**ĐẠI HỌC QUỐC GIA TP. HCM**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO TỔNG KẾT

# ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ SINH VIÊN NĂM 2021

**Tên đề tài tiếng Việt:**

NGHIÊN CỨU PHƯƠNG PHÁP DỊCH MÁY VIỆT - TRUNG DỰA TRÊN CÁC MÔ HÌNH NGÔN NGỮ

**Tên đề tài tiếng Anh:**

AN APPROACH FOR VIETNAMESE - CHINESE MACHINE TRANSLATION BASED ON LANGUAGE MODELS

**Khoa/ Bộ môn:** Khoa Khoa học và Kỹ thuật Thông tin

**Thời gian thực hiện:** 06 tháng

**Cán bộ hướng dẫn:** Đặng Văn Thìn

**Tham gia thực hiện:**

| TT | Họ và tên, MSSV | Chịu trách nhiệm | Điện thoại | Email |
|---|---|---|---|---|
| 1. | Nguyễn Bá Đại<br>21521914 | Chủ nhiệm | 0948798843 | 21521914@gm.uit.edu.vn |
| 2. | Phan Cả Phát<br>21520389 | Tham gia | 0798385868 | 21520389@gm.uit.edu.vn |

# MỤC LỤC


MỤC LỤC 2

2

DANH MỤC HÌNH 3

DANH MỤC BẢNG 6

DANH MỤC TỪ VIẾT TẮT 8

NỘI DUNG 9

1. Tóm tắt 9
2. Giới thiệu 9
3. Nghiên cứu liên quan 11
4. Phương pháp tiếp cận 13

4.1.1. Dữ liệu 13

4.1.2. Chọn mô hình 14

4.1.3. Cài đặt mô hình 18

5. Thí nghiệm và phương pháp đo 19

5.1.1. Dữ liệu thí nghiệm 19

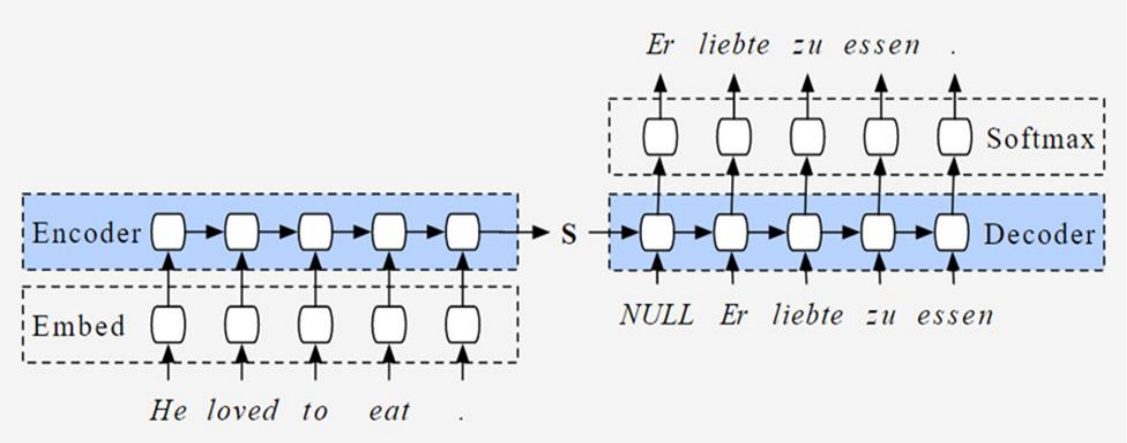
5.1.2 Phương pháp đo 20

5.1.3. Môi trường thực hiện thí nghiệm 21

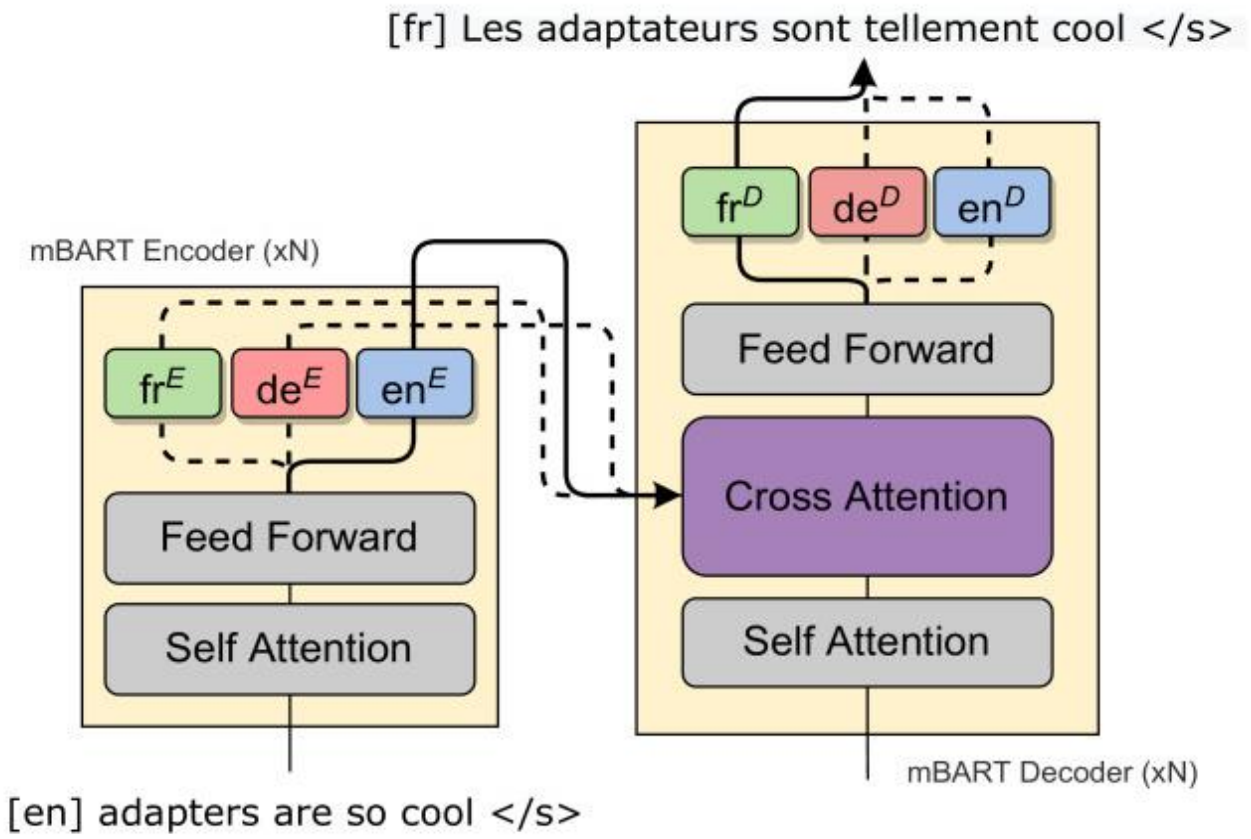
6. Kết quả và phân tích 22
7. Thảo luận 23
8. Kết luận 24
9. Hướng phát triển trong tương lai 25

TÀI LIỆU THAM KHẢO 26

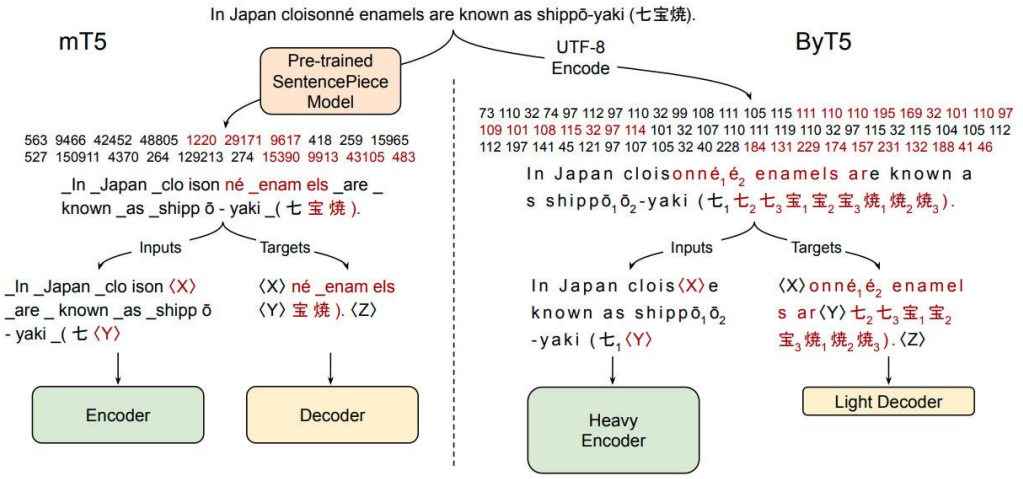
# DANH MỤC HÌNH



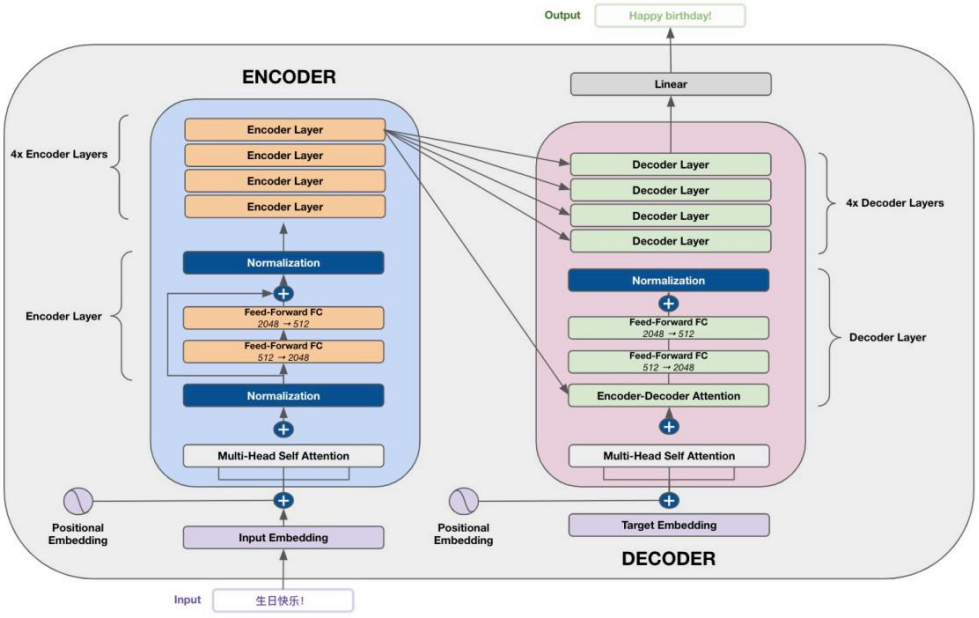

*Hình 1. Mô hình Seq2Seq cho bài toán dịch máy*



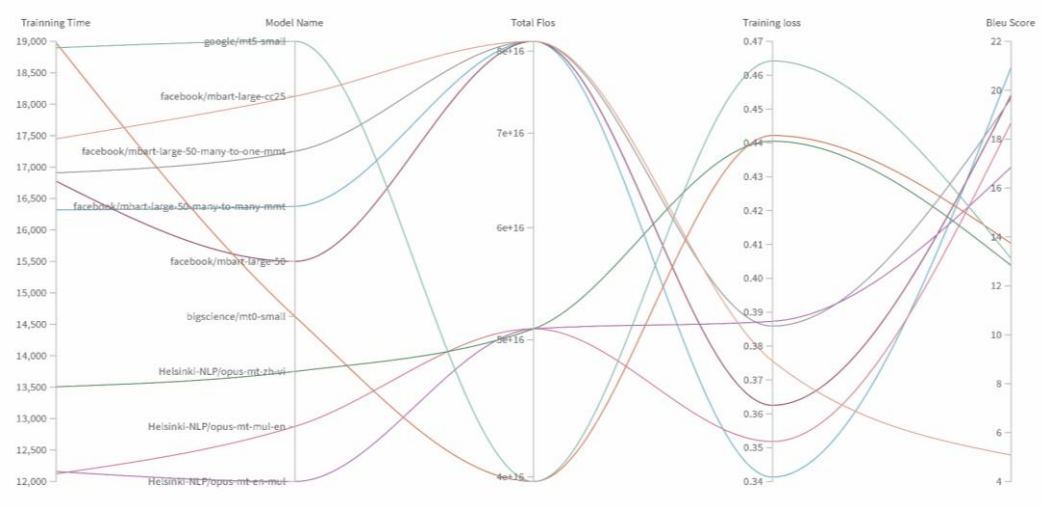

*Hình 2. Mô hình mBart*




*Hình 3. Mô hình MT5 và ByT5*




*Hình 4. Mô hình MarianMT*




*Hình 5. Mối quan hệ giữa các thông số*

# DANH MỤC BẢNG

| | Train | | Test | |
|---|---|---|---|---|
| Size | 300348 | | 1000 | |
| | vi | zh | vi | zh |
| Min Length | 11 | 1 | 35 | 8 |
| Max Length | 3093 | 713 | 402 | 172 |
| Avg Length | 83.04 | 22.85 | 95.00 | 28.83 |
| Min Token | 4 | 1 | 20 | 5 |
| Max Token | 1725 | 490 | 235 | 72 |
| Avg Token | 49.98 | 15.11 | 55.92 | 17.56 |

*Bảng 1. Chi tiết số liệu bộ dữ liệu*

| | Epochs | Learning rate | Weight decay | Batch size | Gradient accumulation steps | Optimization | FP16 |
|---|---|---|---|---|---|---|---|
| facebook/mbart-large-50 | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| facebook/mbart-large-50-many-to-one-mmt | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| facebook/mbart-large-50-many-to-many-mmt | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| facebook/mbart-large-cc25 | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| Helsinki-NLP/opus-mt-zh-vi | 5 | 1E-04 | 0.01 | 64 | 1 | AdamW | TRUE |
| Helsinki-NLP/opus-mt-en-mul | 5 | 1E-04 | 0.01 | 64 | 1 | AdamW | TRUE |
| Helsinki-NLP/opus-mt-mul-en | 5 | 1E-04 | 0.01 | 64 | 1 | AdamW | TRUE |
| google/byt5-small | 1 | 1E-04 | 0.01 | 16 | 1 | AdaFactor | FALSE |
| google/mt5-small | 1 | 1E-04 | 0.01 | 16 | 1 | AdaFactor | FALSE |
| bigscience/mt0-small | 1 | 1E-04 | 0.01 | 16 | 1 | AdaFactor | FALSE |

*Bảng 2. Chi tiết thông số chinh tỉnh mô hình*

| | vi | | zh | |
|---|---|---|---|---|
| | Raw | Pre-process | Raw | Pre-process |
| Size | 300348 | 300270 | 300348 | 300270 |
| Min Length | 11 | 13 | 1 | 2 |
| Max Length | 3093 | 1912 | 713 | 1197 |
| Avg Length | 83.04 | 85.45 | 22.85 | 36.07 |
| Min Token | 4 | 4 | 1 | 2 |
| Max Token | 1725 | 465 | 490 | 502 |
| Avg Token | 49.98 | 21.48 | 15.11 | 15.12 |

*Bảng 3. Thông số bộ dữ liệu Train sau khi tiền xử lý*

| Model Type | Model Name | Bleu Score |
|---|---|---|
| mBart | facebook/mbart-large-50 | 19.8046 |
| mBart | facebook/mbart-large-50-many-to-one-mmt | 19.6683 |
| **mBart** | **facebook/mbart-large-50-many-to-many-mmt** | **20.9445** |
| mBart | facebook/mbart-large-cc25 | 5.0788 |
| MarianMT | Helsinki-NLP/opus-mt-zh-vi | 12.8316 |
| MarianMT | Helsinki-NLP/opus-mt-en-mul | 16.8552 |
| MarianMT | Helsinki-NLP/opus-mt-mul-en | 18.6754 |
| mT5 | google/mt5-small | 13.1134 |
| mT0 | bigscience/mt0-small | 13.7330 |

*Bảng 4. Điểm BLEU của từng mô hình*
*(in nghiêng là tốt nhất cho từng loại mô hình,*
*in đậm là điểm cao nhất trên tất cả)*

| Model Type | Model Name | Trainning Loss |
|---|---|---|
| mBart | facebook/mbart-large-50 | 0.3625 |
| mBart | facebook/mbart-large-50-many-to-one-mmt | 0.3859 |
| mBart | facebook/mbart-large-50-many-to-many-mmt | 0.3413 |
| mBart | facebook/mbart-large-cc25 | 0.3756 |
| MarianMT | Helsinki-NLP/opus-mt-zh-vi | 0.4405 |
| MarianMT | Helsinki-NLP/opus-mt-en-mul | 0.3873 |
| MarianMT | Helsinki-NLP/opus-mt-mul-en | 0.3518 |
| mT5 | google/mt5-small | 0.4642 |
| mT0 | bigscience/mt0-small | 0.4422 |

*Bảng 5. Loss của từng mô hình*

# DANH MỤC TỪ VIẾT TẮT

| Từ viết tắt | Diễn giải |
|---|---|
| NLP | Natural Language Processing |
| MT | Machine Translation |
| NMT | Neural Machine Translation |
| SMT | Statistical Machine Translation |
| LRL | Low-Resource Language |
| VLSP | Vietnamese Language and Speech Processsing |
| Seq2Seq | Sequence to Sequence |
| mT5 | multilingual Text-to-Text Transfer Transformer |
| ByT5 | Byte-to-Byte Text-to-Text Transfer Transformer |
| mBart | multilingual Bidirectional and Auto-Regressive Transformers |
| MarianMT | Marian Machine Translation |
| mul | multilingual |
| en | english |
| zh | chinese |
| vi | vietnamese |
| mmt | multilingual machine translation |
| BPEmb | Byte-Pair Encoding Embeddings |
| FP16 | Floating Point 16-bit |
| UTF-8 | Unicode Transformation Format 8-bit |
| BLEU | Bilingual Evaluation Understudy |
| RAM | Random Access Memory |
| vRAM | virtual Random Access Memory |
| CPU | Central Processing Unit |
| GPU | Graphics Processing Unit |

# NỘI DUNG

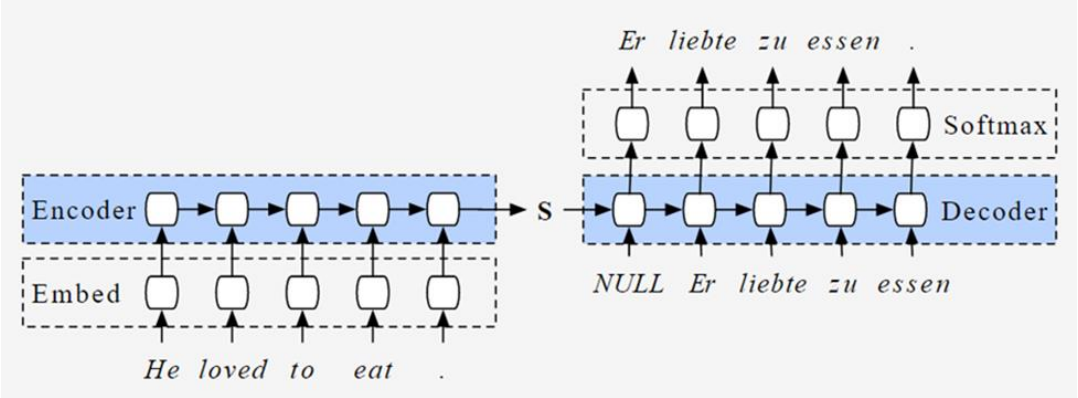
## 1. Tóm tắt

Nghiên cứu này tập trung vào phát triển một phương pháp hiệu quả cho dịch máy Việt-Trung với dữ liệu hạn chế. Chúng tôi đã kết hợp các kỹ thuật tiền xử lý cho văn bản tiếng Việt và tiếng Trung với các mô hình ngôn ngữ hiện đại như BART, Marian và T5. Ngoài ra, chúng tôi cũng đã tối ưu hóa việc sử dụng tài nguyên trong quá trình huấn luyện mô hình. Thông qua thử nghiệm và đánh giá, chúng tôi đã chứng minh tính hiệu quả của phương pháp này. Kết quả cho thấy mô hình mBart vượt trội hơn so với các mô hình khác như T5 và Marian. Ngoài ra, cấu trúc mô hình many-many cũng cho ra hiệu quả tốt hơn so với cấu trúc many-one, one-many và one-one. Để nâng cao hiệu suất, các nghiên cứu tiếp theo có thể tập trung vào mở rộng dữ liệu và khám phá các mô hình thay thế. Kết quả nghiên cứu này đóng góp vào sự phát triển của dịch máy ngôn ngữ hạn chế và giúp giải quyết những thách thức trong dịch máy Việt-Trung. Tổng thể, nghiên cứu này cải thiện dịch máy Việt-Trung thông qua sự sáng tạo trong kỹ thuật và cung cấp những hiểu biết quý giá cho sự phát triển trong tương lai.

## 2. Giới thiệu

Trung Quốc đang là một trong những quốc gia đi đầu trên thế giới về mọi mặt. Đi kèm theo đó là sự phát triển về các mặt hàng, sản phẩm liên quan đến nhiều lĩnh vực khác nhau. Với sự phát triển mạnh mẽ của công nghệ, Internet, các trang mạng, thương mại điện tử của Trung Quốc ngày càng nhiều. Nhu cầu tiếp cận dịch vụ của Trung Quốc cũng ngày một tăng lên. Đối với người sử dụng, họ cần một công cụ có thể chuyển từ ngôn ngữ Trung Quốc sang Việt Nam để có thể dễ dàng tiếp cận một cách dễ dàng hơn. Mặt khác, với sự phát triển của công nghệ thông tin và đa dạng các ngôn ngữ khác nhau trên thế giới như hiện nay thì bài toán dịch máy (Machine Translation - MT) là một trong những chủ đề nghiên cứu nổi bật trong lĩnh vực xử lý ngôn ngữ Tự nhiên (NLP). Mục tiêu của bài toán dịch máy là xây dựng một hệ thống hay chương trình máy tính cho phép chúng ta tạo ra một bản dịch chính xác giữa các ngôn ngữ khác nhau. Tuy nhiên để tạo ra một chương trình dịch máy hiệu quả và phù hợp cần nhiều lượng tài nguyên và dữ liệu huấn luyện để áp dụng các mô hình máy học.

Trong các năm trở lại đây, các tiếp cận giải quyết các bài toán NLP dựa trên hướng tiếp cận tinh chỉnh các mô hình ngôn ngữ đã mang lại hiệu quả và cải tiến trong các bài toán dịch máy và các bài toán nói riêng khác. Các phương pháp này khắc phục các bài toán hạn chế của việc phụ thuộc và mạng nơ ron truyền thống. Hầu hết các mô hình tiếp cận dựa trên các kiến trúc mô hình chuỗi (Sequence2Sequence – Seq2Seq) [1]. Cách tiếp cận Seq2Seq cho phép chúng ta biến đổi từ một chuỗi đầu vào thành một chuỗi đầu ra tùy thuộc vào bài toán mong muốn. Hình 1 mô tả tổng quan kiến trúc mô hình Seq2Seq chuẩn dựa trên kiến trúc Encoder-Decoder.




*Hình 1. Mô hình Seq2Seq cho bài toán dịch máy*

Đối với tiếng Việt, bài toán dịch máy cũng đang rất được quan tâm trong giới cộng đồng nghiên cứu và cộng đồng công nghiệp. Tuy nhiên, các nghiên cứu về tiếng Việt còn khá ít và hầu hết chủ yếu tập trung vào hệ dịch Anh - Việt [2, 3, 4].

Còn đối với các nghiên cứu dịch máy của ngôn ngữ Việt-Trung còn khá ít vì vướng phải sự thiếu hụt dữ liệu huấn luyện. Hiện này, các hệ thống mở dịch tương đối chính xác khi dịch từ Tiếng Anh sang các ngôn ngữ khác, tuy nhiên khi dịch từ Tiếng Việt sang Tiếng Trung còn phải sử dụng Tiếng Anh làm trung gian nên là chất lượng của bản dịch còn thiếu chính xác.

Chính vì lý do đó, trong đề này, chúng tôi tập trung nghiên cứu áp dụng các phương pháp tiền xử lý văn bản trong tiếng Việt và tiếng Trung kết hợp với sự hiệu quả của của các mô hình ngôn ngữ hiện đại như BART [5], T5 [6] để xây dựng một phương pháp hiệu quả cho bài toán dịch máy Việt – Trung trong trường hợp hạn chế dữ liệu. Kết quả nghiên cứu của đề tài là bước tiền đề giúp nhóm nghiên cứu mở rộng hướng phát triển trong tương lai về lĩnh vực này.

Hiện tại, chúng tôi đang sử dụng Bộ dữ liệu thử nghiệm trong đề tài, nhóm sinh viên dự kiến sử dụng bộ dữ liệu VLSP 2022 được công bố cho các đội thi tham gia giải quyết bài toán dịch máy. Bên cạnh bộ dữ liệu mà ban tổ chức đưa ra, chúng tôi sử dụng thang đo BLEU (Bilingual Evaluation Understudy) [7] để đánh giá kết quả của mô hình. Sau khi đánh giá kết quả chúng tôi thực hiện phân tích kết quả, tìm hiểu lý do tại sao mô hình vẫn còn chưa phiên dịch tốt 1 số câu từ đó tìm hướng giải quyết để cải thiện hiệu suất trong tương lai.

Thách thức lớn nhất của nhóm nghiên cứu trong việc thực hiện đề tài bây giờ chính là sự khan hiếm của bộ dữ liệu Việt - Trung cũng như Trung - Việt. Sau khi nghiên cứu một số chương trình liên quan, chúng tôi nhận thấy rằng một số tác giả đã nghiên cứu bài toán này nhưng đa phần là thực hiện trên cặp Việt - Anh hoặc những cặp từ khác và hiện tại bộ dữ liệu Việt - Trung vẫn còn khá

khan hiếm. Điều này khiến cho chúng tôi gặp khó khăn trong việc tìm kiếm dữ liệu cũng như gặp nhiều khó khăn trong việc cải thiện hiệu suất của máy dịch.

Tổng kết lại, bài báo này tìm hiểu và nghiên cứu về phương pháp phân tích khảo sát các mô hình cho bài toán dịch máy giữa hai ngôn ngữ Việt – Trung cũng như Trung – Việt dựa trên kho dữ liệu có sẵn VLSP2022 [8].

## 3. Nghiên cứu liên quan

Vào năm 2016, nhóm nghiên cứu của Phuoc Tran, Dien Dinh và Hien T. Nguyen đã đề xuất một phương pháp mới để dịch từ tiếng Trung Quốc sang tiếng Việt. Phương pháp này kết hợp các ưu điểm của dịch ở cấp độ ký tự và cấp độ từ. Đặc biệt, nó sử dụng một phương pháp kết hợp số liệu thống kê và quy tắc để dịch ở cấp độ từ, đồng thời sử dụng dịch thống kê ở cấp độ ký tự. Kết quả thử nghiệm đã cho thấy phương pháp này đã cải thiện hiệu suất dịch máy so với việc dịch chỉ ở cấp độ ký tự hoặc từ [9]. Phương pháp đề xuất bởi nhóm nghiên cứu đã giải quyết vấn đề dịch giữa tiếng Trung và tiếng Việt, hai ngôn ngữ không sử dụng khoảng trắng để phân tách từ. Bằng cách kết hợp cả dịch từ và dịch ký tự, phương pháp này đưa ra một giải pháp linh hoạt và hiệu quả cho dịch máy giữa hai ngôn ngữ này. Đáng chú ý, việc sử dụng số liệu thống kê và quy tắc để dịch ở cấp độ từ giúp cải thiện chất lượng dịch máy, đồng thời giảm số lượng từ chưa biết được tạo ra trong quá trình dịch. Điều này đặc biệt quan trọng trong việc dịch giữa các cặp ngôn ngữ ít tài nguyên như tiếng Trung và tiếng Việt. Kết quả của các thử nghiệm đã chứng minh rằng phương pháp đề xuất đã tăng cường hiệu suất dịch máy so với việc dịch chỉ ở cấp độ ký tự hoặc từ. Điều này cho thấy tính khả thi và hiệu quả của phương pháp kết hợp này trong việc dịch giữa tiếng Trung và tiếng Việt, đặc biệt là trong cặp ngôn ngữ có tài nguyên thấp như vậy.

Đầu năm 2020, Hongzheng Li và Heyan Huang đã tiến hành một nghiên cứu để khám phá tác động của dịch ngược (BT) đối với các bản dịch giữa các cặp ngôn ngữ châu Á, đặc biệt là giữa tiếng Trung Quốc và tiếng Việt, với tài nguyên rất thấp. Nghiên cứu này đã đánh giá và so sánh tác động của các kích thước dữ liệu tổng hợp khác nhau trên cả hai mô hình Dịch Máy Nơ-ron (NMT) và Dịch Máy Thống kê (SMT) cho cả tiếng Trung sang tiếng Việt và tiếng Việt sang tiếng Trung, sử dụng cả cấu hình dựa trên ký tự và từ [10]. Các kết quả của nghiên cứu này sẽ cung cấp thông tin quan trọng về tác động của BT đối với việc dịch ngôn ngữ châu Á và đặc biệt là giữa tiếng Trung Quốc và tiếng Việt. Nhóm nghiên cứu đã thu thập dữ liệu tổng hợp với các kích thước khác nhau và tiến hành đánh giá hiệu quả của chúng trên các mô hình dịch máy khác nhau. Việc tìm hiểu tác động của BT đối với các ngôn ngữ châu Á là rất cần thiết, vì các nghiên cứu trước đây thường tập trung vào các ngôn ngữ châu Âu. Việc mở rộng nghiên cứu này đến các ngôn ngữ châu Á sẽ giúp ta hiểu rõ hơn về tác động của BT trong các ngữ cảnh nguồn lực thấp. Nghiên cứu này cũng có ý nghĩa quan trọng vì Trung Quốc và Việt Nam là hai quốc gia có mối quan hệ gần gũi và đã có nhiều sự trao đổi trong nhiều lĩnh vực. Việc cải thiện chất lượng và hiệu suất của dịch máy giữa hai ngôn ngữ này sẽ đóng góp quan trọng cho sự phát triển của cả hai quốc gia. Nghiên cứu này cũng là một công trình

đầu tiên và toàn diện trong việc nghiên cứu tác động của BT đối với việc dịch máy giữa tiếng Trung Quốc và tiếng Việt. Những kết quả và nhận định mới được rút ra từ nghiên cứu này sẽ cung cấp thông tin quan trọng và có ích để hiểu sâu hơn về BT.

Nghiên cứu gồm Franck Burlot và François Yvon đã sử dụng các mô hình dịch với hiệu suất khác nhau để tạo ra dữ liệu giả song song với chất lượng khác nhau bằng cách thực hiện dịch ngược, và từ đó đánh giá chất lượng của các bộ dữ liệu này [11]. Mục tiêu chính của nghiên cứu là khám phá tác động của chất lượng dữ liệu giả song ngữ lên hiệu suất của mô hình dịch và chứng minh mối liên hệ giữa chất lượng dữ liệu và hiệu suất. Bằng cách thực hiện quá trình dịch ngược, nghiên cứu đã tạo ra dữ liệu giả song ngữ với một loạt chất lượng khác nhau. Bằng việc so sánh hiệu suất của các mô hình dịch khi sử dụng các bộ dữ liệu giả song ngữ này, nghiên cứu đã có thể đánh giá sự ảnh hưởng của chất lượng dữ liệu lên khả năng cải thiện hiệu suất. Nghiên cứu đã chứng minh rằng chất lượng dữ liệu giả song ngữ đóng vai trò quan trọng trong việc tăng cường hiệu suất của mô hình dịch. Khi chất lượng dữ liệu giả tăng lên, mô hình dịch cũng có khả năng dịch tốt hơn và cải thiện đáng kể về hiệu suất. Nghiên cứu cung cấp một cái nhìn sâu sắc về việc tăng cường hiệu suất của mô hình dịch thông qua việc sử dụng dữ liệu giả song ngữ và mở ra cơ hội cho việc cải thiện quy trình dịch nâng cao chất lượng dịch vụ. Nghiên cứu này cung cấp một cái nhìn sâu sắc và chứng minh mối liên hệ giữa chất lượng dữ liệu giả song ngữ và hiệu suất của mô hình dịch.

Nhóm nghiên cứu gồm JIA Chengxun, LAI Hua, YU Zhengtao, WEN Yonghua, YU Zhiqiang đã tạo ra bộ dữ liệu giả song song Hán – Việt ở cả hai chiều thuận và ngược lại [12]. Mô hình ngôn ngữ của ngôn ngữ đích thu được bằng cách sử dụng một lượng lớn dữ liệu huấn luyện đơn ngữ được hợp nhất vào mô hình dịch máy thần kinh. Hiệu quả dự kiến của việc hợp nhất mô hình ngôn ngữ đích là tích hợp các đặc điểm ngôn ngữ vào việc tạo dữ liệu giả song song thông qua mô hình ngôn ngữ để giúp tạo câu đúng ngữ pháp. Các thử nghiệm trên các tác vụ dịch thuật tiếng Trung-Việt đã chỉ ra rằng so với các phương pháp dịch ngược thông thường, phương pháp dịch máy thần kinh Trung-Việt đã đạt được một cải thiện về giá trị đánh giá ngôn ngữ hai chiều (BLEU) là 1,41 điểm phần trăm. Điều này đã được đạt thông qua việc kết hợp dữ liệu giả song song được tạo bởi mô hình ngôn ngữ. Ngoài ngôn ngữ Trung Quốc ra, chúng tôi cũng xem xét những bảo khảo sát đối với những ngôn ngữ "nghèo tài nguyên".

Năm 2021, một bộ mã nguồn mở dịch máy Neural Machine Translation (NMT) đã được giới thiệu bởi Nguyen Hoang Qua và đồng nghiệp [13]. Bộ mã nguồn này đã nhằm mục tiêu tạo ra một khuôn khổ độc lập, đơn giản và nhất quán, nhằm hỗ trợ các nhiệm vụ dịch máy trên nhiều miền khác nhau. Bộ mã nguồn này đáp ứng nhu cầu ngày càng tăng của cộng đồng nghiên cứu và người dùng trong việc sử dụng các công cụ dịch máy hiệu quả và dễ dàng. Với sự phát triển mạnh mẽ của công nghệ mạng nơ-ron và ứng dụng rộng rãi của nó, các nhiệm vụ xử lý ngôn ngữ tự nhiên nói chung và nhiệm vụ dịch máy nói riêng đã được hưởng lợi từ kiến trúc mới này, mang lại sự tiến bộ đáng kể về chất lượng và

sự trôi chảy trong quá trình dịch. Trong bối cảnh đó, bộ mã nguồn mở NMT này đã tạo ra sự đột phá trong lĩnh vực dịch máy bằng cách tập trung vào việc cung cấp một khuôn khổ linh hoạt và dễ sử dụng cho các nhiệm vụ dịch máy trên nhiều miền khác nhau. Nhờ vào cấu trúc độc lập và nhất quán của bộ mã nguồn, người dùng có thể dễ dàng tùy chỉnh và mở rộng chức năng theo nhu cầu cụ thể của họ. Sự ra đời của bộ mã nguồn này đã mang lại lợi ích rõ rệt cho cả cộng đồng nghiên cứu. Bộ mã nguồn NMT này đã góp phần đáng kể vào sự phát triển và tiến bộ của lĩnh vực dịch máy, đồng thời đáp ứng nhu cầu ngày càng cao của người dùng trong việc nghiên cứu và triển khai các hệ thống dịch máy hiệu quả và linh hoạt trên nhiều miền khác nhau.

Nghiên cứu của Ranathunga và đồng nghiệp đã nhấn mạnh sự phát triển đáng kể và việc sử dụng rộng rãi của Dịch máy Thần kinh (NMT), đồng thời nhận thức về hiệu suất chưa tối ưu của các mô hình NMT trên các cặp ngôn ngữ có tài nguyên thấp so với các đối tác có tài nguyên cao, chủ yếu do thiếu ngữ liệu song song lớn [14]. Để giải quyết vấn đề này, bài báo cáo cung cấp một cái nhìn tổng quan về cảnh quan nghiên cứu LRL-NMT và đề xuất các khuyến nghị để nâng cao công cuộc nghiên cứu trong tương lai. Hơn nữa, báo cáo nhấn mạnh sự cần thiết của các chiến lược hiệu quả để vượt qua những thách thức mà nng trình và tập dữ liệu liên quan đã được khám phá trong lĩnh vực LRL-NMT, làm sáng tỏ về sự tiến bộ đã đạt được và tiềm năng để cải thiện hơn nữa. Bài báo cáo khảo sát này phân tích một cách có hệ thống các kỹ thuật hiện có được sử dụng trong LRL-NMT và đánh giá tính ứng dụng của chúng trong các tình huống thực tế. Nó xác định các yếu tố chính ảnh hưởng đến việc lựa chọn kỹ thuật NMT phù hợp cho các tập dữ liệu ngôn ngữ có tài nguyên thấp cụ thể, bao gồm kích thước và loại dữ liệu có sẵn, cũng như tài nguyên tính toán.Ngoài ra, báo cáo cũng xác định các xu hướng mới nổi và hướng phát triển trong nghiên cứu LRL-NMT, nhấn mạnh tầm quan trọng của việc tạo nguồn lực ngôn ngữ, công khai tài nguyên tính toán và các mô hình đã được huấn luyện, và thúc đẩy cộng đồng nghiên cứu cấp vùng. Những khuyến nghị này nhằm tạo điều kiện thuận lợi cho việc phát triển các kỹ thuật NMT hiệu quả phục vụ cho những thách thức và đặc điểm độc đáo của các ngôn ngữ có tài nguyên thấp.

## 4. Phương pháp tiếp cận

### 4.1.1. Dữ liệu

Để huấn luyện một mô hình dịch máy từ tiếng Việt sang tiếng Trung, chúng tôi đã sử dụng bộ dữ liệu được cung cấp bởi VLSP 2022 - Machine Translation [8]. Bộ dữ liệu này bao gồm các cặp câu song ngữ, tức là mỗi câu trong tiếng Việt tương ứng với một câu trong tiếng Trung. Các câu trong bộ dữ liệu được lưu trữ dưới dạng văn bản UTF-8, và mỗi cặp câu được căn chỉnh sao cho câu 1 trong tiếng Việt đứng trên một dòng, câu 2 trong tiếng Trung đứng trên một dòng, và cách nhau bởi một dòng trống.

Các câu không nhất thiết phải là các câu hoàn chỉnh, mà có thể là các cụm từ hoặc một phần của câu. Điều này cho phép bộ dữ liệu bao gồm đa dạng các loại câu và cấu trúc ngôn ngữ khác nhau.

| | Train | | Test | |
|---|---|---|---|---|
| Size | 300348 | | 1000 | |
| | vi | zh | vi | zh |
| Min Length | 11 | 1 | 35 | 8 |
| Max Length | 3093 | 713 | 402 | 172 |
| Avg Length | 83.04 | 22.85 | 95.00 | 28.83 |
| Min Token | 4 | 1 | 20 | 5 |
| Max Token | 1725 | 490 | 235 | 72 |
| Avg Token | 49.98 | 15.11 | 55.92 | 17.56 |

*Bảng 1. Chi tiết số liệu bộ dữ liệu*

Bộ dữ liệu đã được chia thành hai tập là train và test. Tập train được sử dụng để huấn luyện mô hình, trong khi tập test được sử dụng để đánh giá hiệu suất của mô hình sau khi huấn luyện. Chi tiết thông số bộ dữ liệu được liệt kê như Bảng 1.

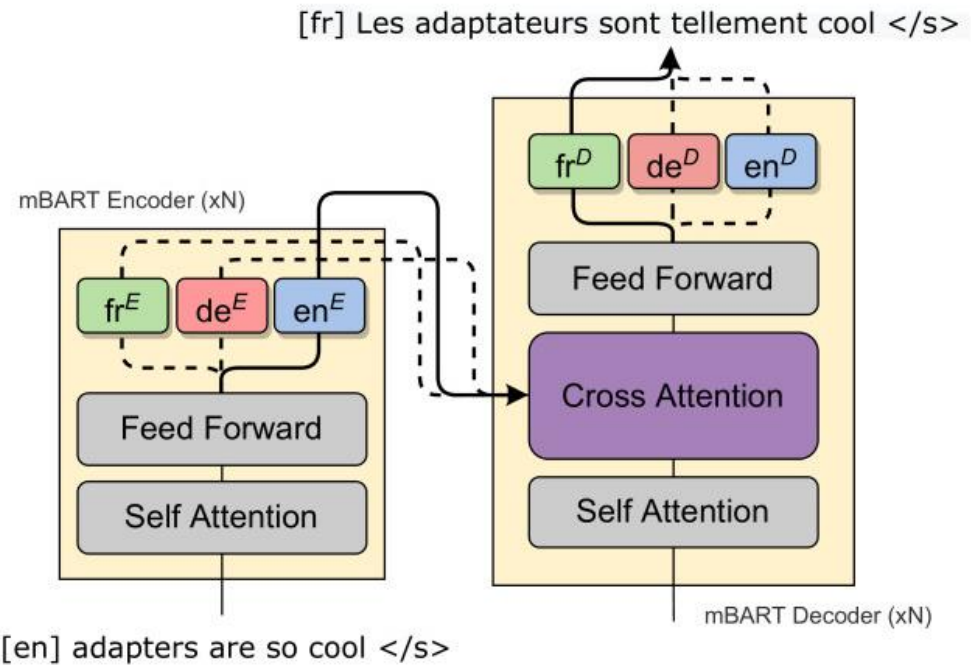
### 4.1.2. Chọn mô hình

Để giải quyết bài toán dịch máy từ tiếng Việt sang tiếng Trung, phương pháp tiếp cận phổ biến là sử dụng các mô hình dịch máy đa ngôn ngữ và kiến trúc transformer [15]. Các mô hình này được huấn luyện bằng cách sử dụng một lượng lớn dữ liệu dịch máy từ nhiều ngôn ngữ khác nhau, bao gồm cả tiếng Việt và tiếng Trung.

Trong phương pháp tiếp cận này, mô hình sử dụng mạng nơ-ron và học máy để học cách biểu diễn các cặp câu từ tiếng Việt sang tiếng Trung. Mô hình transformer, với khả năng xử lý ngôn ngữ tự nhiên mạnh mẽ, được áp dụng để xây dựng mô hình dịch máy đa ngôn ngữ.

Trước khi huấn luyện, mô hình được cung cấp một lượng lớn dữ liệu dịch máy tiếng Việt - tiếng Trung. Qua quá trình huấn luyện, mô hình học cách ánh xạ từ tiếng Việt sang tiếng Trung bằng cách tìm hiểu mối quan hệ ngữ nghĩa và cú pháp giữa các từ và câu trong hai ngôn ngữ.

Các mô hình dịch máy đa ngôn ngữ sử dụng cơ chế attention trong transformer để tập trung vào các từ quan trọng trong câu nguồn khi dịch sang câu đích. Điều này giúp cải thiện khả năng dịch máy bằng cách giữ nguyên ý nghĩa của câu trong quá trình dịch. Mô hình transformer cũng cho phép mô hình học được các đặc trưng ngôn ngữ chung giữa tiếng Việt và tiếng Trung, từ đó cải thiện khả năng dịch máy cho các cặp ngôn ngữ này.




*Hình 2. Mô hình mBart*

Một yếu tố quan trọng trong phương pháp tiếp cận này là việc sử dụng dữ liệu huấn luyện lớn. Dữ liệu dịch máy từ nhiều nguồn và ngôn ngữ khác nhau được sử dụng để huấn luyện mô hình. Việc sử dụng dữ liệu đa ngôn ngữ giúp mô hình học được các khía cạnh ngôn ngữ chung và tăng cường khả năng dịch máy cho các cặp ngôn ngữ mới, như tiếng Việt - tiếng Trung.

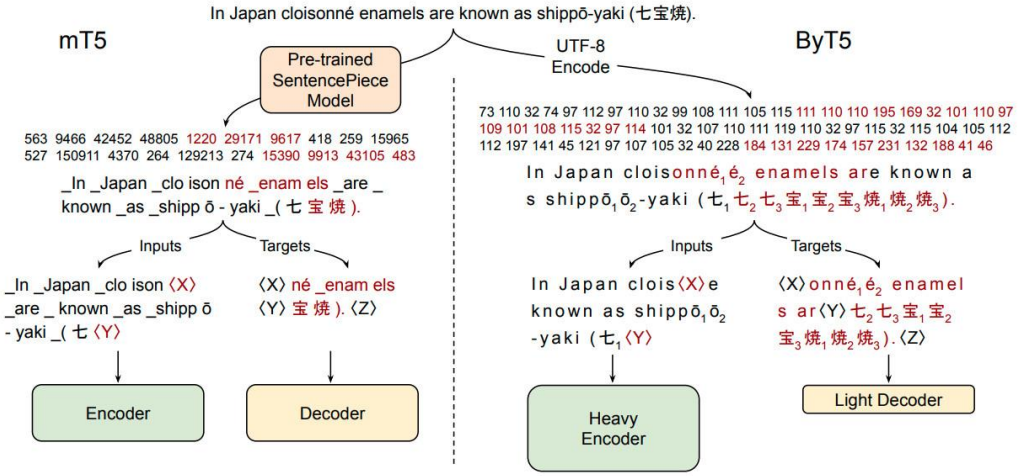
Ngoài ra, mô hình dịch máy đa ngôn ngữ cũng có khả năng thực hiện nhiều tác vụ ngôn ngữ khác nhau. Điều này có nghĩa là mô hình không chỉ dịch máy mà còn có thể thực hiện các tác vụ như nhận diện ngôn ngữ, tóm tắt văn bản, hay phân loại ngôn ngữ. Qua việc chia sẻ kiến thức và đặc trưng học được từ các tác vụ khác nhau, mô hình trở nên đa năng và có khả năng tổng hợp thông tin ngôn ngữ hiệu quả.

Các mô hình dịch máy như mBart [16], mT5 [17], ByT5 [18] và Marina [19] cũng áp dụng phương pháp tiếp cận đa nhiệm (multitask learning) để cải thiện dịch máy và thực hiện nhiều tác vụ ngôn ngữ.

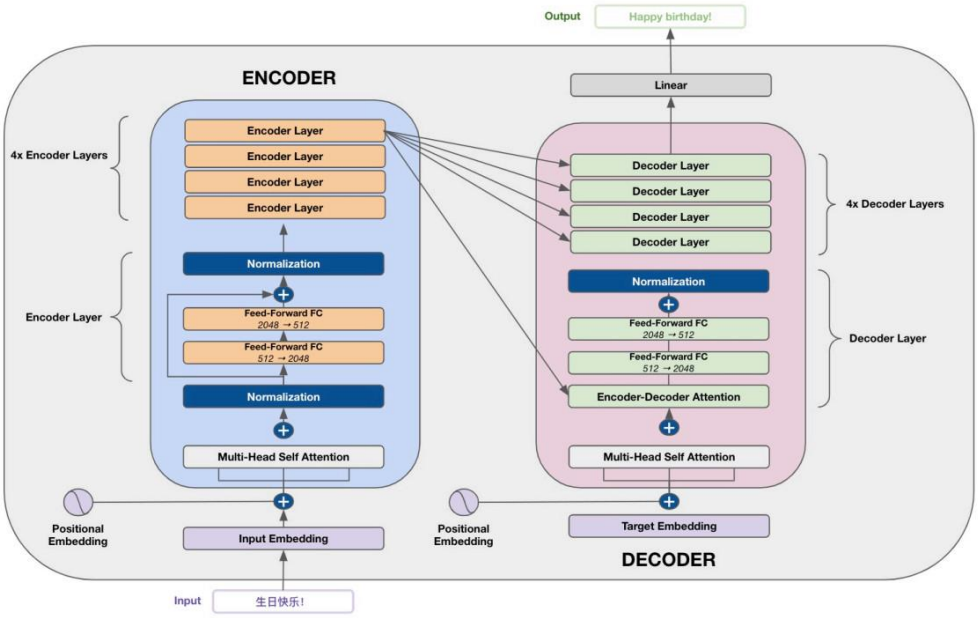
- mBart sử dụng phương pháp "dạy không gian mã hóa" (denoising autoencoder). Mô hình này học cách biểu diễn các cặp câu trong các ngôn ngữ khác nhau bằng cách dịch một câu ngôn ngữ đích thành một câu ngôn ngữ nguồn, sau đó đưa lại vào để phục hồi câu ngôn ngữ đích ban đầu. Qua quá trình này, mBart học được các đặc trưng ngôn ngữ chung và nâng cao khả năng dịch máy giữa các ngôn ngữ.

  mT5 sử dụng phương pháp "đồng nhất hóa mô hình" (unified model) để huấn luyện. Mô hình này được đào tạo trên một tập dữ liệu đa ngôn ngữ lớn và học cách dịch từ một ngôn ngữ sang ngôn ngữ khác. Đặc biệt, mT5 áp dụng phương pháp "đưa vào đầu vào, đưa ra đầu ra" (inputting and

outputting) để xử lý các tác vụ dịch máy đa ngôn ngữ, giúp mở rộng dễ dàng cho việc dịch giữa nhiều ngôn ngữ.




*Hình 3. Mô hình MT5 và ByT5*




*Hình 4. Mô hình MarianMT*

- ByT5 sử dụng phương pháp "nối tiếp" (chaining) để thực hiện các tác vụ ngôn ngữ. Mô hình này được huấn luyện bằng cách liên kết các tác vụ ngôn ngữ khác nhau thông qua việc đưa ra các đầu vào và đầu ra tương ứng. Quá trình này giúp ByT5 học cách xử lý và tổ chức thông tin ngôn ngữ một cách hiệu quả, từ đó nâng cao khả năng dịch máy và thực hiện các tác vụ ngôn ngữ khác.

- Marina sử dụng phương pháp "dịch chuyển" (transfer learning) để xử lý các tác vụ ngôn ngữ. Mô hình này được huấn luyện trên một tập dữ liệu đa ngôn ngữ và sau đó sử dụng kiến thức đã học để dịch máy và thực hiện các tác vụ ngôn ngữ khác. Phương pháp tiếp cận này cho phép Marina chia sẻ và sử dụng lại thông tin đã học từ các ngôn ngữ khác nhau, từ đó tăng cường khả năng dịch máy và đa nhiệm.

Chính vì những ưu điểm trên của mô hình dịch máy trên, chúng tôi đã lựa chọn kỹ lưỡng những mô hình sau:

- facebook/mbart-large-50: Mô hình này được chọn vì nó đã được huấn luyện trên nhiều ngôn ngữ và có khả năng dịch một ngôn ngữ sang nhiều ngôn ngữ khác nhau. Điều này sẽ hỗ trợ trong việc xử lý dịch máy giữa Tiếng Việt và Tiếng Trung.

- facebook/mbart-large-50-many-to-one-mmt: Mô hình này được tối ưu hóa cho tác vụ dịch máy many-to-one (nhiều-đến-một), nghĩa là dịch từ nhiều ngôn ngữ vào một ngôn ngữ cụ thể. Với việc dịch từ nhiều ngôn ngữ khác nhau vào Tiếng Trung, mô hình này có thể đáp ứng yêu cầu dịch máy từ Tiếng Việt sang Tiếng Trung.

- facebook/mbart-large-50-many-to-many-mmt: Mô hình này được tối ưu hóa cho tác vụ dịch máy many-to-many (nhiều-đến-nhiều), nghĩa là dịch từ nhiều ngôn ngữ vào nhiều ngôn ngữ khác nhau. Điều này sẽ hỗ trợ trong việc dịch giữa Tiếng Việt và Tiếng Trung, vì cả hai đều là ngôn ngữ đích và nguồn.

- facebook/mbart-large-cc25: Mô hình này được huấn luyện trên một tập dữ liệu lớn từ "Common Crawl", bao gồm dữ liệu từ hàng triệu trang web. Điều này giúp mô hình có khả năng xử lý nhiều ngôn ngữ và đáng tin cậy cho các tác vụ dịch máy đa ngôn ngữ.

- Helsinki-NLP/opus-mt-zh-vi: Mô hình này được huấn luyện đặc thù trên cặp ngôn ngữ Trung Quốc - Tiếng Việt. Với việc tối ưu hóa cho tác vụ dịch máy giữa hai ngôn ngữ này, mô hình này có khả năng cung cấp kết quả dịch chất lượng cao từ Tiếng Trung sang Tiếng Việt hay từ tiếng Việt qua tiếng Trung.

- Helsinki-NLP/opus-mt-en-mul: Mô hình này được huấn luyện trên cặp ngôn ngữ Tiếng Anh - Nhiều ngôn ngữ. Với khả năng dịch máy từ Tiếng Anh sang nhiều ngôn ngữ khác nhau, mô hình này có thể hữu ích trong việc dịch máy từ Tiếng Việt sang Tiếng Trung.

- Helsinki-NLP/opus-mt-mul-en: Mô hình này được huấn luyện cho tác vụ dịch máy từ nhiều ngôn ngữ sang Tiếng Anh. Với khả năng dịch từ nhiều ngôn ngữ khác nhau vào Tiếng Anh, mô hình này có thể hỗ trợ trong việc dịch máy từ Tiếng Việt sang tiếng Trung.

- google/byt5-small: Mô hình này được dựa trên kiến trúc ByT5 và có hiệu suất cao trên các tác vụ dịch máy đa ngôn ngữ. Với khả năng xử lý đa

ngôn ngữ và hiệu suất cao, mô hình này có thể hỗ trợ trong việc dịch máy giữa Tiếng Việt và Tiếng Trung.

- google/mt5-small: Mô hình này dựa trên kiến trúc mT5 và có khả năng dịch từ nhiều ngôn ngữ vào nhiều ngôn ngữ khác nhau. Với huấn luyện trên một tập dữ liệu đa ngôn ngữ lớn, mô hình này có thể đáp ứng yêu cầu dịch máy từ Tiếng Việt sang Tiếng Trung và ngược lại.

- bigscience/mt0-small: Mô hình này được huấn luyện bởi BigScience trên nhiều ngôn ngữ và có khả năng thực hiện dịch máy đa ngôn ngữ. Với việc xây dựng trên kiến trúc transformer, mô hình này có thể hỗ trợ trong việc dịch máy giữa Tiếng Việt và Tiếng Trung.

Tóm lại, để giải quyết bài toán dịch máy từ tiếng Việt sang tiếng Trung, phương pháp tiếp cận chung là sử dụng các mô hình dịch máy đa ngôn ngữ và kiến trúc transformer. Bằng cách sử dụng dữ liệu huấn luyện đa ngôn ngữ và áp dụng cơ chế attention trong transformer, mô hình có khả năng học và ánh xạ từ ngữ nghĩa và cú pháp của tiếng Việt sang tiếng Trung. Điều này cho phép cải thiện khả năng dịch máy và thực hiện các tác vụ ngôn ngữ đa ngôn ngữ. Các mô hình dịch máy

như mBart, mT5, ByT5 và Marina đã áp dụng thành công phương pháp tiếp cận này để cải thiện dịch máy và đa nhiệm trong việc xử lý ngôn ngữ.

### 4.1.3. Cài đặt mô hình

Theo Bảng 2, các mô hình dịch máy Marian như Helsinki-NLP/opus-mt-zh-vi, Helsinki-NLP/opus-mt-mul-en và Helsinki-NLP/opus-mt-mul-en đã được tinh chỉnh với các thông số sau đây:

- Epochs: 5
- Learning rate: 1e-4
- Weight decay: 0.01

| | Epochs | Learning rate | Weight decay | Batch size | Gradient accumulation steps | Optimization | FP16 |
|---|---|---|---|---|---|---|---|
| facebook/mbart-large-50 | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| facebook/mbart-large-50-many-to-one-mmt | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| facebook/mbart-large-50-many-to-many-mmt | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| facebook/mbart-large-cc25 | 1 | 1E-04 | 0.01 | 8 | 64 | AdamW | TRUE |
| Helsinki-NLP/opus-mt-zh-vi | 5 | 1E-04 | 0.01 | 64 | 1 | AdamW | TRUE |
| Helsinki-NLP/opus-mt-en-mul | 5 | 1E-04 | 0.01 | 64 | 1 | AdamW | TRUE |
| Helsinki-NLP/opus-mt-mul-en | 5 | 1E-04 | 0.01 | 64 | 1 | AdamW | TRUE |
| google/byt5-small | 1 | 1E-03 | 0.01 | 8 | 1 | AdaFactor | FALSE |
| google/mt5-small | 1 | 1E-03 | 0.01 | 8 | 1 | AdaFactor | FALSE |
| bigscience/mt0-small | 1 | 1E-03 | 0.01 | 8 | 1 | AdaFactor | FALSE |

*Bảng 2. Chi tiết thông số chinh tỉnh mô hình*

- Batch size: 64
- Gradient accumulation steps: 1

- Optimization: AdamW

Với các mô hình mBart như facebook/mbart-large-50, facebook /mbart-large-50-many-to-one-mmt, facebook/mbart-large-50-many-to-many- mmt, và facebook /mbart-large-cc25, do các mô hình này có kích thước lớn cũng như do sự thiếu hụt về tài nguyên và lượng lớn dữ liệu trên nên chúng tôi đã giảm Epochs, Batch size, tăng Gradient accumulation steps, và thay hàm Optimiztion thành AdaFactor để tối ưu RAM. Chi tiết thông số như sau:

- Epochs: 1
- Learning rate: 1e-4
- Weight decay; 0.01
- Batch size: 8
- Gradient accumulation steps: 64
- Optimization: AdaFactor

Cả 7 mô hình trên đều được huấn huyện với loại dữ liệu FP16 (floating point 16-bit) nhằm hiệu suất và giảm thời gian huấn luyện xuống, đồng thời cũng để giảm lượng RAM cần thiết để huấn luyện mô hình. Tuy nhiên, các mô hình còn lại như google/byt5-small, google/mt5-small, và bigscience/mt0-small đều không hỗ trợ loại dữ liệu FP16 này nên lượng RAM cần thiết khi huấn luyện đã tăng lên vì vậy nên thông số Batch đều được cài đặt ở 16 và epochs là 1. Chi tiết các thông số như sau;

- Epochs: 1
- Learning rate: 1e-3
- Weight decay; 0.01
- Batch size: 8
- Gradient accumulation steps: 1
- Optimization: AdamW

Tóm lại, các mô hình đã được tinh chỉnh về thông số nhằm có thể cải thiện hiệu suất cũng như tối ưu hoá tài nguyên cần thiết để huấn luyện mô hình.

## 5. Thí nghiệm và phương pháp đo

### 5.1.1. Dữ liệu thí nghiệm

Để có thể cải thiện hiệu suất khi huấn luyện mô hình, chúng tôi đã tiền xử lý dữ liệu tập Train theo các bước sau:

- Chuẩn hoá dấu câu: Chúng tôi sử dụng MosesPunctNormalizer từ thư viện sacremoses để chuẩn hoá các dấu câu. MosesPunctNormalizer sẽ chèn một khoảng trắng trước và sau các dấu câu để đồng nhất chúng trong văn bản. Điều này giúp đảm bảo sự nhất quán và dễ dàng trong việc xử lý ngôn ngữ tự nhiên, bởi vì các biểu thức dấu câu đã được chuẩn hóa và được xử lý một cách thống nhất.

| | vi | | zh | |
|---|---|---|---|---|
| | Raw | Pre-process | Raw | Pre-process |
| Size | 300348 | 300270 | 300348 | 300270 |
| Min Length | 11 | 13 | 1 | 2 |
| Max Length | 3093 | 1912 | 713 | 1197 |
| Avg Length | 83.04 | 85.45 | 22.85 | 36.07 |
| Min Token | 4 | 4 | 1 | 2 |
| Max Token | 1725 | 465 | 490 | 502 |
| Avg Token | 49.98 | 21.48 | 15.11 | 15.12 |

*Bảng 3. Thông số bộ dữ liệu Train sau khi tiền xử lý*

Input: "Xin chào! Bạn có khỏe không?"
Output: "Xin chào ! Bạn có khỏe không ?"

Input: "Thời tiết rất tuyệt vời!!"
Output: "Thời tiết rất tuyệt vời !!"

- Tokenization: Chúng tôi sử dụng BPEmb (Byte-Pair Encoding Embeddings) là một thư viện và công cụ mã nguồn mở được sử dụng để mã hóa từ vựng và nhúng từ sử dụng phương pháp mã hóa Byte-Pair Encoding (BPE) [20]. Nó cung cấp các bộ mã hóa từ vựng tiền huấn luyện trên các tập dữ liệu lớn, nhằm tạo ra các mã hóa nhúng từ vựng phổ biến và đa ngôn ngữ.

- Chúng tôi sử dụng BPemb model đã được huấn luyện với kích thước từ vựng là 50.000 và 200-dimensional embeddings.

Input: "Tôi vừa có kế hoạch thay họ biểu diễn ở bữa tiệc"
Output: "tôi vừa có kế hoạch thay họ biểu diễn ở bữa tiệc"

Input: "正打算让这群废物在派对上表演呢"
Output: "正 打算 让 这 群 废 物 在 派 对 上 表演 呢"

- Cleaning: Chúng tôi làm sạch data bằng cách sử dụng script “clean-corpus-n.perl” của thư viện Moses [21]. Chúng tôi giới hạn độ dài của câu và từ trong một câu, đồng thời xóa các ký tự không thể xuất cũng như các ký tự UTF-8 không hợp lệ. Chi tiết các thông số script được điều chỉnh như sau:

Độ dài tối đa của một từ: 50
Độ dài tối đa của 1 câu: 512

Sau khi tiền xử lý bộ dữ liệu thì bộ dữ liệu đã giảm được 0.0259% lượng dữ liệu. Chi tiết thông số bộ dữ liệu được thể hiện trong Bảng 3.

### 5.1.2 Phương pháp đo

Chúng tôi sử dụng phương pháp đo sacreBLEU [22], đó là một phương pháp

được sử dụng để đánh giá chất lượng của các bộ dịch máy dựa trên độ chính xác của việc dịch các câu hoặc văn bản so với các bản dịch tham chiếu (reference translations). SacreBLEU được xây dựng dựa trên các nguyên tắc của BLEU [7] (Bilingual Evaluation Understudy), một phương pháp phổ biến để đo lường chất lượng dịch máy. Tuy nhiên, sacreBLEU có một số cải tiến và điều chỉnh nhằm cải thiện tính khách quan và tin cậy của kết quả đánh giá. Phương pháp đo này bao gồm các bước sau:

- Tính điểm BLEU: SacreBLEU tính toán chỉ số BLEU dựa trên các n-gram chung giữa dự đoán và các bản dịch tham chiếu. BLEU đo lường độ chính xác của các từ và cụm từ, đồng thời đánh giá sự chính xác của cấu trúc câu. Các bước tính toán BLEU bao gồm tính toán số lượng n-gram chính xác, tính toán giá trị BP (brevity penalty) để đối phó với sự chênh lệch về độ dài câu, và tính toán BLEU tổng hợp dựa trên các thành phần trên.

- Tính điểm SacreBLEU: Sau khi tính toán BLEU, sacreBLEU áp dụng một số điều chỉnh và cải tiến để cải thiện tính khách quan và tin cậy của kết quả. Các điều chỉnh này bao gồm việc xử lý trường hợp đặc biệt (như các từ viết hoa), xử lý đặc biệt với các ký tự Unicode và các ký tự đặc biệt, và áp dụng phân phối chuẩn để tính toán khoảng tin cậy của BLEU.Cuối cùng, kết quả đo sacreBLEU được biểu diễn dưới dạng một giá trị BLEU tổng hợp, thường nằm trong khoảng từ 0 đến 100, thể hiện mức độ chính xác của bộ dịch máy so với các bản dịch tham chiếu. Giá trị càng cao, tức là bộ dịch máy càng chất lượng.

### 5.1.3. Môi trường thực hiện thí nghiệm

Chúng tôi sử dụng môi trường thực hiện thí nghiệm là Google Colab [23] của Google. Trong đó, CPU được cung cấp bao gồm Intel Xeon CPU với 2 vCPUs (virtual CPUs), RAM có dung lượng là 13GB, và GPU được sử dụng là NVIDIA Tesla T4 với 15GB VRAM.

Với tài nguyên này, chúng tôi có thể thực hiện các phép tính phức tạp và tăng tốc quá trình huấn luyện. Đặc biệt, GPU NVIDIA Tesla T4 cung cấp một lượng lớn bộ nhớ đồ họa VRAM, đảm bảo rằng chúng tôi có đủ dung lượng để xử lý mô hình lớn và dữ liệu đầu vào.

Việc sử dụng môi trường Google Colab và tài nguyên tính toán cao cấp này giúp chúng tôi thực hiện các thí nghiệm hiệu quả và đạt được kết quả mong đợi trong nghiên cứu của chúng tôi.

| Model Type | Model Name | Bleu Score |
|---|---|---|
| mBart | facebook/mbart-large-50 | 19.8046 |
| mBart | facebook/mbart-large-50-many-to-one-mmt | 19.6683 |
| ***mBart*** | ***facebook/mbart-large-50-many-to-many-mmt*** | ***20.9445*** |
| mBart | facebook/mbart-large-cc25 | 5.0788 |
| MarianMT | Helsinki-NLP/opus-mt-zh-vi | 12.8316 |
| MarianMT | Helsinki-NLP/opus-mt-en-mul | 16.8552 |
| *MarianMT* | *Helsinki-NLP/opus-mt-mul-en* | *18.6754* |
| mT5 | google/mt5-small | 13.1134 |
| *mT0* | *bigscience/mt0-small* | *13.7330* |

*Bảng 4. Điểm BLEU của từng mô hình*
*(in nghiêng là tốt nhất cho từng loại mô hình,*
*in đậm là điểm cao nhất trên tất cả)*

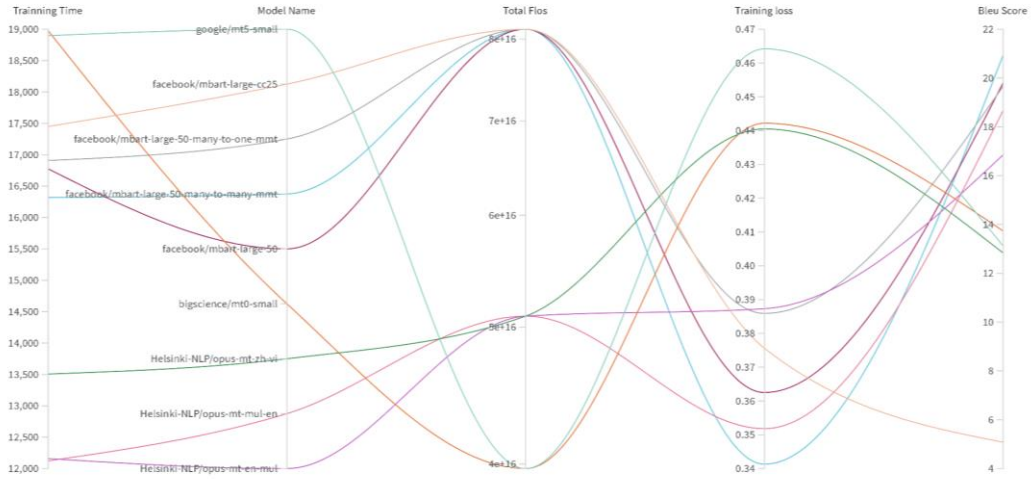
| Model Type | Model Name | Trainning Loss |
|---|---|---|
| mBart | facebook/mbart-large-50 | 0.3625 |
| mBart | facebook/mbart-large-50-many-to-one-mmt | 0.3859 |
| mBart | facebook/mbart-large-50-many-to-many-mmt | 0.3413 |
| mBart | facebook/mbart-large-cc25 | 0.3756 |
| MarianMT | Helsinki-NLP/opus-mt-zh-vi | 0.4405 |
| MarianMT | Helsinki-NLP/opus-mt-en-mul | 0.3873 |
| MarianMT | Helsinki-NLP/opus-mt-mul-en | 0.3518 |
| mT5 | google/mt5-small | 0.4642 |
| mT0 | bigscience/mt0-small | 0.4422 |

*Bảng 5. Loss của từng mô hình*

## 6. Kết quả và phân tích

Theo Bảng 4, ta có thể thấy với mô hình dạng mBart thì mô hình facebook /mbart-large-50-many-to-many-mmt đạt điểm số BLEU cao nhất với 20.9445 điểm, với mô hình dạng MarianMT thì cao nhất là mô hình Helsinki-NLP /opus-mt-mul-en với điểm số là 18.6754, còn mô hình dạng T5 như mT5, mT0 thì mô hình bigscience/mt0-small đạt điểm số cao nhất là 13.7330 điểm.Với bảng điểm trên, có thể dễ dàng nhận biết mô hình dạng mBart cho hiệu suất cao nhất tuy nhiên mô hình facebook/mbart-large-cc25 lại cho điểm số kém nhất trên toàn bộ là 5.0788 điểm.

Với Bảng 5 là hệ số loss của mỗi mô hình sau khi train, ta cũng có thể thấy rằng




*Hình 5. Mối quan hệ giữa các thông số*

tương tự Bảng 4, mô hình dạng mBart cho hệ số loss thấp nhất cho tất cả và mô hình facebook/mbart-large-50-many-to-many-mmt đồng thời cũng cho hệ số loss thấp nhất trong vùng mô hình và tổng quát là 0.3413. Với mô hình dạng MarianMT, thì mô hình Helsinki-NLP /opus-mt-mul-en cũng cho hệ số loss thấp nhất là 0.3518. Tương tự với mô hình dạng T5 như mT0, mT5 thì mô hình bigscience/mt0-small cho hệ số loss thấp nhất là 0.4422.

## 7. Thảo luận

Dựa vào Hình 5, ta có thể thấy mối liên hệ giữa các thông số với nhau. Cụ thể hơn, Với các mô hình mBart thì thời gian cần để huấn luyện là trung bình tuy nhiên lượng tài nguyên cần để huấn luyện lại lớn nhất. Lượng tài nguyên sử dụng lớn cũng phản ảnh cho việc hệ số loss thấp đồng thời kéo theo điểm BLEU cao. Mô hình dạng MarianMT cũng không ngoại lệ và tuân theo quy luật như vậy. Tuy nhiên với mô hình dạng T5 do không thể sử dụng FP16 nên thời gian huấn luyện đã tăng đáng kể so với lượng tài nguyên sử dụng là ít nhất, đồng thời hệ số loss là cao nhất. Qua đó ta có thể phản ánh rằng lượng tài nguyên tiêu thụ khi tính toán càng cao thì loss càng thấp từ đó dẫn tới hệ số điểm BLEU cao.

Chi tiết hơn tại Bảng 4, ta thấy các mô hình có input_target dạng multilingual cho điểm số cao hơn hẳn so với các biến thể khác trong cùng dạng mô hình. Điển hình như mô hình với mô hình dạng mBart thì mô hình facebook /mbart-large-50-many-to-many-mmt cho điểm số cao hơn từ 1.1398 đến 1.2762 tương ứng với mô hình facebook/mbart-large-50 và facebook /mbart-large-50-many-to-one-mmt. Tuy nhiên lại cao hơn mô hình facebook/mbart-large-cc25 nững 15.8656. Tương tự với mô hình dạng MarianMT, mô hình Helsinki-NLP/opus-mt-mul-en cho ra điểm số cao hơn 5.8437 và 1.8202 so với mô hình Helsinki-NLP/opus-mt-zh-vi và mô hình Helsinki-NLP/opus-mt-en-mul. Tuy nhiên ở hai mô hình dạng T5 là google /mt5-small và bigscience/mt0-small lại đều sử

dụng multilingual ở input_target nên hệ số loss không chênh lệch quá nhiều so với các dạng mô hình trên, cụ thể là 0.6196.

Hơn nữa, dựa vào số liệu Bảng 4 trên, ta cũng có thể thấy mô hình mBart về tổng quan cho ra kết quả cao vượt trội so với các mô hình dạng MarianMT hay T5. Thêm vào đó là với mô hình dạng many-to-many sẽ hiệu quả hơn mô hình dạng many-to-one như facebook/mbart-large-50-many-to-many-mmt và facebook /mbart-large-50-many-to-one-mmt với điểm số BLEU lần lượt là 20.944 và 19.6683. Tương tự mô hình many-to-one sẽ hiệu quả hơn one-to-many như Helsinki-NLP/opus-mt-mul-en và Helsinki-NLP /opus-mt-en-mul với điểm số BLEU lần lượt là 18.6754 và 16.8552.

Mặc dù mô hình dạng mBart cho hiệu suất cao hơn thảy so với các mô hình dạng T5 hay MarianMT nhưng mô hình facebook/mbart-large-cc25 lại cho điểm số thấp nhất là 5.0788, chênh lệch rất nhiều so với mô hình cùng loại cũng như các mô hình khác. Điều này có thể dễ hiểu do các mô hình khác đã được huấn luyện sắn cho tác vụ dịch máy trong khi đó facebook/mbart-large-cc25 thì không, từ đó dẫn tới mô hình này cho hiệu suất kém hẳn so với các mô hình khác.

Ở các mô hình dạng MarianMT, ta có thể thấy biên độ điểm BLEU của các mô hình khá cao từ 1.8202 đến 5.8437 điểm khi so với mô hình dạng mBart là từ 1.1398 tới 1.2762. Một lần nữa ta có thể thấy mô hình mBart hoàn toàn vượt trội so với các mô hình khác. Còn lý do tại sao mô hình MarianMT lại có biên độ điểm BLEU cao như vậy thì có lẽ là do sự khác nhau lớn dữ back-translate từ ngôn ngữ zh-vi sang vi-zh của mô hình Helsinki-NLP/opus-mt-zh-vi cũng như sự khác nhau về input_target là multilingual hay không, đồng thời cũng là do sự khác nhau về cấu trúc mô hình theo many-one, one-many, one-one. Qua đó, ta có thể nhận xét mức độ hiệu quả của các kiểu cấu trúc như sau:

- many-many > many-one > one-many > one-one

Thêm nữa dựa vào Bảng 2, ta có thể thấy rằng mặc dù mô hình dạng mBart được huấn luyện với epochs là 1, và gradient accumulation steps là 64. Việc đặt thồn số gradient accumulation steps là 64 thay vì 1 có thể dẫn tới việc hệ số loss cao từ đó giảm điểm số của mô hình. Tuy nhiên mô hình mBart vẫn cho thấy được hiệu quả xuất sắc mặc dù đã bị hạn chế về hiệu năng khi so với mô hình MarianMT đã được huấn luyện với epochs là 5 và gradient accumulation steps là 1.

## 8. Kết luận

Sau khi khảo sát các mô hình dạng mBart, MarianMT, và T5, nhóm chúng tôi đưa ra các kết luận sau:

1. Mô hình mBart cho hiệu suất vượt trội hơn hẳn so với các mô hình MarianMT, và T5 trong bài toán dịch máy.

2. Cấu trúc mô hình many-many cho hiệu quả tốt hơn rất nhiều so với các cấu trúc khác và chúng tôi cũng đưa ra các mức độ hiệu

quả của các cấu trúc như sau:

- many-many > many-one > one-many > one-one

Tóm lại, qua nghiên cứu khoa học này, nhóm chúng tôi đã khảo sát được độ hiệu quả của các mô hình, mối liên hệ giữa các thông số với độ hiệu quả của mô hình, các cấu trúc input_ouput hiệu quả của mô hình cũng như cách để tối ưu tài nguyên sử dụng khi huấn luyện mô hình. Tuy nhiên chúng tôi vẫn chưa thể khảo sát được mô hình google/byt5-small do sự thiếu hụt về tài nguyên nhưng dựa vào những nghiên cứu, khảo sát trên, chúng tôi tin rằng mô hình dạng mBart vẫn sẽ cho hiệu quả cao nhất và vượt trội so với các dạng mô hình khác.

## 9. Hướng phát triển trong tương lai

Dựa vào các kết luận trên, chúng tôi nhận thấy rằng để cải thiện hiệu suất dịch máy Việt - Trung và Trung - Việt, chúng tôi cần tăng cường lượng tài nguyên và dữ liệu huấn luyện, cũng như nghiên cứu các mô hình và phương pháp tiền xử lý văn bản hiện đại.

Trong tương lai, chúng tôi dự định cải thiện lượng tài nguyên sử dụng để có thể tiến hành khảo sát và nghiên cứu chi tiết hơn về các mô hình dịch máy khác như google/byt5-small, cũng như các mô hình lớn hơn như large, xl, xxl. Bằng cách sử dụng các mô hình có dung lượng lớn hơn, chúng tôi hy vọng có thể nắm bắt được sự phức tạp và sự tương quan giữa các ngôn ngữ Việt - Trung và Trung - Việt một cách tốt hơn.

Ngoài ra, chúng tôi cũng dự định tối ưu hóa các thông số khác trong mô hình như số lượng layer, kích thước embedding, và kích thước batch để đạt được kết quả tốt nhất có thể. Thông qua việc tối ưu hóa các thông số này, chúng tôi hy vọng mô hình sẽ có khả năng học và hiểu ngữ cảnh ngôn ngữ tốt hơn, dẫn đến kết quả dịch chính xác và tự nhiên hơn.

Bên cạnh đó, chúng tôi cũng định hướng nghiên cứu các phương pháp tiền xử lý văn bản hiện đại để tăng cường hiệu suất dịch máy. Điều này có thể bao gồm việc áp dụng các kỹ thuật như tách từ, phân đoạn câu, loại bỏ stop words, xử lý chuỗi ký tự đặc biệt và sử dụng các biểu đồ ngữ nghĩa (semantic graph) để cải thiện sự hiểu và chính xác của bản dịch.

Qua việc tăng cường lượng tài nguyên, nghiên cứu các mô hình và phương pháp tiền xử lý văn bản hiện đại, chúng tôi hy vọng có thể đạt được những cải tiến đáng kể trong bài toán dịch máy Việt - Trung và Trung - Việt. Điều này sẽ không chỉ hỗ trợ người dùng trong việc tiếp cận thông tin và giao tiếp giữa hai ngôn ngữ này một cách thuận tiện và chính xác hơn, mà còn đóng góp vào sự phát triển và hợp tác trong lĩnh vực ngôn ngữ và văn hóa giữa Việt Nam và Trung Quốc.

# TÀI LIỆU THAM KHẢO


[1]. Liu, T., Wang, K., Sha, L., Chang, B., & Sui, Z. (2017). Table-to-text Generation by Structure-aware Seq2seq Learning. AAAI Conference on Artificial Intelligence.

[2]. Doan, L., Nguyen, L.T., Tran, N.L., Hoang, T.Q., & Nguyen, D.Q. (2021). PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation. ArXiv, abs/2110.12199.

[3]. Nguyen, T.H., Phung, D., Nguyen, D.T., Tran, H.M., Luong, M., Vo, T., Bui, H.H., Phung, D., & Nguyen, D.Q. (2022). A Vietnamese-English Neural Machine Translation System. Interspeech.

[4]. Phan-Vu, H., Tran, V., Nguyen, V., Dang, H., & Do, P. (2018). Machine Translation between Vietnamese and English: An Empirical Study. ArXiv, abs/1810.12557.

[5]. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Annual Meeting of the Association for Computational Linguistics.

[6]. Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P.J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv, abs/1910.10683.

[7]. Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. Annual Meeting of the Association for Computational Linguistics.

[8]. Nguyen, V.V., Nguyen, H., Le, H.T., Nguyen, T.P., Bui, T.V., Pham, L.N., Phan, A., Nguyen, C., Tran, V., & Tran, A. (2022). KC4MT: A High-Quality Corpus for Multilingual Machine Translation. International Conference on Language Resources and Evaluation.

[9]. Tran, P., Dien, D., & Nguyen, H.T. (2016). A Character Level Based and Word Level Based Approach for Chinese-Vietnamese Machine Translation. Computational Intelligence and Neuroscience, 2016.

[10]. Li, H., & Huang, H. (2020). Evaluating Low-Resource Machine Translation between Chinese and Vietnamese with Back-Translation. ArXiv, abs/2003.02197.

[11]. Burlot, F., & Yvon, F. (2018). Using Monolingual Data in Neural Machine Translation: a Systematic Study. ArXiv, abs/1903.11437.

[12]. Wen, Y., Guo, J., Yu, Z., & Yu, Z. (2023). Chinese-Vietnamese Pseudo-

Parallel Sentences Extraction Based on Image Information Fusion. Inf., 14, 298.

[13]. Quan, N.H., Dat, N.T., Hoang, M., Vinh, N.V., Vinh, N.T., Thai, N.P., & Viet, T.H. (2021). ViNMT: Neural Machine Translation Tookit. ArXiv, abs/2112.15272.

[14]. Ranathunga, S., Lee, E.A., Skenduli, M.P., Shekhar, R., Alam, M., & Kaur, R. (2021). Neural Machine Translation for Low-resource Languages: A Survey. ACM Computing Surveys, 55, 1 - 37.

[15]. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv, abs/1910.03771.

[16]. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics, 8, 726-742.

[17]. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. North American Chapter of the Association for Computational Linguistics.

[18]. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2021). ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. Transactions of the Association for Computational Linguistics, 10, 291-306.

[19]. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H.T., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A., Bogoychev, N., Martins, A.F., & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. ArXiv, abs/1804.00344.

[20]. Heinzerling, B., & Strube, M. (2017). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. ArXiv, abs/1710.02187.

[21]. Koehn, P., Hoang, H.T., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics.

[22]. Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. Conference on Machine Translation.

[23]. Carneiro, T., Medeiros Da Nóbrega, R.V., Nepomuceno, T., Bian, G., de Albuquerque, V.H., & Filho, P. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. IEEE Access, 6, 61677-61685.

| **Cơ quan Chủ trì** | **Chủ nhiệm đề tài** |
| --- | --- |
| *(ký, họ và tên, đóng dấu)* | *(ký, họ và tên)* |
| | Nguyễn Bá Đại |