ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO ĐỒ ÁN NLP

XÂY DỰNG DATASET TỪ DỮ LIỆU THÔ

Môn: Nhập môn xử lý ngôn ngữ tự nhiên

Lớp: 23_22 - Nhóm : 13

MSSV	Họ và tên
23120019	Trương Bảo Thiên Ân
23120035	Phạm Ngọc Duy
23120060	Trần Kim Ngân
23120093	Vũ Duy Thụ

GVHD: PGS.TS Đinh Điền

TP. Hồ Chí Minh, Tháng 12 năm 2025

Mục lục

1. Mô tả cách xử lý đồ án

1.1 Đối với file gốc PDF

Giải quyết bài toán chuyển đổi dữ liệu từ file PDF thành các cặp câu song ngữ Trung – Việt.

Trích xuất thông tin: Sử dụng phương pháp Hybrid OCR, kết hợp 2 model là PaddleOCR và Tesseract để phù hợp với dữ liệu thô bao gồm tiếng Trung (gồm các Hán tự nhiều nét phức tạp) và tiếng Việt (sử dụng bảng chữ cái Latin).
Áp dụng kỹ thuật Layout Analyst bằng OpenCV để tách, cắt từng khối văn bản trước khi đưa vào nhận diện, giúp giảm nhiễu và giữ đúng thứ tự.

Làm sạch, chuẩn hóa dữ liệu thô: Sử dụng Regex để loại bỏ nhiễu và chuẩn hóa Unicode cho tiếng Việt và loại bỏ các ký tự đặc biệt để đưa về dạng văn bản thuần túy.

1.2 Đối với file gốc JSON

Xây dựng công cụ căn chỉnh câu cho cặp ngôn ngữ Trung–Việt, phục vụ xây dựng tập dữ liệu huấn luyện dịch máy.

Hệ thống sử dụng biểu diễn câu do mô hình LaBSE sinh ra để đo độ giống nhau giữa câu tiếng Trung và câu tiếng Việt, sau đó áp dụng thuật toán quy hoạch động Vecalign để quyết định ghép cặp theo các kiểu 1–1, 1–2, 2–1, 2–2,...

Kết quả căn chỉnh được lưu ra nhiều định dạng (tệp song song hai cột, JSON chứa cả độ tương đồng, loại căn chỉnh, nguồn và id), thuận tiện cho hậu xử lý và thống kê.

2. Quy trình thực hiện đồ án

Đồ án thực hiện được chia thành các bước chính như sau:

2.1 Giai đoạn OCR

Bước 1: Chuẩn bị môi trường để thực thi (Tesseract, PaddleOCR)
Bước 2: Chuyển PDF sang ảnh: Sử dụng pdf2image (DPI 400) từ trang start_page đến

end_page.

Bước 3: OCR Layout từng trang: tạo các block, tiền xử lý, nhận diện ký tự là tiếng Trung hay tiếng Việt để đưa vào model tương ứng xử lý.

Bước 4: Trích xuất nội dung thư:

- Đối với tiếng Việt: tìm các dòng bắt đầu bằng cụm "chính mình số X"(cho phép sai lệch nhỏ), bỏ qua các dòng giống phiên âm tiếng Trung, chỉ giữ dòng có dấu tiếng Việt hoặc từ khóa quen thuộc như "là", "và", "của".

- Đối với tiếng Trung: cắt theo cụm tiêu đề "thư thứ X viết cho chính mình", ghép tiêu đề với nội dung phía sau. Ghép đôi nội dung hai ngôn ngữ theo số thư bằng cách ưu tiên dòng có số khớp, dùng hàng đợi xử lý dòng thừa, tạo bảng dữ liệu với cột số thư, nội dung Việt và nội dung Trung.

- Xử lý các trường hợp ngoại lệ : 1 bức thư (tiếng Việt hoặc tiếng Trung) bị chia ra 2 trang hoặc 1 cụm bức thư (gồm bức thư tiếng Việt và tiếng Trung) bị chia ra 2 trang → Đầu tiên sẽ gộp các kết quả OCR thô lại (từ 1 dict → list), sau đó ta sẽ trích xuất ra từng bức thư và align chúng lại theo từng cụm bức thư.

Bước 5: Xuất kết quả

Lưu văn bản thô từng trang vào file csv (số trang, văn bản Trung, văn bản Việt). Gộp toàn bộ, trích xuất thư sạch từ số 1 đến 500, đồng thời, xuất file json với các nội dung tương ứng để chuẩn bị cho bước Align tiếp theo.

2.2 Giai đoạn Align

Bước 1: Chuẩn bị môi trường và thư viện

Thiết lập môi trường Python 3.10, cài đặt và nạp các thư viện cần thiết bao gồm `vecalign` (thư viện chính thuật toán Alignment), `sentence-transformers` (để chạy mô hình LaBSE), `pysbd` (công cụ tách câu), cùng các thư viện xử lý dữ liệu như `numpy`, `pandas`. Mô hình LaBSE được tải lên thiết bị tương ứng như GPU hoặc CPU.

Bước 2: Tiền xử lý và tách câu

Văn bản tiếng Trung và tiếng Việt từ dữ liệu JSON đầu vào được chia thành các câu riêng lẻ sử dụng bộ tách câu `pysbd`. Sau khi tách, hệ thống loại bỏ khoảng trắng thừa và lọc nhiễu: chỉ giữ lại các câu có độ dài lớn hơn 1 ký tự (cho cả tiếng Trung và tiếng Việt) để loại bỏ các ký tự rác hoặc dấu câu đứng một mình.

Bước 3: Mã hóa vector (Embedding)

Sử dụng mô hình LaBSE để chuyển đổi danh sách các câu đơn và các tổ hợp câu (ghép N câu liền kề để xử lý trường hợp N-N) thành ma trận vector. Các vector này được chuẩn hóa (L2 normalization) để phục vụ cho việc tính toán cosine similarity.

Bước 4: Thuật toán căn chỉnh (Vecalign)

Áp dụng thuật toán Vecalign dựa trên Dynamic Programmin. Thuật toán duyệt qua ma trận vector để tìm đường dẫn căn chỉnh tối ưu giữa văn bản nguồn và đích. Chi phí căn chỉnh được tính dựa trên khoảng cách cosine (1 trừ đi độ tương đồng), cho phép các kiểu ghép cặp linh hoạt: 1-1, 1-N, N-1, hoặc N-N (tối đa ghép 4 câu).

Bước 5: Lọc kết quả và đảm bảo chất lượng

Sau khi tìm được đường căn chỉnh tối ưu, các cặp câu kết quả được trích xuất kèm theo điểm số căn chỉnh (score). Hệ thống áp dụng ngưỡng lọc để đảm bảo chất lượng: chỉ giữ lại các cặp câu có độ tương đồng quy đổi đạt từ 0.6 trở lên để đảm bảo chất lượng corpus được tốt nhất có thể.

Bước 6: Xuất dữ liệu

Kết quả cuối cùng được ghi ra file CSV. File bao gồm các trường thông tin: `src_id` (ID nguồn), `zh` (câu tiếng Trung) và `vi` (câu tiếng Việt) đã được căn chỉnh song ngữ.

3. Thống kê dữ liệu đã xử lý

Corpus	json1_1_212 json2_1_1163 pdf1_1_125	json1_213_424 json2_1164_2326 pdf1_126_250	json1_425_636 json2_2327_3489 pdf1_251_375	json1_637_848 json2_3490_4652 pdf1_376_500
Số câu	7625	8795	9333	9630
BLEU	0.37	0.32	0.35	0.35
chrF	0.55	0.51	0.53	0.54
Mean LaBSE	0.8557	0.8548	0.8497	0.8470
Med LaBSE	0.8501	0.8487	0.8420	0.8399
Min LaBSE	0.6458	0.6458	0.6458	0.6458
Số loại căn chỉnh	1-1: 7523 1-2: 101 1-3: 1	1-1: 8616 1-2: 179	1-1: 9147 1-2: 183 1-4: 1 2-2: 2	1-1: 9393 1-2: 237

3.1 Phương pháp đánh giá

Để kiểm chứng độ chính xác và chất lượng của các cặp câu song ngữ sau khi căn chỉnh, nhóm đã thực hiện quy trình đánh giá tự động dựa trên phương pháp so sánh chéo với phần kết quả tham chiếu. Quy trình cụ thể như sau:

1. **Lấy mẫu kiểm thử:** Từ mỗi tập dữ liệu con đã căn chỉnh, chúng tôi chọn ngẫu nhiên

50 cặp câu. Để đảm bảo việc đánh giá có ý nghĩa, nhóm em chỉ chọn những câu tiếng Trung có độ dài từ 10 ký tự trở lên mới được đưa vào tập kiểm thử.

2. **Cơ chế so sánh:** Với mỗi cặp câu $(ZH_{src}, VI_{aligned})$, nhóm em sử dụng Google Translate từ thư viện `googletrans` để dịch câu tiếng Trung gốc sang tiếng Việt, tạo ra một bản dịch tham chiếu giả định (VI_{MT}). Sau đó, hệ thống sẽ so sánh mức độ tương đồng giữa câu tiếng Việt trong bộ dữ liệu $(VI_{aligned})$ và câu do máy dịch (VI_{MT}).

3. **Metrics đánh giá:** Nhóm đã sử dụng kết hợp các chỉ số sau để đo lường chất lượng:

   - **Semantic Similarity:** Sử dụng mô hình LaBSE để mã hóa $VI_{aligned}$ và VI_{MT} thành vector, sau đó tính độ cosine similaritty. Chỉ số này để chúng em biết là 2 câu có cùng ý nghĩa hay không.
   - **BLEU & chrF:** Sử dụng thư viện `sacrebleu` để đo lường mức độ trùng khớp về từ và ký tự.
   - **ROUGE (1, 2, L):** Đánh giá độ bao phủ thông tin (Recall) giữa văn bản căn chỉnh và văn bản dịch máy.

4. **Tính toán phân phối Semantic:** Ngoài ra, chúng em cũng thực hiện tính toán phân phối điểm tương đồng ngữ nghĩa (Mean, Median, Std, Percentiles) trên toàn bộ bộ dữ liệu để thấy được chất lượng của nguyên cả ngữ liệu.

3.2 Nhận xét

Về độ tương đồng ngữ nghĩa (LaBSE) Với việc nhóm chỉ giữ lại những cặp câu có độ tương đồng trên 0.6, thì ngữ liệu của nhóm đã đạt được mức trung bình 0.85 và trung vị 0.84, ngữ liệu này vượt qua ngưỡng chấp nhận được thông thường (0.6 - 0.7). Ngay cả điểm thấp nhất cũng nằm ở mức an toàn (0.6458), chứng tỏ ngữ liệu không có cặp câu nào bị lệch nghĩa hoàn toàn.
Về chất lượng dịch (BLEU/chrF): Điểm BLEU đạt mức 0.32-0.37 kết hợp với điểm chrF 0.53-0.55, cho thấy ngữ liệu không chỉ đúng về ngữ nghĩa mà cấu trúc câu và từ vựng cũng có độ tự nhiên cao.

4. Ví dụ mẫu

4.1 Sử dụng vecalign để dóng hàng

Dùng một vài dòng dữ liệu nhỏ trong các file dữ liệu thô để thử:

File json1: ID 213

Dữ liệu trước khi xử lí là những đoạn văn dài, gồm văn bản tiếng Trung và tiếng Việt. Các câu chưa được tách, ghép để khớp với nhau.

Một vài dòng dữ liệu sau khi xử lý có kết quả như sau:

```
src_id,zh,vi
213,步骤,Các bước
213,方法 1,Phương pháp 1
213,方法 1 的 3:,Phương pháp 1 của 3:
213,使用网络浏览器,Sử dụng trình duyệt web
213,1. 在移动设备上打开微信。,1. Mở WeChat trên thiết bị di động.
213,如果还没有在平板或手机上登录微信，请先登录账户。,"Nếu bạn chưa đăng nhập trên điện thoại hay máy tính bảng, hãy tiến hành ngay."
213,4. 在电脑浏览器中前往 https://web.wechat.com 。,"4. Trên máy tính, bạn điều hướng đến địa chỉ https://web.wechat.com."
213,你可以使用任意浏览器打开页面，如Safari或Chrome浏览器。,Bạn có thể sử dụng bất kỳ trình duyệt web nào như Safari hay Chrome.
213,5. 用移动设备扫描屏幕上的二维码。,5. Sử dụng thiết bị di động của bạn để quét mã QR trên màn hình.
```

⇒ Dễ thấy rằng các cặp câu, cụm câu đã được dóng hàng, các câu có chất lượng thấp bị bỏ đi, sau khi đã tách câu tiếng Trung và tiếng Việt, vecalign sẽ ghép lại thành các cặp 1-1, 1-2, 2-1, 2-2,...

File json2: ID 1164 - 1166

Dữ liệu trước khi xử lí cũng gần giống với file json1, nhưng không còn là cả một đoạn văn dài mà được chia nhỏ thành một hoặc chỉ vài câu ngắn. Các câu vẫn chưa được dóng hàng, tách và ghép để khớp với nhau.

Một vài dòng dữ liệu sau khi xử lý có kết quả như sau:

File Edit View

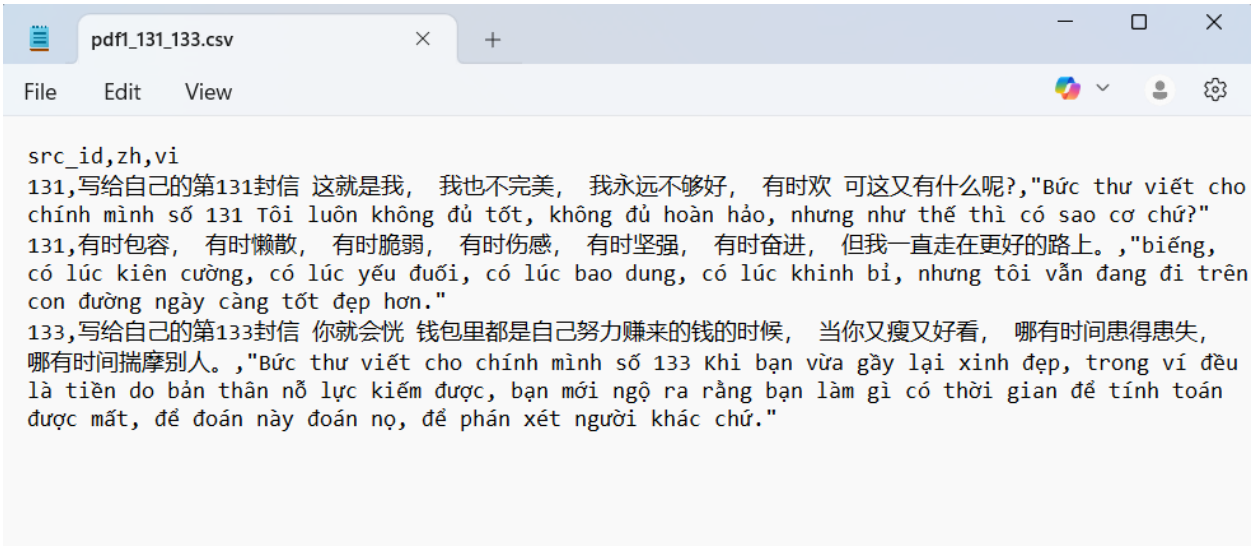
```
src_id,zh,vi
1164,其中一些是在斯坦福大学完成的，但也有一些是在其他地方完成的，即鼻腔微生物群特别擅长清除细菌，,"trong đó một số được thực hiện tại Stanford, nhưng cũng ở những nơi khác, rằng microbiome mũi đặc biệt giỏi trong việc loại bỏ vi khuẩn,"
1165,它可以预防某些类型的感染。,Nó có thể ngăn ngừa một số loại nhiễm trùng.
1166,不要用嘴呼吸。,Không hít thở bằng miệng.
1166,你的鼻子比嘴巴更能有效地过滤病毒和细菌。,Mũi của bạn là bộ lọc tốt hơn nhiều cho vi-rút và vi khuẩn so với miệng.
```

⇒ Các cặp câu đã được dóng hàng, trong 1 câu có thể có nhiều phần được tách ra và chỉ giữ lại các phần có chất lượng xử lý tốt. Câu cũng đã được ghép cặp giống như file json1.

File pdf1: ID 131 - 133

Dữ liệu trước khi xử lý là những đoạn song ngữ ngắn, chưa được dóng hàng, tách, ghép để 2 ngôn ngữ khớp với nhau.

Một vài dòng dữ liệu sau khi xử lý có kết quả như sau:

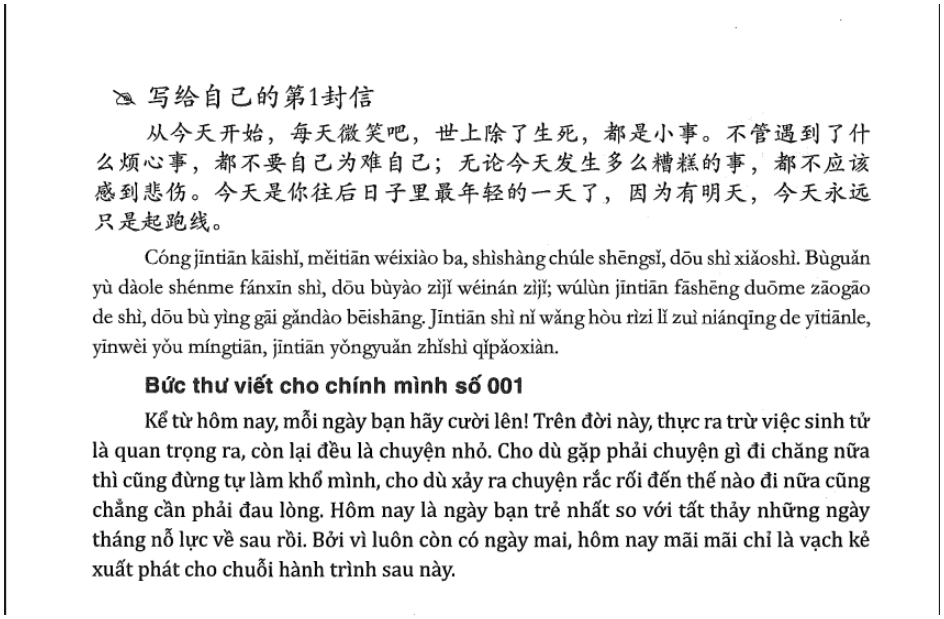


⇒ Các câu đã được xử lý dóng hàng, tách thành nhiều phần nếu cần, và ghép cho đúng nghĩa giữa 2 ngôn ngữ. Các câu (hoặc các phần trong câu) có chất lượng xử lý thấp đã bị bỏ đi.

4.2 Sử dụng PaddleOCR và Tesseract để OCR

Dữ liệu trước khi xử lý là ảnh được lưu dưới dạng file pdf.

Đây là 1 đoạn dữ liệu chưa được xử lý.



✍ 写给自己的第1封信

从今天开始，每天微笑吧，世上除了生死，都是小事。不管遇到了什么烦心事，都不要自己为难自己；无论今天发生多么糟糕的事，都不应该感到悲伤。今天是你往后日子里最年轻的一天了，因为有明天，今天永远只是起跑线。

Cóng jīntiān kāishǐ, měitiān wéixiào ba, shìshàng chúle shēngsǐ, dōu shì xiǎoshì. Bùguǎn yù dàole shénme fánxīn shì, dōu bùyào zìjǐ wéinán zìjǐ; wúlùn jīntiān fāshēng duōme zāogāo de shì, dōu bù yīng gāi gǎndào bēishāng. Jīntiān shì nǐ wǎng hòu rìzi lǐ zuì niánqīng de yītiānle, yīnwèi yǒu míngtiān, jīntiān yǒngyuǎn zhǐshì qǐpǎoxiàn.

Bức thư viết cho chính mình số 001

Kể từ hôm nay, mỗi ngày bạn hãy cười lên! Trên đời này, thực ra trừ việc sinh tử là quan trọng ra, còn lại đều là chuyện nhỏ. Cho dù gặp phải chuyện gì đi chăng nữa thì cũng đừng tự làm khổ mình, cho dù xảy ra chuyện rắc rối đến thế nào đi nữa cũng chẳng cần phải đau lòng. Hôm nay là ngày bạn trẻ nhất so với tất thảy những ngày tháng nỗ lực về sau rồi. Bởi vì luôn còn có ngày mai, hôm nay mãi mãi chỉ là vạch kẻ xuất phát cho chuỗi hành trình sau này.

Kết quả OCR (chưa được làm sạch) sẽ được lưu vào file csv như sau:

```
今天永远
只是起跑线。
写给自己的第2封信
和生生不息的希望。
不管前方的路有
总会有不期而遇的温暖，
人生，
多苦，
不管多么崎岖不平，都比站在原地更接近幸福。
只要走的方向正确，
写给自己的第3封信
时时刻刻要求自己做到百分之百的超出期望值。但
你是个要强的人，
不要太着急，
是苛求并不是个好现象，
请允许自己犯错。
你并不是天才，
你的努力，
时间都会帮你兑现。
","



+
Cóng jintian kaishǏ, měitiān wéixiào ba, shìshàng chúle shengsī, dōu shì xiǍoshì. Bùguǎn

yù dàole shénme fánxin shì, dōu bùyào zǏjš wéinán zǏjš; wúlùn Jintian fasheng duome zaogao
de shì, dou bù yìng gāi gǎndào beishäng, Jintian shì n văng hòu tìzi lǍ Zuì niánging de yItianle,

yinvèi yǒu míngtian, jIntian yóngyuǎn zhǐshì gǐpšoxiǎn.
Bức thư viết cho chính mình số 001
Kể từ hôm nay, mỗi ngày bạn hãy cười lên! Trên đời này, thực ra trừ việc sinh tử
là quan trọng ra, còn lại đều là chuyện nhỏ. Cho dù gặp phải chuyện gì đi chăng nữa
thì cũng đừng tự làm khổ mình, cho dù xảy ra chuyện rắc rối đến thế nào đi nữa cũng
-

chẳng cần phải đau lòng. Hôm nay là n,fDrz`JẤy bạn trẻ nhất so với tất thảy những I.lgÈ""5Ǐ}7J
tháng nỗ lực về sau rồi. Bởi vì luôn còn có ngày mai, hôm nay mãi mãi chỉ là vạch kẻ
xuất phát cho chuỗi hành trình sau này.
```

Phần chữ Hán, pinyin và tiếng Việt được tách ra sau đó bỏ phần pinyin, đánh số thứ tự và ghép cặp các dòng tiếng Trung và tiếng Việt ta được kết quả lưu trong file json như sau:

File Edit View


```
[
    {
        "src_id": 1,
        "src_lang": "Bức thư viết cho chính mình số 001 Kể từ hôm nay, mỗi ngày bạn hãy cười lên! Trên đời này, thực ra
trừ việc sinh tử thì cũng đừng tự làm khổ mình, cho dù xảy ra chuyện rắc rối đến thế nào đi nữa cũng tháng nỗ lực về sau
rồi. Bởi vì luôn còn có ngày mai, hôm nay mãi mãi chỉ là vạch kẻ xuất phát cho chuỗi hành trình sau này.",
        "tgt_lang": "写给自己的第1封信 不管遇到了什事，都不应该 世上除了生死， 都是小事。 每天微笑吧， 从今天开始， 无论今天发生多
么糟糕的事， 都不要自己为难自己； 么烦心事， 因为有明天， 感到悲伤。 今天是你往后日子里最年轻的一天了， 今天永远 只是起跑线。"
    },
    {
        "src_id": 2,
        "src_lang": "Bức thư viết cho chính mình số 002 Đời người luôn có những điều ấm áp không mong mà tới cùng với cả
những hi vọng không ngừng lớn lên. Cho dù con đường phía trước có biết bao khổ ải, chỉ cần hướng đi chính xác thì dù
trên đường đi đó có bao nhiêu chông gai, gập ghềnh cũng còn gần với bến bờ hạnh phúc hơn rất nhiều so với việc chỉ đứng
mãi ở vạch xuất phát.",
        "tgt_lang": "写给自己的第2封信 和生生不息的希望。 不管前方的路有 总会有不期而遇的温暖， 人生， 多苦， 不管多么崎岖不平，都
比站在原地更接近幸福。 只要走的方向正确，"
    },
```