TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Trần Quốc Khương - Nguyễn Thị Hồng Nhung

ĐO LƯỜNG VÀ TĂNG CƯỜNG MỨC ĐỘ SỬ DỤNG NGỮ CẢNH TRONG DỊCH MÁY MẠNG NEURAL ANH-VIỆT

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

Tp. Hồ Chí Minh, tháng 07/2022

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Trần Quốc Khương - 18120427
Nguyễn Thị Hồng Nhung - 18120498

ĐO LƯỜNG VÀ TĂNG CƯỜNG MỨC ĐỘ SỬ DỤNG NGỮ CẢNH TRONG DỊCH MÁY MẠNG NEURAL ANH-VIỆT

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH CHÍNH QUY

GIÁO VIÊN HƯỚNG DẪN
PGS.TS. Đinh Điền

Tp. Hồ Chí Minh, tháng 07/2022

Nhận xét hướng dẫn

Theo bản nhận xét của giảng viên hướng dẫn (có chữ kí) do giáo vụ cung cấp.

Nhận xét phản biện

Theo bản nhận xét của giảng viên phản biện (có chữ kí) do giáo vụ cung cấp.

Lời cam đoan

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng chúng tôi. Các số liệu và kết quả nghiên cứu trong luận văn này là trung thực và không trùng lặp với các đề tài khác.

Lời cảm ơn

Đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến Thầy PGS.TS. Đinh Điền và Thầy Ths. Nguyễn Hồng Bửu Long đã cho phép chúng em thực hiện khóa luận, cũng như tận tình hướng dẫn, động viên và giúp đỡ chúng em trong suốt khoảng thời gian vừa qua. Nếu không có những lời chỉ dạy, những lời động viên của các Thầy thì khóa luận này khó lòng mà thực hiện được.

Chúng con xin cảm ơn cha mẹ đã luôn dành những tình cảm yêu thương nhất, luôn hỗ trợ và dõi theo chúng con trong suốt những năm học vừa qua.

Chúng em cũng gửi lời cảm ơn đến Trung tâm Ngôn ngữ học tính toán (CLC) đã hỗ trợ chúng em để có địa điểm học tập và nghiên cứu hàng tuần.

Chúng em cũng xin gửi lời cảm ơn đến tất cả các Thầy Cô của khoa Công Nghệ Thông Tin đã dạy dỗ, truyền đạt cho chúng em những kiến thức quý báu trong suốt quãng thời gian chúng em theo học tại trường.

Và cuối cùng xin cảm ơn tất cả bạn bè của chúng tôi đã sát cánh chia sẻ những niềm vui, chia sẻ những khó khăn của chúng tôi, cùng chúng tôi giải quyết những khóa khăn suốt khoảng thời gian đại học.

Xin chân thành cảm ơn!

Trần Quốc Khương - Nguyễn Thị Hồng Nhung

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

ĐO LƯỜNG VÀ TĂNG CƯỜNG MỨC ĐỘ SỬ DỤNG NGỮ CẢNH TRONG DỊCH MÁY MẠNG NEURAL ANH-VIỆT

(Measuring and Increasing Context Usage in English-Vietnamese Neural Machine Translation)

THÔNG TIN CHUNG

- **Người hướng dẫn:**
  - PGS. TS. Đinh Điền (Khoa Công nghệ Thông tin)
- **Nhóm sinh viên thực hiện:**
  1. Trần Quốc Khương (MSSV:18120427)
  2. Nguyễn Thị Hồng Nhung (MSSV:18120498)
- **Loại đề tài:** Nghiên cứu
- **Thời gian thực hiện:** Từ *01/2022* đến *07/2022*

NỘI DUNG THỰC HIỆN

Giới thiệu về đề tài

Các nghiên cứu gần đây trong dịch máy mạng neural (Neural Machine Translation - NMT) cho thấy cả sự cần thiết và sự khả thi của việc sử dụng ngữ cảnh liên câu (ngữ cảnh từ những câu xung quanh khác với câu đang được dịch) ở cấp độ văn bản (*Toral và các cộng sự*, 2017 [30]. *Läubli và các cộng sự*, 2018 [16]). Mặc dù trên lý thuyết, có rất nhiều kiến trúc mô hình dịch máy mạng neural có thể sử dụng ngữ cảnh nói trên (*Tiedemann và Scherrer*, 2017 [29]; *Zhang và cộng sự*, 2018) [40] (các mô hình dịch máy mạng neural theo ngữ cảnh - Context-aware NMT), nhưng các mô hình này thường không có khả năng nhận biết lượng ngữ cảnh cần thiết cho quá trong quá trình xử lý.

Để giải quyết vấn đề đo lường lượng thông tin trong dịch máy mạng neural, đề tài nghiên cứu các độ đo để đo lường một cách tường minh lượng ngữ cảnh mà mô hình thực sự cần sử dụng. Ngoài ra khóa luận cũng nghiên cứu phương pháp sử dụng độ đo thông tin ngữ cảnh để cải tiến chất lượng dịch máy bằng cách kiểm soát lượng thông tin ngữ cảnh cần thiết.

Mục tiêu đề tài

- Tìm hiểu và nắm rõ về dịch máy mạng neural nói chung và dịch máy mạng neural theo ngữ cảnh nói riêng.

- Nghiên cứu các độ đo gần đây dùng để đo lượng thông tin ngữ cảnh mà mô hình dịch máy mạng neural cần trong quá trình dịch.

- Thông qua độ đo lượng ngữ cảnh cần thiết, đề tài tìm hiểu phương pháp nâng cao lượng ngữ cảnh khi độ đo lượng ngữ cảnh chưa đủ để cho ra kết quả dịch có chất lượng.

Phạm vi của đề tài

Nội dung nghiên cứu chính

- Đo lường và tăng cường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural cho song ngữ Anh-Việt. Cụ thể, đề tài nghiên cứu các độ đo lượng ngữ cảnh cần thiết ở cấp độ liên câu (cấp độ văn bản) và từ độ đo này, mô hình sẽ quyết định khi nào cần tăng hay giảm lượng ngữ cảnh tương ứng.

Một số giới hạn và ràng buộc

- Mô hình dịch máy mạng neural Anh-Việt theo ngữ cảnh trong đề tài ở cấp độ văn bản và sử dụng kích thước ngữ cảnh từ 0 đến 4. Trong đó, kích thước ngữ cảnh là số lượng câu đứng phía trước câu đang được dịch. Ví dụ, kích thước ngữ cảnh là 2 thì ngữ cảnh sẽ là 2 câu đứng phía trước câu đang được dịch.

Cách tiếp cận dự kiến

Patrick Fernandes, Kayo Yin, Graham Neubig, André F. T. Martins, 2021 [7] đã thực hiện và cho thấy được sự hiệu quả trong việc đo lường mức độ sử dụng ngữ cảnh bằng độ đo "Thông tin tương hỗ chéo có điều kiện" và tăng cường mức độ sử dụng ngữ cảnh thông qua việc sử dụng mô hình "Loại bỏ từ theo ngữ cảnh" trong dịch máy mạng neural theo ngữ cảnh cho hai cặp ngôn ngữ Anh-Pháp và Anh-Đức. Dựa vào đó, nhóm dự kiến tiếp cận đề tài khóa luận này như sau:

- Nghiên cứu và cài đặt mô hình Transformer cho bài toán dịch máy mạng neural Anh-Việt theo ngữ cảnh ở cấp độ văn bản.

- Nghiên cứu và cài đặt độ đo "Thông tin tương hỗ chéo có điều kiện" để đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt.

- Nghiên cứu và cài đặt mô hình "Loại bỏ từ theo ngữ cảnh" để tăng cường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt.

Kết quả dự kiến của đề tài

- Đo lường được mức độ sử dụng ngữ cảnh của mô hình trong quá trình dịch một cách rõ ràng bằng độ đo "Thông tin tương hỗ chéo có điều kiện".

- Mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt được gia tăng thông qua việc sử dụng mô hình "Loại bỏ từ theo ngữ cảnh".

- Công bố được một bài báo có liên quan trên tạp chí (hội nghị) trong nước hoặc quốc tế (nếu có thể).

Kế hoạch thực hiện

Mốc thời gian	Người thực hiện	Công việc
01 - 02/2022	Khương, Nhung	Nhận đề tài và tìm hiểu các kiến thức liên quan.
02 - 03/2022	Khương, Nhung	Nghiên cứu các bài báo khoa học có trong tài liệu tham khảo.
03 - 04/2022	Khương, Nhung	Cài đặt mô hình Transformer.
04 - 05/2022	Khương, Nhung	Cài đặt Conditional Cross-Mutual Information và Context-aware Word Dropout
05 - 06/2022	Khương, Nhung	Viết báo cáo khóa luận
15/07/2022	Khương, Nhung	Hoàn thành báo cáo khóa luận
31/07/2022	Khương, Nhung	Bảo vệ khóa luận trước hội đồng

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày 04 tháng 04 năm 2022
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Mục lục

Danh sách hình

Danh sách bảng

Bảng thuật ngữ

Byte Pair Encoding	Mã hóa cặp byte
Context-aware Machine Translation	Dịch máy mạng neural theo ngữ cảnh
Context-aware Word Dropout	Loại bỏ từ theo ngữ cảnh
Decoder	Bộ giải mã
Document-level Translation	Dịch cấp độ văn bản
Encoder	Bộ mã hóa
Feed Forward	Lan truyền tới
Inside Integration	Tích hợp bên trong
Machine Translation	Dịch máy
Multi-head Attention	Chú ý đa đầu
Neural Machine Translation	Dịch máy mạng neural
Normalization	Bình thường hóa
Outside Integration	Tích hợp bên ngoài
Position Encoding	Mã hóa vị trí
Recurrent Neural Network	Mạng neural hồi quy
Self-Attention Mechanism	Cơ chế tự chú ý
Sentence-level Translation	Dịch cấp độ câu
Word Dropout	Loại bỏ từ

Danh mục từ viết tắt

BLEU	Bilingual Evaluation Understudy
BPE	Byte Pair Encoding
NMT	Neural Machine Translation
PE	Positional Encoding
RNN	Recurrent Neural Network
SentNMT	Sentence Neural Machine Translation

Tóm tắt

Các nghiên cứu gần đây trong dịch máy mạng neural cho thấy sự cần thiết và tính khả thi của việc sử dụng ngữ cảnh liên câu (ngữ cảnh từ những câu xung quanh câu đang được dịch) (*Hassan và các cộng sự, 2018* [9]; *Toral và các cộng sự, 2018* [30]). Theo lý thuyết, có nhiều kiến trúc mô hình có thể sử dụng ngữ cảnh nói trên (*Tiedemann và Scherrer, 2017* [29]; *Zhang và các cộng sự, 2017* [40]). Tuy nhiên, lượng ngữ cảnh được sử dụng bởi các kiến trúc mô hình trên trong quá trình dịch không được đo đạc một cách tường minh.

Trong khóa luận này, chúng tôi tìm hiểu các độ đo để đo lường lượng ngữ cảnh sử dụng trong các mô hình dịch máy mạng neural hiện tại, và sử dụng độ đo "Thông tin tương hỗ chéo có điều kiện" để đo lượng ngữ cảnh được sử dụng bởi các mô hình dịch máy mạng neural Anh-Việt. Sau khi thực hiện các thí nghiệm với độ đo "Thông tin tương hỗ chéo có điều kiện" của *Patrick Fernandes và các cộng sự, 2021* [7], chúng tôi nhận thấy: (1) lượng ngữ cảnh bên phía đích được sử dụng nhiều hơn so với lượng ngữ cảnh bên phía nguồn, (2) lượng ngữ cảnh được sử dụng không tăng đồng đều với kích thước ngữ cảnh và thậm chí có thể dẫn đến giảm lượng thông tin ngữ cảnh được sử dụng.

Từ những kết quả đạt được với độ đo "Thông tin tương hỗ chéo có điều kiện", chúng tôi sử dụng "Loại bỏ từ theo ngữ cảnh" của *Patrick Fernandes và các cộng sự, 2021* [7] để gia tăng lượng ngữ cảnh được sử dụng bởi các mô hình dịch máy mạng neural Anh-Việt. Sau khi thực hiện các thí nghiệm, chúng tôi nhận thấy "Loại bỏ từ theo ngữ cảnh" giúp các

mô hình dịch máy mạng neural Anh-Việt sử dụng nhiều ngữ cảnh hơn trong quá trình dịch, và giúp cải thiện chất lượng dịch của mô hình theo độ đo BLEU của *Papineni và các cộng sự, 2002* [24] và COMET của *Rei và các cộng sự, 2020* [26].

Chương 1

Mở đầu

Trong chương Mở đầu này, chúng tôi sẽ nêu lý do vì sao chọn đề tài này tại phần 1.1, mục tiêu cần đạt được trong đề tài tại phần 1.2, phạm vi nghiên cứu của đề tài tại phần 1.3, ý nghĩa mà đề tài mang lại tại phần 1.4, cấu trúc của khóa luận tại phần 1.5.

1.1 Lý do chọn đề tài

Các hệ thống dịch máy mạng neural (Neural Machine Translation - NMT) có tiềm năng giải quyết nhiều vấn đề của hệ thống dịch thuật dựa trên cụm từ truyền thống và đã được chứng minh là tạo ra các bản dịch chất lượng tốt hơn. Với những tiến bộ nhanh chóng trong những năm gần đây, nhiều hệ thống NMT thậm chí được đánh giá là đạt được tính ngang bằng với các bản dịch của con người trong một số nhiệm vụ và cặp ngôn ngữ, đặc biệt là trên các cặp ngôn ngữ giàu tài nguyên (Hassan và các cộng sự, 2018 [9]), tuy nhiên các hệ thống NMT tiêu chuẩn được xây dựng để dịch ở cấp độ câu, không thể xem xét được sự phụ thuộc giữa các câu trong cùng tài liệu, điều này khiến các mô hình NMT tiêu chuẩn không còn chính xác nếu đánh giá bản dịch ở cấp độ văn bản (*Toral và các cộng sự, 2018* [30]; *Läubli và các cộng sự, 2018* [16]) nhiều tuyên bố cho rằng NMT cấp câu tạo ra các lỗi cấp tài liệu, ví dụ như bản dịch không nhất

quán về từ vựng hay giải quyết các đại từ đảo ngữ (Guillou và các cộng sự, 2018 [8]; *Läubli và các cộng sự, 2018* [16]).

Hệ thống dịch máy mạng neural cấp câu chỉ mã hóa các câu nguồn mà không sử dụng ngữ cảnh. Để giải quyết thách thức trên, các nghiên cứu gần đây trong dịch máy mạng neural cho thấy cả sự cần thiết và sự khả thi của việc sử dụng ngữ cảnh liên câu (ngữ cảnh từ những câu xung quanh khác với câu đang được dịch) ở cấp độ văn bản (*Toral và cộng sự, 2017* [30]. *Läubli và cộng sự, 2018* [16]) và đã cho thấy kết quả đầy hứa hẹn trong việc tạo ra các bản dịch nhất quán và mạch lạc (*Zhang và các cộng sự, 2018* [40]; *Voita và các cộng sự, 2018* [36]; *Kim và các cộng sự, 2019* [12]; *Bawden và các cộng sự, 2018* [5]; *Miculicich và các cộng sự, 2018*; *Maruf and Haffari, 2018* [20]; *Maruf và các cộng sự, 2019* [21]). Mô hình dịch máy mạng neural có thể sử dụng ngữ cảnh nói trên bằng cách kết hợp các câu ngữ cảnh xung quanh (trong một trong hai hoặc cả hai phía nguồn và phía đích) làm đầu vào bổ sung cho mô hình NMT, một số đề xuất phổ biến như sử dụng các bộ mã hóa khác nhau cho ngữ cảnh của *Zhang và các cộng sự, 2018* [40], sử dụng bộ nhớ cache-based để mã hóa ngữ cảnh của *Tu và các cộng sự, 2018a* [31], hoặc sử dụng các mô hình có cơ chế chú ý phân cấp của *Miculicich và các cộng sự, 2018* [22]; *Maruf và các cộng sự, 2019a* [21]),... nhằm mô hình hóa ngữ cảnh để xem xét mối quan hệ phức tạp giữa các câu. Tuy nhiên, ngay cả khi điểm số về chất lượng dịch thuật (ví dụ như BLEU) có cải thiện đáng kể nhưng sự cải thiện này chỉ giới hạn ở các bộ dữ liệu cụ thể tương đối nhỏ do dữ liệu mức độ văn bản (document corpora) đã được tinh chỉnh ở quy mô lớn không dễ dàng có sẵn (*Wang và các cộng sự, 2017* [38]; *Kuang và Xiong, 2018* [13]; *Zhang và các cộng sự, 2018* [40]; *Tu và các cộng sự, 2018* [31]). Nhiều nghiên cứu gần đây đã chỉ ra rằng trên các bộ dữ liệu lớn, mặc dù chi phí trong mô hình hóa cao, các hệ thống NMT theo ngữ cảnh chỉ cho chất lượng bản dịch tương đương với mô hình cơ sở (*Li và các cộng sự, 2020* [17]; *Lopes và các cộng sự, 2020* [19]). Giả thuyết được đưa ra nhằm giải thích cho sự cải thiện mờ nhạt trên là do thực tế các mô hình NMT có kiến trúc đủ

mạnh để có thể mô hình hóa ngữ cảnh xuyên suốt không nhất thiết phải học cách làm như vậy khi được đào tạo với các mô hình đào tạo hiện có.

Như lý thuyết ở trên thì có rất nhiều kiến trúc mô hình NMT có thể sử dụng ngữ cảnh bổ sung (*Tiedemann và Scherrer, 2017* [29]; *Zhang và cộng sự, 2018* [40]) (các mô hình dịch máy mạng neural theo ngữ cảnh - Context-aware NMT), nhưng các mô hình này thường không có khả năng nhận biết lượng ngữ cảnh cần thiết trong quá trình xử lý. Nhiều nghiên cứu gần đây cho thấy các mô hình NMT theo ngữ cảnh hiện tại thường không sử dụng ngữ cảnh một cách đầy đủ và có ý nghĩa (*Kim và cộng sự, 2019* [12]) tuyên bố rằng những cải tiến theo mô hình nhận thức ngữ cảnh chủ yếu là từ việc chính thức hóa bằng cách đặt các tham số cho đầu vào ngữ cảnh. Vậy lượng ngữ cảnh mà mô hình thật sự đã sử dụng cho quá trình dịch là bao nhiêu? Thực tế việc định lượng ngữ cảnh sử dụng vẫn đang là một thách thức và chưa có một bộ công cụ cụ thể để đánh giá.

Vì vậy nhằm giải quyết vấn đề đo lường lượng thông tin trong dịch máy mạng neural, chúng tôi tiến hành nghiên cứu các độ đo để đo lường một cách tường minh lượng ngữ cảnh mà mô hình thực sự sử dụng. Đồng thời nghiên cứu phương pháp sử dụng độ đo thông tin ngữ cảnh để cải tiến chất lượng dịch máy bằng cách kiểm soát lượng thông tin ngữ cảnh cần thiết.

1.2 Mục tiêu đề tài

- Tìm hiểu và nắm rõ về dịch máy mạng neural nói chung và dịch máy mạng neural theo ngữ cảnh nói riêng.

- Nghiên cứu các độ đo gần đây dùng để đo lượng thông tin ngữ cảnh mà mô hình dịch máy mạng neural cần trong quá trình dịch.

- Thông qua độ đo lượng ngữ cảnh cần thiết, tìm hiểu phương pháp nâng cao lượng ngữ cảnh khi độ đo lượng ngữ cảnh chưa đủ để cho ra kết quả dịch có chất lượng.

1.3 Phạm vi nghiên cứu

Trong khóa luận này chúng tôi mong muốn có thể đo lường và tăng cường mức độ sử dụng ngữ cảnh của mô hình trong dịch máy mạng neural cho song ngữ Anh-Việt. Cụ thể, đề tài nghiên cứu các độ đo lượng ngữ cảnh cần thiết ở cấp độ liên câu (cấp độ văn bản) và từ độ đo này, mô hình sẽ quyết định khi nào cần tăng hay giảm lượng ngữ cảnh tương ứng. Độ đo này áp dụng cho bất kỳ mô hình dịch máy theo ngữ cảnh. Sau đó, chúng tôi thực hiện tiến hành các phân tích thực nghiệm nghiêm ngặt về CXMI giữa ngữ cảnh phía đích - câu đích và giữa ngữ cảnh phía nguồn - câu nguồn theo kích thước ngữ cảnh từ 0 đến 4. Trong đó, kích thước ngữ cảnh là số lượng câu đứng phía trước câu đang được dịch.

Dựa trên các kết quả đó, chúng tôi xem xét làm thế nào để giúp các mô hình dịch máy mạng neural Anh-Việt tăng cường sử dụng nhiều ngữ cảnh hơn cho quá trình dịch. Cụ thể, chúng tôi nghiên cứu cài đặt "Loại bỏ từ theo ngữ cảnh" (CoWord Dropout) của *Patrick Fernandes và cộng sự, 2021* [7] - một biến thể đơn giản nhưng hiệu quả của "Loại bỏ từ" của *Sennrich và các cộng sự, 2016* [27]. Ngoài ra, chúng tôi cũng tiến hành đánh giá so sánh chất lượng bản dịch Anh-Việt được tạo ra từ mô hình dịch máy mạng neural trong trường hợp sử dụng và không sử dụng "Loại bỏ từ theo ngữ cảnh" dựa trên điểm BLEU và COMET. Các bước tiếp cận chính trong khóa luận này của chúng tôi bao gồm:

- Nghiên cứu và cài đặt mô hình Transformer cho bài toán dịch máy mạng neural Anh-Việt theo ngữ cảnh ở cấp độ văn bản.
- Nghiên cứu và cài đặt độ đo "Thông tin tương hỗ chéo có điều kiện" để đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt.
- Nghiên cứu và cài đặt mô hình "Loại bỏ từ theo ngữ cảnh" để tăng cường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt.

1.4 Ý nghĩa

- Định lượng chính xác lượng ngữ cảnh mà mô hình sử dụng trong dịch máy mạng neural cho song ngữ Anh-Việt, từ đó điều chỉnh đầu vào thích hợp cho mô hình để cải thiện chất lượng dịch.

- Cung cấp thêm một thước đo giúp định lượng và thúc đẩy các nghiên cứu tương lai trong dịch máy mạng neural sử dụng ngữ cảnh đối với các cặp ngôn ngữ liên quan tới tiếng Việt.

1.5 Cấu trúc khóa luận

Các phần tiếp theo của khóa luận sẽ được trình bày theo cấu trúc sau:

- Chương 2: Tổng quan về dịch máy
- Chương 3: Đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt
- Chương 4: Tăng cường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural Anh-Việt
- Chương 5: Kết luận

Chương 2

Tổng quan về dịch máy

Dịch máy (Machine Translation) là một lĩnh vực trong Xử lý ngôn ngữ tự nhiên, nghiên cứu việc dạy cho máy có thể học để dịch một cách tự động từ một ngôn ngữ này (gọi là ngôn ngữ nguồn) sang một ngôn ngữ khác (gọi là ngôn ngữ đích) mà không có sự can thiệp của con người. Để có được cái nhìn rõ hơn về những gì được thực hiện trong khóa luận, trong chương này, sẽ giới thiệu sơ lược về lịch sử của dịch máy tại phần 2.1.1, giới thiệu về dịch máy mạng neural nói chung tại phần 2.2.1, kiến trúc Transformer tại phần 2.2.2 và dịch máy mạng neural theo ngữ cảnh nói riêng tại phần 2.2.6, 2.2.8 và một số chỉ số đánh giá chất lượng dịch tại phần 2.3

2.1 Sơ lược về dịch máy

2.1.1 Lịch sử dịch máy

Các ý tưởng về một hệ thống dịch máy được cho là bắt đầu xuất hiện vào thế kỉ thứ 17. Tuy nhiên, mãi về thế kỉ thứ 20 thì ý tưởng cụ thể về một hệ thống dịch máy mới được công nhận. Năm 1933, George Artsrouni đã tạo ra một thiết bị lưu trữ trên băng giấy có thể dùng để tra nghĩa của bất kì từ nào trong ngôn ngữ khác. Cũng trong năm đó, Petr Petrovich Troyanskii đã phát minh ra một từ điển cơ khí cho phép dịch từ ngôn ngữ này sang ngôn ngữ khác.

Warren Weaver và Andrew D.Booth là những người đầu tiên đề xuất về việc sử dụng máy tính để triển khai một hệ thống dịch tự động. Ý tưởng dựa trên lý thuyết thông tin, sự thành công của việc phá vỡ mật mã Enigma trong thế chiến II và sự suy đoán các nguyên tắc nền tảng tổng quát của ngôn ngữ tự nhiên. Vào những năm 1950, các nghiên cứu về dịch máy diễn ra mạnh mẽ tại Liên Xô và Hoa Kỳ và đạt được một vài thành tựu nổi bật như công ty IBM hợp tác cùng trường đại học Georgetown đã xây dựng một hệ thống dịch Nga-Anh thu hút được quan tâm của công chúng và chính phủ.

Tuy nhiên, vào năm 1960, nhà ngôn ngữ học Bar-Hillel đã chỉ ra một số rào cản của dịch máy ở thời điểm lúc bấy giờ và đưa ra kiến nghị rằng dịch máy chỉ nên được áp dụng với phạm vi hẹp. Năm 1966, theo báo cáo của ALPAC (Automatic Language Processing Advisory Committee), dịch máy tỏ ra chậm chạm, thiếu chính xác, tốn kém và tỏ ra không có triển vọng. Do đó, các hoạt động nghiên cứu về dịch máy trong giai đoạn lúc bấy giờ diễn ra chậm lại.

Vào những năm 1980, dịch máy bắt đầu hồi sinh trở lại. Số lượng các hệ thống dịch ngày càng tăng chủ yếu dựa vào máy tính lớn (mainframe). Năm 1984, phương pháp dịch máy dựa trên ví dụ được đề xuất bởi Makoto Nagao và các cộng sự. Vào những năm cuối thập niên 1980 đầu thập niên 1990, phương pháp dịch máy dựa vào thống kê được giới thiệu bởi công ty IBM. Và hai phương pháp này được nghiên cứu chủ yếu trong những năm sau 2000.

Và giai đoạn 2014 trở về sau này, các mô hình dịch máy mạng neural xuất hiện giúp cho chất lượng dịch ngày càng được cải thiện.

Đối với dịch máy cho tiếng Việt, việc dịch tự động được bắt đầu nghiên cứu từ những năm 1960 để phục vụ cho mục đích quân sự. Năm 1969, Bernard E. Scott thành lập công tu Logos với mục tiêu nghiên cứu hệ thống dịch tự động từ tiếng Anh sang tiếng Việt. Vào đầu thập niên 1970, một dự án khác về xây dựng hệ thống dịch tự động từ tiếng Anh ra tiếng Việt đã được tiến hành tại Tập đoàn viễn thông Xyzyx. Hệ dịch máy

Anh-Việt được sử dụng rộng rãi tại Việt nam đầu tiên là EVTRAN năm 1997, sau đó là EVTRAN 2.0 năm 1999 với hơn 200.000 từ và cụm từ. Từ năm 2006, bản EVTRAN 3.0 (được gọi là Ev-Shuttle) biên dịch văn bản hai chiều Anh-Việt và Việt-Anh (với hơn 500.000 mục từ vựng).

2.1.2 Vấn đề nhập nhằng trong việc dịch văn bản

Khó khăn lớn nhất trong việc thiết kế hệ thống dịch máy là làm sao để khử được tính nhặp nhằng của ngôn ngữ. Tính nhập nhằng là sự không rõ ràng của ngôn ngữ, một từ (câu) có thể mang nhiều ý nghĩa khác nhau tùy vào ngữ cảnh hay mục đích của người nói/viết.

Việc nhập nhằng xảy ra khi sử dụng từ đồng âm hoặc quan hệ đồng tham chiếu ở cấp độ từ.

Ví dụ về nhập nhằng do từ đồng âm, chúng ta có hai câu sau:

1. *"Con **đường** này thật là đẹp".*

2. *"Ly cà phê này cần thêm **đường**."*

Trong hai câu đều sử dụng từ **đường** nhưng có hai ý nghĩa khác nhau. Câu 1 là nói về đường đi. Còn câu hai nói về gia vị đường.

Ví dụ về nhập nhằng do đồng tham chiếu, chúng ta có câu sau:

*"Tôi được ba mẹ đưa đến công viên nước, **nó** thật sự rất tuyệt vời."*

Trong câu trên, từ **nó** có thể là chỉ đến công viên nước hoặc là chỉ đến việc được ba mẹ đưa đến công viên nước.

Ngoài ra, các thành ngữ cũng là một nhập nhằng của ngôn ngữ, ví dụ *"Rain cats and dogs"* được dịch sang tiếng Việt là *"Mưa tầm tã"*

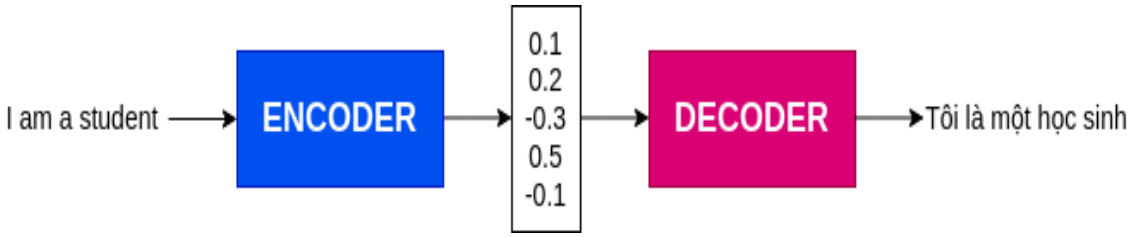
2.2 Dịch máy mạng neural

2.2.1 Giới thiệu dịch máy mạng neural

Dịch máy mạng neural (Neural Machine Translation) là một cách tiếp cận dịch máy sử dụng mạng neural nhân tạo lớn để dự đoán chuỗi từ được dịch bằng cách mô hình hóa toàn bộ các câu văn trong một mạng neural nhân tạo duy nhất.

Cấu trúc của một hệ thống dịch máy mạng neural gồm hai phần như trong hình 2.1:

- Bộ mã hóa (Encoder): nhận thông tin từ câu cần được dịch rồi trả về một vector đại diện.
- Bộ giải mã (Decoder): xử lý vector đầu vào và trả về bản dịch của ngôn ngữ đích.



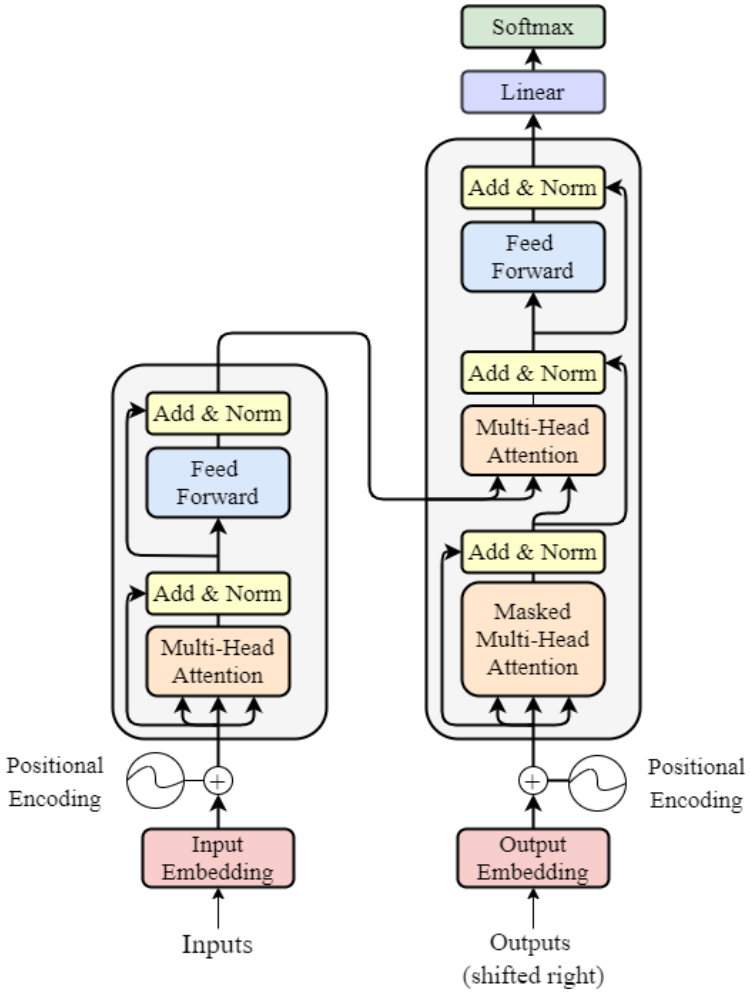
Hình 2.1: Kiến trúc Encoder-Decoder

Dịch máy mạng neural sử dụng toàn bộ câu khi dịch, vì vậy lấy được những thông tin liên quan giữa các từ với nhau như trật tự từ, loại từ. Đây là một điểm hay khác biệt của dịch máy mạng neural so với dịch máy theo phương pháp thống kê.

Bộ mã hóa và bộ giải mã có thể lựa chọn nhiều mô hình kiến trúc khác nhau. Và mô hình kiến trúc được sử dụng phổ biến đó là Transformer của *Vaswani và các cộng sự, 2017* [33] và sẽ được giới thiệu trong 2.2.2 cũng như được sử dụng trong khóa luận này.

2.2.2 Giới thiệu về Transformer

Trước khi có sự xuất hiện của Transformer, hầu như mọi tác vụ trong Xử lý ngôn ngữ tự nhiên, và đặc biệt là dịch máy đều sử dụng kiến trúc mạng neural hồi quy (Recurrent Neural Networks - RNN). Do phải xử lý tuần tự nên nhược điểm lớn nhất của mạng neural hồi quy là tốc độ chậm và hạn chế trong việc biểu diễn các phụ thuộc xa giữa các từ trong câu. Và khi xuất hiện, Transformer đã giải quyết được các vấn đề của mạng neural hồi quy gặp phải nhờ vào khả năng tính toán song song và cơ chế tự chú ý (self-attention).



Hình 2.2: Kiến trúc Transformer [33]

Mã hóa vị trí (Positional Encoding - PE)

Do tất cả các vector biểu diễn từ được đưa song song vào mô hình nên sẽ phát sinh vấn đề về trật tự của các từ trong câu, nên cần một cơ chế để ghi nhớ lại vị trí của các từ trong câu. Do đó, vị trí của các từ trong câu sẽ được mã hóa bằng một vector có kích thước bằng với vector biểu diễn từ (Input Embedding/Output Embedding) và được cộng trực tiếp với vector biểu diễn từ (Input Embedding/Output Embedding) để làm đầu vào cho mô hình.

Vector biểu diễn vị trí của các từ trong câu được tính theo công thức sau:

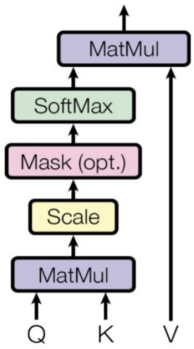
$$PE_{(pos,2i)} = sin(pos, 10000^{2i/d_{model}}) \tag{2.1}$$

$$PE_{(pos,2i+1)} = cos(pos, 10000^{2i/d_{model}}) \tag{2.2}$$

Trong đó, *pos* là vị trí của từ trong câu, i là số chiều (dimension) của vector biểu diễn vị trí.

Cơ chế tự chú ý (Self-Attention)

Cơ chế tự chú ý là cơ chế giúp bộ mã hóa nhìn vào các từ khác trong lúc mã hóa một từ cụ thể. Do đó, Transformer có thể hiểu được mối liên hệ giữa các từ trong câu, cho dù các từ có nằm ở vị trí xa nhau.



Hình 2.3: Cơ chế tự chú ý [33]

Dựa vào hình 2.3, cơ chế tự chú ý gồm có các bước sau:

- Với mỗi từ, cần tạo ra ba vector là vector truy vấn (query) kí hiệu là Q, vector khóa (key) kí hiệu là K, vector giá trị (value) kí hiệu là V bằng cách nhân ma trận biểu diễn các từ đầu vào với ma trận trọng số tương ứng mà chúng ta sử dụng trong quá trình huấn luyện. Ba vector này đóng các vai trò khác nhau và đều quan trọng trong cơ chế tự chú ý.

- Với mỗi từ, ta cần tính vector trọng số chú ý của các từ khác trong câu đối với từ này. Các vector trọng số chú ý này giúp xác định từ nào sẽ cần được chú ý và được chú ý bao nhiêu trong quá trình mã hóa một từ. Điểm được tính bằng tích vô hướng giữa vector truy vấn Q của từ đang xét với từng vector khóa K của các từ trong câu. Tiếp theo, chia cho căn bậc hai của số chiều vector khóa K. Tiếp theo, chuẩn hóa vector trọng số chú ý về giá trị từ 0 đến 1 bằng hàm softmax. Giá trị càng gần 1, có nghĩa là khóa và truy vấn có độ tương đồng cao, được chú ý nhiều, và ngược lại cho giá trị càng gần 0. Sau đó, nhân vector giá trị V của từ đang xét với các vector trọng số chú ý vừa tính được ở trên để được các vector đầu ra. Quá trình này có thể được tính bằng công thức như sau:

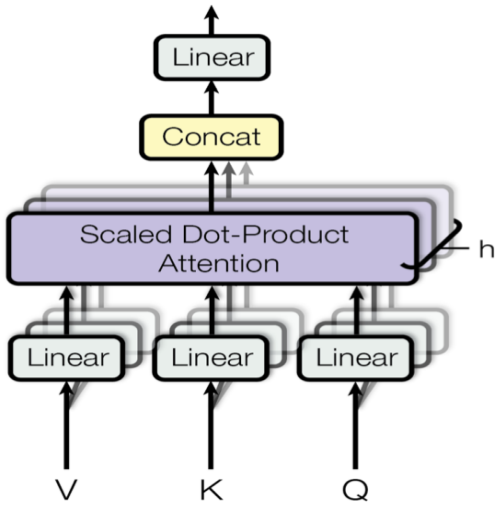
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.3}$$

  Sau đó, cộng các vector đầu ra lại với nhau.

Lớp mặt nạ (Mask Layer) sẽ được sử dụng cho bộ giải mã. Cơ chế tự chú ý khi sử dụng thêm mặt nạ (Mask) khác với cơ chế chú ý bình thường là chỉ được phép dùng các từ được giải mã ở các bước trước đó và che đi các từ chưa được giải mã đến.

Cơ chế chú ý đa đầu (Multi-Head Attention)

Để mô hình có thể học được nhiều kiểu mối quan hệ giữa các từ trong câu với nhau thì cần thêm nhiều cơ chế tự chú ý, mỗi cơ chế tự chú ý sẽ học một kiểu mối quan hệ khác nhau. Bây giờ, các vector sẽ có thêm một chiều nữa đó là chiều sâu h.



Hình 2.4: Cơ chế chú ý đa đầu [33]

Chú ý đa đầu (Multi-Head Attention) sẽ được tính như sau:

$$MultiHead(Q, K, V) = Concat(head_1, ...head_h)W^O \tag{2.4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Trong đó, $W_i^Q \in \mathbb{R}^{d_{model}*d_k}$; $W_i^K \in \mathbb{R}^{d_{model}*d_k}$; $W_i^V \in \mathbb{R}^{d_{model}*d_v}$; $W^O \in \mathbb{R}^{d_{model}*hd_v}$; d_v là chiều của vector giá trị V; d_k là chiều của vector khóa K.

Kiến trúc Transformer

Bộ mã hóa (Encoder) của Transfomer sẽ gồm nhiều lớp mã hóa (Encoder Layer) tương tự nhau. Trong mỗi lớp mã hóa (Encoder Layer) sẽ bao gồm hai phần chính là lớp chú ý đa đầu (Multi-Head Attention Layer) và lớp lan truyền tới (Feed Forward Layer). Ngoài ra còn có bỏ qua kết nối

(Skip Connection) và lớp bình thường hóa (Normalization Layer) giúp cho quá trình huấn luyện hội tụ được nhanh hơn và tránh mất mát thông tin trong quá trình huấn luyện. Vector biểu diễn từ sẽ được cộng với vector biểu diễn vị trí (Positional Encoding). Tiếp theo, vector này sẽ được đưa vào lớp chú ý đa đầu (Multi-Head Attention). Sau đó, được đưa vào lớp lan truyền tới (Feed Forward Layer) để có được đầu ra bên phía bộ mã hóa làm đầu vào cho bộ giải mã. Tại mỗi lớp đều có bỏ qua kết nối (Skip Connection) và lớp bình thường hóa (Normalization).

Bộ giải mã (Decoder) sẽ nhận đầu ra của bộ mã hóa, cụ thể đó là vector khóa K và vector giá trị V để giải mã câu nguồn thành câu đích. Kiến trúc của bộ giải mã cũng tương đối giống với kiến trúc của bộ mã hóa, ngoại trừ việc bộ giải mã có thêm một lớp chú ý đa đầu (Multi-Head Attention Layer) nằm ở giữa dùng để học mối liên hệ giữa các từ đang được dịch và các từ ở câu nguồn. Ngoài ra, còn một điểm khác so với bộ mã hóa là bộ lớp chú ý đa đầu đầu tiên của bộ giải mã là lớp chú ý đa đầu có mặt nạ (Mask).

2.2.3 Bản dịch cấp câu (Sentence-level Translation)

Các mô hình NMT thường mô hình hóa bản dịch cấp độ câu (SentNMT) dựa trên kiến trúc bộ mã hóa-giải mã (*Bahdanau và các cộng sự, 2015* [4]; *Vaswani và các cộng sự, 2017* [33]). Các mô hình SentNMT tối đa hóa xác suất có điều kiện $\log p(y|x;\theta)$ của câu đích $y = \{y_1, y_2, .., y_T\}$ khi biết một câu nguồn $x = \{x_1, x_2, ..., x_T\}$ từ tập ngữ liệu song ngữ dồi dào $D_s = \{x^{(m)}, y^{(m)}\}_{m=1}^{M}$ với M là số lượng câu, mỗi câu là một chuỗi các từ, khi đó:

$$\mathrm{L}(D_s;\theta) = \sum_{m=1}^{M} \log p(y^{(m)}|x^{(m)};\theta). \tag{2.5}$$

2.2.4 Bản dịch cấp tài liệu (Document-level Translation)

Cho tập dữ liệu song ngữ cấp tài liệu $D_d = \{X^{(m)}, Y^{(m)}\}_{m=1}^{M}$, M là số lượng văn bản, văn bản nguồn $X^{(M)} = \{< x_k^{(m)} >\}_{k=1}^{n}$ và $Y^{(M)} = \{< y_k^{(m)} > \}_{k=1}^{n}$ là văn bản đích, trong đó n là số lượng câu trong mỗi văn bản,khi đó các mô hình NMT cấp độ tài liệu tối đa hóa xác suất cuả tài liệu Y cho trước tài liệu nguồn X:

$$\mathrm{L}(D_d;\theta) = \sum_{m=1}^{M} \log p(Y^{(m)}|X^{(m)};\theta) = \sum_{m=1}^{M}\sum_{k=1}^{n} \log p(y_k^{(m)}|y_{<k}^{(m)}, x_k^{(m)}, x_{-k}^{(m)}, \theta). \tag{2.6}$$

Trong đó $y_{<k}^{(m)}$ biểu thị các câu đích phía trước câu đích hiện tại $y_k^{(m)}$ đã được dịch, $x_{-k}^{(m)}$ là phần còn lại của các câu nguồn khác với câu nguồn hiện tại $x_k^{(m)}$.

2.2.5 Tại sao chúng ta cần dịch máy mạng neural theo ngữ cảnh?

Để dịch một văn bản, chúng ta có thể chia chúng thành các câu, dịch từng câu độc lập với nhau. Sau đó, chúng ta gộp các câu được dịch lại với nhau để có một văn bản được dịch hoàn chỉnh. Tuy nhiên, đây không phải là một cách hợp lý, vì khi chúng ta ghép các câu được dịch độc lập lại với nhau chúng sẽ không mang lại một ý nghĩa nhất quán. Vậy những lý do nào gây ra vấn đề trên?

Trực chỉ (Deixis)

Theo [1], "Trực chỉ về thực chất là một hiện tượng nằm trong phạm vi quy chiếu". Cách gọi trực chỉ bắt nguồn từ những *hành động chỉ xuất* ngoài ngôn ngữ. Vì vậy, trực chỉ được dùng để áp dụng cho những phương tiện ngôn ngữ thực hiện chức năng quy chiếu. Nói cách khác, trực chỉ chỉ

ra và đồng nhất quy chiếu bằng cách trực tiếp dựa ngay vào những mốc do hành động phát ngôn của người nói tạo ra. Những mốc cơ bản là:

- Người nói
- Lúc nói
- Nơi nói

Các phương tiện thực hiện chức năng trực chỉ tương ứng:

- Tôi, Tao, Mày,...
- Hôm qua, Hôm kia, Ngày mai,...
- Đây, Kia,...

Ví dụ 1: **"Tôi ở đây để giới thiệu về nó đến cho bạn."** Trong câu này không biết **Tôi** và **bạn** là người nào, **nó** là cái gì, **ở đây** là ở đâu. Cần phải có ngữ cảnh phía trước để hiểu được những từ trên, nên các trực chỉ này làm cho mô hình dịch cho kết quả không tốt.

Ví dụ 2: "Chào mừng quý khách đến với Thảo cầm viên. Tôi là Minh, hướng dẫn viên của Thảo cầm viên. Tôi ở đây để giới thiệu về nó cho bạn." Với các ngữ cảnh cho trước thì chúng ta mới có thể hiểu được các trực chỉ trên nói về gì. **Tôi** chỉ **Minh**, **ở đây** và **nó** chỉ **Thảo cầm viên**, **bạn** chỉ **quý khách**

Hiện tượng tỉnh lược (Ellipsis)

Chúng ta thường lược bỏ từ hoặc cụm từ để tránh lặp đi lặp lại từ hoặc cụm từ đó. Và một số ví dụ gây khó khăn cho việc dịch từ tiếng Anh sang tiếng Việt:

Tỉnh lược những từ đứng cuối cụm danh từ

Ví dụ 3: **Do you want large cakes? No, I will have <u>small</u>.** Nếu không có ngữ cảnh, sẽ không hiểu được cái gì nhỏ (small).

Ví dụ 4: **My bike isn't working. I will use Minh's.** Nếu không có ngữ cảnh, sẽ không biết cái gì của Minh.

Ví dụ 5: **Which shoes are you going to wear? These.** Nếu không có ngữ cảnh, sẽ không hiểu được chúng (These) là cái gì.

Tỉnh lược những từ đứng cuối cụm động từ

Trong tiếng Anh, các trợ động từ thường được dùng một mình thay vì viết đầy đủ cả động từ.

Ví dụ 6: **She hasn't do homework. I haven't either.** Nếu không có ngữ cảnh, sẽ không biết chưa làm điều gì.

Ví dụ 7: **I was planning go to HaNoi next month. But I can't.** Nếu không có ngữ cảnh, sẽ không biết không thể làm cái gì.

Ví dụ 8: **I thought she would be happy. But she wasn't** Nếu không có ngữ cảnh, sẽ không biết cô ấy như thế nào.

Tỉnh lược động từ nguyên thể

Dùng **"to"** thay vì lặp lại cả cụm động từ nguyên thể.

Ví dụ 9: **Are you passing the exam? I hope to.** Nếu không có ngữ cảnh sẽ không biết được hi vọng điều gì.

Liên kết ngữ vựng (Lexical Cohension)

Có hai loại liên kết từ vựng đó là lặp đi lặp lại một từ để nhấn mạnh (reiteration) thực thể quan trọng và liên kết từ (collocation). Tuy nhiên, một từ có thể có nhiều bản dịch đúng khác nhau, điều này làm cho sai đi ý muốn lặp đi lặp lại của tác giả.

Ví dụ 10:

EN ***Mike are from USA. Jane are from USA.***
VI **Mike đến từ Mỹ. Jane đến từ Hoa Kì.**

Về cơ bản, bản dịch tiếng Việt vẫn đúng. Tuy nhiên, nếu trong trường hợp này người viết muốn nhấn mạnh bằng cách lặp đi lặp lại **USA** thì bản

dịch cũng nên được lặp đi lặp lại một trong hai từ **Mỹ** hoặc **Hoa Kì** thì tốt hơn.

Qua những ví dụ về những yếu tố gây ảnh hưởng đến chất lượng dịch khi dịch ở mức độ văn bản, đã giúp chúng ta thấy được sự cần thiết để sử dụng mô hình dịch máy mạng neural theo ngữ cảnh.

2.2.6 Giới thiệu dịch máy mạng neural theo ngữ cảnh

Ngữ cảnh bên ngoài câu hiện tại rất quan trọng đối với việc dịch máy (*Bawden và các cộng sự, 2018* [5]; *Läubli và các cộng sự, 2018* [16]; *Müller và các cộng sự ,2018* [23]; *Voita và các cộng sự, 2018* [36]; *Voita và các cộng sự, 2019b* [35]) cho thấy rằng nếu không có quyền truy cập vào ngữ cảnh cấp độ văn bản, NMT có khả năng không duy trì được sự nhất quán về từ vựng, thì, dấu chấm và dấu chấm lửng, giải quyết các đại từ đảo ngữ và các đặc điểm diễn ngôn khác. Hầu hết các mô hình NMT cấp độ văn bản hiện tại có thể được phân thành hai loại chính, mô hình NMT sử dụng ngữ cảnh và mô hình xử lý sau: Các mô hình xử lý sau giới thiệu một module bổ sung nhằm học cách tinh chỉnh các bản dịch được tạo ra bởi các hệ thống NMT bất khả tri theo ngữ cảnh để gắn kết diễn ngôn mạch lạc hơn *Xiong và các cộng sự, 2019* [39]; *Voita và các cộng sự, 2019a* [34]. Mặc dù kiểu tiếp cận này dễ triển khai, nhưng quá trình tạo hai giai đoạn có thể dẫn đến tích lũy lỗi. Trong khóa luận này, chúng tôi chủ yếu tập trung vào các mô hình theo ngữ cảnh (Context-aware Neural Machine Translation), trong khi các phương pháp xử lý sau có thể được kết hợp và tạo điều kiện cho bất kỳ kiến trúc NMT nào.

2.2.7 Lý thuyết dịch máy mạng neural theo ngữ cảnh

Nhiệm vụ của của các mô hình dịch máy mạng neural theo ngữ cảnh là dịch các văn bản (tập hợp của nhiều câu có liên quan đến nhau) giữa hai ngôn ngữ với nhau. Cụ thể, cho một tập ngữ liệu song ngữ ở cấp độ văn bản $D = \{D_1, D_2, ..., D_N\}$, N là số lượng văn bản trong

ngữ liệu. Mỗi văn bản là một chuỗi các câu nguồn và câu đích, $D_i = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(K)}, y^{(K)})\}$, $i \in \{1..N\}$, trong đó D_i là văn bản thứ i trong tập ngữ liệu, K là số lượng cặp câu song ngữ có trong văn bản D_i.

Cho một hệ thống dịch máy mạng neural q_θ tham số hóa với θ. Xác suất để dịch từ một câu nguồn $x^{(i)}$ sang câu đích $y^{(i)}$ có cho trước ngữ cảnh $C^{(i)}$ là:

$$q_\theta(y^{(i)}|x^{(i)}, C^{(i)}) = \prod_{t=1}^{T} q_\theta(y_t^{(i)}|x^{(i)}, y_{<t}^{(i)}, C^{(i)}) \tag{2.7}$$

Trong đó, $y_t^{(i)}$ đại diện cho token thứ t của câu $y^{(i)}$. Ngữ cảnh $C^{(i)}$ có thể có rất nhiều dạng. Có trường hợp không có ngữ cảnh được được đưa vào, tức là $C^{(i)} = \emptyset$, lúc này vấn đề sẽ quay trở lại mô hình dịch máy ở cấp độ câu. Và cũng sẽ có trường hợp, ngữ cảnh là tất các câu bên phía ngôn ngữ nguồn nguồn và tất cả các câu đã được tạo ra bên phía ngôn ngữ đích, tức là $C^{(i)} = \{x^{(1)}, ..., x^{(K)}, y^{(1)}, ..., y^{(i-1)}\}$.

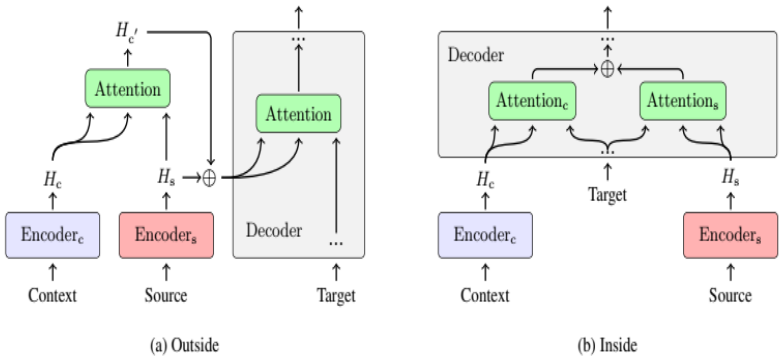
2.2.8 Cách tiếp cận mô hình dịch máy theo ngữ cảnh

Có hai cách chủ yếu để đưa thông tin ngữ cảnh vào hệ thống NMT. Đó là đơn mã hóa và đa mã hóa.

Tiếp cận đơn mã hóa

Đầu vào của hệ thống mã hóa đơn là sự kết hợp của các câu ngữ cảnh và câu hiện tại, token đặc biệt <CONCAT> được chèn vào để phân biệt chúng (*Tiedemann và Scherrer, 2017* [29]; *Agrawal và các cộng sự, 2018* [3]). Sau đó, câu ghép này được đưa vào mô hình Transformer tiêu chuẩn. Các hệ thống này tương đối đơn giản và dễ thực hiện nhưng có thể phải đối mặt với thách thức mã hóa đầu vào là cực kỳ dài, dẫn đến việc tính toán không hiệu quả.

Tiếp cận đa mã hóa



Hình 2.5: Tổng quan về hai hệ thống đa mã hóa

Hình 2.5 cho thấy hai phương pháp tích hợp ngữ cảnh vào NMT theo hướng tiếp cận đa mã hóa. Hầu hết các phương pháp tiếp cận đa mã hóa (Voita và các cộng sự, 2018 [36]; *Zhang và các cộng sự, 2018* [40]) đều hoạt động với cơ chế sau đây:

- Tích hợp bên ngoài (Outside Integration): Như được thể hiện trong hình 2.5a, ngữ cảnh và câu hiện tại được tổng hợp thành một vector đại diện mới thông qua một mạng lưới chú ý (Attention). Sau đó, vector này và câu nguồn hiện tại được hợp nhất tại một cổng tổng hợp.

- Tích hợp bên trong (Inside Integration): Bộ giải mã có thể tham gia vào hai bộ mã hóa tương ứng hình 2.5b. Sau đó cơ chế gating bên trong bộ giải mã được sử dụng để tạo ra vector tổng hợp. Một số các chiến lược đa mã hóa đã được thử nghiệm như sử dụng một đại diện của câu trước đó để khởi tạo bộ mã hóa chính hoặc bộ giải mã (*Wang và các cộng sự , 2017* [38]) và sử dụng nhiều cơ chế chú ý, với các chiến lược khác nhau để kết hợp các kết quả vectơ ngữ cảnh, chẳng hạn như liên kết (*Zoph và Knight, 2016* [41]), sự chú ý phân cấp (*Libovicky và Helcl, 2017* [18]) và gating (*Jean và các cộng sự, 2017* [10]).

Trong khóa luận này, chúng tôi thực hiện với cách tiếp cận tích hợp bên ngoài nên sẽ giải thích rõ hơn phần tích hợp bên ngoài. Các mô hình đa mã hóa lấy các câu xung quanh làm ngữ cảnh và sử dụng một mạng neural bổ sung để mã hóa ngữ cảnh, tức là chúng ta sẽ có một bộ mã hóa câu nguồn và một bộ mã hóa ngữ cảnh. Mô hình mã hóa câu trước bằng cách sử dụng một bộ mã hóa riêng biệt (với các tham số riêng biệt) để tạo ra một vectơ ngữ cảnh làm đầu vào bổ sung song song với câu hiện tại. Hai vectơ ngữ cảnh $c_i^{(1)}$ và $c_i^{(2)}$ thu được sau đó sẽ được kết hợp và trải qua biến đổi tuyến tính để tạo thành một vectơ ngữ cảnh duy nhất c_i có kích thước như ban đầu và được sử dụng để giải mã (*Zoph và Knight, 2016* [41]).

$$c_i = W_c[c_i^{(1)}; c_i^{(2)}] + b_c$$

Một cổng r_i giữa hai vectơ ngữ cảnh này được tạo để cung cấp tầm quan trọng khác nhau về các yếu tố trong mỗi vectơ ngữ cảnh (*Wang và các cộng sự, 2017* [38])

$$r_i = \tanh(W_r c_i^{(1)} + W_s c_i^{(2)} + b_r$$

$$c_i = r_i \odot (W_t c_i^{(1)}) + (1 - r_i) \odot (W_u c_i^{(2)})$$

Một cơ chế chú ý bổ sung (phân cấp) (*Libovicky và Helcl, 2017* [18]) để gán trọng số cho mỗi vectơ ngữ cảnh của bộ mã hóa (được thiết kế cho một số bộ mã hóa tùy ý).

$$e_i^{(k)} = \{v_b^T \tanh(W_{b^z(i-1)} + U_b^k c_i^{(k)}) + b_r\}$$

$$\beta_i^{(k)} = \frac{\exp\left(e_i^{(k)}\right)}{\sum_{k'=1}^{K} \exp\left(e_i^{(k')}\right)}$$

$$c_i = \sum_{k=1}^{K} \beta_i^{(k)} U_c^{(k)} c_i^{(k)}$$

2.2.9 Một số vấn đề khó khăn trong dịch máy mạng neural theo ngữ cảnh

Dịch máy mạng neural nói chung và dịch máy mạng neural theo ngữ cảnh nói riêng để đạt được kết quả cao đòi hỏi mô hình phải được huấn luyện trên một lượng dữ liệu song ngữ rất lớn. Tuy nhiên, dữ liệu song ngữ Anh-Việt ở cấp độ văn bản rất ít, hiện tại có tập dữ liệu IWSLT15 (khoảng 133000 cặp câu) và tập dữ liệu TED2020 v1 (khoảng 300000 cặp câu).

Và một thách thức nữa trong dịch máy mạng neural theo ngữ cảnh là hiện tại chưa có nhiều độ đo để có thể đo một cách tường minh lượng ngữ cảnh thực sự được sử dụng trong quá trình dịch. Và vấn đề này sẽ được nói rõ hơn trong chương 3

2.3 Đánh giá chất lượng dịch

Đánh giá chất lượng dịch dựa vào các chỉ số đánh giá tự động là công việc sử dụng chương trình máy tính để đánh giá độ tương đồng giữa bản dịch do máy thực hiện và bản dịch tham chiếu của con người. Trong phần này, chúng tôi giới thiệu về một số độ đo như Precision, Recall, F, BLEU và COMET.

Bản tham chiếu	UNIVERSITY OFFICIALS ARE RESPONSIBLE FOR STUDENTS SECURITY
Bản dịch máy 1	UNIVERSITY OFFICIALS RESPONSIBILITY OF STUDENTS SAFETY
Bản dịch máy 2	STUDENTS SECURITY UNIVERSITY OFFICIALS ARE RESPONSIBLE

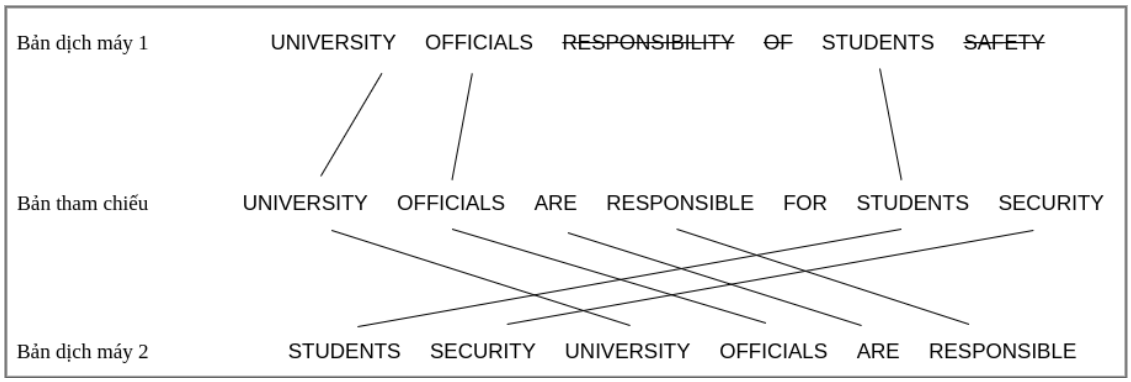
Bảng 2.1: Ví dụ để đánh giá chất lượng dịch

2.3.1 Precision, Recall và F

$$Precision = \frac{\text{số lượng từ dịch đúng}}{\text{tổng số từ trong bản máy dịch}} \tag{2.8}$$

$$Recall = \frac{\text{số lượng từ dịch đúng}}{\text{tổng số từ trong bản tham khảo}} \tag{2.9}$$

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.10}$$



Hình 2.6: Biểu diễn ví dụ bảng 2.1

Độ đo	Bản dịch máy 1	Bản dịch máy 2
Precision	$3/6 = 0.5$	$6/6 = 1$
Recall	$3/7 = 0.43$	$6/7 = 0.86$
F	$(2 * 0.5 * 0.43)/(0.5 + 0.3) = 0.46$	$(2 * 1 * 0.86)(1 + 0.86) = 0.92$

Bảng 2.2: Precision, Recall, F cho ví dụ hình 2.6

2.3.2 BLEU

Độ đo BLEU (Bilingual Evaluation Understudy), được đề xuất bởi *Papineni và các cộng sự, 2002* [24]. BLEU tính $Precision_n$ cho n-grams (kích thước từ 1 đến 4). Thêm brevity penalty cho những bản dịch quá ngắn. BLEU có công thức như sau:

$$\text{BLEU} = \text{brevity penaty} * (\prod_{i=1}^{4} Precision_i)^{1/4} \tag{2.11}$$

$$\text{brevity penaty} = \min(1, \frac{\text{tổng số từ trong bản máy dịch}}{\text{tổng số từ trong bản tham khảo}})$$

Độ đo BLEU chủ yếu là sử dụng cho toàn bộ ngữ liệu, không phải câu riêng lẻ.

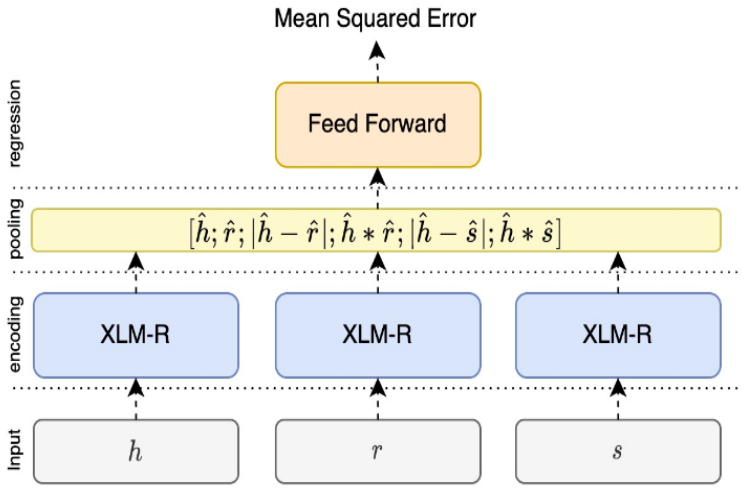
Độ đo	Bản dịch máy 1	Bản dịch máy 2
Precision (1-gram)	3/6	6/6
Precision (2-gram)	1/5	4/5
Precision (3-gram)	0/4	2/4
Precision (4-gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0	0.52

Bảng 2.3: BLEU cho ví dụ ở bảng 2.1

2.3.3 COMET

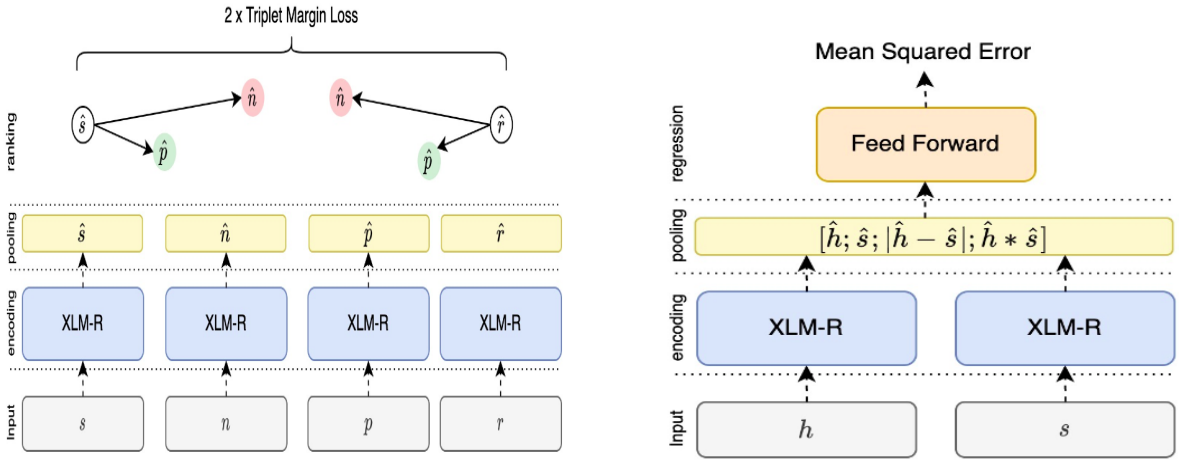
COMET là một neural framework để đánh giá chất lượng dịch của các mô hình dịch của nhiều ngôn ngữ. Theo hình 2.7, các câu h (hypothesis là các câu do được dịch bởi các mô hình dịch), r (reference là các câu tham chiếu), s (source là các câu nguồn) sẽ được mã hóa độc lập bằng XML-R của *Conneau và các cộng sự, 2019* [15]. Sau đó các vector biểu diễn từ sẽ truyền cho lớp pooling để tạo ra các vector biểu diễn câu, các vector biểu diễn câu sẽ được nối lại với nhau thành một vector duy nhất để truyền

vào lớp lan truyền tới (Feed-Forward). Sau đó, toàn bộ mô hình sẽ được tối thiểu hóa với "Sai số toàn phương trung bình" (Mean Square Error - MSE).



Hình 2.7: Kiến trúc mô hình COMET [32]

Ngoài ra, COMET còn một số kiến trúc mô hình khác như hình 2.8, nhưng kiên trúc được giới thiệu ở hình 2.7 cho kết quả tốt nhất và được sử dụng trong khóa luận này.



(a) (b)

Hình 2.8: Kiến trúc khác của COMET [32]

2.3.4 Đánh giá các độ đo

Các chỉ số đánh giá tự động có lợi ích là chi phí thấp, có thể tinh chỉnh và nhất quán. Tuy nhiên, để các độ đo này có được độ "tin tưởng" cao thì cần có sự tương quan giữa chúng với sự đánh giá của con người. Và một trong các hệ số đánh giá sự tương quan đó là hệ số Pearson. Cho hai biến x là điểm tự động, y là sự đánh giá của con người. Từ đó ta có tập hợp các điểm số $\{(x_1; y_1), (x_2; y_2), ...\}$. Hệ số tương quan Pearson được tính như sau:

$$r_{xy} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{(n-1) * s_x * s_y} \tag{2.12}$$

Trong đó:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} (x_i)$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} (y_i)$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1} n(x_i - \overline{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1} n(y_i - \overline{y})^2$$

Các chỉ số tự động là công cụ cần thiết cho sự phát triển của hệ thống dịch máy. Tuy nhiên, các chỉ số đánh giá này vẫn còn mở ra nhiều thách thức trong lĩnh vực dịch máy.

Chương 3

Lý thuyết

Đo lường và tăng cường mức độ sử dụng ngữ cảnh có ý nghĩa quan trọng trong việc cải thiện chất lượng dịch của các mô hình dịch máy mạng neural. Để có cái nhìn rõ ràng hơn về việc đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural, trong chương này sẽ trình bày tổng quan về vấn đề đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural tại phần 3.1, độ đo "Thông tin tương hỗ chéo có điều kiện" tại phần 3.1.3 và sử dụng "Loại bỏ từ theo ngữ cảnh" tại phần 3.2 để tăng cường mức độ sử dụng ngữ cảnh của mô hình.

3.1 Đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural

3.1.1 Tổng quan về đo lường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural

Các mô hình dịch máy mạng neural theo ngữ cảnh ra đời nhằm cho phép mô hình sử dụng ngữ cảnh trong quá trình dịch, tuy nhiên các mô hình dịch máy mạng neural theo ngữ cảnh hiện nay không đảm bảo thông tin ngữ cảnh được sử dụng một cách chính xác (mô hình có thể chỉ dựa vào câu nguồn hiện tại đang được dịch hoặc các từ đã được giới thiệu trước đó

trong cùng câu đích để tạo đầu ra mà không phải sử dụng hoàn toàn ngữ cảnh được cung cấp)

Chưa có nhiều phương pháp dùng để đánh giá khả năng của các mô hình dịch máy mạng neural trong việc nắm bắt ngữ cảnh. Trong khi các mô hình NMT cấp độ văn bản có thể được so sánh, đánh giá bằng cách sử dụng các thước đo hiệu suất tự động như BLEU (*Papineni và các cộng sự, 2002* [24]), các độ đo này không cung cấp một bức tranh rõ ràng về việc liệu lý do mô hình đang hoạt động tốt hơn là từ những cải thiện trong việc xử lý, tích hợp ngữ cảnh vào mô hình hay do lý do khác. Một phương pháp phổ biến khác là "Đánh giá mâu thuẫn" (Contrastive evaluation), trong đó các mô hình được đánh giá dựa trên khả năng phân biệt các bản dịch chính xác với các bản dịch tương phản, phương pháp này thường được dùng để đánh giá về khả năng trong việc nắm bắt các hiện tượng diễn ngôn cụ thể theo ngữ cảnh liên tâm chẳng hạn như từ trung tính (*Muller và các cộng sự , 2018* [23]) và sự gắn kết từ vựng của mô hình, (*Bawden và các cộng sự, 2018* [5]). Tuy nhiên phương pháp này chỉ gián tiếp cung cấp thước đo về việc sử dụng ngữ cảnh của mô hình đối với một số hiện tượng cụ thể và có thể không nắm bắt và sử dụng được cho các hiện tượng khác trong dịch máy, đồng thời phương pháp đánh giá này không chỉ ra được liệu mô hình có thể đang sử dụng ngữ cảnh trong quá trình dịch hay không. Và phương pháp này không cho thấy được sự khả thi trong dịch máy mạng neural Anh-Việt khi không có tập dữ liệu để đánh giá giống các song ngữ khác như tập ContraPro của *Müller và các cộng sự, 2018* [23] cho song ngữ Anh Đức.

Kim và các cộng sự, 2019 [12] cho thấy rằng hầu hết các cải tiến đối với chất lượng dịch thuật trong dịch máy mạng neural là do các yếu tố phi ngữ cảnh, như đưa vào một bộ điều chỉnh là bộ mã hóa hoặc giải mã đóng vai trò như là yếu tố gây nhiễu thông tin, và nếu thay thế bộ mã hóa ngữ cảnh bằng một bộ gây nhiễu khác như Gauss thì vẫn thu được kết quả tương đương. Điều này làm cho vấn đề định lượng chính xác lượng thông tin ngữ cảnh mà mô hình sử dụng ngày càng khó khăn hơn bởi thực tế ta

không có định nghĩa rõ ràng về ngữ cảnh được sử dụng là gì, thông tin nào trong ngữ cảnh bổ sung là cần thiết cho mô hình.

Khắc phục các hạn chế của các phương pháp nêu trên, *Patrick Fernandes và các cộng sự, 2021* [7] đã giới thiệu độ đo "Thông tin tương hỗ chéo có điều kiện" để có thể đo lường một cách tường mình lượng ngữ cảnh mô hình sử dụng trong quá trình dịch.

3.1.2 Các cơ sở lý thuyết liên quan

Trước khi tìm hiểu về độ đo "Thông tin tương hỗ chéo có điều kiện", chúng ta cần tìm hiểu rõ kiến thức về *entropy*, *cross-entropy* và độ đo *"Thông tin tương hỗ" (Mutual Information - MI).*

Entropy

Nếu ta có $\mathbb{X} = \{x_1, ..., x_n\}$ là các thông tin mà biến ngẫu nhiên X có thể nhận và $p(x_i)$ là xác suất X nhận giá trị x_i với $i \in \{1..n\}$. Entropy được định nghĩa như sau:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i)\mathrm{log}p(x_i) \tag{3.1}$$

Entropy càng cao thì càng khó rút ra kết luận nào từ thông tin đó. Nói cách khác, entropy là độ đo về sự khó đoán của thông tin.
Ví dụ, khi tung đồng xu có phân phối xác suất $(sp, nga) = (p_1, p_2) = (0.5, 0.5)$ lúc này $p_1 = p_2 = 0.5$ thì Entropy sẽ cực đại (Entropy = 1), thì sẽ rất khó đoán lần tung tiếp theo sẽ là sấp hay ngửa. Tuy nhiên, khi $(p_1, p_2) = (0.2, 0.8)$ thì lúc này Entropy sẽ thấp hơn rất nhiều và có thể dễ dàng dự đoán lần tiếp theo là ngửa.

Cross-Entropy

Nếu ta có $\mathbb{X} = \{x_1, ..., x_n\}$ là các thông tin mà biến ngẫu nhiên X có thể nhận và $p(x_i)$ và $q(x_i)$ là xác suất X nhận giá trị x_i với $i \in \{1..n\}$.Khi đó ta có phân phối xác suất $P = \{p_1, ..., p_n\}$ và $Q = \{q_1, ..., q_n\}$. Cross-entropy được định nghĩa như sau:

$$\mathrm{H}(P, Q) = -\sum_{i=1}^{n} p(x_i)\log q(x_i) \tag{3.2}$$

Cross-entropy đạt giá trị cực tiểu khi $P = Q$. Giá trị của p_i càng khác với q_i thì Cross-entropy càng tăng. Dựa vào tính chất này để áp dụng vào việc tối ưu hóa các mô hình trong các bài toán học máy (để có đầu ra càng giống mục tiêu càng tốt).

Thông tin tương hỗ

Thông tin tương hỗ (Mutual Information - MI) là một thước đo sự phụ thuộc lẫn nhau giữa hai biến ngẫu nhiên. Nó giúp định lượng "lượng thông tin" thu được về một biến ngẫu nhiên bằng cách quan sát biến ngẫu nhiên kia. Khái niệm thông tin tương hỗ được liên kết mật thiết với khái niệm entropy của một biến ngẫu nhiên. MI giữa hai biến ngẫu nhiên X và Y có thể được phát biểu như sau:

$$\mathrm{MI}(X, Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) \tag{3.3}$$

Trong đó MI là thông tin tương hỗ của X và Y, $\mathrm{H}(Y)$ là entropy của Y, $\mathrm{H}(Y|X)$ là entropy có điều kiện của Y cho trước X. Trực quan, nếu entropy $\mathrm{H}(Y)$ được coi là thước đo độ không chắn của một biến ngẫu nhiên, khi đó $\mathrm{H}(Y|X)$ là thước đo về những gì mà X không cung cấp về Y hay chính là "lượng không chắc chắn còn lại về Y" sau khi X được biết. Điều này chứng thực ý nghĩa trực quan của "Thông tin tương hỗ" là lượng thông tin mà việc biết đến một trong hai biến sẽ cung cấp cho biến kia.

3.1.3 Độ đo "Thông tin tương hỗ chéo có điều kiện"

Bugliarello và các cộng sự, 2020 [6] đã đề xuất độ đo "Thông tin tương hỗ chéo" để đo "độ khó" của việc dịch giữa hai ngôn ngữ ở mức câu trong dịch máy mạng neural. Cho một mô mình ngôn ngữ q_{LM} cho câu đích Y và một mô hình dịch q_{MT} để dịch từ câu nguồn X sang câu đích Y, "Thông tin tương hỗ chéo" được tính như sau:

$$\mathrm{XMI}(X \to Y) = \mathrm{H}_{q_{LM}}(Y) - \mathrm{H}_{q_{MT}}(Y|X) \tag{3.4}$$

Trong đó $\mathrm{H}_{q_{LM}}$ là cross-entropy của câu đích Y dưới mô hình ngôn ngữ q_{LM}, q_{MT} là cross-entropy có điều kiện của câu đích Y cho trước X dưới mô hình dịch q_{MT}. Công thức này cho chúng ta biết về lượng thông tin mà câu nguồn cung cấp về câu đích. Dựa trên ý tưởng của độ đo "Thông tin tương hỗ chéo" (Cross-Mutual Information - XMI) được đề xuất bởi *Bugliarello và các cộng sự, 2020* [6] để đo "độ khó" của việc dịch giữa hai ngôn ngữ ở mức độ câu trong dịch máy mạng neural, *Patrick Fernandes và các cộng sự, 2021* [7] đã đề xuất độ đo "Thông tin tương hỗ chéo có điều kiện" (Conditional Cross-Mutual Information - CXMI), đây là một độ đo mới để đánh giá sự ảnh hưởng của ngữ cảnh đến sự dự đoán của mô hình. Sự khác biệt của "Thông tin tương hỗ chéo có điều kiện" (CXMI) so với "Thông tin tương hỗ chéo" (XMI) là chúng ta sẽ bổ sung thêm một biến ngữ cảnh C và đo lượng thông tin mà ngữ cảnh C cung cấp cho câu đích Y khi cho biết trước câu nguồn X.
Công thức tính "Thông tin tương hỗ chéo có điều kiện" như sau:

$$\mathrm{CXMI}(C \to Y|X) = \mathrm{H}_{q_{MT_A}}(Y|X) - \mathrm{H}_{q_{MT_C}}(Y|X, C) \tag{3.5}$$

Trong đó $\mathrm{H}_{q_{MT_A}}$ là entropy của mô hình dịch máy không theo ngữ cảnh và $\mathrm{H}_{q_{MT_C}}$ là entropy của mô hình dịch máy theo ngữ cảnh.

Gọi S là một biến ngẫu nhiên trên câu nguồn, T là một biến ngẫu nhiên trên câu đích, C là một biến ngẫu nhiên trên ngữ cảnh. Giả sử các biến

ngẫu nhiên này có một phân phối "đúng" là $p(\mathbf{s,t,c})$. Cross-entropy giữa phân phối "đúng" p và xác suất dịch của mô hình dịch máy mạng neural theo ngữ cảnh $q_{MT_C}(\mathbf{t}|\mathbf{s,c})$được tính như sau:

$$\mathrm{H}_{q_{MT_C}}(T|S,C) = -\sum_{\mathbf{s}\in V_S^*}\sum_{\mathbf{t}\in V_T^*}\sum_{\mathbf{c}\in V_C^*} p(\mathbf{s,t,c})\log q_{MT_C}(\mathbf{s,t,c}) \tag{3.6}$$

Trong đó V_S^*, V_T^*, V_C^* lần lượt là không gian có thể của câu nguồn, câu đích và ngữ cảnh. Vì chúng ta không thể tính toán chính xác cụ thể được phân phối "đúng" p. Tuy nhiên, khi có một mẫu dữ liệu $\{(\mathbf{s}^{(i)}, \mathbf{t}^{(i)}, \mathbf{c}^{(i)})\}_{i=0}^{N}$, chúng ta có thể tính phân phối p bằng cách sử dụng mô phỏng Monte Carlo, lúc này công thức 3.6 trở thành:

$$\mathrm{H}_{q_{MT_C}}(T|S,C) \approx -\frac{1}{N}\sum_{i=0}^{N}\log q_{MT_C}(\mathbf{s}^{(i)}, \mathbf{t}^{(i)}, \mathbf{c}^{(i)}) \tag{3.7}$$

Nếu chúng ta cân nhắc $p(\mathbf{s,t}) = \sum_{\mathbf{c}\in V_C^*} p(\mathbf{s,t,c})$, chúng ta cũng có thể tính cho mô hình dịch máy không theo ngữ cảnh như sau:

$$\mathrm{H}_{q_{MT_A}}(T|S) \approx -\frac{1}{N}\sum_{i=0}^{N}\log q_{MT_A}(\mathbf{s}^{(i)}, \mathbf{t}^{(i)}) \tag{3.8}$$

Từ 3.5, 3.7, 3.8, ta được công thức sau:

$$\mathrm{CXMI}(C \to Y|X) \approx -\frac{1}{N}\sum_{i=1}^{N}\log\frac{q_{MT_A}(y^{(i)}|x^{(i)})}{q_{MT_C}(y^{(i)}|x^{(i)}, C^{(i)})} \tag{3.9}$$

Theo lý thuyết, q_{MT_C} và q_{MT_A} có thể là bất kì mô hình nào. Tuy nhiên,các tác giả nhận thấy nếu sử dụng hai mô hình khác nhau cho q_{MT_C} và q_{MT_A} sẽ gây ra những khác biệt trong việc tính toán. Vì vậy, việc đào tạo một mô hình duy nhất q_{MT} vừa có khả năng dịch như mô hình bất khả tri ngữ cảnh và cả mô hình sử dụng ngữ cảnh sẽ đảm bảo rằng các giá trị CXMI khác 0 là do bối cảnh mà không phải các yếu tố khác.

3.2 Tăng cường mức độ sử dụng ngữ cảnh trong dịch máy mạng neural

3.2.1 Giới thiệu

Với động lực từ hạn chế trong việc sử dụng ngữ cảnh của mô hình huấn luyện. Câu hỏi đặt ra lúc này là "Liệu có thể điều chỉnh phương pháp huấn luyện để tăng cường lượng ngữ cảnh được sử dụng của mô hình?". Dựa vào kĩ thuật bình thường hóa được sử dụng phổ biến trong dịch máy của mức độ câu của *Sennrich và các cộng sự, 2016* [27] đó là "Loại bỏ từ" (Word Droput). *Patrick Fernandes và các cộng sự, 2021* [7] đã ứng dụng vào trong mô hình dịch máy có ngữ cảnh và tạo ra "Loại bỏ từ theo ngữ cảnh" (Context-aware Word Dropout - CoWord Dropout).

3.2.2 Cơ sở lý thuyết liên quan

Trước tiên chúng ta cần tìm hiểu rõ kiến thức về phân phối Bernoulli để phục vụ cho việc tìm hiểu kiến thức về "Loại bỏ từ theo ngữ cảnh" sau này.

Phân phối Bernoulli

Phân phối Bernoulli là một phân phối xác suất rời rạc của biến ngẫu nhiên chỉ nhận hai giá trị là 0 và 1, trong đó giá trị 1 đạt được với xác suất p (gọi là xác suất thành công) và giá trị 0 đạt được với xác suất $1-p$ (gọi là xác suất thất bại). Nếu X là một biến ngẫu nhiên với phân phối Bernoulli, kí hiệu $X \sim Bernoulli(p)$, ta sẽ có:

$$P(X = 1) = p$$

$$P(X = 0) = 1 - P(X = 1) = 1 - p$$

Một ví dụ điển hình của phân phối Bernoulli, gọi X là kết quả của việc tung đồng xu (có thể đồng chất hoặc không), đồng xu xuất hiện mặt ngửa tương ứng với giá trị $X = 1$, đồng xu xuất hiện mặt sấp tương ứng với giá trị $X = 0$. Đồng xu xuất hiện mặt ngửa với xác suất p và mặt sấp với xác suất $1 - p$. Khi đó, ta có hàm xác suất của phần phối Bernoulli như sau:

$$f(k) = P(X = k) = \begin{cases} p & \text{nếu } k = 1 \\ 1 - p & \text{nếu } k = 0 \end{cases}$$

Hoặc:

$$f(k) = p^k(1 - p)^{1-k}, \text{ k} \in \{0, 1\}$$

3.2.3 Loại bỏ từ theo ngữ cảnh

Sau khi đã tìm hiểu rõ kiến thức về phân phối Bernoulli, chúng ta sẽ tìm hiểu về lý thuyết của "Loại bỏ từ theo ngữ cảnh". Ý tưởng chính của "Loại bỏ từ theo ngữ cảnh" là mô hình hóa xác suất dịch giữa câu nguồn $x^{(i)}$ và câu đích $y^{(i)}$ như sau:

$$p_\theta(y^{(i)}|x^{(i)}) = \prod_{t=1}^{T} p_\theta(y_t^{(i)}|\tilde{x}^{(i)}, y_{<t}^{(i)}, C^{(i)}) \tag{3.10}$$

Trong đó, $\tilde{x}^{(i)}$ là một phiên bản bị gây nhiễu của câu nguồn $x^{(i)}$ hiện tại, được tạo ra bằng cách loại bỏ ngẫu nhiên các token và thay thế chúng bằng một mask token với xác suất p cho trước như sau:

$$r_t^{(i)} \sim Bernoulli(p) \tag{3.11}$$

$$\tilde{x}_t^{(i)} = \begin{cases} < MASK > & \text{nếu } r_t^{(i)} = 1 \\ x_t^{(i)} & \text{nếu ngược lại} \end{cases} \tag{3.12}$$

Trong trường hợp không có ngữ cảnh được đưa vào mô hình thì "Loại bỏ từ theo ngữ cảnh" sẽ quay trở về "Loại bỏ từ" tương tự như ở mức độ câu.

Chương 4

Thực nghiệm

4.1 Dữ liệu

Chúng tôi thực hiện các thực nghiệm ở mức độ văn bản trên bộ dữ liệu IWSLT15 Anh-Việt (khoảng 133000 cặp câu), sử dụng các tập tst2010, tst2011, tst2012 làm tập xác nhận (validate) và tập tst2013 làm tập thử nghiệm.

Đối với các tập tiếng Việt, chúng tôi tiến hành tách từ thông qua "Bộ công cụ xử lý ngôn ngữ tự nhiên Tiếng Việt" VnCoreNLP của *Vu và các cộng sự, 2018* [37].

Toàn bộ dữ liệu sẽ được mã hóa/vector hóa với mã hóa cặp byte (Byte Pair Encoding) của *Sennrich và các cộng sự, 2016* [28] thông qua sử dụng framework SentencePiece của *Kudo và Richardson, 2018* [14]. Trong khóa luận này, chúng tôi sẽ dụng kích thước từ vựng là 20000 từ chia sẻ giữa nguồn và đích.

4.1.1 Mã hóa cặp byte (Byte Pair Encoding - BPE)

Theo [2], các không gian vector được huấn luyện từ Word2Vec hay Glove là những cách biểu diễn từ phổ biến được sử dụng nhiều trong lĩnh vực Xử lý ngôn ngữ tự nhiên. Trong nhiều năm, chúng là cách biểu diễn đáng tin cậy trong việc huấn luyện các mô hình học máy trong xử lý ngôn

ngữ tự nhiên khi không có nhiều dữ liệu. Mặc dù vậy, chúng không phải là công cụ toàn năng khi đối mặt với những từ hiếm xuất hiện. Các từ này được thay thế bởi tokens <unk> khi cài đặt mô hình.

Để giải quyết vấn đề từ hiếm, các nhà nghiên cứu tại Đại học Edinburgh phát triển phương pháp sử dụng "phụ từ" (subword) với mã hóa cặp byte (Byte Pair Encoding).

Theo [2], Mã hóa cặp byte là một thuật toán nén dữ liệu hoạt động bằng cách thay thế các cặp byte liên tiếp có tần suất lớn bằng một byte không tồn tại trong dữ liệu. Để thực hiện mã hóa sử dụng phụ từ (subword), mã hóa cặp byte đã sửa đổi lại. Các cặp phụ từ (subword) thường xuyên xuất hiện được hợp nhất với nhau chứ không bị thay thế. Thông qua việc biểu diễn các từ bằng các ký tự, mã hóa cặp byte sẽ thống kê toàn bộ các cặp kí tự xuất hiện nhiều nhất và bổ sung phụ từ (subword) và tập từ vựng. Quá trình này lặp lại liên tục đến khi kích thước tập từ vựng đạt được kích thước mà chúng ta mong muốn.

Ví dụ tại [11]: Giả sử chúng ta có một ngữ liệu gồm các từ **old, older, highest, lowest**. Và chúng ta đếm tần suất xuất hiện của những từ này trong ngữ liệu như sau:

{"old":7, "older":3, "finest":9, "lowest":4}

Tiếp theo, thêm một token kết thúc **"</w>"** vào cuối mỗi từ để thuật toán có thể biết được giới hạn kết thúc của từ:

{"old</w>":7, "older</w>":3, "finest</w>":9, "lowest</w>":4}

Tiếp theo, chúng ta sẽ tách từng kí tự và đếm số lần xuất hiện của chúng như bảng 4.1 sau:

Token (T)	<\w>	o	l	d	e	r	f	i	n	s	t	w
Tần suất (F)	23	14	14	10	16	3	9	9	9	13	13	4

Bảng 4.1: Tần suất xuất hiện của các token (1)

Tiếp theo, sẽ hợp cặp byte xuất hiện nhiều nhất trong ngữ liệu đó là **e** và **s** với tần suất xuất hiện là 13. Bây giờ, sẽ hợp thành token **es** với tần suất là 13, giảm tần suất của token **e** và **s** đi 13 lần, được bảng 4.2 bên

dưới:

T	<\w>	o	l	d	e	r	f	i	n	s	t	w	es
F	23	14	14	10	16-13 =3	3	9	9	9	13-13 =0	13	4	13

Bảng 4.2: Tần suất xuất hiện của các token (2)

Tương tự như phía trên, lúc này ta thấy tần suất xuất hiện của cặp byte **es** và **t** là cao nhất với 13 lần. Lúc này, chúng ta sẽ bắt đầu hợp nhất **es** và **t** thành **est** với tần suất xuất hiện là 13, đồng thời trừ đi tần suất xuất hiện của **es** và **t** đi 13 lần, ta được kết quả như bảng 4.3 bên dưới:

T	<\w>	o	l	d	e	r	f	i	n	s	t	w	es	est
F	23	14	14	10	3	3	9	9	9	0	13-13 =0	4	13-13 =0	13

Bảng 4.3: Tần suất xuất hiện của các token (3)

Tiếp theo, chúng ta thấy cặp byte **est** và **</w>** xuất hiện nhiều nhất với 13 lần. Lúc này chúng ta hợp **est** và **</w>** thành **est</w>** với tần suất là 13, ta được bảng 4.4 bên dưới:

T	<\w>	o	l	d	e	r	f	i	n	s	t	w	es	est	est<\w>
F	23-13 =10	14	14	10	3	3	9	9	9	0	0	4	0	13-13 =0	13

Bảng 4.4: Tần suất xuất hiện của các token (4)

Sau đó, có token có tần suất là 0 sẽ được số khỏi bảng như bảng 4.5. Lúc này, **</w>** thông báo cho biết kết thúc của từ để chuyển sang từ khác:

T	<\w>	o	l	d	e	r	f	i	n	w	est<\w>
F	10	14	14	10	3	3	9	9	9	4	13

Bảng 4.5: Tần suất xuất hiện của các token (5)

Tiếp tục cho từ khác, thực hiện cho **o** là **l**, tạo thêm token **ol** với tần suất là 10 và trừ tần suất của **o** và l đi 10, ta được bảng 4.6:

T	<\w>	o	l	d	e	r	f	i	n	w	est<\w>	ol
F	10	14-10 =4	14-10 =4	10	3	3	9	9	9	4	13	10

Bảng 4.6: Tần suất xuất hiện của các token (6)

Tiếp tục, thực hiện cho **ol** và **d**, tạo ra token **old** với tần suất là 10, đồng thời trừ đi tần suất của **ol** và **d** đi 10, ta được bảng 4.7 bên dưới:

T	<\w>	o	l	d	e	r	f	i	n	w	est<\w>	ol	old
F	10	4	4	10-10 =0	3	3	9	9	9	4	13	10-10 =0	10

Bảng 4.7: Tần suất xuất hiện của các token (7)

Chúng ta, có thể thấy rằng **f**, **i**, **n** có tần suất là 9, nhưng chúng ta không hợp chúng lại vì chúng chỉ có một từ với các ký từ này trong ngữ liệu. Tiếp tục, chúng ta làm lần lượt như vậy đến khi nào đủ kích thước từ vựng mà chúng ta mong muốn rồi dừng lại. Trong ví dụ này sẽ dừng lại ở đây.

Sau đó, loại bỏ các token có tần suất là 0 ra khỏi bảng, ta được bảng 4.8:

T	<\w>	o	l	e	r	f	i	n	w	est<\w>	old
F	10	4	4	3	3	9	9	9	4	13	10

Bảng 4.8: Tần suất xuất hiện của các token (8)

4.2 Mô hình và tối ưu hóa

Chúng tôi huấn luyện mô hình theo kiến trúc Transformer có kích thước ẩn (Hidden Size) là 512, kích thước lan truyền tới (Feed Foward Size) là 2048, 6 lớp (Layer), 8 đầu chú ý (8 Attention Head). Sử dụng trình tối ưu hóa Adam với $\beta_1 = 0.9$ và $\beta_2 = 0.98$ và sử dụng bộ lập tỷ lệ học căn bậc hai nghịch đảo (inverse square root learning rate scheduler) với giá trị bắt đầu là 10^{-4}, khởi động tuyến tính (linear warm-up) trong 4000 bước đầu tiên. Mô hình tự động dừng sớm trong quá trình huấn luyện với validation perplexity.

4.3 Thực nghiệm với độ đo "Thông tin tương hỗ chéo có điều kiện"

Để đánh giá tầm quan trọng tương đối của các kích thước ngữ cảnh khác nhau ở cả phía nguồn và phía đích, chúng tôi thực hiện hai mô hình, một cho phía nguồn và một cho phía đích. Mô hình nhận ngữ cảnh C có kích thước k, ngữ cảnh cho phía nguồn là $C_{(i)} = \{x^{(i-k)}, ..., x^{(i-1)}\}$ và ngữ cảnh cho phía đích là $C_{(i)} = \{y^{(i-k)}, ..., y^{(i-1)}\}$. Trong thực nghiệm này k sẽ nằm trong khoảng từ 1 đến 4.

Mô hình sẽ sử dụng ngữ cảnh theo hướng đơn mã hóa, kết hợp của các câu ngữ cảnh và câu hiện tại, token đặc biệt <CONCAT> được chèn vào để phân biệt chúng như của *Tiedemann và Scherrer, 2017* [29]. Sau đó, cho vào mô hình Transformer với các thông số như ở phần 4.2.

4.4 Thực nghiệm với "Loại bỏ từ theo ngữ cảnh"

4.4.1 Đánh giá mức độ sử dụng ngữ cảnh sau khi sử dụng "Loại bỏ từ theo ngữ cảnh"

Tương tự như phần thực nghiệm trong phần 4.3, khóa luận tiếp tục thực hiện huấn luyện mô hình *Transformer* trên tập dữ liệu IWSLT15 Anh-Việt với thông số huấn luyện như 4.2.

Tuy nhiên, dựa vào kết quả có được từ thực nghiệm ở phần 4.3 cho thấy rằng việc sử dụng ngữ cảnh ở phía ngôn ngữ nguồn không mang lại kết quả tốt, nhóm cân nhắc chỉ huấn luyện với ngữ cảnh phía ngôn ngữ đích thay vì huấn luyện trên cả hai phía ngôn ngữ như trước đó với kích thước ngữ cảnh k từ 1 đến 4. Đồng thời, nhóm sẽ kết hợp với "Loại bỏ từ theo ngữ cảnh", xác suất loại bỏ p lần lượt 0.0, 0.1, 0.2 và 0.3 để huấn luyện mô hình.

Sau đó, nhóm sẽ dùng độ đo "Thông tin tương hỗ chéo có điều kiện" để đánh giá tính hiệu quả của "Loại bỏ từ theo ngữ cảnh" trong việc tăng cường mức độ sử dụng ngữ cảnh.

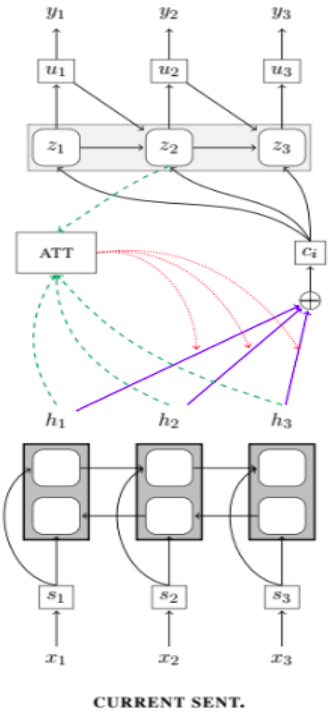
4.4.2 Đánh giá chất lượng dịch của mô hình sau khi sử dụng "Loại bỏ từ theo ngữ cảnh"

Để đáng giá việc tăng cường mức độ sử dụng ngữ cảnh của "Loại bỏ từ theo ngữ cảnh" có mối liên hệ như thế nào đến việc cải thiện chất lượng dịch của mô hình. Dựa vào những thực nghiệm ở chương 4.3 và 4.4.1, nhóm thực hiện huấn luyện các mô hình với kích thước ngữ cảnh cố định, sử dụng cả đơn mã hóa và đa mã hóa cho ngữ cảnh như sau:

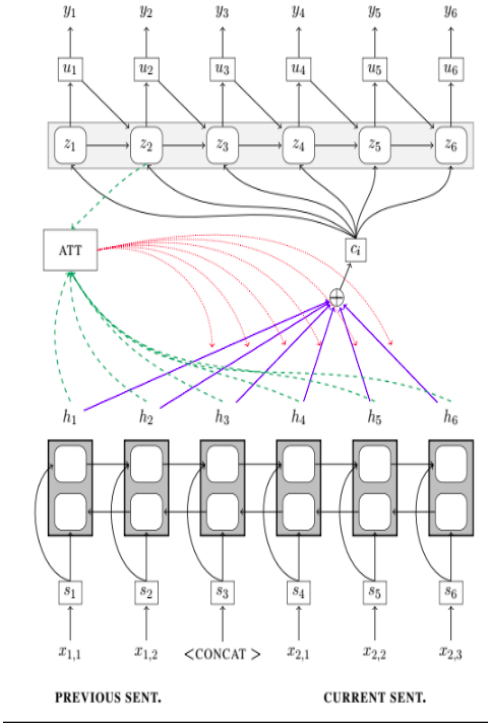
- Chúng tôi huấn luyện 3 mô hình đơn mã hóa, một mô hình cơ sở (baseline) (Hình 4.1a) và hai mô hình sử dụng ngữ cảnh (Hình 4.1b), cụ thể như sau:
  - Mô hình **baseline** là mô hình cơ sở dịch các câu độc lập với nhau hay nói cách khác, ngữ cảnh bằng không $C^{(i)} = \emptyset$.
  - Mô hình **one-to-two** là mô hình có ngữ cảnh là một câu phía trước của câu đang được dịch bên phía ngôn ngữ đích, tức là $C^{(i)} = \{y^{(i-1)}\}$.
  - Mô hình **two-to-two** là mô hình có ngữ cảnh là một câu phía trước của câu đang được dịch bên phía ngôn ngữ nguồn và một câu phía trước của câu đang được dịch bên phía ngôn ngữ đích, tức là $C^{(i)} = \{x^{(i-1)}, y^{(i-1)}\}$.
- Để khám phá ra lợi ích của "Loại bỏ từ theo ngữ cảnh" trong nhiều mô hình kiến trúc khác nhau, nhóm huấn luyện mô hình đa mã hóa (Hình 4.1c) **one-to-two multi-encoder** theo hướng tích hợp bên ngoài. Ngữ cảnh sẽ được mã hóa bằng một bộ mã hóa *Transformer* riêng biệt, sau đó được tính hợp vào mô hình.

Lưu ý, với tất cả những mô hình có sử dụng ngữ cảnh bên phía ngôn ngữ đích, sử dụng những câu được giải mã trước đó bên phía ngôn ngữ đích làm ngữ cảnh.

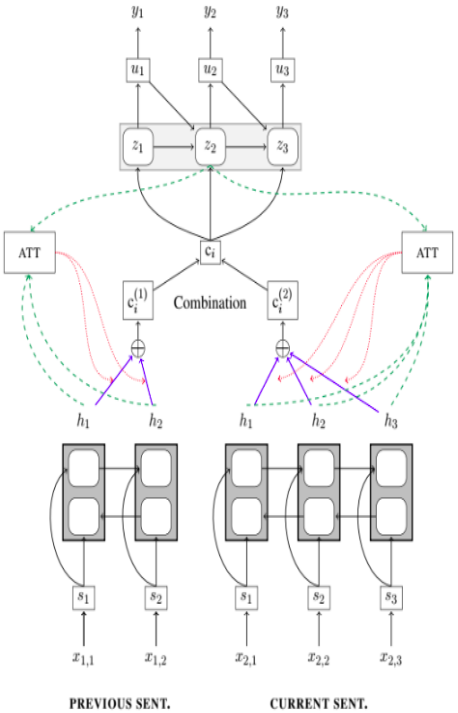
Để đánh giá chính lượng dịch của mô hình, nhóm sử dụng độ đo BLEU chuẩn của *Papineni và các cộng sự, 2002* [24] thông qua sacreBLEU của *Post, 2018* [25] và độ đo COMET của *Rei và các cộng sự, 2020* [26]



(a) Mô hình baseline



(b) Đơn mã hóa ngữ cảnh



(c) Đa mã hóa ngữ cảnh

Hình 4.1: Các chiến lược sử dụng ngữ cảnh

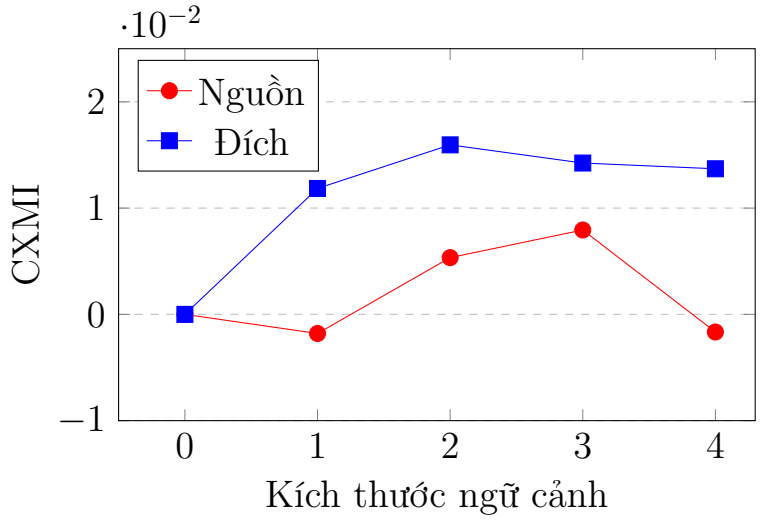
4.5 Kết quả và phân tích thực nghiệm "Thông tin tương hỗ chéo có điều kiện"

4.5.1 Kết quả

Sau khi thực hiện thực nghiệm 4.3 để khảo sát mức độ sử dụng ngữ cảnh với "Thông tin tương hỗ chéo có điều kiện" với các kích thước ngữ cảnh từ 1 đến 4 ở cả phía nguồn và đích, thu được kết quả như ở 4.9 và đồ thị hình 4.2:

Kích thước ngữ cảnh	1	2	3	4
Phía nguồn	-0.00180	0.00534	**0.00794**	-0.00166
Phía đích	0.01185	**0.01595**	0.01424	0.01376

Bảng 4.9: Kết quả CXMI với các kích thước ngữ cảnh khác nhau ở cả phía nguồn và phía đích



Hình 4.2: CXMI với các kích thước ngữ cảnh khác nhau ở cả phía nguồn và phía đích

4.5.2 Phân tích

Dựa vào kết quả ở bảng 4.9 và đồ thị hình 4.2, nhóm có các nhận xét như sau:

- Lượng ngữ cảnh bên phía đích được sử dụng nhiều hơn so với lượng ngữ cảnh bên phía nguồn.
- Lượng ngữ cảnh được sử dụng có bước nhảy vọt khi từ kích thước ngữ cảnh 0 lên 1 đối với phía đích. Đối với phía nguồn thì từ kích thước ngữ cảnh 1 lên 2.
- Lượng ngữ cảnh được sử dụng không tăng đồng đều với kích thước ngữ cảnh và thậm chí có thể dẫn đến giảm lượng thông tin ngữ cảnh được sử dụng (như với kích thước ngữ cảnh 3 và 4 bên phía đích và kích thước ngữ cảnh 4 bên phía nguồn).
- Đối với phía đích, lượng ngữ cảnh được sử dụng nhiều nhất với kích thước ngữ cảnh là 2. Trong khi đó, đối với nguồn lượng ngữ cảnh được sử dụng nhiều nhất với kích thước ngữ cảnh là 3.

Kết luận: Lượng ngữ cảnh được sử dụng bên phía đích nhiều hơn đáng kể so với bên phía nguồn. Lượng ngữ cảnh được sử dụng không tăng đồng đều với kích thước ngữ cảnh và thậm chí có thể dẫn đến giảm lượng thông tin ngữ cảnh được sử dụng. Sử dụng từ 1 đến 3 câu phía trước làm ngữ cảnh cho cả 2 phía sẽ có lợi cho mô hình cho quá trình dịch.

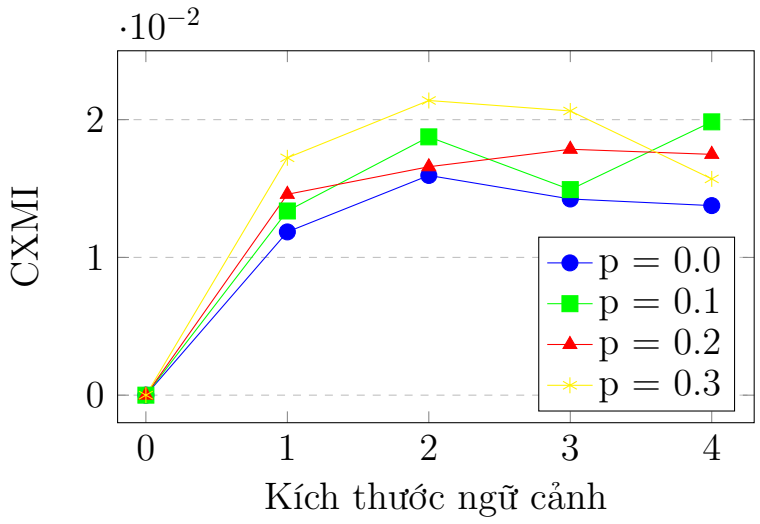
4.6 Kết quả và phân tích thực nghiệm "Loại bỏ từ theo ngữ cảnh"

4.6.1 Kết quả thực nghiệm 4.4.1

Sau khi thực hiện thực nghiệm 4.4.1 để đánh giá về việc tăng cường mức độ sử dụng ngữ cảnh thông qua việc sử dụng "Loại bỏ từ theo ngữ cảnh", thu được kết quả tại 4.10 và đồ thị hình 4.3:

KTNC \ p	0.0	0.1	0.2	0.3
1	0.01185	0.01337	0.01458	**0.01723**
2	0.01595	0.01875	0.01658	**0.02139**
3	0.01424	0.01492	0.01785	**0.02063**
4	0.01376	**0.01984**	0.01748	0.01570

Bảng 4.10: Kết quả CXMI khi thực hiện "Loại bỏ từ theo ngữ cảnh" với kích thước ngữ cảnh (KTNC) khác nhau bên phía đích



Hình 4.3: CXMI khi thực hiện "Loại bỏ từ theo ngữ cảnh" kích thước ngữ cảnh (KTNC) khác nhau bên phía đích

4.6.2 Phân tích thực nghiệm 4.4.1

Dựa vào kết quả ở bảng 4.10 và đồ thị hình 4.3 nhóm có một số nhận xét như sau:

- Việc sử dụng "Loại bỏ từ theo ngữ cảnh" giúp gia tăng mức độ sử dụng ngữ cảnh của mô hình một cách đáng kể.
- Ở tất cả các kích thước ngữ cảnh được xét, khi tăng giá trị xác suất loại bỏ từ p từ 0.0 lên 0.1 đều cho thấy sự gia tăng mức độ sử dụng ngữ cảnh.
- Tuy nhiên, khi tăng giá trị xác suất loại bỏ từ p từ 0.1 lên 0.2 thì mức độ sử dụng ngữ cảnh lại giảm ở kích thước ngữ cảnh 2 và 4. Từ đó cho thấy, sự gia tăng mức độ sử dụng ngữ cảnh không tỉ lệ thuận với giá trị của xác suất loại bỏ từ p.
- Với các mô hình sử dụng kích thước ngữ cảnh là 1, 2 và 3 cho kết quả tốt nhất với giá trị xác suất loại bỏ từ $p = 0.3$. Trong khi đó, mô hình sử dụng kích thước ngữ cảnh là 4 cho kết quả tốt nhất ở $p = 0.1$

Kết luận: Việc sử dụng "Loại bỏ từ theo ngữ cảnh" mang lại kết quả tốt trong việc tăng cường mức độ sử dụng ngữ cảnh của mô hình dịch máy theo ngữ cảnh. Tuy nhiên, sự gia tăng mức độ sử dụng ngữ cảnh không tỉ lệ thuận với giá trị của xác suất loại bỏ từ p. Vì vậy để có được kết quả tốt nhất như mong muốn, cần phải thực hiện các thực nghiệm để có thể chọn ra được giá trị xác suất loại bỏ từ p phù hợp cho các mô hình với các kích thước ngữ cảnh khác nhau.

4.6.3 Kết quả thực nghiệm 4.4.2

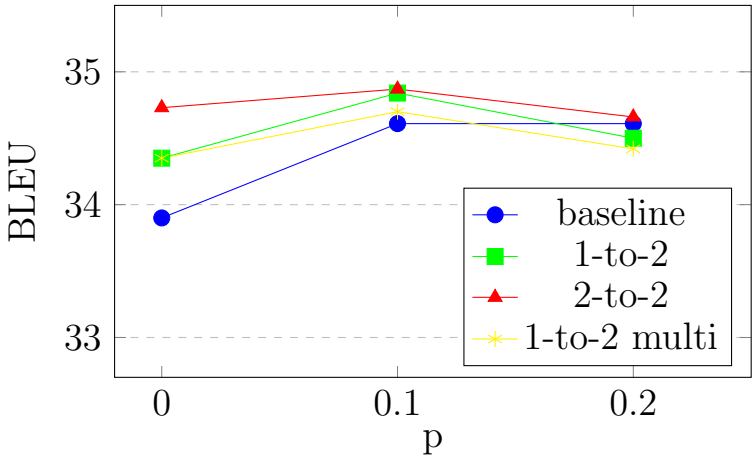
Sau khi thực hiện huấn luyện trên 3 mô hình **baseline**, **one-to-two**, **two-to-two** với xác suất loại bỏ từ lần lượt là 0.0, 0.1 và 0.2, thu được kết quả như bảng 4.11, hình 4.4 và huấn luyện với mô hình **one-to-two multi-encoder** với xác suất loại bỏ từ lần lượt 0.0, 0.1, 0.2, thu được kết quả như bảng 4.12, hình 4.5:

Mô hình	p	BLEU	COMET
baseline	0.0	33.90	0.2474
	0.1	**34.61**	0.2720
	0.2	**34.61**	**0.2865**
one-to-two	0.0	34.35	0.2616
	0.1	**34.84**	**0.2816**
	0.2	34.50	0.2712
two-to-two	0.0	34.73	0.2775
	0.1	**34.87**	0.2862
	0.2	34.66	**0.2890**

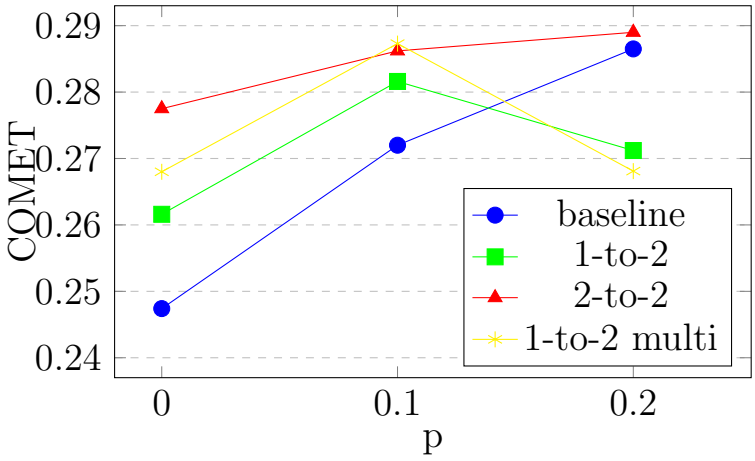
Bảng 4.11: Kết quả BLEU và COMET sau khi thực hiện "Loại bỏ từ theo ngữ cảnh" với ba mô hình đơn mã hóa

Mô hình	p	BLEU	COMET
one-to-two	0.0	34.35	0.2616
	0.1	**34.84**	**0.2816**
	0.2	34.50	0.2712
one-to-two multi-encoder	0.0	34.35	0.2680
	0.1	**34.70**	**0.2873**
	0.2	34.42	0.2681

Bảng 4.12: Kết quả BLEU và COMET sau khi sử dụng "Loại bỏ từ theo ngữ cảnh" với mô hình **one-to-two multi-encoder**



Hình 4.4: Kết quả BLEU sử dụng"Loại bỏ từ theo ngữ cảnh" với các mô hình đơn và đa mã hóa



Hình 4.5: Kết quả COMET sử dụng "Loại bỏ từ theo ngữ cảnh" với các mô hình đơn và đa mã hóa

4.6.4 Phân tích thực nghiệm 4.4.2

Dựa vào kết quả ở bảng 4.11, 4.12 và hình 4.4,4.5 nhóm có những nhận xét như sau:

- Tất cả các mô hình khi được sử dụng "Loại bỏ từ theo ngữ cảnh" ($p > 0$) đều cho thấy được sự cải thiện chất lượng dịch so với việc không sử dụng "Loại bỏ từ theo ngữ cảnh" ($p = 0$). Ngoại trừ điểm BLEU của mô hình **two-to-two** với $p = 0.2$. Và điều đó cũng đó cho mô hình **baseline** khi thực hiện ở mức độ câu.

- Tất cả các mô hình đều cho kết quả tốt khi được sử dụng "Loại bỏ từ theo ngữ cảnh" với giá trị xác suất loại bỏ $p = 0.1$. Tất cả các điểm BLEU của các mô hình này đều cho kết quả nhất trong ba giá trị xác suất loại bỏ.

- "Loại bỏ từ theo ngữ cảnh" cũng giúp cho mô hình **multi-encoder** cải thiện được chất lượng dịch như các mô hình trong bảng 4.12. Đều này cho thấy rằng "Loại bỏ từ theo ngữ cảnh" có tác dụng cải thiện chất lượng dịch cho nhiều kiến trúc dịch theo ngữ cảnh khác nhau.

- Mô hình two-to-two cho kết quả điểm BLEU cũng như là điểm COMET cao nhất trong cả hai bảng 4.11 và 4.12. Từ đó cho thấy việc kết hợp ngữ cảnh ở cả hai phía nguồn và đích sẽ mang lại kết quả cao hơn so với việc chỉ sử dụng ngữ cảnh từ một phía.

Kết luận: "Loại bỏ từ theo ngữ cảnh" giúp ích cho việc cải thiện chất lượng dịch của nhiều mô hình kiến trúc dịch máy theo ngữ cảnh khác nhau. Việc sử dụng ngữ cảnh từ cả phía ngôn ngữ nguồn và đích sẽ mang lại kết quả tốt hơn so với việc chỉ sử dụng ngữ cảnh từ một phía.

4.6.5 Một số ví dụ minh họa

source context	I find that Americans see the fragility in changes.
source	I fear that these changes will not last much beyond the U.S. troops' withdrawal.
reference context	Tôi thấy rằng người Mỹ thấy sự yếu_ớt, dễ vỡ trong những thay_đổi.
reference	Tôi sợ rằng tất_cả những thay_đổi đó sẽ không kéo_dài hơn sau khi quân_đội Mỹ rút đi.
target context p=0	Tôi thấy rằng người Mỹ thấy sự mỏng_manh trong sự thay_đổi.
target p=0	Tôi e rằng những thay_đổi này sẽ không_chỉ dừng lại xa những quân_đội Mỹ.
target context p=0.1	Tôi thấy rằng người Mỹ thấy sự mỏng_manh trong sự thay_đổi.
target p=0.1	Tôi e rằng những thay_đổi này sẽ không còn nhiều hơn cả quân_đội Hoa_Kỳ rút_lui.
ΔCXMI	0.1933
ΔBLEU	6.5436

Bảng 4.13: Ví dụ cho thấy mô hình sử dụng "Loại bỏ từ theo ngữ cảnh" sử dụng nhiều ngữ cảnh hơn những mô hình không sử dụng (mô hình sử dụng ngữ cảnh là 1 bên phía đích) và giúp cải thiện chất lượng dịch của mô hình

source context	I couldn't offer them any direct help.
source	I couldn't give them money, nothing.
reference context	Tôi đã không_thể trực_tiếp giúp gì cho họ.
reference	Tôi chẳng cho họ tiền được, không gì cả.
target context p=0	Tôi không_thể cho họ bất_cứ sự giúp_đỡ nào.
target p=0	Tôi không_thể trả tiền cho họ, không có gì cả.
target context p=0.1	Tôi không_thể cho họ bất_kỳ sự giúp_đỡ trực_tiếp nào.
target p=0.1	Tôi không_thể cho họ tiền, không gì cả.
ΔCXMI	0.1642
ΔBLEU	10.6958

Bảng 4.14: Ví dụ cho thấy mô hình sử dụng "Loại bỏ từ theo ngữ cảnh" sử dụng nhiều ngữ cảnh hơn những mô hình không sử dụng (mô hình sử dụng ngữ cảnh là 1 bên phía đích) và giúp cải thiện chất lượng dịch của mô hình

Chương 5

Kết luận

5.1 Đóng góp của khóa luận

Trong khóa luận này chúng tôi đã nghiên cứu các kiến thức liên quan đến độ đo "Thông tin tương hỗ chéo có điều kiện" và áp dụng để đo mức độ sử dụng ngữ cảnh của các mô hình dịch máy theo ngữ cảnh cho song ngữ Anh-Việt. Ngoài ra, chúng tôi còn nghiên cứu "Loại bỏ từ theo ngữ cảnh" để khuyến khích mô hình dịch máy mạng neural Anh-Việt sử dụng nhiều ngữ cảnh hơn trong quá trình dịch. Sau khi thực hiện các thí nghiệm chúng tôi có những đóng góp như sau:

- Lượng ngữ cảnh được sử dụng bên phía đích (tiếng Việt) nhiều hơn đáng kể so với bên phía nguồn (tiếng Anh).
- Lượng ngữ cảnh được sử dụng không tăng đồng đều với kích thước ngữ cảnh và thậm chí có thể dẫn đến giảm lượng thông tin ngữ cảnh được sử dụng. Sử dụng từ 1 đến 3 câu phía trước làm ngữ cảnh cho cả 2 phía sẽ có lợi cho mô hình cho quá trình dịch.
- "Loại bỏ từ theo ngữ cảnh" giúp ích cho việc tăng cường mức độ sử dụng ngữ cảnh của các mô hình dịch máy mạng neural Anh-Việt.
- Sự gia tăng mức độ sử dụng ngữ cảnh không tỉ lệ thuận với giá trị xác suất loại bỏ từ.

- Việc sử dụng ngữ cảnh và "Loại bỏ từ theo ngữ cảnh" giúp ích cho việc cải thiện chất lượng dịch của nhiều mô hình kiến trúc dịch máy theo ngữ cảnh khác nhau.
- Việc sử dụng ngữ cảnh từ cả phía ngôn ngữ nguồn và đích sẽ mang lại kết quả tốt hơn so với việc chỉ sử dụng ngữ cảnh từ một phía.

5.2 Các hướng trong phát triển

Các công việc trong tương lai có thể phát triển cho khóa luận này:

- Xây dựng bộ dữ liệu lớn ở mức độ văn bản cho song ngữ Anh-Việt.
- Xây dựng bộ dữ liệu để có thể thử nghiệm phương pháp "Đánh giá mâu thuẫn" như bộ dữ liệu ContraPro của *Müller và các cộng sự, 2018* [23] cho song ngữ Anh-Đức
- Xây dựng thêm những cách mã hóa-giải mã khác để sử dụng ngữ cảnh vào mô hình dịch máy mạng neural Anh-Việt
- Tìm ra những độ đo khác để đo mức độ sử dụng ngữ cảnh của các mô hình dịch máy mạng neural Anh-Việt.
- Tìm ra những kỹ thuật khác để khuyến khích mô hình sử dụng nhiều ngữ cảnh hơn trong quá trình dịch.

Tài liệu tham khảo

Tiếng Việt

[1] EnglishSemantics. *Ngữ dụng học - Trực chỉ - deixis*. URL: `http://english-semantics.blogspot.com/2013/05/ngu-dung-hoc-truc-chi-deixis.html/` (visited on 06/20/2022).

[2] trituenhantao.io. *BPE – Byte Pair Encoding – Vũ khí bí mật của NLP hiện đại*. URL: `https://trituenhantao.io/kien-thuc/byte-pair-encoding-vu-khi-bi-mat-cua-nlp-hien-dai/` (visited on 06/20/2022).

Tiếng Anh

[3] Agrawal, Ruchit, Turchi, Marco, and Negri, Matteo. "Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides". In: 2018.

[4] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473 (2015).

[5] Bawden, Rachel et al. "Evaluating Discourse Phenomena in Neural Machine Translation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1304–1313. DOI: 10.18653/v1/N18-1118. URL: https://aclanthology.org/N18-1118.

[6] Bugliarello, Emanuele et al. "It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1640–1649. DOI: 10.18653/v1/2020.acl-main.149. URL: https://aclanthology.org/2020.acl-main.149.

[7] Fernandes, Patrick et al. "Measuring and Increasing Context Usage in Context-Aware Machine Translation". In: *CoRR* abs/2105.03482 (2021). arXiv: 2105.03482. URL: https://arxiv.org/abs/2105.03482.

[8] Guillou, Liane et al. "A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 570–577. DOI: 10.18653/v1/W18-6435. URL: https://aclanthology.org/W18-6435.

[9] Hassan, Hany et al. "Achieving Human Parity on Automatic Chinese to English News Translation". In: *CoRR* abs/1803.05567 (2018). arXiv: 1803.05567. URL: http://arxiv.org/abs/1803.05567.

[10] Jean, Sebastien et al. *Does Neural Machine Translation Benefit from Larger Context?* 2017. DOI: 10.48550/ARXIV.1704.05135. URL: https://arxiv.org/abs/1704.05135.

[11] Khanna, Chetna. *Byte-Pair Encoding: Subword-based tokenization algorithm*. URL: https://towardsdatascience.com/byte-pair-

encoding-subword-based-tokenization-algorithm-77828a70bee0 (visited on 06/20/2022).

[12] Kim, Yunsu, Tran, Duc Thanh, and Ney, Hermann. "When and Why is Document-level Context Useful in Neural Machine Translation?" In: *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 24–34. DOI: 10.18653/v1/D19-6503. URL: https://aclanthology.org/D19-6503.

[13] Kuang, Shaohui et al. "Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 596–606. URL: https://aclanthology.org/C18-1050.

[14] Kudo, Taku and Richardson, John. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://aclanthology.org/D18-2012.

[15] Lample, Guillaume and Conneau, Alexis. *Cross-lingual Language Model Pretraining*. 2019. DOI: 10.48550/ARXIV.1901.07291. URL: https://arxiv.org/abs/1901.07291.

[16] Läubli, Samuel, Sennrich, Rico, and Volk, Martin. "Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4791–4796. DOI: 10.18653/v1/D18-1512. URL: https://aclanthology.org/D18-1512.

[17] Li, Bei et al. “Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 3512–3518. DOI: 10.18653/v1/2020.acl-main.322. URL: https://aclanthology.org/2020.acl-main.322.

[18] Libovický, Jindřich and Helcl, Jindřich. “Attention Strategies for Multi-Source Sequence-to-Sequence Learning”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 196–202. DOI: 10.18653/v1/P17-2031. URL: https://aclanthology.org/P17-2031.

[19] Lopes, António et al. “Document-level Neural MT: A Systematic Comparison”. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 225–234. URL: https://aclanthology.org/2020.eamt-1.24.

[20] Maruf, Sameen and Haffari, Gholamreza. “Document Context Neural Machine Translation with Memory Networks”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1275–1284. DOI: 10.18653/v1/P18-1118. URL: https://aclanthology.org/P18-1118.

[21] Maruf, Sameen, Martins, André F. T., and Haffari, Gholamreza. “Selective Attention for Context-aware Neural Machine Translation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019,

pp. 3092–3102. DOI: 10.18653/v1/N19-1313. URL: https://aclanthology.org/N19-1313.

[22] Miculicich, Lesly et al. "Document-Level Neural Machine Translation with Hierarchical Attention Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2947–2954. DOI: 10.18653/v1/D18-1325. URL: https://aclanthology.org/D18-1325.

[23] Müller, Mathias et al. "A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 61–72. DOI: 10.18653/v1/W18-6307. URL: https://aclanthology.org/W18-6307.

[24] Papineni, Kishore et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[25] Post, Matt. "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. DOI: 10.18653/v1/W18-6319. URL: https://aclanthology.org/W18-6319.

[26] Rei, Ricardo et al. "COMET: A Neural Framework for MT Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2685–2702. DOI: 10.

18653/v1/2020.emnlp-main.213. URL: https://aclanthology.org/2020.emnlp-main.213.

[27] Sennrich, Rico, Haddow, Barry, and Birch, Alexandra. "Edinburgh Neural Machine Translation Systems for WMT 16". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 371–376. DOI: 10.18653/v1/W16-2323. URL: https://aclanthology.org/W16-2323.

[28] Sennrich, Rico, Haddow, Barry, and Birch, Alexandra. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

[29] Tiedemann, Jörg and Scherrer, Yves. "Neural Machine Translation with Extended Context". In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 82–92. DOI: 10.18653/v1/W17-4811. URL: https://aclanthology.org/W17-4811.

[30] Toral, Antonio et al. "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 113–123. DOI: 10.18653/v1/W18-6312. URL: https://aclanthology.org/W18-6312.

[31] Tu, Zhaopeng et al. "Learning to Remember Translation History with a Continuous Cache". In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 407–420. DOI: 10.1162/tacl_a_00029. URL: https://aclanthology.org/Q18-1029.

[32] Unbabel. *COMET Metrics*. URL: https://unbabel.github.io/COMET/html/models.html#comet-metrics (visited on 06/20/2022).

[33] Vaswani, Ashish et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[34] Voita, Elena, Sennrich, Rico, and Titov, Ivan. "Context-Aware Monolingual Repair for Neural Machine Translation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 877–886. DOI: 10.18653/v1/D19-1081. URL: https://aclanthology.org/D19-1081.

[35] Voita, Elena, Sennrich, Rico, and Titov, Ivan. "When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1198–1212. DOI: 10.18653/v1/P19-1116. URL: https://aclanthology.org/P19-1116.

[36] Voita, Elena et al. "Context-Aware Neural Machine Translation Learns Anaphora Resolution". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1264–1274. DOI: 10.18653/v1/P18-1117. URL: https://aclanthology.org/P18-1117.

[37] Vu, Thanh et al. "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. New Orleans, Louisiana: Association for Computa-

tional Linguistics, June 2018, pp. 56–60. DOI: 10.18653/v1/N18-5012. URL: https://aclanthology.org/N18-5012.

[38] Wang, Longyue et al. "Exploiting Cross-Sentence Context for Neural Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2826–2831. DOI: 10.18653/v1/D17-1301. URL: https://aclanthology.org/D17-1301.

[39] Xiong, Hao et al. "Modeling Coherence for Discourse Neural Machine Translation". In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 7338–7345. ISBN: 978-1-57735-809-1. URL: https://aaai.org/ojs/index.php/AAAI/article/view/4721.

[40] Zhang, Jiacheng et al. "Improving the Transformer Translation Model with Document-Level Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 533–542. DOI: 10.18653/v1/D18-1049. URL: https://aclanthology.org/D18-1049.

[41] Zoph, Barret and Knight, Kevin. "Multi-Source Neural Translation". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 30–34. DOI: 10.18653/v1/N16-1004. URL: https://aclanthology.org/N16-1004.