

GVHD

PGS.TS.Nguyễn Lưu Thùy Ngân



TS.Nguyễn Văn Kiệt



ThS.Nguyễn Đức Vũ



ThS.Lưu Thanh Sơn



NHÓM 12

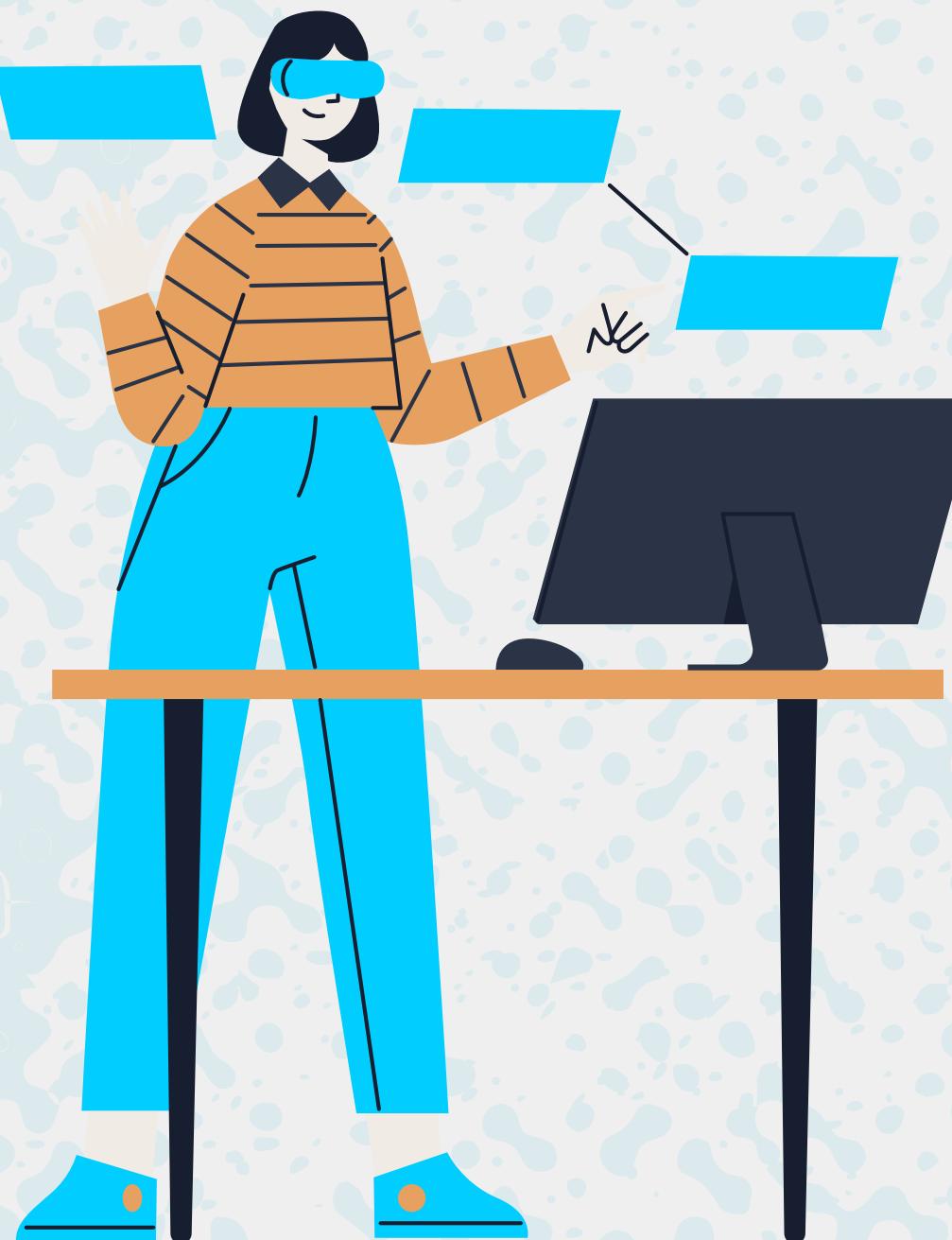
22521207 Trương Nhật Quang

22521548 Đỗ Tuấn Trực

22521277 Nguyễn Đức Tài

TABLE TO TEXT QA ON HITAB DATASET

INTRODUCTION



01

Table-to-text generation

Xử lý ngôn ngữ tự nhiên đối với thông tin từ Bảng

02

Question answering (QA)

Trả lời được các câu hỏi dựa trên dữ liệu trong bảng

03

**Bộ dữ liệu và mô hình dùng cho nghiên cứu
HiTab**

TAPAS và TAPEX

04

Đóng góp

**Khai thác giá trị từ dữ liệu bảng phân cấp trong các
tác vụ QA.**

HITAB

| Province | Farm operators | | | |
|---------------------------|-----------------------------|---------------|-------------------|------------------|
| | Immigrated between 2011 and | | Other immigrat | Non- immigrat |
| | China | United States | | |
| percent | | | | |
| Newfoundland and Labrador | 0.0 | 0.0 | 0 | 0.2 |
| Prince Edward Island | 0.0 | 10.7 | 0.6 | 0.7 |
| Nova Scotia | 0.0 | 0.0 | 1.7 | 1.7 |
| New Brunswick | 8.9 | 0.0 | 0.9 | 1.1 |
| Quebec | 0.0 | 4.5 | 6.6 | 16.3 |
| Ontario | 58.7 | 13.1 | 34.6 | 25.2 |
| Manitoba | 0.0 | 4.6 | 6.7 | 7.4 |
| Saskatchewan | 8.2 | 11.2 | 4.0 | 17.9 |
| Alberta | 0.0 | 17.0 | 18.1 | 21.4 |
| British Columbia | 24.3 | 39.0 | 26.8 | 8.1 |

| region or country of ancestry | women | | men | |
|----------------------------------|-----------------|-------------------|-----------------|-------------------|
| | employment rate | unemployment rate | employment rate | unemployment rate |
| | percent | | | |
| region of ancestry | | | | |
| caribbean and latin america | 78.3 | 8.0 | 77.3 | 10.8 |
| africa | 75.8 | 9.6 | 73.9 | 11.3 |
| other regions | 75.9 | 8.1 | 77.1 | 9.6 |
| country of ancestry | | | | |
| jamaica | 76.1 | 9.0 | 75.5 | 11.3 |
| haiti | 81.1 | 7.1 | 76.2 | 11.6 |
| trinidad and tobago | 78.6 | 7.4 | 80.4 | 10.0 |
| barbados | 81.9 | 5.6 | 82.6 | 8.3 |
| guyana | 78.0 | 7.5 | 79.5 | 11.1 |
| saint vincent and the grenadines | 78.8 | 6.6 | 77.3 | 13.3 |
| grenada | 76.7 | 10.6 | 81.5 | 8.8 |
| ghana | 75.5 | 10.3 | 79.4 | 8.6 |
| nigeria | 76.9 | 7.3 | 78.4 | 8.3 |
| united states | 72.0 | 9.9 | 76.9 | 9.8 |
| united kingdom | 79.1 | 7.6 | 76.7 | 9.6 |
| canada | 66.7 | 9.3 | 70.0 | 11.8 |



INPUT

- **Bảng dữ liệu:** Một bảng phân cấp có nhiều hàng và cột, chứa thông tin cần được truy vấn. Bảng có thể bao gồm các giá trị số, văn bản hoặc hỗn hợp cả hai
- **Câu hỏi** có thể yêu cầu tìm kiếm trực tiếp giá trị trong bảng hoặc yêu cầu suy luận từ nhiều phần của bảng.

OUTPUT

- **Câu trả lời:** Một câu trả lời ngắn gọn, chính xác dựa trên câu hỏi và bảng dữ liệu đầu vào. Câu trả lời có thể là một giá trị cụ thể (số hoặc văn bản) hoặc một kết quả tính toán từ dữ liệu.

HITAB



Tính phân cấp (Hierarchical Structure)



Đa dạng ngữ cảnh, câu hỏi, thông tin phức tạp



Hỗ trợ tốt cho QA và NLG

TAPAS AND TAPEX



Mô hình tiên tiến cho câu hỏi và trả lời dựa trên bảng dữ liệu



Khả năng suy luận và truy xuất thông tin

Bộ dữ liệu gốc



Bộ dữ liệu dạng JSON
gồm thông tin và file các
table tương ứng

Những cột cần thiết
trong nghiên cứu:

table_id, question, answer,
reference_cells_map



- Dataframe gồm các cột: (10672 dòng)
- **id, table_id, table_source, sentence_id, sub_sentence_id, sub_sentence, question, answer, aggregation, linked_cells, answer_formulas, reference_cells_map**
- Table dạng csv: (3597 tables)

Aggregation:

- 7536 (NONE)
- 3136 (KHÁC NONE)

Khác NONE:

pair-argmax, div, opposite, argmax, sum, pair-argmin, diff, argmin, topk-argmax,



Tiền xử lý

Encode

Trích xuất tọa độ tham
chiếu sẽ có dạng (x, y)

Table

- Điền giá trị trống
- Thêm 1 hàng tiêu đề vào cột dữ liệu

Chuẩn hóa câu trả lời:

- Danh sách cách nhau bằng dấu ",";
- số nguyên ≥ 1000 định dạng dấu phẩy ($1000.0 \rightarrow 1,392$)
- số thực không có phần thập phân
chuyển thành số nguyên ($51.0 \rightarrow 51$).

Tiền xử lý & Encode

Xử lý Tokenize cho Tapex:

Do Tapex chỉ giới hạn Tokenize ở 1024
8649 dòng nhỏ hơn hoặc bằng 1024

Mã hóa bảng và câu hỏi

Quá trình này chuyển đổi thông tin
thành các tensor như input_ids, đảm
bảo dữ liệu phù hợp với mô hình.

Mã hóa labels và padding:

Nhãn (labels) được mã hóa thành
tensor bằng tokenizer, sau đó thực
hiện padding để đồng nhất kích thước.

| <i>Trước</i> | <i>Sau</i> |
|---|----------------------------------|
| [51.0] | 51 |
| [8189.0] | 8,189 |
| ['cameroon', "cote d'ivoire", 'senegal'] | cameroon, cote d'ivoire, senegal |
| [4.581961] | 4.581961 |
| [298.07] | 298.07 |
| [77.1, 69.8] | 77.1, 69.8 |

TOKENIZE 1024



Độ đa dạng
Tổng 8649 dòng

Table and question

Label padding

- (none): 6254 (71.96%)
- (pair-argmax): 533 (6.13%)
- (div): 354 (4.07%)
- (opposite): 290 (3.34%)
- (argmax): 246 (2.83%)
- (sum): 189 (2.17%)
- (pair-argmin): 180 (2.07%)
- (diff): 161 (1.85%)
- (argmin): 81 (0.93%)
- (max): 59 (0.68%)
- (topk-argmax): 52 (0.60%)
- (greater_than): 47 (0.54%)
- (min): 34 (0.39%)
- input_ids shape: torch.Size([1, 1024])
- attention_mask shape: torch.Size([1, 1024])
- Mẫu 8648: label shape: torch.Size([3])
- Mẫu 8649: label shape: torch.Size([5])
- Kích thước padded labels: torch.Size([8649, 53])

TAPAS & TAPEX

Fine-tuned WTQ

| | TAPAS Fine-tuned WTQ | TAPEX Fine-tuned WTQ |
|--------------------------|---|--|
| Mục tiêu chính | Hỏi-đáp cơ bản trên bảng. | Truy vấn bảng với logic và phép toán phức tạp. |
| Dữ liệu đầu vào (encode) | Bảng + Câu hỏi | Bảng + Câu hỏi |
| Tiền xử lý dữ liệu | Kết hợp thông tin hàng, cột, giá trị vào embedding. | Mã hóa bảng và câu hỏi thành đầu vào BART. |
| Đầu ra | Văn bản câu trả lời kèm tọa độ của ô. | Văn bản câu trả lời chứa kết quả phép toán hoặc logic. |
| Ví dụ | Answer: 655 (4, 4) | Answer: 655 |
| Ứng dụng | Hỏi-đáp trực tiếp từ dữ liệu bảng. | Tổng hợp, tính toán, hoặc truy vấn logic phức tạp. |

Fine-tune TAPEX

CẤU HÌNH MÔ HÌNH:

- microsoft/tapex-base
- AdamW với tốc độ học $5e-5$, batch size là 8.

QUY TRÌNH HUÂN LUYỆN

- Train: 70%, Dev: 15%, Test: 15%.
- Forward pass: Mã hóa bảng và câu hỏi, tính toán lỗi (loss).
- Backward pass: Cập nhật trọng số bằng AdamW.

ĐÁNH GIÁ:

Tính loss trung bình trên tập phát triển (Dev) sau mỗi epoch.

KẾT QUẢ



**Kết quả Dự đoán trên tập dữ liệu
Test(1298 dòng).**

| Phương pháp | Tỷ lệ đúng (%) |
|------------------------|----------------|
| Tapex (fine-tuned WTQ) | 29.04 |
| Tapas (fine-tuned WTQ) | 24.50 |
| Tapex (tự fine-tuned) | 45.07 |

Phân tích lỗi

| Câu hỏi | Câu trả lời dự đoán | Câu trả lời đúng |
|--|---------------------------|--|
| List any cups that Queen of the South F.C play for? | League Cup, League Cup | Challenge Cup, Scottish Cup, League Cup |
| What is the decline of the rate of persons charged in criminal incidents in general has declined between 2004 and 2014? | -19.7 | 19.7 |
| How many percent of adults under correctional supervision in the provinces and territories in 2013/2014 were in custody? | 87 | 0.186393 |

Nguyên nhân:

- Dữ liệu huấn luyện chưa đầy đủ.
- Xử lý ngữ nghĩa chưa tốt.
- Giới hạn khả năng tính toán.

Conclusion

Kết quả đạt được:

- TAPEX tự fine-tune đạt 49.97%
- Tăng hiệu suất đáng kể khi áp dụng dữ liệu cụ thể.

Hạn chế:

- Lắp kết quả trong câu trả lời.
- Chưa tối ưu hóa suy luận ngữ nghĩa.
- Giới hạn trong xử lý phép toán.

Future Work

- Bổ sung dữ liệu huấn luyện đa dạng hơn.
- Tăng cường khả năng tính toán và hiểu ngữ nghĩa.

- Loại bỏ lỗi lặp và cải thiện suy luận ngữ cảnh.
- Tối ưu hóa cho câu trả lời



THANK YOU FOR YOUR ATTENTION