xây dựng ngữ liệu huấn luyện
  - Thu thập văn bản chữ Nôm
  - Cần chuyên viên để xác định ngữ nghĩa của từng từ, xác định lỗi ngữ pháp, …
  - Tìm giải pháp cho các từ lạ, chưa được đăng ký trên hệ thống Unicode

Thanks

Do you have any questions?