| LOGO DHCNTT -hinh.jpg | ĐẠI HỌC QUỐC GIA TP. HCM  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN** | Ngày nhận hồ sơ |  |
| --- | --- | --- | --- |
| *(Do CQ quản lý ghi)* | |

**THUYẾT MINH**

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2022

# **THÔNG TIN CHUNG**

**A1. Tên đề tài**

- Tên tiếng Việt (IN HOA):

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

- Tên tiếng Anh (IN HOA):

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

## **A**2**. Thời gian thực hiện**

**..06..** tháng (kể từ khi được duyệt).

## **A**3**. Tổng kinh phí**

*(Lưu ý tính nhất quán giữa mục này và mục B8. Tổng hợp kinh phí đề nghị cấp)*

Tổng kinh phí: …**6**.. triệu đồng,gồm

* Kinh phí từ Trường Đại học Công nghệ Thông tin: ..**6**.. triệu đồng

## **A**4**. Chủ nhiệm**

Họ và tên: **Nguyễn Bá Đại** **.**

Ngày, tháng, năm sinh: 04/03/2003 .

Giới tính (Nam/Nữ): Nam . .

Số CMND: 038203029792; Ngày cấp: 31/5/2021 **.**

Nơi cấp: Cục trưởng Cục Cảnh sát quản lý hành chính về trật tự xã hội .

Mã số sinh viên: 21521914 **.**

Số điện thoại liên lạc: 0948798843 **.**

Đơn vị (Khoa): KH&KTTT **.**

Số tài khoản: 50210000553025 .

Ngân hàng: BIDV .

## **A**5**.** Thành viên đề tài

| **TT** | **Họ tên** | **MSSV** | **Khoa** |
| --- | --- | --- | --- |
| 1 | Nguyễn Bá Đại | 21521914 | KH&KTTT |
| 2 | Phan Cả Phát | 21520389 | KH&KTTT |

# 

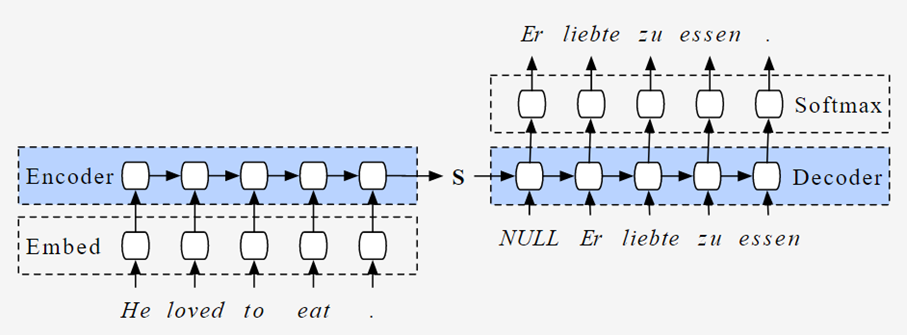
# **MÔ TẢ NGHIÊN CỨU**

## **B1. Giới thiệu về đề tài**

### B1.1. Lý do chọn đề tài

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 – Seq2Seg). 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 1.1 mô tả tổng quan kiến trúc mô hình Seq2Seq chuẩn dựa trên kiến trúc Encoder-Decoder.



Hình 1.1. Mô hình Seq2Seq cho bài toán dịch máy[[1]](#footnote-0)

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 **[1, 2, 3].**

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, nhóm sinh viên 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 **[4]**, T5 **[5]** 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

### B1.2. Tình hình nghiên cứu ngoài nước

Bài toán dịch máy Việt Trung đã được nhiều nhóm nghiên cứu thực hiện trong những năm gần đây bởi sự liên hệ của chương trình. Cụ thể năm 2019, Shengxiang Gao, Jihao Huang và các cộng sự của mình **[6]** do việc khan hiếm dữ liệu, họ đã đề xuất phương pháp dịch máy dựa trên song ngữ, phương pháp này giúp tăng độ chính xác đáng kể mà không làm giảm đi hiệu suất của Model. Vào năm 2020, Hongzheng Li, Jiu Sha và Can Shi **[7]** đã sử dụng Dịch máy bằng nơ-ron ( Neural Machine Translation - NMT ) cùng với phương pháp Revisiting Back-Translation thay vì SMT như những năm về trước. Cùng với năm 2020, nhóm của Wanjin Che và các cộng sự **[8]** để xử lý việc thiếu dữ liệu, họ đã sử dụng việc phân loại từ vựng để xử lý những từ mà không có trong Model. Việc này đã giúp Model tăng hiệu suất và độ chính xác, đặc biệt càng hiệu quả hơn khi trong môi trường thiếu dữ liệu. Đến năm 2022, Zhiqiang Yu, Zhengtao Yu và các cộng sự **[9]** đã tận dụng sự khác biệt về 2 ngôn ngữ Việt - Trung, điều này đã giúp khắc phục sự khan hiếm về dữ liệu, đồng thời cũng đã tăng được hiệu suất và độ chính xác của Model.

### B1.3. Tình hình nghiên cứu trong nước

Năm 2016, nhóm của Phuoc Tran, Dien Dinh và Long H. B. Nguyen **[10]** đã áp dụng phương pháp chỉnh lại phân đoạn từ khiến cho độ dài các câu giữa Việt và Trung đối xứng hơn khiến nó gần hơn với dạng 1 - 1 điều này khiến cho hiệu suất của máy được tăng đáng kể. Cùng với năm đó nhóm họ cũng tạo ra cách tiếp cận mới kết hợp giữa mức độ chữ cái và mức độ từ **[11]**. Năm 2018, nhóm của Phuoc Tran , Le Nguyen và Dien Dinh **[12]** đã sử dụng quan hệ phụ thuộc giữa tiếng Trung và DE (的) trong Trung để khiến các từ của Trung phù hợp với trật tự từ trong tiếng Việt. Năm 2021, Nguyen Hoang Qua và các đồng nghiệp **[13]** đã giới thiệu một bộ mã nguồn mở cho dịch máy NMT cùng với nhiều cải tiến khác nhằm tạo ra một khuôn khổ độc lập, đơn giản để sử dụng, nhất quán và có thể sử dụng trên nhiều miền khác nhau.

### B1.4. Thử thách của đề tài

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.

## **B2. Mục tiêu, nội dung, kế hoạch nghiên cứu**

### **B2.1 Mục tiêu**

Mục tiêu của chúng tôi trong đề tài này là tìm hiểu và nghiên cứu một phương pháp dịch máy Việt – Trung với hạn chế dữ liệu dựa trên các mô hình ngôn ngữ khác nhau. Để hoàn thành được mục tiêu này, chúng tôi có các nội dung và phương pháp nghiên cứu như sau.

### **B2.2 Nội dung và phương pháp nghiên cứu**

**Nội dung 1:** Tiền xử lý bộ dữ liệu Việt - Trung thu thập được

Sau khi thu thập được bộ dữ liệu thô Việt - Trung, chúng tôi sẽ bắt đầu tạo các cặp câu Việt - Trung đối xứng với độ dài từ tương ứng với nhau. Đồng thời sẽ sửa lại các lỗi cú pháp, font chữ, ký hiệu cũng như các bước tiền xử lý cơ bản[[2]](#footnote-1)

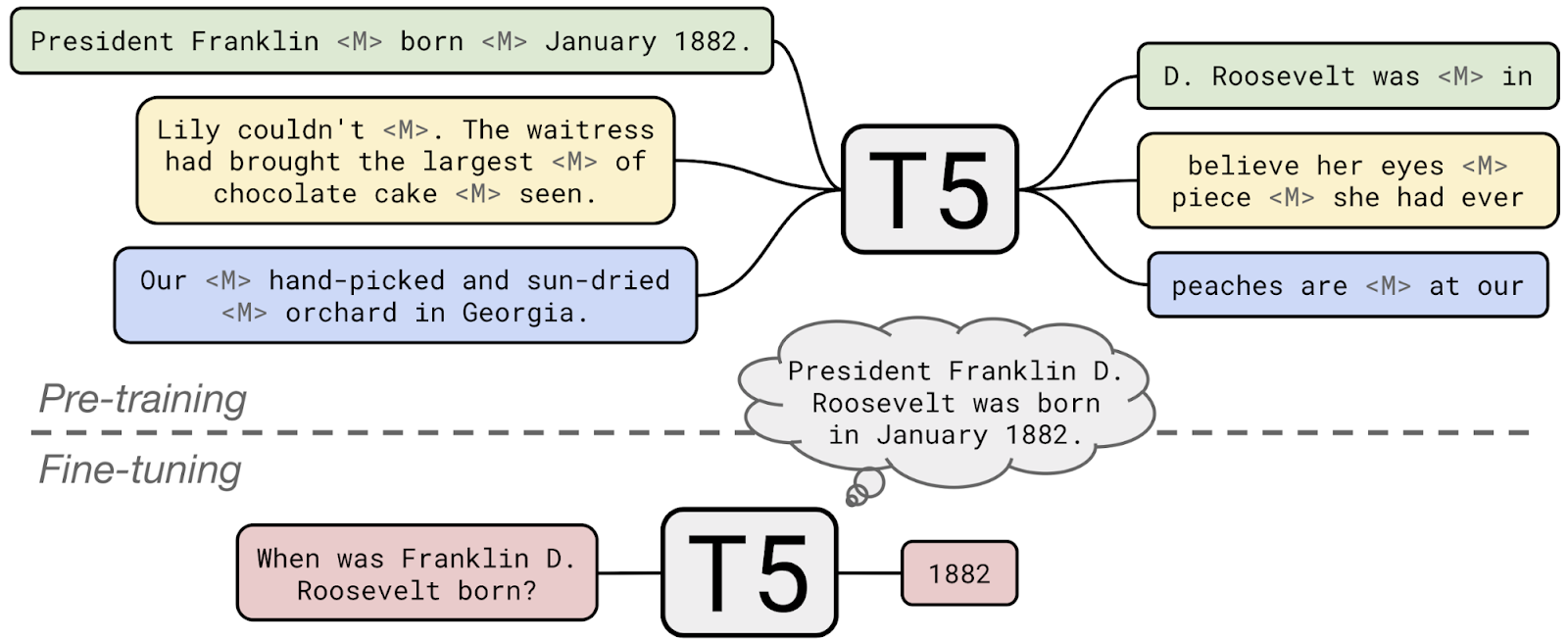
1. Lowercasing ( Đưa chữ hoa về chữ thường). Tại vì mô hình có thể nhận diện chữ cái in hoa như là từ bắt đầu của một câu, điều này có thể ảnh hưởng đến hiệu suất của mô hình.
2. Normalization ( Chuẩn hóa lại các từ ngữ ). Vì trong dữ liệu có thể có các từ viết tắt như ( đc - được, ko - không ) hoặc các từ kéo dài thể hiện mức độ cảm xúc như ( hahahah - ha, tốttttt - tốt ). Mô hình có thể dùng các từ này như 1 từ mới dù nó có y nghĩa y hệt các từ gốc của nó. Điều này khiến cho mô hình có thể giảm hiệu suất.
3. Tokenization ( Tách từ ). Trong dữ liệu thuộc dạng câu vì thế ta phải tách các câu đó ra thành các từ để có thể nối các cặp từ Việt - Trung lại và huấn luyện cho máy.
4. POS ( Phân loại các từ trong câu ) sau khi chúng ta đã làm bước tách từ, tiếp theo chúng ta sẽ gán các loại từ cho câu như ( động từ, danh từ, tính từ, … ) việc này có thể làm tăng mối quan hệ giữa các từ với nhau và từ đó có thể tăng được hiệu suất cũng như độ chính xác của mô hình.

Và sau tất cả những bước xử lý trên chúng tôi sẽ có một bộ dữ liệu có thể đáp ứng được bài toán.

**Nội dung 2**: Nghiên cứu các phương pháp học máy cho bài toán dịch máy Việt - Trung

Các phương pháp tiếp cận cho bài toán của chúng tôi bao gồm:

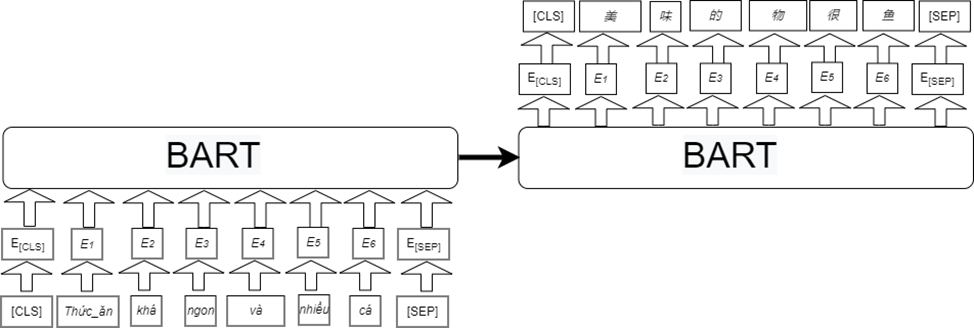
Trong đề tài này, nhóm sinh viên dự kiến sử dụng hai mô hình ngôn ngữ hiện này đang tốt nhất cho dạng bài toán Seq2Seq là mô hình T5 **[5]** và mô hình BART **[4]**. Trong đó, T5 là một mô hình mã hóa-giải mã và chuyển đổi tất cả các vấn đề NLP thành định dạng văn bản thành văn bản. T5 đủ linh hoạt để có thể dễ dàng sửa đổi để áp dụng cho nhiều nhiệm vụ, thường mang lại thành công lớn. Nghiên cứu có hệ thống của chúng tôi so sánh các mục tiêu tiền đào tạo, kiến trúc, bộ dữ liệu không được gắn nhãn, cách tiếp cận chuyển giao và các yếu tố khác trên hàng chục nhiệm vụ hiểu ngôn ngữ. Một cách để sử dụng khung chuyển văn bản thành văn bản là đối với các vấn đề đọc hiểu, trong đó mô hình được cung cấp một số ngữ cảnh cùng với một câu hỏi và được đào tạo để tìm câu trả lời của câu hỏi từ ngữ cảnh. Trong quá trình đào tạo trước, T5 học cách điền vào các khoảng văn bản bị bỏ trống (ký hiệu là <M>) từ các tài liệu trong C4. Để áp dụng T5 cho câu trả lời câu hỏi dạng sách đóng, chúng tôi đã tinh chỉnh nó để trả lời các câu hỏi mà không cần nhập thêm bất kỳ thông tin hoặc ngữ cảnh nào. Điều này buộc T5 phải trả lời các câu hỏi dựa trên “kiến thức” mà nó đã tiếp thu trong quá trình đào tạo trước.T5 làm tốt nhiệm vụ này một cách đáng ngạc nhiên. Mô hình đầy đủ 11 tỷ tham số tạo ra văn bản chính xác của câu trả lời lần lượt là 50,1%, 37,4% và 34,5% trên TriviaQA, WebQuestions và Natural Question.



Hình 2.2. Tổng quan mô hình T5 và ứng dụng tinh chỉnh mô hình T5.[[3]](#footnote-2)

Mô hình MBart đã được trình bày trong công trình nghiên cứu của tác giả Yinhan Liu và các cộng sự **[14]**. mBart là một mô hình bộ mã hóa-giải mã đa ngôn ngữ (chuỗi thành chuỗi) chủ yếu dành cho nhiệm vụ dịch thuật. Vì mô hình đa ngôn ngữ nên nó mong đợi các trình tự ở một định dạng khác. Mã thông báo id ngôn ngữ đặc biệt được thêm vào cả văn bản nguồn và văn bản đích. MBart-50 cho phép các mô hình dịch đa ngôn ngữ có thể được tạo ra thông qua việc tinh chỉnh đa ngôn ngữ. Thay vì tinh chỉnh theo một hướng, một mô hình được điều chỉnh trước được tinh chỉnh trên nhiều hướng cùng một lúc. Nó chứng tỏ rằng các mô hình được đào tạo trước có thể được mở rộng để kết hợp các ngôn ngữ bổ sung mà không làm giảm hiệu suất. Định dạng văn bản cho mBart-50 hơi khác so với mBart. Đối với MBart-50, mã thông báo id ngôn ngữ được sử dụng làm tiền tố cho cả văn bản nguồn và văn bản đích.

Chúng tôi có thể dịch với mô hình mBART-50 bằng cách sử dụng thư viện Huggingface và một vài dòng đơn giản của mã Python mà không cần sử dụng bất kỳ API hoặc dịch vụ đám mây trả phí nào. Dễ dàng dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác.



Hình 2.2. Mô hình Seq2Seq dựa trên kiến trúc BART cho bài toán dịch máy Việt Trung.

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 độ đo mà ban tổ chức đưa ra, đề tài cũng sẽ đánh giá dựa trên các mô hình khác như là BLEU **[15]** ( Bilingual Evaluation Understudy ), METEOR **[15]** ( Metric for Evaluation of Translation with Explicit Ordering ), WER **[15]** ( Word error rate ) 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.

**Nội dung 3**: Báo cáo tổng kết

Thực hiện báo cáo tổng kết chi tiết đề tài bao gồm các nội dung như tìm hiểu các công trình nghiên cứu liên quan, cơ sở lý thuyết, chi tiết cài đặt và kết quả nghiên cứu của đề tài.

**B2.3 Kế hoạch nghiên cứu**.

Chúng tôi thực hiện nghiên cứu đề tài Bản dịch máy từ tiếng Việt sang tiếng Trung và từ tiếng Trung sang tiếng Việt trong vòng 6 tháng, kế hoạch thực hiện chi tiết được trình bày như trong Bảng 1:

**Bảng 1.** Kế hoạch thực hiện đề tài

| **Công việc** | **T1** | **T2** | **T3** | **T4** | **T5** | **T6** |
| --- | --- | --- | --- | --- | --- | --- |
| Thu thập dữ liệu cho bài toán |  |  |  |  |  |  |
| Tiền xử lý bộ dữ liệu |  |  |  |  |  |  |
| Nghiên cứu các phương pháp |  |  |  |  |  |  |
| Cài đặt, thử nghiệm mô hình |  |  |  |  |  |  |
| Báo cáo tổng kết |  |  |  |  |  |  |

## **B3. Kết quả nghi**ên cứudự kiến

Với đề tài Bản dịch máy từ tiếng Việt sang tiếng Trung và từ tiếng Trung sang tiếng Việt,

chúng tôi đặt ra những kết quả nghiên cứu được trình bày như trong Bảng 2 dưới đây:

**Bảng 2.** Kết quả nghiên cứu dự kiến của đề tài

| **Nội dung** | **Kết quả nghiên cứu** |
| --- | --- |
| 1 | * Tiền xử lý lại bộ dữ liệu. * Sản phẩm dự kiến: Bộ dữ liệu Việt - Trung đã được xử lý và báo cáo tổng kết |
| 2 | * Nghiên cứu các mô hình dịch máy như RBMT, SMT, NMT, mBart và mBart-50 cho bài toán dịch máy. * Sản phẩm dự kiến: Báo cáo tổng kết |
| 3 | * Cài đặt và thử nghiệm một số mô hình dịch máy dựa trên bộ dữ liệu đã thu thập. Sau đó so sánh và phân tích kết quả đạt được. * Sản phẩm dự kiến: Mã nguồn ( Source code ) và báo cáo tổng kết |
| 4 | * Thực hiện báo cáo tổng kết cho đề tài * Sản phẩm dự kiến: Báo cáo tổng kết |

## 

## B4. Tài liệu tham khảo

[1]Doan, L., Tran, N. L., Hoang, T., & Nguyen, D. Q. (2021, November). PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 4495-4503).

[2]Nguyen, T. H., Nguyen, T. D. H., Phung, D., Nguyen, D. T. C., Tran, H. M., Luong, M., ... & Nguyen, D. Q. (2022). A Vietnamese-English Neural Machine Translation System. INTERSPEECH.

[3]Phan-Vu, H. H., Tran, V. T., Dang, H. V., & Do, P. T. (2019). Neural machine translation between Vietnamese and English: an empirical study. Journal of Computer Science and Cybernetics, 35(2), 147-166.

[4]M. Lewis et al. (2019) "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". arXiv preprint arXiv:1910.13461

[5]L. Xue et al. (2020) "mT5: A massively multilingual pre-trained text-to-text transformer". arXiv preprint arXiv:2010.11934

[6]S. Gao, J. Huang, M. Xue, Z. Yu, Z. Wang and Y. Zhang. (2019). "Syntax-Based Chinese-Vietnamese Tree-to-Tree Statistical Machine Translation with Bilingual Features | ACM Transactions on Asian and Low-Resource Language Information Processing", ACM Journals.

[7]Li, H., Sha, J., & Shi, C. (2020). Revisiting back-translation for low-resource machine translation between chinese and vietnamese. IEEE Access, 8, 119931-119939.

[8]W. Che, Z. Yu, Z. Yu, Y. Wen and J. Guo. (2019) "Towards Integrated Classification Lexicon for Handling Unknown Words in Chinese-Vietnamese Neural Machine Translation | ACM Transactions on Asian and Low-Resource Language Information Processing", ACM Journals.

[9]Z. Yu, Z. Yu, Y. Xian, Y. Huang and J. Guo. (2022) "Improving Chinese-Vietnamese Neural Machine Translation with Linguistic Differences | ACM Transactions on Asian and Low-Resource Language Information Processing", ACM Journals.

[10]P. Tran, D. Dinh and L. Nguyen. (2016) "Word Re-Segmentation in Chinese-Vietnamese Machine Translation | ACM Transactions on Asian and Low-Resource Language Information Processing", ACM Journals.

[11]Tran, P., Dinh, D., & Nguyen, H. T. (2016). A character level based and word level based approach for Chinese-Vietnamese machine translation. Computational intelligence and Neuroscience, 2016.

[12]P. Tran, L. Nguyen and D. Dinh, "PRE-ORDERING OF “DE PHRASE” IN CHINESE-VIETNAMESE MACHINE TRANSLATION.", ICIC Express, 2018.

[13]Quan, N. H., Dat, N. T., Cong, N. H. M., Van Vinh, N., Vinh, N. T., Thai, N. P., & Viet, T. H. (2021). ViNMT: Neural Machine Translation Tookit. arXiv preprint arXiv:2112.15272.

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

[15]"Evaluation of machine translation - Wikipedia" (2021).

| *Ngày 15 tháng 09 năm 2022*  **Giảng viên hướng dẫn**  (Ký và ghi rõ họ tên) | *Ngày 15 tháng 09 năm 2022*  **Chủ nhiệm đề tài**  (Ký và ghi rõ họ tên) |
| --- | --- |
| Đặng Văn Thìn | Nguyễn Bá Đại |
|  |  |
|  |  |

1. Nguồn: https://miro.medium.com/max/942/1\*KtWwvLK-jpGPSnj3tStg-Q.png [↑](#footnote-ref-0)
2. K. Ganesan, "All you need to know about text preprocessing for NLP and Machine Learning - KDnuggets", KDnuggets, 2019. [↑](#footnote-ref-1)
3. Nguồn: https://1.bp.blogspot.com/-89OY3FjN0N0/XlQl4PEYGsI/AAAAAAAAFW4/knj8HFuo48cUFlwCHuU5feQ7yxfsewcAwCLcBGAsYHQ/s1600/image2.png [↑](#footnote-ref-2)