TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN CÔNG NGHỆ TRI THỨC

HOÀNG HUY LỊCH

XÂY DỰNG KHO NGỮ LIỆU SONG NGỮ CHO KHU VỰC ĐÔNG NAM Á TỪ WIKIPEDIA

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN

CHƯƠNG TRÌNH CHÍNH QUY

Tp. Hồ Chí Minh, tháng 03/2022

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN CÔNG NGHỆ TRI THỨC

HOÀNG HUY LỊCH

XÂY DỰNG KHO NGỮ LIỆU SONG NGỮ CHO KHU VỰC ĐÔNG NAM Á TỪ WIKIPEDIA

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN

CHƯƠNG TRÌNH CHÍNH QUY

GIÁO VIÊN HƯỚNG DẪN

PGS.TS ĐINH ĐIỀN

Th.S LÊ THÀNH NGUYÊN

Tp. Hồ Chí Minh, tháng 03/2022

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Tp. Hồ Chí Minh, ngày .... tháng .... năm ....

PGS. TS. Đinh Điền

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Tp. Hồ Chí Minh, ngày...., tháng...., năm ..

Lời cảm ơn

Đầu tiên, em xin gửi lời cảm ơn sâu sắc đến **Thầy PGS.TS Đinh Điền** – người đã cho phép em thực hiện đề tài này cũng như đã tận tình hướng dẫn, truyền đạt cho em những tri thức nền tảng vô cùng quý báu về lĩnh vực Xử lý Ngôn Ngữ Tự nhiên cũng như đã luôn khích lệ, động viên hỗ trợ em mỗi khi khó khăn trong thời gian thực hiện khóa luận tốt nghiệp. Cảm ơn Thầy vì đã tạo mọi điều kiện tốt nhất cho em được tự do phát triển ý tưởng, được lựa chọn làm những chủ đề lý thú về xử lý ngôn ngữ tự nhiên nói riêng và khoa học máy tính nói chung.

Em xin gửi lời cảm ơn đặc biệt đến **ThS. Lê Thành Nguyên**, hiện đang nghiên cứu tại trung tâm đã hướng dẫn và theo sát quá trình thực hiện khóa luận cũng như định hướng nghiên cứu trong tương lai.

Xin chân thành cảm ơn quý **Thầy Cô khoa Công Nghệ Thông Tin** – trường đại học Khoa Học Tự Nhiên, những người đã ân cần giảng dạy, trang bị cho chúng em nền tảng kiến thức vững chắc để làm hành trang bước vào đời.

Em cũng xin gửi lời cảm ơn đến **Khoa Công Nghệ Thông Tin**, trường đại học Khoa Học Tự Nhiên đã tạo môi trường thuận lợi cho sinh viên chúng em học tập, trau dồi kiến thức để hoàn thành khóa luận tốt nghiệp này.

Cuối cùng, con xin cảm ơn **Ba Mẹ** đã sinh thành, nuôi dưỡng, dạy bảo và tạo mọi điều kiện thuận lợi để con có thể học tập và trưởng thành như ngày hôm nay. **Ba Mẹ và Anh Chị** luôn là chỗ dựa vững chắc và là nguồn động lực lớn lao để con có thể vượt qua mọi khó khăn và thử thách trong cuộc sống.

TP Hồ Chí Minh, tháng 3/2022.

Hoàng Huy Lịch

Mục Lục

Danh mục hình ảnh

Danh mục bảng

CHƯƠNG 1. GIỚI THIỆU

Chúng tôi trình bày một phương pháp để trích các cặp câu song ngữ tự động cho một số ngôn ngữ khu vực Đông Nam Á thông qua không gian vectơ câu chung từ các bài viết trên Wikipedia.

Sau đó để đánh giá các cặp câu song ngữ, chúng tôi đã huấn luyện mô hình dịch máy thống kê và đánh giá mô hình này thông qua điểm BLEU và cho kết quả khả quan.

1.1 ĐỘNG LỰC

Trong tình hình hội nhập sâu và rộng như hiện nay, nhu cầu về giao lưu thương mại , văn hóa và giáo dục giữa các nước ta với nước ngoài và đặc biệt là các nước trong khu vực Đông Nam Á ASEAN ngày càng tăng. Dẫn đến nhu cầu về dịch tài liệu song ngữ đa lĩnh vực ngày càng tăng. Cấp thiết cần có một hệ thống dịch chất lượng cao đa ngữ, có khả năng dịch tài liệu đa lĩnh vực. Nhưng một điểm hạn chế hiện nay đó là khuyết thiếu một kho ngữ liệu song ngữ đủ lớn, đã lĩnh vực có chất lượng cao.

Để giải quyết vấn đề này thì tôi đề xuất một phương pháp tự động khai thác cơ sở ngữ liệu Wikipedia nhằm xây dựng một kho ngữ liệu song ngữ cho một số ngôn ngữ của khu vực Đông Nam Á thuộc nhiều lĩnh vực khác nhau. Điều này sẽ giúp giảm được rất nhiều chi phí khi so với việc xây dựng kho ngữ liệu thủ công.

Chúng tôi chọn Wikipedia cho mục đích xây dựng kho ngữ liệu song ngữ đa lĩnh vực. Vì Wikipedia được biết đến rộng rãi là bách khoa toàn thư mở lớn nhất trên internet hiện nay với bài viết trải rộng trên nhiều lĩnh vực khác nhau. Wikipedia chính thức bắt đầu vào ngày 15 tháng 1 năm 2001 nhờ hai người sáng lập Jimmy Wales và Larry Sanger cùng với vài người cộng tác và chỉ có phiên bản tiếng Anh. Tính tới ngày 6/2/2021 thì Wikipedia đã có bài viết bằng 321 thứ tiếng. Wikipedia có khoảng 6,38 triệu bài viết tiếng Anh, khoảng 1,28 triệu bài viết bằng tiếng Việt và khoảng 1,2 triệu bài viết bằng tiếng Trung trải rộng trên nhiều lĩnh vực nên phù hợp cho việc xây dựng kho ngữ liệu song ngữ cho một số ngôn ngữ phổ biến ở khu vực Đông Nam Á (tiếng Việt, tiếng Anh, tiếng Trung, tiếng Indonesia và tiếng Malaysia).

1.2 ẢNH HƯỞNG CỦA CHẤT LƯỢNG KHO NGỮ LIỆU TỚI CHẤT LƯỢNG DỊCH MÁY

Dịch tự động hay còn goi là dịch máy nghiên cứu việc sử dụng phần mềm để dịch văn bản từ một ngôn ngữ này sang ngôn ngữ khác, chẳng hạn như dịch một văn bản từ tiếng Anh sang tiếng Việt. Phần mềm dịch máy tự động áp dụng các thuật toán hoạt động dựa trên cơ sở tổng hợp và xử lý các tri thức từ ngôn ngữ tự nhiên, chẳng hạn như thông qua từ điển, các cặp câu song ngữ, các luật ngữ pháp ...

Vì vậy có thể thấy rằng để có thể có một hệ thống dịch tự động chất lượng cao, cần có hai yếu tố then chốt là kho ngữ liệu chất lượng và phương pháp dịch.

1.2.1 KHO DỮ LIỆU SONG NGỮ

Kho ngữ liệu chất lượng phải đáp ứng được các yêu cầu sau:

- Có chất lượng tốt, nghĩa là dữ liệu phải chính xác, ngữ nghĩa không nhập nhằng...
- Có số lượng lớn, nghĩa là có đầy đủ các luật ngữ pháp, có số lượng cặp câu song ngữ đủ lớn, bao phủ đầy đủ các lĩnh vực, có đầy đủ các từ, cụm từ trong ngôn ngữ tự nhiên.

Những kho ngữ liệu song ngữ thường cần rất nhiều thời gian và chi phí để xây dựng nên hiện chỉ tồn tại cho một số ngôn ngữ đặc biệt là các ngôn ngữ giàu tài nguyên (Anh, Pháp, Đức, Hoa…). Đối với các ngôn ngữ thuộc khu vực Đông Nam Á, các kho ngữ liệu song ngữ hầu như không tồn tại hoặc có số lượng cặp câu rất ít.

Kho ngữ liệu song ngữ được gióng hàng mức câu là một dạng tài nguyên ngôn ngữ quan trọng được sử dụng trong nhiều ứng dụng của xử lý ngôn ngữ tự nhiên, như: nghiên cứu ngôn ngữ học so sánh, tìm kiếm thông tin xuyên ngữ, xây dựng từ điển song ngữ. Đặc biệt trong lĩnh vực dịch máy, chất lượng và độ lớn của kho ngữ liệu song ngữ có vai trò quyết định đến chất lượng dịch. Hiện nay các kho ngữ liệu song ngữ hiện có cho tiếng Việt đều có kích thước hạn chế và thường chỉ phục vụ cho một lĩnh vực nhất định. Nên nhu cầu về một kho ngữ liệu song ngữ cho tiếng Việt và các ngôn ngữ lớn thuộc khu vực Đông Nam Á là rất bức thiết.

1.2.2 ẢNH HƯỞNG CỦA KHO NGỮ LIỆU TỚI CHẤT LƯỢNG DỊCH MÁY

Các phương pháp dịch phổ biến hiện nay là phương pháp dịch máy thống kê và dịch máy sử dụng mạng nơ ron. Công trình của Nguyễn Văn Bình và Huỳnh Công Pháp(2021)[1] đã thực nghiệm xây dựng kho ngữ liệu song ngữ Anh - Việt tổng hợp từ nhiều nguồn và đánh giá ảnh hưởng của chất lượng, kích thước của kho ngữ liệu tới chất lượng của hai phương pháp dịch máy phổ biến hiện nay.

Kết quả cho thấy vai trò quan trọng của khối lượng và chất lượng của kho ngữ liệu ảnh hưởng tới chất lượng của kết quả hệ thống dịch máy tự động tiếng Việt. Khối lượng kho ngữ liệu càng lớn, chất lượng dịch sẽ càng tốt. Chính vì vậy, vấn đề nâng cao chất lượng và khối lượng của các kho ngữ liệu tiếng Việt và các kho ngữ liệu song ngữ, xuyên ngữ trong khu vực Đông Nam Á cần được quan tâm nghiên cứu góp phần xây dựng các hệ thống dịch mà sản phẩm có thể áp dụng vào thực tiễn.

1.3 MỤC TIÊU

Xây dựng được kho ngữ liệu song song cho một số ngôn ngữ khu vực Đông Nam Á một cách tự động để phục vụ cho việc dịch tài liệu đa lĩnh vực hướng tới xây dựng kho ngữ liệu xuyên ngữ.

1.4 CÁCH TIẾP CẬN DỰ KIẾN

Quá trình trích xuất kho ngữ liệu song ngữ từ Wikipedia của chúng tôi thông qua các bước sau:

- Bước 1: Trích xuất nội dung (loại bỏ meta token, hình ảnh, ...).
- Bước 2: Chia tách các câu thành đoạn.
- Bước 3: Loại bỏ các câu trùng.
- Bước 4: Loại bỏ các câu thuộc ngôn ngữ khác (thường là trích dẫn, hoặc liên kết tới ngôn ngữ khác).
- Bước 5: Đưa câu vào không gian nhúng câu chung.
- Bước 6: Sử dụng tiêu chuẩn cận biên và khoảng cách giữa các câu để tìm ra các câu tương đồng.

1.5 ĐÓNG GÓP CỦA KHÓA LUẬN

Xây dựng được kho ngữ liệu song ngữ chất lượng cao cho một số ngôn ngữ phổ biến(tiếng Việt, tiếng Anh, tiếng Trung, tiếng Indonesia và tiếng Malaysia)

1.6 ỨNG DỤNG

1.6.1 ỨNG DỤNG CỦA KIẾN TRÚC SỬ DỤNG

Huấn luyện bằng cách nhúng đa ngôn ngữ vào không gian vectơ câu chung trên nhiều ngôn ngữ cùng một lúc cũng có lợi thế mà các ngôn ngữ tài nguyên thấp có thể được hưởng lợi từ sự tương tự với ngôn ngữ khác trong cùng một ngữ hệ. Ví dụ, chúng tôi có thể khai thác dữ liệu song ngữ cho một số ngôn ngữ mặc dù chúng không được sử dụng để huấn luyện không gian vectơ nhúng câu chung **LASER**. Dẫn đến, chúng tôi có thể khai thác ngữ liệu song ngữ cho một số ngôn ngữ thiểu số nhóm Việt-Mường, ngành Môn-Khmer như tiếng Mường, tiếng Khmer (đó có thể là hướng phát triển trong tương lai)

1.6.2 ỨNG DỤNG CỦA KHO NGỮ LIỆU SONG NGỮ

Kho ngữ liệu song ngữ xây dựng được có thể được sử dụng cho vô vàn các ứng dụng khác nhau của bài toán xử lý ngôn ngữ tự nhiên. Ví dụ truy vấn văn bản xuyên ngữ, xây dựng từ điển song ngữ, dịch máy... Trong đó bài toán dịch máy và xây dựng từ điển song ngữ là hai trong những bài toán có nhu cầu và ứng dụng thực tế cao, phục vụ cho nhu cầu học tập, đào tạo và dịch thuật đang lên cao hiện nay.

1.7 BỐ CỤC

Chương 2: Dịch máy mạng NEURAL

Chương 3: Đặc điểm ngôn ngữ

Chương 4: Phương hướng tiếp cận

Chương 5: Quá trình thực hiện

Chương 6: Kết quả

Chương 7: Kết luận và phương hướng phát triển

CHƯƠNG 2. DỊCH MÁY MẠNG NEURAL

Dịch máy (machine translation) là một lĩnh vực nhỏ của ngành Ngôn ngữ học tính toán (computational linguistics) nghiên cứu việc sử dụng phần mềm máy tính để dịch văn bản hoặc lời nói từ ngôn ngữ này sang ngôn ngữ tự nhiên khác. Đây là lĩnh vực kết hợp nhiều ý tưởng và các kỹ thuật với nhau: từ ngôn ngữ học, khoa học máy tính, xác suất thống kê và trí tuệ nhân tạo. Mục tiêu của dịch máy là phát triển một hệ thống cho phép tạo ra bản dịch giống với ngôn ngữ tự nhiên của con người nhất. Dịch máy có một lịch sử lâu đời từ thế kỉ 17 khi hai nhà triết học Leibniz và Descartes đề xuất một hệ mã có khả năng kết nối các ngôn ngữ dù với mục đích nghiên cứu triết học. Cho đến đầu những năm đầu tiên của thập niên 1950 thì hệ thống dịch máy đầu tiên mới ra đời. Đáng chú ý nhất là đề xuất của Warren Weaver (1949) đã đánh dấu sự khởi đầu cho sự phát triển của dịch máy ở mỹ. Những đề xuất của ông giúp giải quyết vấn đề nhập nhằng bằng cách kết hợp tri thức về thống kê, mã hóa và lý thuyết thông tin cũng như dựa trên các phán đoán về nguyên lý cơ bản của ngôn ngữ tự nhiên. Kể từ đó, dịch máy đã trải qua nhiều giai đoạn phát triển. Trong chương này sẽ trình các kiến thức liên quan đến dịch máy và mô hình được sử dụng trong khóa luận này như lịch sử dịch máy (phần 2.1), mạng neural hồi quy(RNNs) (phần 2.2), Mạng LSTM (phần 2.3), dịch máy mạng Neural (phần 2.4).

2.1 LỊCH SỬ DỊCH MÁY

Dịch máy nói chung bắt đầu từ những thập niên 50-60 của thế kỉ 20, nhưng chủ yếu là thay thế từng từ một dựa vào từ điển song ngữ. Thí nghiệm Georgetown-IBM năm 1954 xây dựng một hệ thống dịch Nga-Anh tuy còn nhiều hạn chế nhưng đã thu hút được sự quan tâm lớn của công chúng và sự đầu tư của chính phủ. Kết quả là trong những năm 1950 và 1960, nhiều hệ thống đã được cài đặt và hoạt động.

Tuy nhiên, một đòn giáng mạnh vào các nghiên cứu dịch máy trong năm 1966 là bản báo cáo ALPAC. Bản báo cáo kết luận rằng máy dịch tốn kém hơn, không chính xác và chậm hơn con người và mặc dù đắt đỏ, chất lượng bản dịch không có vẻ gì sẽ đạt đến chất lượng của bản dịch của con người trong tương lai gần. Điều đó được coi là

mùa đông của dịch máy, ngăn cản sự phát triển của dịch máy trong vài thập niên. Các nghiên cứu sâu hơn về máy dịch đã được tiến hành cho đến cuối những năm 1980, khi các hệ thống máy dịch thống kê đầu tiên được phát triển.

Tính đến những năm 1980, hầu hết các hệ thống xử lý ngôn ngữ tự nhiên dựa trên các bộ quy tắc viết tay phức tạp. Tuy nhiên, bắt đầu từ cuối những năm 1980, đã có một cuộc cách mạng về xử lý ngôn ngữ tự nhiên với việc giới thiệu các thuật toán máy học để xử lý ngôn ngữ. Điều này là do sự gia tăng đều đặn về sức mạnh tính toán (xem định luật Moore) và giảm dần sự thống trị của các lý thuyết ngôn ngữ học Chomskyan (ví dụ ngữ pháp chuyển đổi), có nền tảng lý thuyết không khuyến khích các ngôn ngữ học tập thể hiện cách tiếp cận máy học để xử lý ngôn ngữ. Một số thuật toán máy học được sử dụng sớm nhất, chẳng hạn như cây quyết định (decision trees), hệ thống các quy tắc cứng hard if-then rules tương tự như các quy tắc viết tay hiện có. Tuy nhiên, part-of-speech tagging giới thiệu việc sử dụng các mô hình Markov để xử lý ngôn ngữ tự nhiên và càng nghiên cứu tập trung vào mô hình thống kê, làm cho xác suất quyết định dựa trên việc gắn giá trị thực vào các tính năng đầu vào của dữ liệu. Các mô hình ngôn ngữ bộ nhớ cache mà nhiều nhận dạng giọng nói các hệ thống hiện nay dựa vào là các ví dụ về các mô hình thống kê như vậy. Các mô hình như vậy thường mạnh mẽ hơn khi đưa vào đầu vào không quen thuộc, đặc biệt là đầu vào có lỗi (rất phổ biến cho dữ liệu trong thế giới thực) và tạo ra kết quả đáng tin cậy hơn khi tích hợp vào một hệ thống lớn hơn bao gồm nhiều nhiệm vụ phụ.

Nhiều thành công ban đầu đáng chú ý đã được gặt hái trong lĩnh vực máy dịch, đặc biệt là tại IBM Research, nơi các mô hình thống kê phức tạp liên tục được phát triển. Những hệ thống này có thể tận dụng lợi thế của các thư viện đa văn bản hiện có đã được Nghị viện Canada và Liên minh châu Âu sản xuất, chúng là kết quả của các luật kêu gọi dịch thuật tất cả các thủ tục tố tụng của chính phủ sang tất cả các ngôn ngữ chính thức của các hệ thống chính phủ tương ứng. Tuy nhiên, hầu hết các hệ thống khác phụ thuộc vào hoàn toàn hệ thống được phát triển cho nhiệm vụ chuyên biệt, đó là một hạn chế lớn trong sự thành công của các hệ thống này. Kết quả là, rất nhiều nghiên cứu đã đi vào các phương pháp học tập hiệu quả hơn từ số lượng dữ liệu hạn chế.

Nghiên cứu gần đây đã ngày càng tập trung vào các thuật toán học tập không giám sát và bán giám sát. Trong những năm 2010, representation learning và deep neural network-style trở nên phổ biến trong xử lý ngôn ngữ tự nhiên, một phần do một loạt các kết quả cho thấy rằng kỹ thuật như vậy có thể đạt được kết quả tiên tiến trong nhiều tác vụ ngôn ngữ tự nhiên, ví dụ như trong mô hình hóa ngôn ngữ, phân tích cú pháp, và nhiều thứ khác. Các kỹ thuật phổ biến bao gồm việc sử dụng các từ nhúng để nắm bắt các thuộc tính ngữ nghĩa của các từ và tăng cường học tập từ đầu đến cuối của nhiệm vụ cấp cao hơn (ví dụ trả lời câu hỏi) thay vì dựa vào một đường ống nhiệm vụ trung gian riêng biệt (ví dụ: gắn thẻ và phụ thuộc từng phần phân tích cú pháp)

Trong một số lĩnh vực, sự thay đổi này đã dẫn đến những thay đổi đáng kể về cách thức các hệ thống NLP được thiết kế, như vậy các phương pháp dựa trên mạng nơron sâu (deep neural network) có thể được xem như một mô hình mới khác với xử lý ngôn ngữ tự nhiên thống kê. Ví dụ, thuật ngữ máy dịch thần kinh (neural machine translation – NMT) nhấn mạnh thực tế là phương pháp học tập dựa trên việc máy dịch trực tiếp trình tự (sequence-to-sequence transformations), loại bỏ các bước trung gian như sự liên kết văn bản và xây dựng mô hình ngôn ngữ được sử dụng trong thống kê máy dịch (statistical machine translation – SMT).

Có nhiều cách tiếp cận đối với dịch máy:

- Dịch máy dựa trên luật (Rule-Based MT)
- Dịch máy dựa trên thống kê (SMT)
- Dịch máy dựa trên cơ sở tri thức (KBMT)
- Dịch máy dựa trên ví dụ (Example-Based MT)
- Dịch máy dựa trên ngữ liệu (Corpus-Based MT)
- Dịch máy hỗn hợp (Hybrid MTS)

2.2 MẠNG NEURAL HỒI QUY (RNNS)

Mạng neural hồi quy (Recurrent Neural Networks-RNNs) là kiến trúc cực kì hiệu quả và mạnh mẽ trong việc lý dữ liệu tuần tự đặc biệt là dữ liệu ngôn ngữ nhờ khả năng lưu giữ thông tin là kết quả của lần thực thi trước - bộ nhớ (memory). Ví dụ: khi phân

tích một câu từng từ từng từ một thì RNNs có khả năng lưu trữ thông tin của từ đầu tiên trong khi đang xử lý từ cuối cùng.

Các kiến trúc RNNs chứa vòng lặp giữa các nút (node) của nó. Điều này cho phép thông tin lưu lại trong mô hình trong thời gian dài hơn. Bởi vì điều này, đầu ra từ mô hình trở thành cả một dự đoán và một bộ nhớ, sẽ được sử dụng khi bit tiếp theo của văn bản theo trình tự được chuyển qua mô hình.

2.2.1 ỨNG DỤNG CỦA RNNS

Mặc dù, mạng RNNs hoạt động tốt với dữ liệu tuần tự như tệp văn bản, âm thanh và video clips. RNNs cũng có một vài ứng dụng khác cho các vấn đề thực tế để hiểu tại sao RNNs phát triển mạnh mẽ như hiện nay.

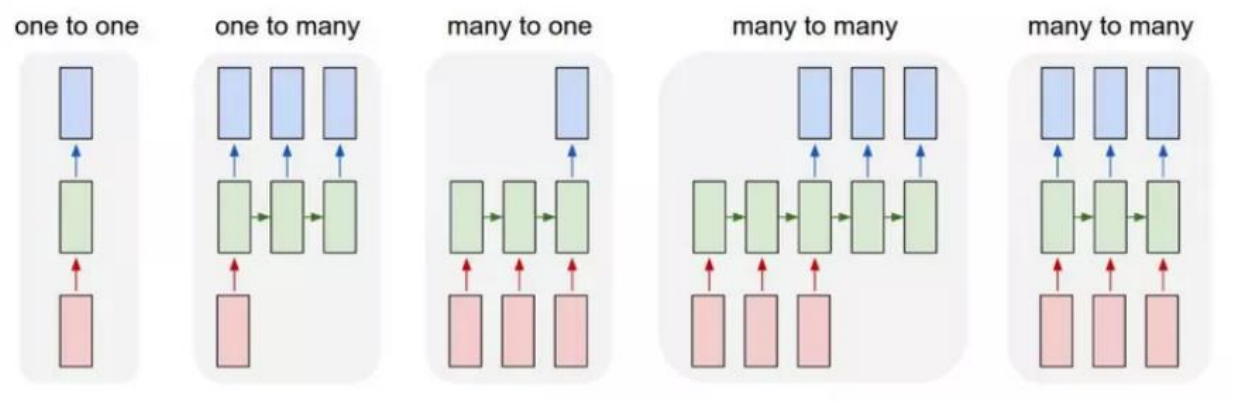
- Xử lý ngôn ngữ tự nhiên (NLP): Sinh văn bản, tạo chatbot,
- Nhận dạng tiếng nói: Tương tự với NLP thì nhận dạng tiếng nói cũng cố để hiểu và tái hiện lại ngôn ngữ tự nhiên nhưng thay vì đầu vào là văn bản thì sử dụng tệp âm thanh, tệp video clip.
- Dịch máy: Với sự phát triển của toàn cầu hóa và nhu cầu trao đổi, đi lại tăng cao nên nhu cầu về một công cụ hỗ trợ dịch tốt hơn luôn cao. Đây là một lĩnh vực luôn phát triển với tốc độ cao.
- Dự đoán thời gian thực (time-series forcasting): được hưởng lợi từ việc lưu trữ thông tin trong quá khứ để dự đoán nhu cầu, thu nhập... trong tương lai.
- Nhận dạng hình ảnh(image recognization): huấn luyện nhận diện nhãn, mô tả cho hình ảnh.

2.2.2 CÁC LOẠI KIẾN TRÚC RNNS

Hiện nay RNNs có một số dạng như sau:

- Kiến trúc một-một: mẫu bài toán cho Neural Network (NN) và Convolutional Neural Network (CNN), một đầu vào và một đầu ra, ví dụ với bài toán phân loại ảnh MNIST đầu vào là ảnh và đầu ra là ảnh đấy là số nào.

- Kiến trúc một-nhiều: bài toán có một đầu vào nhưng nhiều đầu ra, ví dụ với bài toán caption cho ảnh, đầu vào là 1 ảnh nhưng đầu ra là nhiều chữ mô tả cho ảnh đấy, dưới dạng một câu.
- Kiến trúc nhiều-một: bài toán có nhiều đầu vào nhưng chỉ có một đầu ra, ví dụ bài toán phân loại hành động trong video, đầu vào là nhiều ảnh (frame) tách ra từ video, ouptut là hành động trong video.
- Kiến trúc nhiều-nhiều: bài toán có nhiều đầu vào và nhiều đầu ra, ví dụ bài toán dịch từ tiếng Anh sang tiếng Việt, đầu vào là một câu gồm nhiều chữ: "I love Vietnam" và đầu ra cũng là một câu gồm nhiều chữ "Tôi yêu Việt Nam". Để ý là độ dài dữ liệu tuần tự của đầu vào và đầu ra có thể khác nhau.



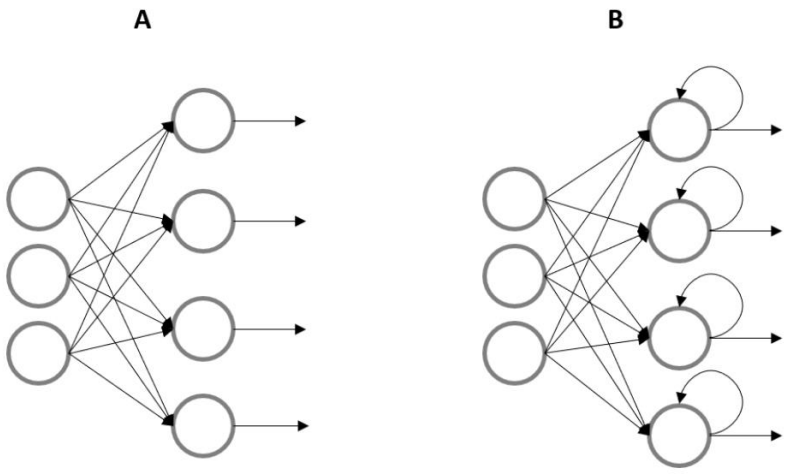
HÌNH 2.1: MỘT SỐ LOẠI HÌNH RNNS

2.2.3 KIẾN TRÚC RNNS

Đầu tiên chúng ta cần đề cập tới điểm khác biệt giữa mạng neural thông thường và mạng neural hồi quy. Mạng neural thông thường hay còn được gọi là mạng chuyển tiếp vì thông tin chỉ đi theo một chiều từ đầu vào tới đầu ra mà không đi qua một nút hai lần để có thể dự đoán kết quả. Vì vậy, chúng không lưu trữ được dữ liệu trong quá khứ, làm cho chúng kém hiệu quả trong việc dự đoán điều gì sẽ xảy ra tiếp theo.

Mặt khác, trong RNNs, các chu trình thông tin sử dụng các vòng lặp, do đó mọi dự đoán đều được thực hiện khi xem xét cả đầu vào và bộ nhớ từ các dự đoán trước đó. Nó hoạt động bằng cách sao chép kết quả đầu ra của từng dự đoán và chuyển nó trở lại

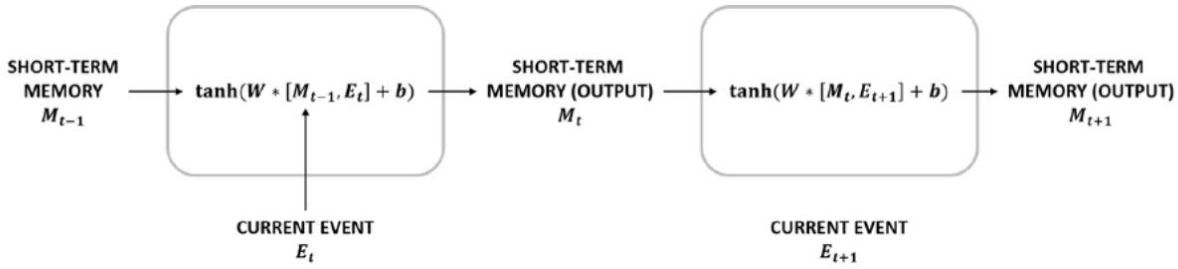
mạng để dự đoán tiếp theo. Theo cách này, RNNs có hai đầu vào: giá trị hiện tại và thông tin quá khứ:



HÌNH 2.2: A: KIẾN TRÚC CỦA MẠNG CHUYỂN TIẾP B:KIẾN TRÚC MẠNG RNNS

Bộ nhớ nội bộ của mô hình RNNs là bộ nhớ ngắn hạn (Short-Term Memory) nên k thể lưu trữ dữ liệu trong thời gian dài. Phần sau chúng tôi sẽ giới thiệu một kiến trúc sở hữu cả bộ nhớ dài hạn (Long-Term Memory) và bộ nhớ ngắn hạn.

Bằng cách sử dụng thông tin từ các dự đoán trước đó, mạng được huấn luyện với dữ liệu tuần tự cho phép nó dự đoán bước sau. Điều này đạt được bởi sự kết hợp thông tin hiện tại với kết quả từ bước trước thành một hoạt động (như hình 1.2). Đầu ra từ hoạt động này sẽ trở thành dự đoán cũng như một phần đầu vào cho dự đoán tiếp theo.



HÌNH 2.3:TÍNH TOÁN TẠI MỖI NÚT CỦA RNNS

Như bạn có thể thấy, hoạt động xảy ra bên trong một nút là hoạt động của bất kỳ nút nào khác trong mạng neural. Ban đầu, dữ liệu được chuyển qua một hàm tuyến tính. Các tham số thì được cập nhật trong quá trình huấn luyện. Tiếp theo, sử dụng hàm kích hoạt đối với tuyến tính của đầu ra. Trong trường hợp này (hình 2.3), đây là hàm tanh, như một số nghiên cứu đã chỉ ra rằng nó đạt được kết quả tốt hơn cho hầu hết các vấn đề về dữ liệu:

$$M_t = \partial(W*[M_{t-1}, E_t] + b)$$

Ở đây, M_{t-1} đóng vai trò là bộ nhớ được rút ra từ dự đoán trước. W và b là các tham số còn E_t đề cập tới sự kiện hiện tại. ∂ là hàm kích hoạt, thường sử dụng hàm tanh để đạt kết quả tốt.

Trong bài toán dịch máy, tại thời điểm thứ t, RNN đưa ra kết quả đầu ra y_t nhờ vào một phân phối xác suất p trên bộ từ vựng của ngôn ngữ đích Y :

$$s_t = W_0 * M_t$$
$$s_t = \text{softmax}(s_t)$$

Trong đó $W_0 \in \mathbb{R}^{|Y|*d}$, với d là số chiều của trạng thái ẩn RNN.

2.3 MẠNG LSTM (LONG SHORT-TERM MEMORY)

Mạng Long Short-Term Memory (LSTM) là một loại RNNs có thể giữ cả bộ nhớ dài hạn và bộ nhớ ngắn hạn, đặc biệt hữu ích cho các chuỗi dữ liệu dài, chẳng hạn như video clip.

Như đã đề cập trước đây, RNNs chỉ lưu trữ bộ nhớ ngắn hạn. Đây là một vấn đề khi xử lý các chuỗi dữ liệu dài, trong đó mạng sẽ gặp khó khăn khi mang thông tin từ các bước trước đó đến những bước cuối cùng.

Các mạng RNNs truyền thống không có khả năng lưu trữ bộ nhớ dài hạn là do vấn đề được gọi là bùng nổ và mất mát đạo hàm. Bùng nổ là khi đạo hàm trở nên rất lớn khi tiến hành lan truyền ngược làm cho quá trình huấn luyện không ổn định. Mất mát đạo hàm xảy ra khi các đạo hàm giảm nhanh và trở nên vô cùng nhỏ gần về **0** khiến chúng

không còn đóng góp vào quá trình học tập của mạng. Cái này thường xảy ra trong các lớp đầu tiên của mạng, làm cho mạng quên những gì nó đã thấy một thời gian trước đây.

Do đó, mạng LSTM đã được phát triển để giải quyết vấn đề còn tồn đọng của RNNs. Mạng LSTM có thể nhớ thông tin trong một khoảng thời gian dài vì chúng lưu trữ bộ nhớ trong của chúng trong các cổng, điều cho phép chúng đọc, viết và xóa thông tin khi cần thiết. Các cổng này giúp mạng quyết định thông tin nào cần lưu giữ và thông tin nào cần xóa khỏi bộ nhớ (có mở cổng hay không), dựa trên mức độ quan trọng nó đã gán cho mỗi bit thông tin.

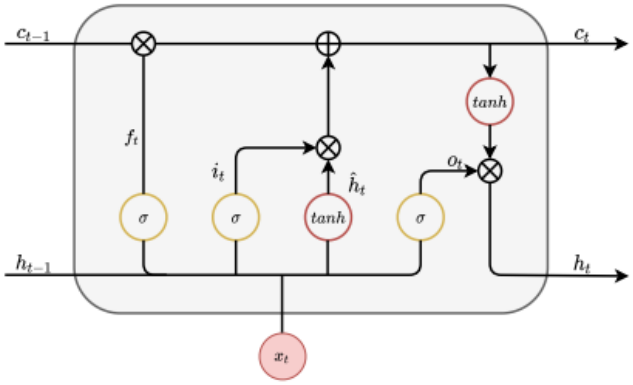
Điều này cực kỳ hữu ích vì nó không chỉ cho phép có thêm thông tin được lưu trữ dưới dạng bộ nhớ dài hạn, nhưng nó cũng giúp loại bỏ thông tin vô ích có thể thay đổi kết quả của một dự đoán, ví dụ mạo từ trong một câu.

2.3.1 ỨNG DỤNG CỦA LSTM

Bên cạnh những ứng dụng của RNNs, khả năng lưu trữ bộ nhớ dài hạn của mạng LSTM đã cho phép chúng ta thực hiện những tác vụ phức tạp hơn, tận dụng được dữ liệu tuần tự có kích thước lớn. Sau đây là một số ứng dụng của LSTM:

- Tự tạo văn bản: Văn bản đầu ra được tạo ra có phong cách giống với những văn bản được chọn làm đầu vào để huấn luyện mô hình. Ví dụ: mô hình GPT-3 của OpenAI
- Tự tạo âm thanh:
- Tạo ra chữ viết tay và nhận dạng chữ viết tay:

2.3.2 KIẾN TRÚC LSTM



HÌNH 2.4: MÔ TẢ CẤU TRÚC MỘT CELL CỦA LSTM

Ở trạng thái thứ t của mô hình LSTM:

• Output: h_t, c_t; ta gọi c là cell state, h là hidden state.

• Đầu vào: h_{t-1}, c_{t-1}, x_t. Trong đó xt là đầu vào ở state thứ t của model. ct−1;ht−1 là đầu ra của lớp trước. h đóng vai trò khá giống như s ở RNN, trong khi c là điểm mới của LSTM.

Trong hình 2-4, ô tròn với kí hiệu σ (sigma) là cổng. Chúng sẽ quyết định bao nhiêu thông tin từ nút trước trong mạng được giữ lại. Sau đó, các cổng chuyển thông tin này đến hàm tanh của chúng tôi. Các hàm tanh chịu trách nhiệm cập nhật trọng số của các nút trong mô hình.

Cuối cùng, LSTM xuất ra hai bộ thông tin: bộ nhớ dài hạn(long-term memory) tương ứng với c_t và bộ nhớ ngắn hạn tương ứng với h_t.

Bên cạnh đó, ta cũng cần hiểu rõ các cổng của LSTM:

- f_t (forget gate)- cổng quên: quyết định xem thông tin nào trong bộ nhớ dài hạn được giữ lại (thông tin từ nút phía trước)
- i_t (input gate)- cổng vào: quyết định bao nhiêu thông tin trong bộ nhớ ngắn hạn được giữ lại. Phần được giữ lại là đầu vào cho hàm tanh.

- o_t (output gate)-cổng ra: Sử dụng thông tin từ cả bộ nhớ dài hạn và ngắn hạn, quyết định xem bao nhiêu thông tin được sử dụng làm đầu ra cho bộ nhớ ngắn hạn tương ứng với lớp ẩn h_t (hidden state)

Các cổng được tính toán như sau:

$$i_t = sigmoid(W_{xi}x_t + W_{hi}h_t)$$
$$f_t = sigmoid(W_{xf}x_t + W_{hf}h_t)$$
$$o_t = sigmoid(W_{xo}x_t + W_{ho}h_t)$$
$$\hat{h}_t = tanh(W_{xh}x_t + W_{hh}h_t)$$

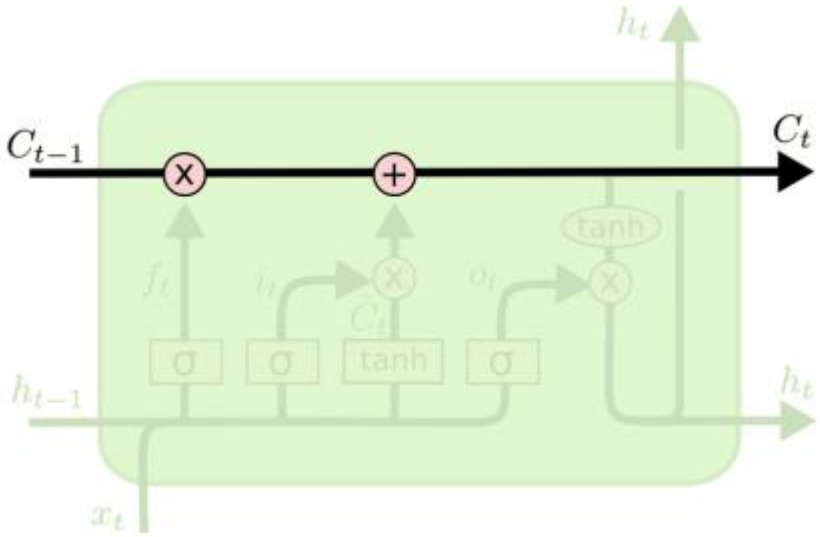
Tất cả chúng đều được tính toán bằng cách sử dụng hàm sigmoid để chúng luôn xuất ra các giá trị từ 0 đến 1. Nếu một cổng tạo ra thứ gì đó gần 1 hơn, nó được coi là mở (dữ liệu có thể được giữ lại và đi qua cổng) và nếu nó đưa ra giá trị gần 0 hơn, thông tin đó sẽ bị bỏ qua.

Bộ nhớ trong LSTM được định nghĩa như sau:

$$c_t = f_t * c_{t-1} + i_t * \hat{h}_t$$
$$h_t = o_t * \tanh(c_t)$$

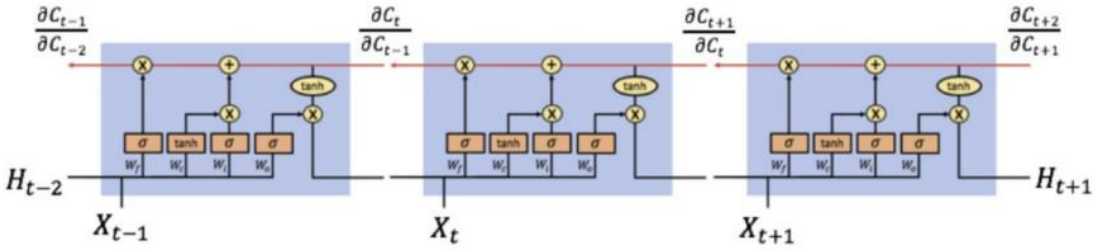
h_t ; $\tilde{c}_t$ khá giống với RNN nên mô hình có bộ nhớ ngắn hạn. Trong khi đó c_t giống như một băng chuyền trên RNN vậy, thông tin nào quan trọng và dùng ở sau sẽ được gửi và dùng khi cần nên mang thông tin được đi xa hơn. Vì vậy, LSTM có bộ nhớ dài hạn. LSTM có cả bộ nhớ dài hạn và bộ nhớ ngắn hạn.

Bằng cách sử dụng 3 cổng này để điều chỉnh luồng thông tin, chúng ta giảm thiểu được những vấn đề liên quan đến mất mát đạo hàm.



HÌNH 2.5: CELL STATE TRONG LSTM

2.3.3 LSTM CHỐNG MẤT MÁT ĐẠO HÀM



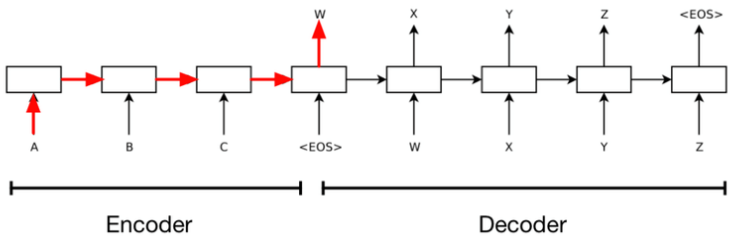
HÌNH 2.6: MÔ HÌNH LSTM

Ta áp dụng thuật toán backpropagation through time cho LSTM. Thành phần chính gây mất mát đạo hàm trong RNN là $\frac{\partial s_{t+1}}{\partial s}$ =(1-s_t^2)*W, trong đó s_t, W <1. Tương tự trong LSTM ta quan tâm đến $\frac{\partial c_t}{\partial c_{t-1}}$ =f_t. Do 0 < f_t. < 1 nên về cơ bản thì LSTM vẫn bị vanishing gradient nhưng bị ít hơn so với RNN. Hơn thế nữa, khi mang thông tin trên cell state thì ít khi cần phải quên giá trị cell cũ, nên ft ≈ 1 giúp ta tránh được mất mát đạo hàm

2.4 DỊCH MÁY MẠNG NEURAL

Mặc dù dịch máy thống kê (SMT) đạt được nhiều thành công khi áp dụng vào các hệ thống thương mại, chúng lại không hoạt động tốt do vấp phải hai vấn đề chính. Đầu tiên là do SMT chỉ tiến hành dịch cụm theo cụm và cũng do vậy các thông tin phụ thuộc xa thường bị bỏ qua. Hơn nữa, toàn bộ hệ thống SMT sẽ càng phức tạp hơn khi tích hợp càng nhiều đặc trưng. Nhiều thành phần trong SMT cần được tinh chỉnh một cách độc lập với các thành phần còn lại (ví dụ như mô hình ngôn ngữ, mô hình dịch, v.v) điều này gây khó khăn cho việc kết hợp chúng lại với nhau.

Dịch máy mạng neural (NMT) là hướng tiếp cận mới nhằm giải quyết các vấn đề đã nêu ở trên. Thứ nhất, NMT là một mạng neural lớn với hàng triệu neuron (tham số) được thiết kế cho việc mô hình hóa quá toàn bộ quá trình dịch máy. NMT cần một lượng tối thiểu tri thức về lĩnh vực nào đó, đơn giản là chỉ cần ngữ liệu song ngữ của các câu trong ngôn ngữ nguồn và ngôn ngữ đích, tương tự với SMT, nhưng ít bước tiền xử lý hơn. Đáng chú ý nhất là NMT có thể được huấn luyện trực tiếp từ dữ liệu mà không cần thêm bất kỳ thành phần con nào (mô hình ngôn ngữ, mô hình dịch như của SMT). Thêm vào đó, quá trình dịch của NMT cũng tương đối đơn giản: một bộ mã hóa đọc câu đầu vào cho ra một vectơ với chiều dài cố định, biểu diễn ý nghĩa của câu; một bộ giải mã xử lý vectơ này và cho ra một bản dịch với ngôn ngữ đích. Kiến trúc này thường được gọi là kiến trúc mã hóa–giải mã. Bằng phương pháp này, NMT giải quyết vấn đề dịch cụm theo cụm của SMT. Thay vào đó NMT kết hợp thông tin từ toàn bộ câu nguồn trước khi dịch, nhờ đó có thể học được thông tin biểu diễn phụ thuộc xa trong ngôn ngữ như: giới tính, thứ tự của chủ ngữ, động từ, v.v.



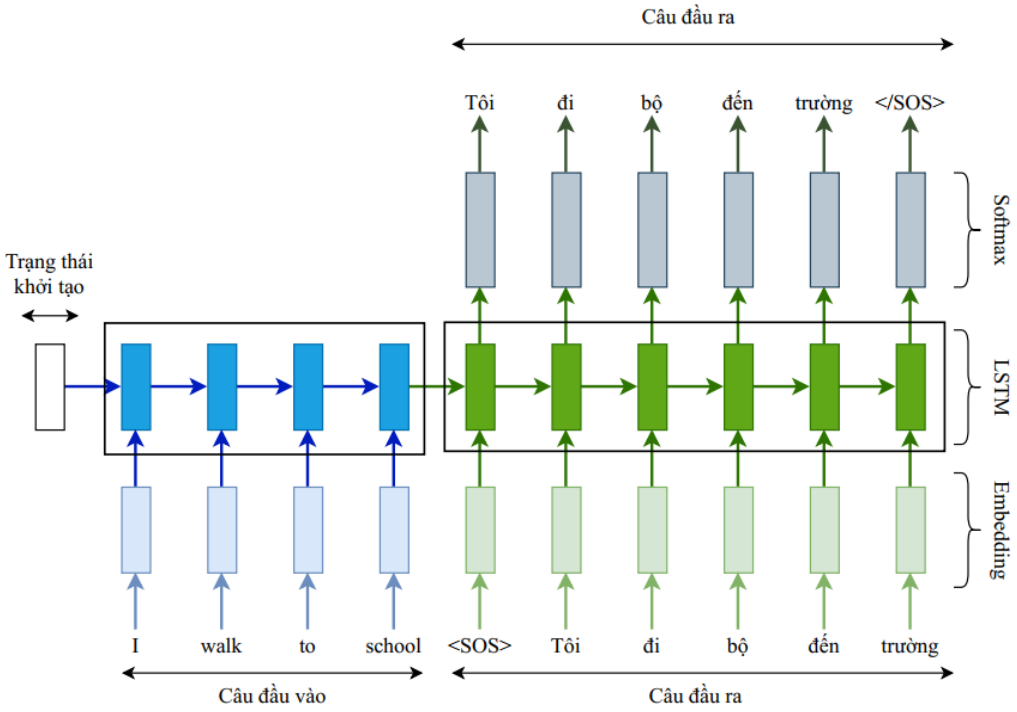
HÌNH 2.7: KIẾN TRÚC ENCODER -DECODER

Tuy nhiên các kiến trúc này lại vấp phải những khó khăn khi xử lý câu dài và nhập nhằng, cơ chế attention ra đời nhằm giải quyết phần nào vấn đề này và cũng đạt được kết quả đầy hứa hẹn, tiêu biểu là công trình của [2] và [3].

Hiện nay có ba kiến trúc dịch máy mạng neural chính là: sử dụng mạng neural hồi quy (seq2seq), mạng neural tích chập (Conv2Seq) và Transformer. Mô hình Seq2Seq cũng là mô hình cơ sở cho khóa luận này.

2.4.1 MẠNG SEQSEQ

Seq2Seq dường như là sự lựa chọn hàng đầu đối với các bài toán NLP đặc biệt là dịch máy. Mô hình seq2seq gồm một bộ mã hóa và bộ giải mã đều sử dụng RNN. Khóa luận này khảo sát mô hình seq2seq với LSTM hai chiều.(mô hình được sử dụng được xây dựng dựa trên kiến trúc Seq2Seq sẽ được trình bày ở mục 4.2).



HÌNH 2.8: MÔ HÌNH SEQ2SEQ SỬ DỤNG LSTM

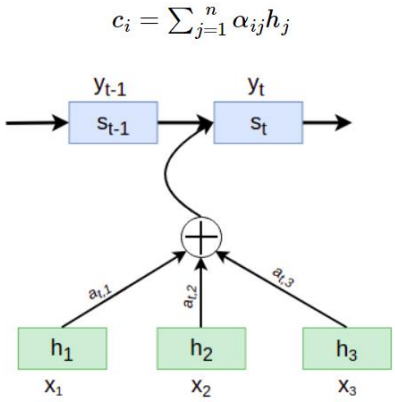
Quá trình huấn luyện mô hình dịch từ câu tiếng Anh “I walk to school” sang câu tiếng Việt “Tôi đi bộ đến trường”. Mô hình bao gồm 2 mạng neural hồi quy như đã trình bày ở

2.2, cụ thể là LSTM: bộ mã hóa LSTM xử lý câu tiếng Anh (không đưa ra dự đoán), bộ giải mã xử lý câu tiếng Việt trong khi vừa dự đoán từ kế tiếp.

Chi tiết hơn, bộ mã hóa và giải mã LSTM nhận giá trị đầu vào là câu tiếng Anh, sau đó là ký tự <SOS> - cho biết bắt đầu quá trình giải mã, và câu tiếng Việt. Với các từ cụ thể, mô hình tìm biểu diễn của các từ đó thông qua lớp embedding của tiếng Anh và tiếng Việt. Trọng số của lớp embedding khác nhau với các ngôn ngữ khác nhau, và được học thông qua quá trình huấn luyện. Ngoài ra, ta có thể sử dụng các mô hình embedding đã được huấn luyện trước như: word2vec [4], Glove [5].

2.4.2 CƠ CHẾ ATTENTION

NMT có thể đạt được kết quả rất tốt trong những tác vụ với dữ liệu lớn như dịch Anh - Đức và thậm chí là với dữ liệu ít như Anh-Việt. Tuy nhiên, NMT phải đối mặt với thách thức lớn chính là xử lý các câu dài. Một cách giải quyết hiệu quả cho vấn đề này là sử dụng cơ chế attention cho phép mô hình học các sự liên kết giữa các thể thức khác nhau như: tiếng nói và văn bản trong bài toán nhận diện tiếng nói, các đặc trưng của một bức ảnh với phần mô tả bức ảnh trong bài toán thêm mô tả ảnh với các ngôn ngữ khác nhau. Đối với bài toán NMT, Bahdanau và cộng sự(2015)[2] đã áp dụng thành công cơ chế attention kết hợp việc dịch và liên kết các từ. Luong và cộng sự(2016)[3] đề xuất hai mô hình dịch máy dựa trên cơ chế attention: cơ chế attention toàn cục (global attention) "chú ý" (attend) đến toàn bộ từ trong câu đầu vào và cục bộ (local attention) chỉ "chú ý" đến một số từ trong phạm vi nhất định.



HÌNH 2.9: CƠ CHẾ ATTENTION

Trong đó:

- α_{ij} là trọng số của thời điểm j của bộ giải mã và timestep i của bộ mã hóa. Nói cách khác, đầu ra thứ j của bộ giải mã nên chú ý một lượng α_{ij} đến đầu vào thứ i của bộ mã hóa.
- h_j là trạng thái ẩn tại thời điểm i của bộ mã hóa.
- n là chiều dài của chuỗi tuần tự đầu vào.

α_{ij} được tính bằng cách lấy Softmax của Attention Score (e_{ij}):

$$\alpha_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k=1}^{m} exp(e_{ik}}$$

$$e_{ij} = f(s_{i-1}, h_j) = AlignScore(s_{i-1}, h_j)$$

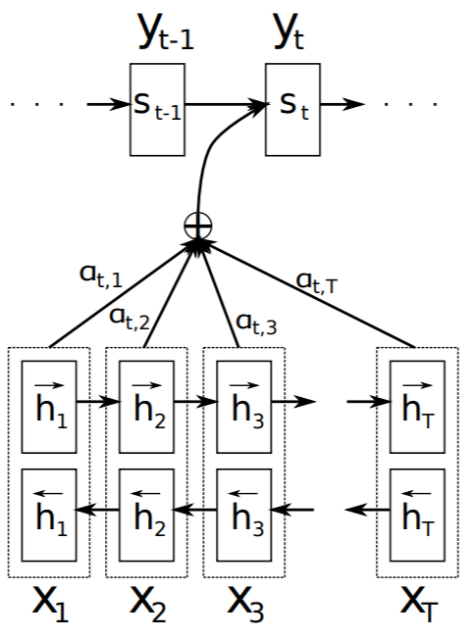
Trong đó:

- h_{i-1} là trạng thái ẩn tại thời điểm i-1 của bộ giải mã.
- s_j là trạng thái ẩn tại thời điểm j của bộ mã hóa.

Vectơ ngữ cảnh (context vector) c_{ij} sau đó được sử dụng để Decoder tính ra đầu ra y_i:

Cơ chế attention của Bahdanau: còn được gọi là Additive Attention do nó thực hiện phép kết hợp tuyến tính (phép cộng) giữa các trạng thái của bộ mã hóa và bộ giải mã, được tạo ra bởi Dzmitry Bahdanau trong bài báo vào năm 2014. Mục tiêu của nó là cải thiện hiệu năng của mô hình Seq2Seq bằng cách thay đổi đầu vào của bộ giải mã với các thông tin từ Input Sequence. Cụ thể ý tưởng như sau: toàn bộ các trạng thái ẩn của bộ mã hóa, h, và trạng thái ở bước $t - 1$ của bộ giải mã, s_{t-1}, được dùng để tính giá trị của vectơ ngữ cảnh c_t thay vì chỉ dùng duy nhất trạng thái ẩn cuối cùng, h_n, như mô hình seq2seq đã trình bày. Cơ chế attention liên kết chuỗi đầu vào và đầu ra bởi một hệ số liên kết, e_{tj}, được tham số hóa bởi một mạng biến đổi tuyến tính, a. Hệ số này giúp mô hình chú ý đến

các phần thông tin liên quan với đầu vào hiện tại trong câu. Sau đó, mô hình dự đoán kết quả tiếp theo y_t dựa trên vectơ ngữ cảnh c_t ứng với đầu vào xi và kết quả trước đó y_{t-1}



HÌNH 2.10: BAHDANAU ATTENTION

- Tạo bộ mã hóa cho lớp ẩn
- Tính toán điểm số liên kết (Alignment Score): hệ số/điểm liên kết e_{ij} biểu thị mức độ khớp giữa bộ giải mã trước s_{t-1} với mỗi lớp ẩn của bộ mã hóa h_t

$$e_{tj}=a\ (s_{t-1},h_t)$$

- Tính toán softmax của điểm số liên kết e_{tj}, ta thu được trọng số attention α_{tj}.

$$\alpha_{tj}=\frac{\exp(e_{tj})}{\sum_{k=1}^{n}\exp(e_{tk})}$$

Hàm **softmax** cho kết quả là một phân phối xác suất với tổng là 1. Bằng việc chuẩn hóa với hàm **softmax** giúp mô hình biểu diễn mức độ ảnh hưởng của mỗi từ trong chuỗi đầu vào. Trọng số attention tại vị trí nào càng cao (xác suất càng lớn) sẽ càng quan trọng trong việc dự đoán từ kế tiếp.

- Tính toán vectơ ngữ cảnh: nhân các bộ mã hóa lớp ẩn với điểm số liên kết tương ứng của nó.

$$c_t = \sum_{j=1}^{n} \alpha_{tj} h_j$$

Do hàm softmax ở bước trước, nếu trọng số tại một vị trí k ∈ [1, n] càng gần với 1 thì ảnh hưởng có nó lên việc đưa ra kết quả y_t sẽ được khuếch đại. Ngược lại, khi trọng số gần bằng 0 thì ảnh hưởng của nó sẽ giảm và bị vô hiệu hóa.

Cuối cùng, vectơ ngữ cảnh c_t được kết hợp với vectơ biểu diễn của kết quả trước đó, y_{t-1}, thông qua phép CONCAT, và truyền vào khối giải mã RNN (có thể là LSTM hoặc GRU) cho ra trạng thái s_t của bộ giải mã. Để có thể đưa ra dự đoán y_t, trạng thái mới s_t phải qua một lớp biến đổi tuyến tính đóng vai trò như một bộ phân lớp cho ra các giá trị xác suất của từ kế tiếp y_t.

Cơ chế attention của Luong: Luong và cộng sự, (2016)[3] đã đề xuất một cơ chết attention mới hay còn được gọi là Multiplicative Attention, kế thừa từ Bahdanau Attention. Dựa trên số lượng phần tử đầu vào thì Luong attention được chia làm: attention toàn cục và attention cục bộ.

Hai điểm khác biết chủ yếu giữa Luong Attention và Bahdanau Attention là có đến 2 cơ chế để tính điểm số liên kết so với chỉ 1 trong Bahdanau attention và vị trí của attention có sự khác biệt so với Bahdanau attention, còn lại cơ bản là giống nhau. Mục tiêu vẫn là nhận được vectơ ngữ cảnh c_t. Cụ thể, với trạng thái ẩn h_t của bộ giải mã và vectơ ngữ cảnh c_t. Trạng thái ẩn mới kết hợp vectơ ngữ cảnh được tính như sau:

$$\hat{S}_t = \tanh(W_c[c_t; h_t])$$

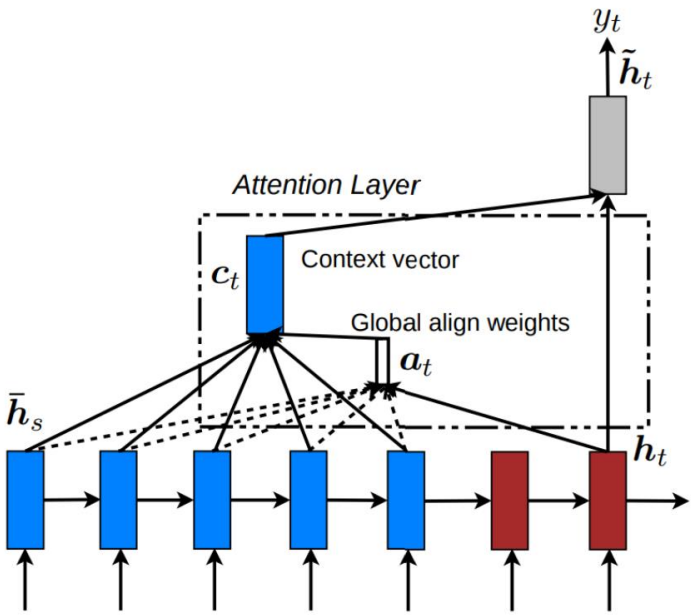
Sau đó $\hat{S}_t$ được truyền vào mạng biến đổi tuyến tính và hàm softmax để dự đoán y_t:

$$P(y_t | t<i, x) = \text{softmax}(W_s \hat{S}_t)$$

- Cơ chế attention toàn cục(Global/soft attention): Ý tưởng là tận dụng toàn bộ trạng thái của bộ mã hóa khi tính vectơ ngữ cảnh c_t (hình 2.10). Trọng số liên kết toàn cục a_t với kích thước bằng với số từ trong câu nguồn, là kết quả của phép so sánh giữa trạng thái của bộ giải mã h_t với từng trạng thái ẩn của bộ mã hóa h_s.

$$a_t(s) = align(h_t, \bar{h}_s) = \frac{exp(score(h_t, \bar{h}_s))}{exp(\sum_{s'} score(h_t, \bar{h}_{s'})}$$

$$score(h_t, \bar{h}_s) = \begin{Bmatrix} h_t^T \bar{h}_s & dot \\ h_t^T W_a \bar{h}_s & general \\ v_a^T tanh(W_a[h_t; \bar{h}_s] & concat \end{Bmatrix}$$



HÌNH 2.11: CƠ CHẾ ATTENTION TOÀN CỤC

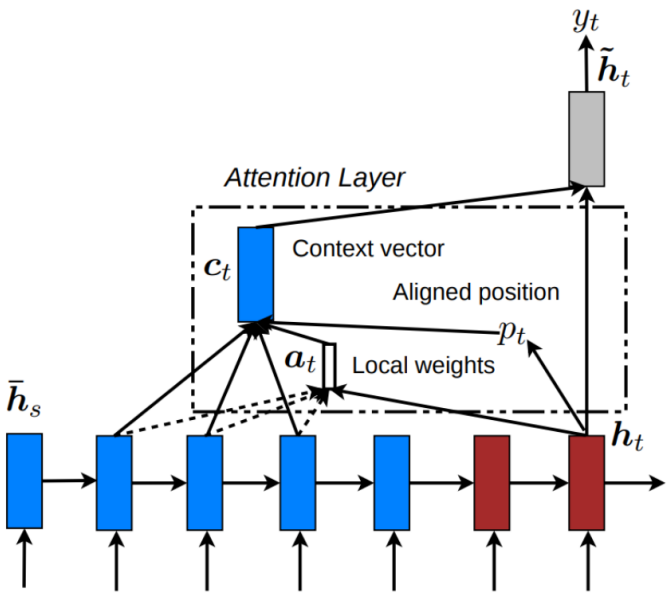
Gọi các vectơ liên kết a là các trọng số attention. Vectơ ngữ cảnh c_t được tính như sau:

$$c_t = \frac{1}{|a|} \sum_{j=1} a_{tj} y_j$$

- <u>Cơ chế attention cục bộ(local/hard attention):</u> Attention toàn cục có một nhược điểm là nó yêu cầu tài nguyên tính toán khá lớn, nhất là đối với các bài toán mà Input Sequence có chiều dài lớn. Đó chính là lý do Attention cục bộ ra đời. Nó giải quyết vấn đề của Attention toàn cục bằng cách chỉ sử dụng một số lượng nhất định bộ mã hóa của lớp ẩn thay vì tất cả (thông qua cửa sổ quét với kích thước D – kích thước của các từ trong câu đầu vào mỗi lần dự đoán).

Đầu tiên, mô hình phát sinh một vị trí kiên kết p_t cho mỗi từ đầu ra tại thời điểm t. Vectơ ngữ cảnh ci được tính bởi trung bình trọng số trên các trạng thái ẩn trong phạm vi $[p_t - D; p_t + D]$, D là kích thước cửa sổ quét. Khi đó, vectơ liên kết cục bộ ai là vectơ với số chiều cố định, $a_t \in R^{2D+1}$ và được tính như sau:

$$a_t = \text{align}(h_t, h_s)\exp\left(-\frac{(t-p_t)^2}{2\sigma^2}\right)$$



HÌNH 2.12: CƠ CHẾ ATTENTION CỤC BỘ

Trong đó $\sigma = \frac{D}{2}$ và p_t được tính bằng một trong hai cách.

- Monotonic alignment (local m): Gán $p_t = t$ với giả thuyết là câu đầu vào và đầu ra được sắp xếp một cách đơn điệu. Khi đó a_t được tính như attention toàn cục.
- Predictive alignment (local p) được tính như sau:

$$p_t = S.\ \text{sigmoid}(v_p^T \tanh(W_p h_t))$$

Trong đó S là chiều dài của câu đầu vào, W_p và v_p là tham số của mô hình.

CHƯƠNG 3. MỘT SỐ NGÔN NGỮ KHU VỰC ĐÔNG NAM Á

Trước khi đi vào trích xuất và xử lý các ngôn ngữ lớn thuộc khu vực Đông Nam Á, cụ thể các ngôn ngữ được lựa chọn ở đây là tiếng Việt, tiếng Anh, tiếng Trung, tiếng Indonesia và tiếng Malaysia. Ta cần hiểu tiêu chuẩn để lựa chọn các ngôn ngữ để làm kho ngữ liệu và sau đó ta cần hiểu về tính chất, đặc điểm các ngôn ngữ để thuận tiện cho quá trình xử lý.

3.1 LỰA CHỌN NGÔN NGỮ

Tiêu chuẩn đầu tiên để lựa chọn ngôn ngữ đó là ngôn ngữ đó phải là ngôn ngữ lớn trong khu vực Đông Nam Á (có số lượng người sử dụng lớn, hoặc có nhu cầu về dịch thuật trong khu vực lớn). Nếu xét về điều này thì những ngôn ngữ sau có thể được xét tới. Đầu tiên là tiếng Anh khi nó là ngôn ngữ toàn cầu (với khoảng 1,26 tỷ người sử dụng như ngôn ngữ mẹ đẻ và ngoại ngữ). Bên cạnh đó, tiếng Anh cũng được coi là ngôn ngữ được sử dụng chính thức tại Philippin và Singapore. Thứ hai, ta phải xét đến tiếng Trung Quốc khi Trung Quốc là thị trường lớn với hơn 1,4 tỷ dân (số liệu 2021), là đối tác thương mại lớn với các nước trong khu vực Đông Nam Á nên nhu cầu về dịch thuật là rất lớn. Hơn thế nữa, tiếng Trung Quốc cũng đang trở nên phổ biến tại một số nước như Myanmar, Campuchia và là một ngôn ngữ chính thức tại Singapore. Nếu xét về các ngôn ngữ chính tại các quốc gia mà có số lượng người sử dụng đông đảo (tính theo số liệu 2020) thì chúng ta có thể kể đến: tiếng Indonesia (199 triệu), Philippin(90 triệu), Việt Nam (97 triệu), tiếng Malaysia(77 triệu), tiếng Thái Lan (61 triệu), và tiếng Myanmar(32 triệu).

Tiêu chuẩn thứ hai cần xét đến ở đây là số lượng bài viết có trên Wikipedia phải lớn. Điều này thì đa phần các ngôn ngữ được đề cập ở trên bao gồm: tiếng Anh, tiếng Trung, tiếng Indonesia, tiếng Việt, tiếng Malaysia, Tiếng Thái Lan đều thỏa mãn khi có số lượng bài viết lớn (tối thiểu là tiếng Malaysia với 300 ngàn bài viết). Tuy nhiên đối với tiếng Philippin thì số lượng bài viết được viết bằng tiếng tagalog (ngôn ngữ chính thức tại Philippin, bên cạnh tiếng Anh) chỉ có 70 ngàn bài viết. Tiếng Burmese (ngôn ngữ chính thức tại Myanmar) thì chỉ có 40 ngàn bài viết. Số lượng bài viết hạn chế dẫn đến

số lượng bài viết chung sẽ ít và số lượng cặp câu song ngữ trích xuất được cũng sẽ hạn chế.

Cuối cùng ta phải xem xét xem có hạn chế, khó khăn nào đối với việc tiền xử lý ngôn ngữ đó hay không? Ở đây ta bắt buộc phải loại bỏ tiếng Thái ra khỏi danh sách các ngôn ngữ được lựa chọn vì một vài lý do như khó khăn trong việc xác định ranh giới câu, khuyết thiếu các công cụ, thư viện có sẵn để tiền xử lý câu tiếng Thái.

Sau khi trải qua quá trình xác định ngôn ngữ phục vụ cho quá trình khai thác thì chúng ta cần hiểu rõ đặc điểm của ngôn ngữ phục vụ cho việc tiền xử lý dữ liệu được tốt hơn.

3.2 PHÂN LOẠI VỀ NGUỒN GỐC

- Tiếng Việt là ngôn ngữ thuộc nhóm Việt Mường, ngành Môn-Khmer, dòng Nam Á của ngữ hệ Nam Phương.
- Tiếng Trung là ngôn ngữ thuộc dòng Hán của ngữ hệ Hán Tạng.
- Tiếng Anh là ngôn ngữ thuộc ngữ hệ Ấn Âu.
- Tiếng Indonesia và Malaysia là những ngôn ngữ thuộc ngữ hệ Nam Đảo.

3.3 PHÂN LOẠI VỀ LOẠI HÌNH.

- Ngôn ngữ đơn lập (isolate): tiếng Việt và tiếng Trung.
- Ngôn ngữ hòa kết (flexional): tiếng Anh.
- Ngôn ngữ chắp dính (agglutinate): tiếng Indonesia và tiếng Malaysia.

3.4 PHÂN LOẠI VỀ CẤU TRÚC CÂU:

3.4.1 ĐIỂM CHUNG:

- Tất cả các ngôn ngữ trên đều theo cấu trúc S - V - O (chủ ngữ - động từ - tân ngữ).

3.4.2 ĐIỂM PHÂN BIỆT VỀ TRẬT TỰ TỪ:

- Vị trí giới từ:
  - Tiếng Trung: cụm giới từ theo ngay sau cụm danh từ làm chủ ngữ.
  - Tiếng Anh, tiếng Việt: cụm giới từ đứng đầu câu trước danh từ hay sau động từ ở cuối câu.

- Tiếng Malaysia: giới từ đứng trước danh từ chính, tính từ và danh từ phụ đứng sau danh từ chính.
- Tiếng Indonesia: giới từ đứng trước danh từ chính, tính từ đứng sau danh từ.

- Cấu trúc của cụm danh từ:
  - Tiếng Trung: có nhiều cách để tạo nên cụm giới từ và danh từ tương tự tiếng Anh.
  - Tiếng Việt : mọi thay đổi đều diễn ra sau danh từ chính.

3.4.3 BẢNG CHỮ CÁI

- Tiếng Anh, tiếng Việt, tiếng Malaysia và tiếng Indonesia đều sử dụng bảng chữ cái latin.
- Tiếng Trung lại sử dụng bảng chữ cái tiếng Hán (Hán tự).

3.4.4 DẤU THANH

- Tiếng Việt có 6 thanh (ngang, sắc, huyền, hỏi, ngã và nặng).
- Tiếng Trung có 4 thanh điệu chính và một thanh điệu nhẹ.
- Tiếng Anh, tiếng Malaysia và tiếng Indonesia không sử dụng thanh diệu.

3.5 ĐẶC ĐIỂM CỦA TIẾNG VIỆT

Tiếng Việt được xếp vào loại hình đơn lập (isolate) hay còn gọi là loại hình phi hình thái, không biến hình, đơn tiết với những đặc điểm chính như sau:

- Trong hoạt động ngôn ngữ, từ không biến đổi hình thái. Ý nghĩa ngữ pháp nằm ở ngoài từ. Ví dụ: “Tôi nhìn anh ấy” và “Anh ấy nhìn tôi”.
- Phương thức ngữ pháp chủ yếu là: trật tự từ và từ hư. Ví dụ Gạo xay và Xay gạo.
- Tồn tại một loại đơn vị đặc biệt là “hình tiết” mà vỏ ngữ âm của chúng trùng khít với âm tiết, và đơn vị đó cũng chính là “hình vị tiếng Việt” hay còn gọi là “tiếng” (tiếng Việt sử dụng khoảng 10000 tiếng)
- Ranh giới từ không được xác định mặc nhiên như các thứ tiếng biến hình khác. Ví dụ: “Học sinh học sinh học”. Điều này dẫn đến việc phân tích hình thái (tách từ) tiếng Việt trở nên khó khăn. Việc nhận diện ranh giới từ là quan trọng và là

tiền đề cho các xử lý tiếp theo: kiểm lỗi chính tả, gán nhãn từ loại, thống kê tần suất từ, ...

- Tồn tại loại từ đặc biệt "từ chỉ loại" (Classifier) hay còn gọi là phó danh từ đi kèm với danh từ, như : cái bàn, cuốn sách, con chó ...
- Về mặt ngữ âm học, các âm tiết tiếng Việt đều mang một trong 6 thanh điệu (ngang, sắc, huyền, hỏi, ngã, nặng). Đây là âm vị siêu đoạn tính.
- Có hiện tượng láy từ trong tiếng Việt, như: lấp lánh, lung linh... Ngoài ra còn có hiện tượng nói lái (do mối liên kết giữa phụ âm đầu và phần vần trong âm tiết là lỏng lẻo), như: hiện đại → hại điện, thầy giáo → tháo giầy, ...

3.6 ĐẶC ĐIỂM CỦA TIẾNG ANH

Tiếng Anh được xếp vào loại hình biến cách (flexion) hay còn gọi là loại hình khuất chiết với những đặc điểm chính như sau:

- Trong hoạt động ngôn ngữ, từ có biến đổi hình thái. Ý nghĩa ngữ pháp nằm ở trong từ. Ví dụ: I see him và He sees me.
- Phương thức ngữ pháp chủ yếu là: phụ tố. Ví dụ: learning và learned.
- Hiện tượng cấu tạo từ bằng cách ghép thêm phụ tố (affix) vào gốc từ là rất phổ biến. Ví dụ: anticomputerizational (anti-comput-er-ize-ation-al)
- Kết hợp giữa các hình vị là chặt chẽ. Ranh giới giữa các hình vị là khó xác định.
- Ranh giới từ được nhận diện bằng khoảng trắng hoặc dấu câu.

3.7 ĐẶC ĐIỂM CỦA TIẾNG TRUNG

Tiếng Trung được xếp vào loại hình đơn lập (isolate) với những đặc điểm chính như sau:

- Hiện tượng biến hóa về hình thái không phát triển, không phổ biến. (động từ không bị biến hóa về hình thái).
- Thứ tự từ và từ hư rất được coi trọng.
- Danh từ muốn nhắc tới luôn phải đặt phía sau câu.
- Có 4 thanh điệu chính và một thanh điệu nhẹ.

- Từ trong tiếng Trung không được chia tách rõ ràng và không có khoảng trắng giữa các từ. Tiền xử lý câu tiếng Trung cũng là một vấn đề khó giải quyết.

3.8 ĐẶC ĐIỂM CỦA TIẾNG INDONESIA

Tiếng Indonesia được xếp vào loại hình chắp dính thuộc ngữ hệ Nam Đảo với những đặc điểm như sau:

- Các tính từ, đại từ chỉ định và đại từ sở hữu theo sau danh từ mà chúng xác định.
- Trật tự từ thường là chủ ngữ, động từ rồi đến tân ngữ (S – V – O).
- Những từ mới thường được hình thành thông qua ba phương pháp: phụ tố hóa (thêm các phụ tố lên từ gốc), hình thành một từ ghép (tổ hợp của hai hoặc nhiều từ riêng biệt), hay phép lặp lại (lặp lại các từ hay các phần của từ).
- Tiếng Indonesia sử dụng một hệ thống phức tạp các phụ tố (tiền tố, trung tố, hậu tố, và phụ tố tình huống (confix, circumfix)). Các phụ tố được áp dụng với quy tắc nhất định phụ thuộc vào chữ cái khởi đầu của từ cơ sở.
- Những từ mới có thể được tạo thành bằng cách nối hai hoặc nhiều từ cơ sở. Các từ ghép, khi chúng tồn tại tự do trong một câu, thường được viết rời. Các từ ghép chỉ được ghép lại với nhau khi chúng được giới hạn bởi confix hoặc khi chúng đã được coi như những từ bền vững. Ví dụ, rumah (ngôi nhà) và makan (ăn), được ghép lại và tạo thành một từ mới rumah makan (nhà hàng, nhà ăn).
- Ranh giới từ được nhận diện bằng khoảng trắng hoặc dấu câu.

3.9 ĐẶC ĐIỂM CỦA TIẾNG MALAYSIA

Tiếng Malaysia được xếp vào loại hình chắp dính thuộc ngữ hệ Nam Đảo với những đặc điểm tương tự về ngữ pháp như tiếng Indonesia. Nhưng về số lượng và ý nghĩa từ vựng lại có nhiều điểm khác biệt. Từ vựng của tiếng Malaysia chủ yếu nguồn gốc từ tiếng Java, tiếng Hà Lan và tiếng Mã Lai vùng đảo. Còn tiếng Malaysia bị ảnh hưởng bởi tiếng tiếng Tây Ban Nha và tiếng gốc Mã Lai.

CHƯƠNG 4. PHƯƠNG HƯỚNG TIẾP CẬN

Trong khi dịch máy mạng neural (NMT) có được những cải tiến đột phá nhưng vẫn đặc biệt nhạy cảm với quy mô và chất lượng của dữ liệu dành cho huấn luyện (Koehn và Knowles, 2017 [7]; Khayrallah và Koehn, 2018 [8]). Trong bối cảnh này, các cách tiếp cận hiệu quả để khai thác và lọc kho ngữ liệu song song là rất quan trọng để áp dụng NMT trong các môi trường thực tế.

Ý tưởng cơ bản của cách tiếp cận khai thác được sử dụng trong nghiên cứu này là trước tiên học cách nhúng câu đa ngôn ngữ, tức là một không gian nhúng trong đó các câu tương tự về mặt ngữ nghĩa gần giống nhau mà không phụ thuộc vào ngôn ngữ mà chúng được viết.

Điều này có nghĩa là khoảng cách trong không gian đó có thể được sử dụng như một chỉ báo cho biết hai câu có phải là bản dịch lẫn nhau hay không. Sử dụng một ngưỡng tuyệt đối về khoảng cách cosin đã được chứng minh là đạt được một kết quả tốt (Schwenk, 2018 [9]).

Tuy nhiên, người ta đã khảo sát và thấy rằng một ngưỡng tuyệt đối về khoảng cách cosin là không phù hợp đối với tất cả các cặp ngôn ngữ, ví dụ: (Guo và cộng sự, 2018 [10]).

Khó khăn để chọn một ngưỡng toàn cục được nhấn mạnh trong cài đặt của chúng tôi vì chúng tôi đang khai thác các câu song song cho nhiều cặp ngôn ngữ khác nhau.

4.1 TIÊU CHUẨN BIÊN

Các phương pháp sử dụng NMT lấy cảm hứng từ kiến trúc mã hóa giải mã encoder - decoder để huấn luyện không gian nhúng câu đa ngôn ngữ hiện có, sau đó được áp dụng trực tiếp để truy xuất và lọc các câu song song mới sử dụng truy xuất các câu hàng xóm gần nhất qua độ tương tự cosine với ngưỡng cứng (Espana-Bonet và cộng sự, 2017[11]; Hassan và cộng sự, 2018[12]; Schwenk, 2018[9]).

$$similarity(A, B) = \cos(x, y) = \frac{A.B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Độ tương đồng cosin của hai vectơ A, B được tính là cosin góc tạo bởi hai vectơ A, B n chiều trong không gian n chiều đó. Điểm độ tương đồng được tính bằng tích vô hướng của hai vectơ chia cho độ dài của hai vectơ đó (công thức 1). Giá trị của độ tương đồng sẽ nằm trong khoảng từ -1 tương ứng với hoàn toàn khác nhau và 1 tương ứng với hoàn toàn tương đồng với nhau.

Trong nghiên cứu này, chúng tôi lập luận rằng phương pháp truy xuất này có tỷ lệ gặp phải độ tương đồng cosine không nhất quán trên toàn bộ ngữ liệu. Ví dụ, một số câu không có bất kỳ bản dịch chính xác nào có điểm cosine tổng thể cao, làm cho chúng xếp hạng cao hơn các câu khác với một bản dịch chính xác. Vấn đề này cũng đã chỉ ra bởi (Guo và cộng sự, 2018 [10]). Chúng tôi đề xuất phương pháp giải quyết vấn đề này bằng cách coi biên của hai câu là khoảng cách cosin giữa hai câu và khoảng cách cosin của một câu với từng câu hàng xóm gần nhất của nó trong ngôn ngữ khác.

Chúng tôi coi biên độ giữa khoảng cách cosin của một câu ứng cử viên đã cho và cosine trung bình của k câu hàng xóm gần nhất của nó theo cả hai hướng như sau:

$$Score(x, y) = margin(cos(x, y), \frac{\sum_{z \in NNk(x)} cos(x, z) + \sum_{z \in NNk(y)} cos(y, z)}{2k}). \quad (2)$$

trong đó NNk(x) biểu thị k câu hàng xóm duy nhất gần nhất của x trong ngôn ngữ khác, và tương tự với NNk(y). Để lấy trung bình khoảng cách cosin cả hai hướng với mỗi hướng k câu, chúng tôi đã chia tổng kết quả cho 2k. Chúng tôi đã sử dụng k = 4 trong tất cả các thí nghiệm. Công thức (2) có một vài biến thể như sau:

- Tuyệt đối: margin(a,b) = a, điều này tương đương với khoảng cách cosin và là cơ sở cho nghiên cứu của chúng tôi.
- Khoảng cách: margin(a,b) = a-b, khoảng cách cosin trừ đi trung bình khoảng cách cosin từ hai câu đến các câu hàng xóm gần nhất. Điều này được lấy cảm hứng từ điểm CSLS (Conneau và cộng sự, 2018 [13]) để tránh xảy ra vấn đề một

vài điểm xuất hiện thường xuyên hơn so với các điểm còn lại khi tìm kiếm K điểm lân cận khi gia tăng kích thước dữ liệu (vấn đề **hubness**) trong nhiệm vụ cảm ứng từ vựng song ngữ (Bilingual Lexicon Induction - BLI) qua nhúng câu đa ngôn ngữ.

- Tỷ lệ : margin = $\frac{a}{b}$, tỷ lệ giữa khoảng cách cosin 2 câu và trung bình khoảng cách giữa hai câu ứng viên tới các hàng xóm gần nhất của nó.

Chất lượng dóng hàng đã được chứng minh hiệu quả hơn khi sử dụng tiêu chuẩn biên thay vì tiêu chuẩn ngưỡng tuyệt đối (Artetxe và Schwenk, 2018a [14]).

Trong nghiên cứu này, chúng tôi đã chọn biên của hai câu x và y là tỷ lệ khoảng cách cosin giữa 2 câu x,y và trung bình cộng khoảng cách hàng xóm gần nhất theo cả hai hướng:

$$Margin(x,y) = \frac{2k * cos(x,y)}{\sum_{z \in NNk(x)} cos(x,z) + \sum_{z \in NNk(y)} cos(y,z)}. \qquad (3)$$

Khi khai thác các câu song song, chúng tôi khám phá các chiến lược sau đây để tạo ra các câu ứng viên:

- Tiến tới: Mỗi câu nguồn đều được căn chỉnh với chính xác một câu đích cho điểm tốt nhất. Một số câu đích có thể được căn chỉnh với nhiều câu nguồn hoặc không có câu nào.
- Quay lùi : Tương đương với chiến lược tiến tới, nhưng đi theo hướng ngược lại. Một câu đích chỉ căn chỉnh với một câu nguồn nhưng một câu nguồn thì có thể căn chỉnh với nhiều câu đích.
- Giao nhau: Giao của các ứng cử viên tiến và lùi, loại bỏ các câu có sự liên kết không nhất quán.
- Điểm số tối đa :Sự kết hợp của các ứng viên tiến và các ứng cử viên lùi, nhưng thay vì loại bỏ tất cả các liên kết không nhất quán, nó sẽ chọn những ứng viên có điểm số cao nhất.

Chúng tôi tuân theo chiến lược "điểm số tối đa" như được mô tả trong (Artetxe và Schwenk, 2018a [14]): tiêu chuẩn biên trước tiên được tính theo cả hai hướng cho tất cả các câu trong ngôn ngữ L1 và L2. Sau đó, chúng tôi kết hợp những câu ứng viên tiến và lùi này. Câu ứng viên được sắp xếp và bắt cặp với các câu nguồn hoặc đích đã được sử dụng sẽ bị bỏ qua.

Chúng tôi sau đó áp dụng một ngưỡng (threshold) về điểm biên để xác định xem hai câu có phải là bản dịch lẫn nhau hay không, cũng như đạt được kết quả và kích thước kho ngữ liệu tốt hơn. Lưu ý rằng với kỹ thuật này, chúng tôi luôn có được các câu được căn chỉnh giống nhau, độc lập với hướng khai thác, ví dụ: tìm kiếm bản dịch của tiếng Việt trong ngữ liệu tiếng Anh và ngược lại.

Độ phức tạp của phương pháp khai thác dựa trên khoảng cách là O (N × M), trong đó N và M là số lượng câu trong mỗi ngữ liệu đơn ngữ. Điều này làm cho cách tiếp cận vét cạn với tính toán khoảng cách lớn trở nên khó thực hiện đối với kho ngữ liệu lớn. Khai thác dựa trên tiêu chuẩn biên đã được chứng minh là hoạt động tốt hơn đáng kể, đạt kết quả cao trên share-task của hội thảo về xây dựng và sử dụng kho ngữ liệu so sánh được - Building and Using Comparable Corpora (BUCC) (Artetxe và Schwenk, 2018a [14]). Kho ngữ liệu BUCC khá nhỏ: nhiều nhất là 567 nghìn câu.

Ví dụ đối với cặp tiếng Anh (150 triệu câu) và tiếng Trung(18 triệu câu) thì cần 2,7 *10^15 phép tính toán để có thể dóng hàng câu. Điều này cần lượng tính toán quá lớn, yêu cầu dung lượng ram và thời gian tính toán quá lâu. Giải pháp cho vấn đề này sẽ được trình bày trong mục 4.3.

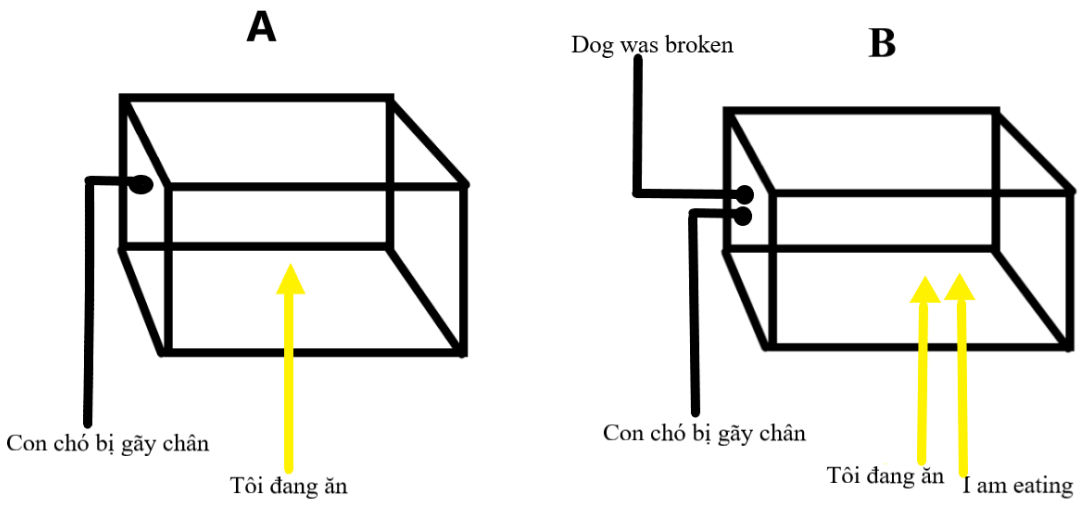
4.2 KHÔNG GIAN VECTƠ CÂU CHUNG

Phương pháp khai thác này dựa chủ yếu vào không gian vectơ câu chung cho tất cả các ngôn ngữ.

Ban đầu, người ta cố gắng huấn luyện một không gian vectơ câu chung cho mỗi cặp ngôn ngữ (minh họa không gian nhúng câu đơn ngữ và song ngữ ở hình 4.1), ví dụ: (Guo và cộng sự, 2018 [10],Espana- Bonet và cộng sự, 2017[11]; Hassan và cộng sự, 2018 [12];]; Yang và cộng sự, 2019[15]), nhưng điều này rất khó mở rộng đến hàng

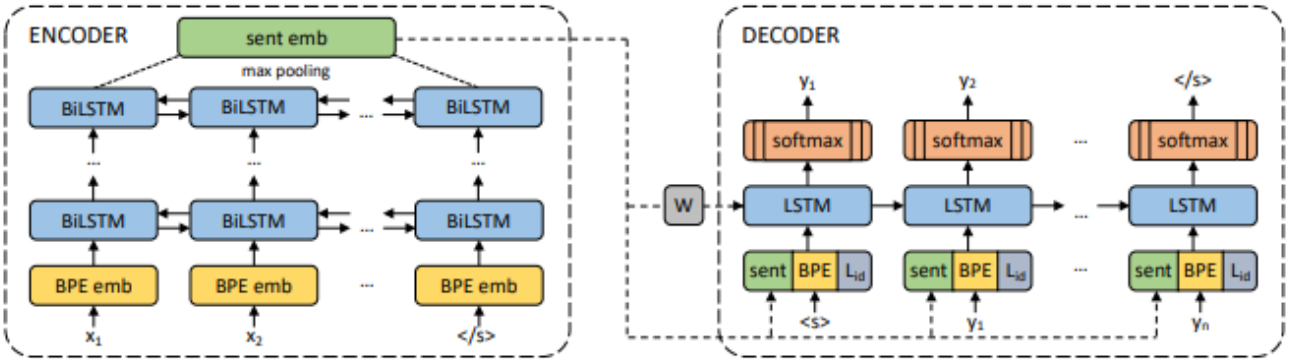
nghìn cặp ngôn ngữ có trong Wikipedia. Thay vào đó, chúng tôi chọn sử dụng một không gian vectơ nhúng câu đa ngôn ngữ cho tất cả mọi ngôn ngữ, cụ thể là ngôn ngữ được đề xuất bởi bộ công cụ LASER mã nguồn mở (Artetxe và Schwenk, 2018b[16]).

Huấn luyện bằng cách nhúng đa ngôn ngữ vào không gian vectơ câu chung trên nhiều ngôn ngữ cùng một lúc cũng có lợi thế mà các ngôn ngữ tài nguyên thấp có thể được hưởng lợi từ sự tương tự với ngôn ngữ khác trong cùng một ngữ hệ. Ví dụ, chúng tôi có thể khai thác dữ liệu song ngữ cho một số ngôn ngữ mặc dù chúng không được sử dụng để huấn luyện không gian vectơ nhúng câu chung **LASER**. Hướng phát triển tương lai, chúng tôi có thể khai thác ngữ liệu song ngữ cho một số ngôn ngữ thiểu số nhóm Việt - Mường, ngành Môn - Khmer như tiếng Mường, tiếng Khmer.



HÌNH 4.1: A: KHÔNG GIAN NHÚNG CÂU ĐƠN NGỮ, B KHÔNG GIAN NHÚNG CÂU ĐA NGỮ

Ý tưởng cơ bản của LASER là huấn luyện hệ thống tuần tự (sequence-to-sequence) trên nhiều cặp ngôn ngữ cùng một lúc sử dụng chung một bộ từ vựng BPE và một bộ mã hóa cho mọi ngôn ngữ. Vectơ đại diện cho câu đạt được thông qua tổng hợp tối đa (max-pooling) trên đầu ra của bộ mã hóa.



HÌNH 4.2: MÔ HÌNH SỬ DỤNG (THAM KHẢO)

Các nhúng câu này được sử dụng để khởi tạo bộ giải mã LSTM thông qua một phép biến đổi tuyến tính và cũng được nối với các lần nhúng đầu vào của nó ở mỗi thời điểm t. Lưu ý rằng không có kết nối nào khác giữa bộ mã hóa và bộ giải mã, vì chúng tôi muốn tất cả thông tin liên quan của trình tự đầu vào được chụp bởi nhúng câu.

Các câu sau khi được nhúng sẽ được sử dụng để khởi tạo bộ giải mã LSTM thông qua một phép biến đổi tuyến tính. Bên cạnh đó chúng cũng được nối lại với đầu vào của bộ giải mã mỗi thời gian thực thi t. Lưu ý là không còn kết nối nào khác giữa bộ mã hóa và bộ giải mã với mong muốn mọi thông tin liên quan sẽ được lưu trữ trong nhúng câu (sentence embeddings).

Chúng tôi sử dụng một bộ mã hóa và bộ giải mã duy nhất trong hệ thống của mình, được dùng chung cho tất cả các ngôn ngữ có liên quan. Vì mục đích đó, chúng tôi đã xây dựng một bộ 50k từ vựng BPE (Byte-Pair Encoding), được huấn luyện trên tất cả các cặp đào tạo. Bằng cách này thì bộ mã hóa sẽ học mà không có tín hiệu rõ ràng về rõ ràng về ngôn ngữ đầu vào là gì, khuyến khích nó để học các biểu diễn độc lập với ngôn ngữ. Ở chiều ngược lại thì bộ giải mã chỉ nhận một bộ mã hóa để xác định ngôn ngữ (LID), được nối với đầu vào và nhúng câu ở mọi bước thời gian.

Trong lúc đào tạo mô hình không gian vectơ câu chung đa ngôn ngữ. Mở rộng quy mô lên đến gần một trăm ngôn ngữ yêu cầu một bộ mã hóa có đủ dung lượng. Ở trong nghiên cứu của tác giả mô hình, tác giả giới hạn nghiên cứu của mình ở một BiLSTM xếp chồng lên nhau với 1 đến 5 lớp, mỗi lớp 512 chiều. Vectơ đại diện cho câu là 1024 chiều sau khi lấy tổng hợp tối đa trên đầu ra của bộ mã hóa. Bộ giải mã có một lớp LSTM 2048 chiều, LID là 32 chiều và nhúng câu của câu đầu vào bộ giải mã là 320 chiều.

4.3 TÌM KIẾM CÂU TƯƠNG ĐỒNG VÀ DÓNG HÀNG CÂU

Tìm kiếm độ tương đồng quy mô lớn một cách nhanh chóng là lĩnh vực nghiên cứu lớn. Thông thường, ứng dụng chủ yếu của tìm kiếm độ tương đồng là tìm kiếm hình ảnh, nhưng các thuật toán tìm kiếm lại có thể được áp dụng cho bất kỳ loại vectơ nào.

Trong nghiên cứu này, chúng tôi sử dụng thư viện **FAISS** mã nguồn mở, triển khai các thuật toán hiệu quả cao để thực hiện tìm kiếm tương đồng trên hàng tỷ vectơ (Johnson và cộng sự, 2017 [17]). Một lợi thế nữa là FAISS có hỗ trợ chạy trên nhiều GPUs cùng lúc. Biểu diễn câu của chúng tôi là 1024 chiều. Điều này có nghĩa là nhúng câu của tất cả các câu tiếng Anh yêu cầu 150x10^6 × 1024 × 4 = 572GB RAM bộ nhớ.

Do đó, giảm kích thước và nén dữ liệu là điều cần thiết để tìm kiếm hiệu quả. Trong nghiên cứu này, chúng tôi đã chọn một cách nén khá mạnh dựa trên nén về dạng (64-bit product quantizer) (Jegou và cộng sự, 2011 [18]), và đưa về không gian tìm kiếm 32 ngàn ô. Điều này tương ứng với loại chỉ mục "OPQ 64, IVF 32768, PQ 64"[19] trong chỉ số **FAISS**. Bên cạnh đó còn nhiều kiểu nén khác, điều này sẽ được thảo luận thêm trong nghiên cứu sau. Chúng tôi xây dựng và huấn luyện một chỉ số **FAISS** cho mỗi ngôn ngữ.

Chỉ số FAISS được nén cho tiếng Anh chỉ yêu cầu 9,2 GB, tức là hơn năm mươi lần nhỏ hơn các câu gốc nhúng. Điều này giúp bạn có thể tải toàn bộ chỉ mục trên một GPU tiêu chuẩn và để chạy tìm kiếm một cách rất hiệu quả trên nhiều GPU song song, không cần thiết phải chia nhỏ chỉ mục.

CHƯƠNG 5. QUÁ TRÌNH THỰC HIỆN

Đối với mỗi bài viết Wikipedia, chúng ta có thể lấy liên kết đến bài viết tương ứng là một phần bản dịch của bài viết gốc. Điều này có thể được sử dụng để khai thác các câu giới hạn cho các bài báo tương ứng. Cách khai thác ngữ liệu Wikipedia này có một số lợi thế như sau:

- Khai thác rất nhanh vì mỗi bài báo thường chỉ có khoảng vài trăm câu.
- Dễ dàng để tìm thấy bản dịch của một câu trong bài báo liên kết so với tìm kiếm trên toàn bộ các bài viết Wikipedia.

Mặt khác, chúng tôi giả thuyết rằng tiêu chí cận biên sẽ là kém hiệu quả hơn vì một bài báo thường có ít câu giống nhau. Điều này có thể dẫn đến kho ngữ liệu khai thác được sẽ xuất hiện các cặp câu tương đồng có dạng : “NAME was born on DATE in CITY”, “BUILDING is a monument in CITY built on DATE”, ... Mặc dù, những cặp câu đó có thể được dóng hàng đúng. Chúng tôi giả thuyết rằng chúng nên được hạn chế để sử dụng trong việc huấn luyện mô hình dịch máy trong trường hợp chúng xuất hiện quá thường xuyên trong kho ngữ liệu. Nói chung là có quá nhiều rủi ro khi sử dụng quá nhiều câu có sự tương đồng về cấu trúc và nội dung để huấn luyện mô hình dịch máy.

Một lựa chọn khác là xem xét toàn bộ nội dung Wikipedia cho mỗi ngôn ngữ: đối với mỗi câu trong ngôn ngữ nguồn, chúng tôi khai thác tất cả các câu trong ngôn ngữ đích. Cách khai thác này có một vài lợi thế tiềm năng:

- Chúng ta có thể dóng hàng hai ngôn ngữ mà có ít bài viết chung thông qua việc sử dụng không gian nhúng câu chung.
- Một vài câu ngắn chỉ khác nhau bởi tên thực thể (named entities) thì có thể bị loại bỏ nhờ tiêu chuẩn biên.

Tuy nhiên, hạn chế của cách khai thác này là nguy dóng hàng sai có thể tăng lên và tỷ lệ thu hồi (recall) thấp.

Trong nghiên cứu này, chúng tôi đã lựa chọn cách khai thác toàn bộ ngữ liệu Wikipedia cho từng ngôn ngữ. Điều này cho phép chúng tôi mở rộng cách khai thác này

đối với những kho ngữ liệu khác, giàu tiềm năng hơn mà ở đó cách dóng hàng và khai thác theo cấp độ tài liệu là không dễ dàng... ví dụ Common Crawl.

Bên cạnh đó, cách tiếp cận thông qua khai thác toàn bộ kho ngữ liệu sẽ làm cho các ngôn ngữ thiểu số có cùng nguồn gốc với các ngôn ngữ lớn (ví dụ là các ngôn ngữ của đồng bào dân tộc ít người) được hưởng lợi trong việc xây dựng kho ngữ liệu cho riêng mình.

5.1 TIỀN XỬ LÝ DỮ LIỆU

Cấu hình máy sử dụng (colab pro): GPU Nvidia K80s, RAM 25GB.

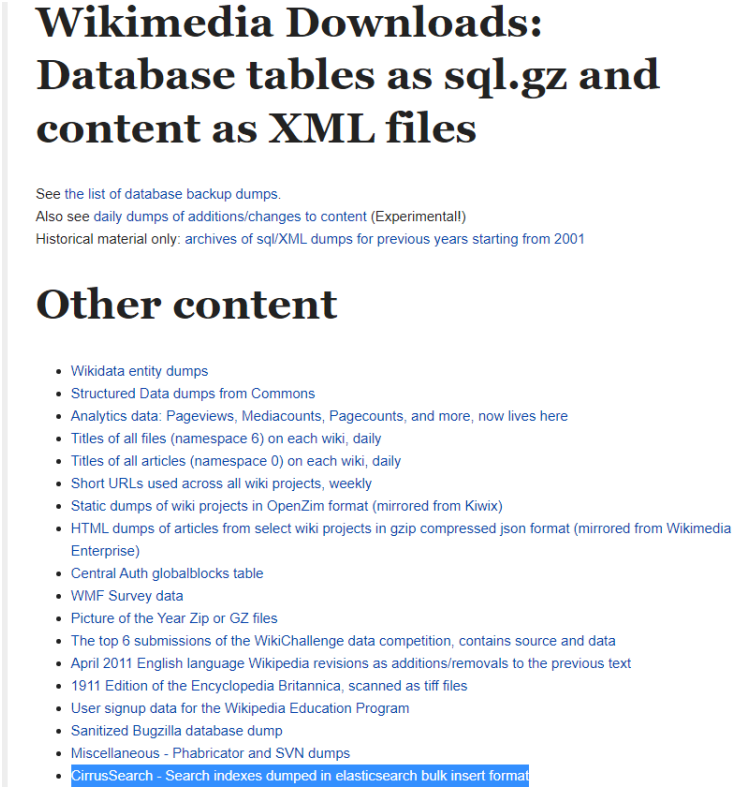
Quy trình tiền xử lý dữ liệu chúng tôi đã sử dụng như sau:

- Trích xuất nội dung
- Chia tách nội dung thành các câu.
- Loại bỏ các câu trùng.
- Loại bỏ các câu thuộc ngôn ngữ khác (thường là trích dẫn, hoặc liên kết tới ngôn ngữ khác)

Đầu tiên, chúng tôi cần trích xuất nội dung từ các bài báo Wikipedia của từng ngôn ngữ. Đây cũng không phải là một công việc dễ dàng. Ví dụ: loại bỏ tất cả các bảng, hình ảnh, dấu chân trang, meta token... Có nhiều cách tiếp cận để có thể trích xuất nội dung như sau:

- Sử dụng một số API có sẵn của Wikipedia để trích xuất.
- Sử dụng một số kho ngữ liệu được trích xuất sẵn từ Wikipedia.

Trong nghiên cứu này, chúng tôi đã sử dụng **CirrusSearch dumps[20]**, vì chúng được sao lưu định kì trực tiếp từ Wikipedia và chỉ giữ lại nội dung sau khi loại bỏ toàn bộ các thẻ meta token. Chúng tôi sử dụng **CirrusSearch dumps[20]** sao lưu từ tháng 9 năm 2021.



Wikimedia Downloads: Database tables as sql.gz and content as XML files

See the list of database backup dumps.
Also see daily dumps of additions/changes to content (Experimental!)
Historical material only: archives of sql/XML dumps for previous years starting from 2001

Other content

- Wikidata entity dumps
- Structured Data dumps from Commons
- Analytics data: Pageviews, Mediacounts, Pagecounts, and more, now lives here
- Titles of all files (namespace 6) on each wiki, daily
- Titles of all articles (namespace 0) on each wiki, daily
- Short URLs used across all wiki projects, weekly
- Static dumps of wiki projects in OpenZim format (mirrored from Kiwix)
- HTML dumps of articles from select wiki projects in gzip compressed json format (mirrored from Wikimedia Enterprise)
- Central Auth globalblocks table
- WMF Survey data
- Picture of the Year Zip or GZ files
- The top 6 submissions of the WikiChallenge data competition, contains source and data
- April 2011 English language Wikipedia revisions as additions/removals to the previous text
- 1911 Edition of the Encyclopedia Britannica, scanned as tiff files
- User signup data for the Wikipedia Education Program
- Sanitized Bugzilla database dump
- Miscellaneous - Phabricator and SVN dumps
- CirrusSearch - Search indexes dumped in elasticsearch bulk insert format

HÌNH 5.1: CIRRUSSEARCH DUMPS

Tách đoạn thành các câu cũng là một nhiệm vụ khó khăn, với nhiều ngoại lệ và có những luật đặc biệt cho riêng từng ngôn ngữ. Ví dụ: mỗi ngôn ngữ lại có một danh sách từ viết tắt riêng. Mỗi ngôn ngữ có những đặc điểm ngữ pháp riêng nên việc sử dụng chung một quy tắc tách câu là không thể. Trong khi đó, một số ngôn ngữ lại không có dấu hiệu phân tách câu một cách không rõ ràng, ví dụ như tiếng Thái Lan. Bên cạnh đó, vì chúng tôi cũng không tìm thấy một công vụ đủ mạnh mẽ và tin cậy để có thể thực hiện tác vụ phân tách câu cho tiếng Thái nên dù là một ngôn ngữ lớn với lượng người sử dụng lớn trong khu vực Đông Nam Á, chúng tôi cũng đành loại bỏ tiếng Thái ra khỏi nghiên cứu lần này.

Chúng tôi đã sử dụng Segtok (một công cụ miễn phí được viết bằng ngôn ngữ Python) cho tác vụ tách câu và tách từ cho Tiếng Anh. Tách câu sử dụng ngắt câu bằng dấu câu (regular expression) được sử dụng cho tiếng Trung, tiếng Indonesia và tiếng Malaysia. Riêng tiếng Việt, chúng tôi đã sử dụng công cụ UnderTheSea – một công cụ miễn phí mã nguồn mở, phục vụ cho các tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt. Tách

câu thành từ trong tiếng Indonesia và Malaysia sử dụng khoảng trắng, với tiếng Trung thì sử dụng công cụ Jieba.

Tiêu chuẩn biên cho khai thác các cặp câu song ngữ thì nhạy cảm với sự trùng lặp các câu. Nên chúng ta cần phải loại bỏ các câu trùng lặp, điều này giúp giảm khoảng 25% số lượng các câu.

Cách nhúng câu vào không gian vectơ câu chung thì hoàn toàn không phụ thuộc vào ngôn ngữ đầu vào. Điều này dẫn đến một tác dụng không mong muốn đó là các câu bằng ngôn ngữ khác (thường là trích dẫn hoặc câu danh ngôn...) thì thường gần hơn với câu gốc hơn là một bản dịch tiềm năng trong ngôn ngữ đích. Ví dụ: câu 1 trong ngôn ngữ nguồn L1 là bản dịch của câu 1 của ngôn ngữ đích L2 nhưng L2 lại chứa một câu danh ngôn tương đồng với câu danh ngôn bằng ngôn ngữ L2 trong L1. Nên câu 1 trong L2 lại được dóng hàng với câu 2 của L1 (minh họa ở bảng 1).

L1:Tiếng Việt Câu 1 Câu 2	 Idioms: “ easy come, easy goes”. Anh ấy nói:” cái gì dễ đến thì dễ đi”.
L2: Tiếng Anh	He said: “easy come, easy goes”.

BẢNG 1: MINH HỌA VIỆC MỘT CÂU SAI NGÔN NGỮ ẢNH HƯỞNG TỚI CHẤT LƯỢNG DÓNG HÀNG.

Bảng 1 chứa một câu tiếng Anh và hai câu tiếng Việt (trong đó một câu là bản dịch hoàn hảo của câu tiếng Anh và một câu chứa câu trích dẫn trong ngôn ngữ gốc là tiếng Anh). Dù câu 2 tiếng Việt là bản dịch hoàn hảo nhưng câu thuộc ngôn ngữ khác (câu 1 tiếng việt) đã làm ảnh hưởng tới chất lượng dóng hàng khi câu 1 được đánh giá cao hơn khi gióng hàng với câu trong tiếng Anh.

Mặc dù chênh lệch không lớn nhưng đã ảnh hưởng tới chất lượng dóng hàng. Để tránh điều này thì chúng tôi thực hiện nhận dạng ngôn ngữ (**LID**) và loại bỏ những câu không cùng ngôn ngữ.

LID được nhận dạng sử dụng **fasttext** (Joulin và cộng sự, 2016 [21]). **Fasttext** không hỗ trợ hết toàn bộ các ngôn ngữ trên Wikipedia nhưng nó vẫn có thể nhận dạng

được 5 ngôn ngữ sử dụng trong nghiên cứu này, bao gồm tiếng Việt, tiếng Anh, tiếng Trung, tiếng Indonesia và tiếng Malaysia. Bên cạnh đó, khi sử dụng fasttext để nhận dạng câu thuộc ngôn ngữ Malaysia thì có một số lượng không nhỏ (khoảng 1 triệu câu) được nhận dạng là thuộc ngôn ngữ Indonesia. Do hai ngôn ngữ này có cùng nguồn gốc, dẫn đến trong ngôn ngữ Malaysia có một số lượng lớn câu được nhận dạng là tiếng Indonesia. Để đảm bảo chất lượng dóng hàng câu thì chúng tôi đã loại bỏ hoàn toàn các câu được nhận dạng là tiếng Indonesia trong ngôn ngữ Malaysia.

Ngôn ngữ	Việt	Anh	Trung	Indonesia	Malaysia	Tổng
Kích thước	6907	150170	18335	5551	1185	182148

BẢNG 2: KÍCH THƯỚC KHO NGỮ LIỆU ĐƠN NGỮ SAU TIỀN XỬ LÝ (ĐƠN VỊ: NGÀN CÂU)

Kết quả thu được sau tất cả các bước tiền xử lý trên là 182 triệu câu cho 5 ngôn ngữ lớn ở khu vực Đông Nam Á, bao gồm 150 Triệu câu tiếng Anh, 18 triệu câu tiếng Trung, 7 triệu cậu tiếng Việt, 5,5 triệu câu tiếng Indonesia và 1,2 triệu câu tiếng Malaysia.

5.2 TỐI ƯU NGƯỠNG

Artetxe và Schwenk (2018a [14]) đã tối ưu hóa cách tiếp cận khai thác cho từng cặp ngôn ngữ trên một tập hợp các liên kết vàng được cung cấp từ bộ dữ liệu xuyên ngữ tham chiếu của XNLI (cross-lingual natural language inference).

Trong công việc này, chúng tôi sử dụng một giao thức đánh giá lấy cảm hứng từ WMT shared task on parallel corpus filtering for low resource conditions (Koehn và cộng sự,2019 [22]): huấn luyện một hệ thống dịch máy trên ngữ liệu được trích xuất trên những ngưỡng khác nhau, và đánh giá kết quả dựa trên điểm BLEU. Vì giới hạn tài nguyên nên chúng tôi đã thay thế việc huấn luyện một mô hình dịch máy bằng cách đưa chúng vào mô hình dịch máy thống kê tự động Moses[23] và cũng đánh giá chất lượng bản dịch từ mô hình dịch máy dựa trên điểm BLEU.

Ở đây, BLEU là viết tắt của Bilingual Evaluation Understudy, là một chỉ số đánh giá chất lượng hệ thống dịch được đề xuất bởi (Kishore và cộng sự,2002)[24]. Điểm

số BLEU có giá trị từ 0 (sai lệch tuyệt đối) đến 1 (khớp tuyệt đối). Điểm số BLEU càng cao thì hệ thống dịch càng đạt chất lượng tốt. Ý tưởng chính của phương pháp là so sánh kết quả bản dịch tự động bằng máy với một bản dịch chuẩn làm bản đối chiếu (cụ thể là câu song ngữ vừa khai thác được). Quá trình so sánh được thực hiện thông qua việc thống kê sự trùng khớp của các từ trong hai bản dịch có tính đến thứ tự của chúng trong câu (phương pháp n-grams theo từ).

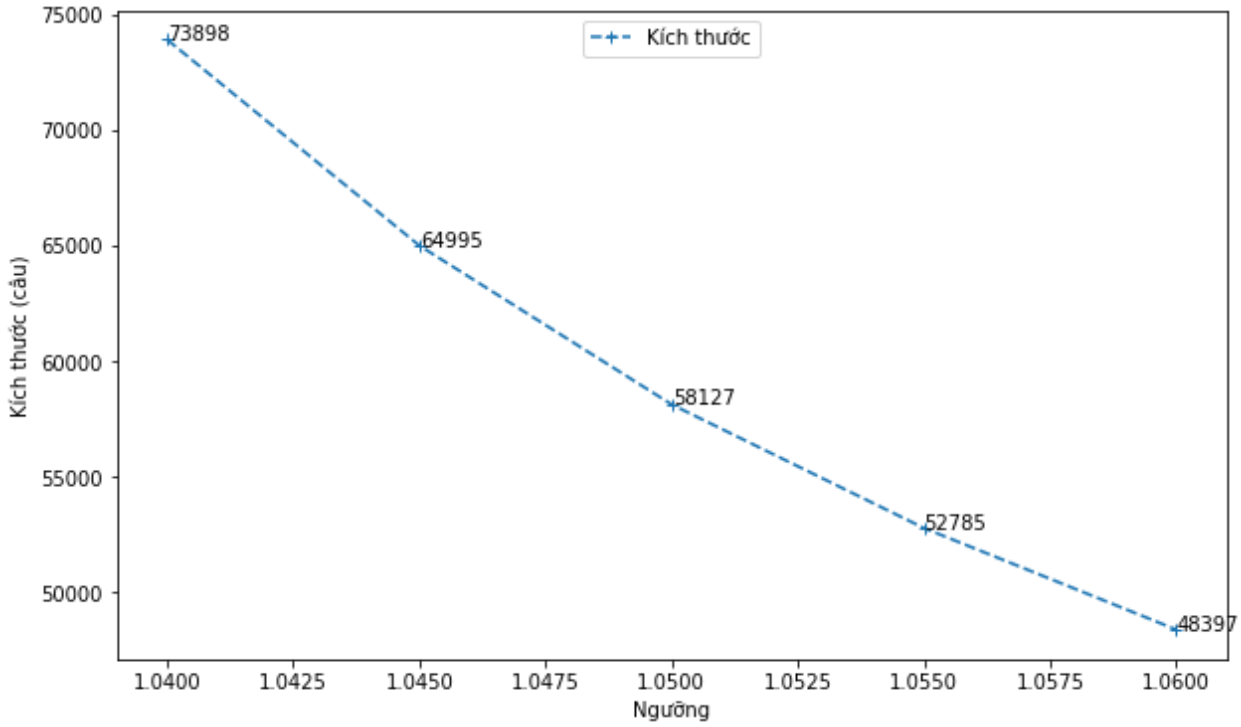
Chúng tôi đã thực hiện đánh giá ngưỡng (threshold) từ 1,04 đến 1,06 cho hai hướng từ tiếng Anh sang tiếng Việt và ngược lại để đánh giá về điểm BLEU và kích thước cho ngữ liệu khai thác được.

Vì thời gian hạn hẹp nên chúng tôi chỉ đánh giá trên tệp kích thước 73898 câu (một phần dữ liệu trích xuất được với ngưỡng 1,04). Tại sao chúng tôi lại quyết định ngưỡng 1,04 vì đây là ngưỡng được lấy chung cho toàn bộ các ngôn ngữ khu vực châu Á trong bài báo WikiMatrix mining (Holger Schwenk và cộng sự, 2019 [25]). Sau khi thực hiện đánh giá tối ưu ngưỡng, ta thu được kết quả của mô hình dịch được trình bày ở bảng 3 và bảng 4:

Ngưỡng	1,04	1,045	1,05	1,055	1,06
Kích thước	73898	64995	58127	52785	48379

BẢNG 3:KẾT QUẢ TỐI ƯU NGƯỠNG CHO KÍCH THƯỚC (ĐƠN VỊ :CÂU)

Từ bảng 3, ta có thể thấy được sự tương quan giữa ngưỡng trích xuất và kích thước kho ngữ liệu thu được. Với ngưỡng càng lớn (từ 1,04 đến 1,06) thì kích thước ngữ liệu song ngữ Anh - Việt khai thác được càng giảm và nhỏ dần (từ 74 ngàn câu giảm xuống còn 48 ngàn câu)(minh họa biểu đồ hình 5.2).



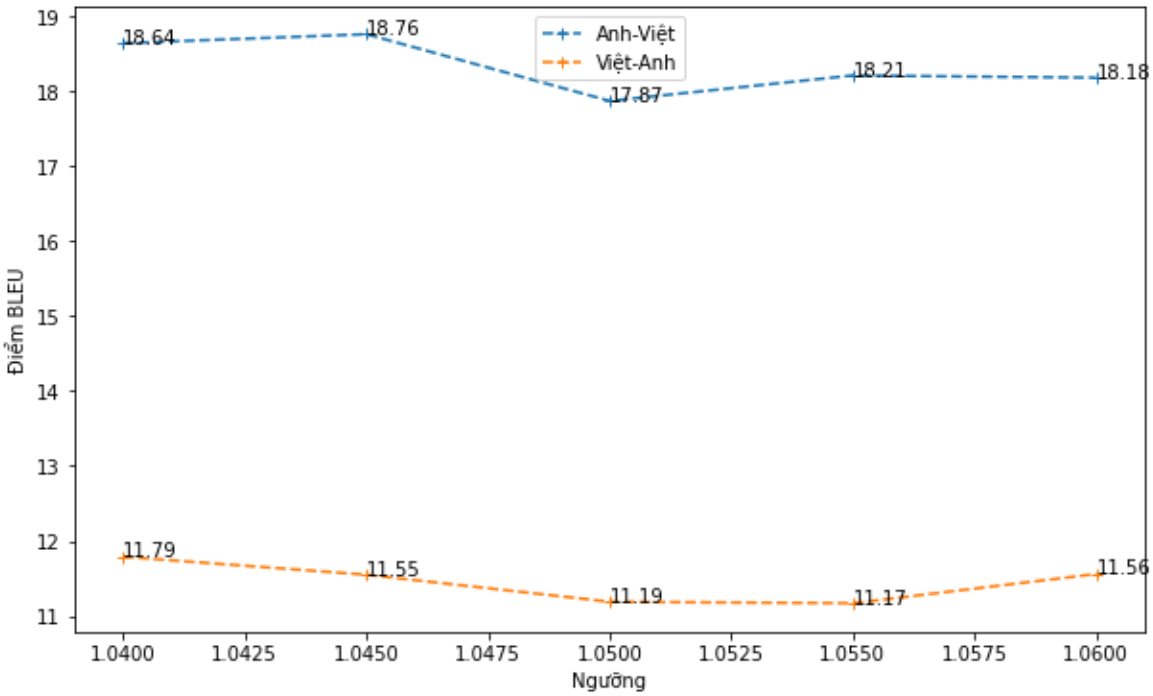
HÌNH 5.2: TỐI ƯU NGƯỠNG CHO KÍCH THƯỚC (ĐƠN VỊ: CÂU)

Dưới đây là bảng 4 thể hiện kết quả tối ưu ngưỡng cho chất lượng dịch máy (điểm BLEU)

Ngưỡng	1,04	1,045	1,05	1,055	1,06
Anh - Việt	18,64	18,76	17,87	18,21	18,18
Việt - Anh	11,79	11,55	11,19	11,17	11,56

BẢNG 4: KẾT QUẢ TỐI ƯU NGƯỠNG CHO ĐIỂM BLEU (ĐƠN VỊ: ĐIỂM)

Từ bảng trên, ta thấy được sự tương quan giữa ngưỡng trích xuất và chất lượng của mô hình dịch máy thống kê (thông qua điểm số BLEU) cho cả hai hướng dịch từ tiếng Anh sang tiếng Việt và ngược lại. Chất lượng mô hình dịch (đánh giá qua điểm BLEU) có sự thay đổi không đồng đều và có xu hướng giảm khi tăng ngưỡng. Chúng ta có thể quan sát rõ hơn sự thay đổi này thông qua biểu đồ dưới đây (hình 5.3).



HÌNH 5.3: BIỂU ĐỒ TỐI ƯU NGƯỠNG CHO ĐIỂM BLEU(ĐƠN VỊ: ĐIỂM)

Biểu đồ thể hiện mối liên hệ giữa ngưỡng và điểm số BLEU cho hai cặp ngữ liệu Anh→Việt và Việt→Anh. Đối với cặp Anh→Việt thì điểm BLEU có tăng nhỏ và sau đó giảm xuống tại ngưỡng 1,05 (18,64 → 18,76 → 17,87), sau đó tăng lại nhưng không đáng kể. Đối với cặp Việt - Anh thì điểm BLEU giảm liên tục từ ngưỡng từ 1,04 → 1,06 (11,79 → 11,55 → 11,19 → 11,17) và sau đó tăng nhẹ nhưng không đáng kể (11,17 → 11,56).

Kết luận, nhìn chung từ ngưỡng 1,04 → 1,06 để đạt kết quả tối ưu về kích thước và chất lượng điểm BLEU thì chúng tôi đã quyết định sử dụng ngưỡng 1,04 cho cặp song ngữ Anh - Việt. Bên cạnh đó, vì lý do hạn hẹp về tài nguyên thì chúng tôi đã sử dụng ngưỡng 1,04 cho tất cả các cặp còn lại. Đây cũng là ngưỡng chung cho tất cả các cặp ngôn ngữ thuộc khu vực Châu Á được (Holger Schwenk và cộng sự, 2019 [25]) đề xuất sử dụng.

CHƯƠNG 6. KẾT QUẢ

Chúng tôi chạy quá trình căn chỉnh cho năm tổ hợp ngôn ngữ lớn trong khu vực Đông Nam Á. Điều này mang lại cho chúng tôi 10 kho ngữ liệu song ngữ mà trong số đó đạt ít nhất 100 ngàn cặp câu cho mỗi cặp. Các kho ngữ liệu song ngữ khai thác được theo hướng từ ngôn ngữ nguồn L1 sang ngôn ngữ đích L2 giống hệt như kết quả khai thác từ L2 sang L1 và chỉ được tính một lần duy nhất.

Kết quả các kho ngữ liệu song ngữ được đánh giá qua hai phương diện. Đầu tiên, chúng tôi đánh giá kho ngữ liệu thông qua số lượng cặp câu song ngữ thu được (mục 6.1). Sau đó, chúng tôi đánh giá chất lượng của kho ngữ liệu bằng cách huấn luyện mô hình dịch máy thống kê tự động đối với từng cặp ngôn ngữ khai thác được và đánh giá chất lượng ngữ liệu song ngữ thông qua mô hình dịch máy thu được (mục 6.2).

6.1 ĐÁNH GIÁ ĐỊNH LƯỢNG

Quá trình khai thác kho ngữ liệu song ngữ cho năm ngôn ngữ lớn của khu vực Đông Nam Á đã thu được 10 cặp ngữ liệu song ngữ mà tất cả các cặp ngôn ngữ đều đạt được ít nhất 100 ngàn cặp câu (kết quả khai thác được dưới ngưỡng 1,04). Kết quả thu được được trình bày ở bảng 5:

Ngôn ngữ	Anh	Trung	Indonesia	Malaysia	Tổng
Việt	**1354**	231	240	113	1939
Anh		**850**	**1123**	**620**	**3948**
Trung			150	240	1472
Indonesia				**519**	2033
Malaysia					1493

BẢNG 5: KÍCH THƯỚC KHO NGỮ LIỆU KHAI THÁC ĐƯỢC THEO TỪNG CẶP CÂU KHAI THÁC ĐƯỢC (ĐƠN VỊ: NGÀN CẶP CÂU)

Từ bảng trên ta thấy được kích thước kho ngữ liệu khai thác được theo các cặp câu (đơn vị ngàn cặp câu) ở đây là khá lớn. Tuy nhiên thì ngữ liệu khai thác được chỉ được

ghi nhận theo một hướng. Vì chúng tôi sử dụng chiến lược "điểm số tối đa", kết hợp kết quả của hai hướng khai thác và chọn những cặp câu có điểm số cao nhất, nên kết quả cho cả hai hướng từ ngôn ngữ nguồn L1 sang ngôn ngữ đích L2 và ngược lại là như nhau.

Chúng ta cũng thấy được dù kết quả cả hai hướng là như nhau nhưng kết quả về kích thước kho ngữ liệu khai thác được vẫn cho kết quả tốt. Cụ thể, tổng cộng có 5,5 triệu cặp câu đã khai thác được. Trong số đó có gần 4 triệu cặp câu đã được dóng hàng với tiếng Anh (1,35 triệu cặp câu Anh - Việt; 850 ngàn cặp câu Anh - Trung; 1,1 triệu cặp câu Anh - Indonesia và 650 ngàn cặp câu Anh - Malaysia). Bên cạnh đó, cũng có một cặp cho kích thước kho ngữ liệu khả quan là cặp Indonesia - Malaysia (519 ngàn cặp câu). Các ngữ liệu được dóng hàng với các ngôn ngữ đều đạt kết quả khả quan với ít nhất là 1,5 triệu cặp câu đối với tiếng Malaysia và tiếng Trung. Đối với tiếng Việt và tiếng Indonesia thì số lượng cặp câu được gióng hàng lên tới gần 2 triệu cặp câu.

Có rất nhiều yếu tố ảnh hưởng đến số lượng câu khai thác được từ kho ngữ liệu Wikipedia. Rõ ràng rằng, nếu kích thước của kho ngữ liệu đơn ngữ (bảng 2) càng lớn thì khả năng cao là số lượng cặp câu tương đồng song ngữ mà khai thác được càng lớn. Điều dễ thấy là số cặp câu tương đồng song ngữ sẽ lớn hơn khi mà một trong hai ngôn ngữ là tiếng Anh (kho ngữ liệu đơn ngữ có 150 triệu câu).

Rõ ràng rằng những ngôn ngữ có nguồn gốc gần nhau như Indonesia vs Malaysia sẽ có số lượng cặp câu tương đồng được khai thác, trích xuất lớn (519 ngàn câu). Rõ ràng rằng kết quả sẽ đạt tốt hơn vì hai ngôn ngữ này có sự tương đồng lớn về mặt ngôn ngữ và số bài viết chung là bản dịch của nhau sẽ lớn. Trừ khi được bắt cặp với tiếng Anh thì kết quả các cặp câu khai thác được của Indonesia và Malaysia khi bắt cặp với tiếng Việt và tiếng trung đều cho kết quả chỉ từ 100 ngàn cặp câu tới 250 ngàn cặp câu (113 ngàn cặp câu Việt - Malaysia, 240 ngàn cặp câu Việt - Indonesia, 150 ngàn cặp câu Trung - Indonesia và 240 ngàn cặp câu Trung - Malaysia). Nhận định chủ quan rằng khả năng số bài viết chung giữa tiếng Trung và Tiếng Malaysia nhiều hơn số bài viết chung của tiếng Trung và tiếng Indonesia.

Bên cạnh đó, theo (Holger Schwenk và cộng sự, 2019 [25]) thì những ngôn ngữ có nhiều bài viết, kích thước kho dữ liệu được trích xuất từ Wikipedia rất lớn nhưng số lượng cặp câu khai thác được lại rất thấp, ví dụ : Cebuano (ceb) lại có số lượng cặp câu song ngữ được trích xuất rất thấp. Vì đa phần các bài viết của ngôn ngữ này trên Wikipedia được tạo ra (dịch lại) từ một con bot tự động nên dẫn đến độ tương đồng cũng như chất lượng các bài viết không được cao, khó tìm được số lượng cặp câu tương đồng lớn với các ngôn ngữ khác.

6.2 ĐÁNH GIÁ CHẤT LƯỢNG

Cấu hình máy dùng để đánh giá : CPU core i5-7800, RAM 16GB, ổ đia 500GB.

Để đánh giá chất lượng các cặp câu song ngữ trong điều kiện hạn hẹp về tài nguyên thì chúng tôi sử dụng 90% lượng dữ liệu của các cặp câu để huấn luyện mô hình dịch máy thống kê tự động. Mô hình dịch máy thống kê tự động được sử dụng ở đây là Moses[23]. Từ mô hình dịch máy thống kê nhận được, chúng tôi sử dụng 10% số lượng cặp câu song ngữ còn lại để đánh giá chất lượng của hệ thống dịch máy. Một câu trong cặp song ngữ dùng làm bản gốc đưa vào mô hình dịch máy từ ngôn ngữ nguồn sang ngôn ngữ đích và một câu tham chiếu trong ngôn ngữ đích. Chất lượng dịch máy được đánh giá thông qua điểm BLEU. Kết quả được trình bày trong bảng 6 dưới đây.

Ngôn ngữ đích / Ngôn ngữ nguồn	Việt	Anh	Trung	Indonesia	Malaysia
Việt		13,64	10,1	13,5	13,14
Anh	**20,77**		14,05	**21,04**	**19,02**
Trung	9,8	10,1		6,4	5.1
Indonesia	15,75	16,5	6,6		**17,38**
Malaysia	15,28	15,1	5,5	**16,77**	

BẢNG 6: KẾT QUẢ CHẤT LƯỢNG CÁC CẶP CÂU NGUỒN - ĐÍCH ĐÁNH GIÁ ĐƯỢC (ĐƠN VỊ ĐIỂM BLEU)

Như trong bảng 6 đã trình bày, kết quả của hệ thống dịch máy được tính riêng cho cả hai hướng mặc dù sử dụng chung một bộ ngữ liệu song ngữ. Một số cặp đạt chất lượng cao: Anh - Việt 20,77 điểm BLEU (tăng so với cặp Anh - Việt từ bộ dữ liệu WikiMatrix với 20,02 điểm BLEU đánh giá trên SMT Moses), Anh - Indonesia 21,04 điểm BLEU, Anh - Malay 19,02 điểm BLEU. Đối với cặp Việt - Anh thì kết quả tăng khoảng 0,5 điểm BLEU (từ 13 lên 13,5 điểm BLEU). Một cặp không được gióng hàng với tiếng Anh nhưng cũng đạt kết quả điểm số BLEU cao đó là cặp Indonesia và Malaysia với (16,77 điểm BLEU đối với hướng từ Malaysia tới Indonesia và 17,38 điểm BLEU đối với hướng dịch từ Indonesia tới Malaysia).

Chúng ta dễ dàng thấy được rằng đa phần các hệ thống dịch máy mà một trong hai ngôn ngữ là tiếng Anh thì có chất lượng khả quan. Một phần là nhờ kích thước kho ngữ liệu khai thác được (kết quả trình bày ở bảng 5) lớn dẫn đến số lượng thông tin hệ thống dịch máy có thể bắt được cũng lớn hơn, giúp hệ thống dịch máy đem lại kết quả tốt hơn.

Từ kết quả trên chúng ta cũng có thể rút ra được nhận xét, khi xây dựng hệ thống dịch máy với kho ngữ liệu càng lớn thì chất lượng dịch càng tăng. Thông qua kết quả của hệ thống dịch máy khi tối ưu ngưỡng và khi đánh giá chất lượng của kho ngữ liệu, ta có thể thấy chất lượng của kho ngữ liệu (số lượng và chất lượng của các cặp câu song ngữ) ảnh hưởng tới chất lượng của hệ thống dịch máy. Quan sát thực tế đánh giá (phần tối ưu ngưỡng), khi số lượng cặp câu song ngữ làm dữ liệu đầu vào cho mô hình dịch máy càng ít, thì kết quả dịch máy sẽ không đầy đủ và nhiều từ không được dịch, vì vậy chất lượng dịch sẽ giảm.

CHƯƠNG 7. KẾT LUẬN

Chúng tôi đã trình bày một phương pháp để khai thác kho ngữ liệu song ngữ từ Wikipedia một cách tự động cho một vài cặp ngôn ngữ lớn của khu vực Đông Nam Á. Chúng tôi đã sử dụng phương pháp được đề xuất dựa trên một không gian vectơ nhúng câu đa ngôn ngữ chung (Artetxe và Schwenk, 2018b[16]) và tiêu chuẩn biên (Artetxe và Schwenk, 2018a [14]). Cách tiếp cận sẽ tương tự đối với mọi cặp ngôn ngữ sau này mà không cần tối ưu hay tinh chỉnh riêng cho từng ngôn ngữ.

7.1 ĐÓNG GÓP CỦA KHÓA LUẬN

Cuối cùng, chúng tôi đạt được 5,5 triệu cặp câu song ngữ cho một số ngôn ngữ lớn trong khu vực Đông Nam Á. Trong số đó có gần 4 triệu cặp câu được dóng hàng với tiếng Anh và 1,9 triệu cặp câu được dóng hàng với tiếng Việt. Sau khi đánh giá chất lượng bản dịch thì đạt được một số kết quả khả quan: Anh - Việt 20,77 điểm BLEU, Anh - Indonesia 21,04 điểm BLEU và Anh - Malay 19,02 điểm BLEU, Indonesia - Malaysia 17,38 điểm BLEU và Malaysia - Indonesia 16,77 điểm BLEU. Bên cạnh đó, đa phần các cặp câu được gióng hàng với tiếng Anh dù là hướng từ ngôn ngữ nguồn sang ngôn ngữ đích là tiếng Anh thì đa phần đều đạt kết quả cao.

7.2 HƯỚNG NGHIÊN CỨU TƯƠNG LAI

Nghiên cứu này đã mở ra cơ hội khai thác ngữ liệu song ngữ cho một số ngôn ngữ thiểu số nhóm Việt - Mường, ngành Môn - Khmer như tiếng Mường, tiếng Khmer.

Bên cạnh đó do khuyết thiếu ngữ liệu trong khi huấn luyện bộ từ vựng BPE. Dẫn đến kết quả chưa được cải thiện nhiều. Trong tương lai, chúng tôi sẽ xây dựng một bộ từ vựng BPE riêng cho các ngôn ngữ được sử dụng khi khai thác kho ngữ liệu.

Trong tương lai, đối với các ngôn ngữ có nhu cầu cao trong xã hội thì chúng tôi cũng mong muốn sẽ xây dựng được kho ngữ liệu song ngữ cho các ngôn ngữ này.

Bên cạnh đó, chúng tôi mong rằng kho ngữ liệu xây dựng được có thể góp phần xây dựng nên những hệ thống nhúng câu và dịch máy tốt hơn.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Bình và Huỳnh Công Pháp, "Đánh Giá Vai Trò Của Kho Ngữ Liệu Đối Với Chất Lượng Dịch Tự Động Tiếng Việt". *ISSN 1859-1531-Tạp Chí Khoa Học Và Công Nghệ- ĐẠI HỌC ĐÀ NẴNG* , VOL. 19,NO. 1, **2021**.

[2] Bahdanau D., Cho K. and Bengio Y, (2014), "Neural Machine Translation by Jointly Learning to Align and Translate". In: *ArXiv* 1409 (Sept. 2014).

[3] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1412–1421

[4] Mikolov T., Sutskever I., Chen K., et al., "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems* 26 (Oct. 2013).

[5] Pennington J., Socher R., and Manning C., "Glove: Global Vectors for Word Representation". In: *EMNLP* 14 (Jan. 2014), pp. 1532–1543.

[6] Đinh Điền (2006), *giáo trình xử lý ngôn ngữ tự nhiên*, nhà xuất bản Đại Học Quốc Gia TP Hồ Chí Minh, tr.11-13.

[7] Koehn P. and Knowles R., (2017), Six Challenges for Neural Machine Translation. In *WNMT*, pages 28–39.

[8] Khayrallah H. and Koehn P., (2018), On the Impact of Various Types of Noise on Neural Machine Translation. In *WNMT*, pages 74–83.

[9] Schwenk H., (2018), Filtering and Mining Parallel Data in a Joint Multilingual Space. In *ACL*, pages228–234.

[10] Guo M., Shen Q., Yang Y., et al., (2018). Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. In *WMT*, pages 165–176.

[11] Bonet E. C., Varga A., Genabith V. J., (2017), An ˜ Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, pages 1340–1348.

[12] Hassan H., Aue A., Chen C., et al., (2018), Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv*:1803.05567.

[13] Conneau A., Lample G., Ranzato A. M., et al., (2018). ´Word Translation Without Parallel Data. In *ICLR*.

[14] Artetxe M. and Schwenk H., (2018a), Marginbased Parallel Corpus Mining with Multilingual Sentence Embeddings. https://arxiv.org/abs/1811.01136.

[15] Yang Y., Gustavo Hernandez ´ Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *https://arxiv.org/abs/1902.08564*.

[16] Artetxe M. and Schwenk H. (2018b). Massively multilingual sentence embeddings for zeroshot cross-lingual transfer and beyond. In *https://arxiv.org/abs/1812.10464*

[17] Johnson J., Douze M., and J ´ egou H., (2017), ´Billion-scale similarity search with GPUs. *ArXiv preprint arXiv*:1702.08734.

[18] Jegou H., Douze M. and Schmid C., (2011), Product quantization for nearest neighbor search. *IEEE Trans. PAMI*, 33(1):117–128.

[19] https://github.com/facebookresearch/faiss/wiki/The-index-factory; 2021.

[20] https://dumps.wikimedia.org/other/cirrussearch; 2021.

[21] Joulin A., Grave E., Bojanowski P., et al. (2016). Bag of tricks for efficient text classification. *https://arxiv.org/abs/1607.01759.*

[22] Chaudhary V., Schwenk H., Koehn P., et al., (2019), Lowresource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (WMT).*

[23] https://www.statmt.org/moses; 2021.

[24] Kishore, P; Salim, R; Todd, W and Wei-Jing Zhu. 'BLEU: a Method for Automatic Evaluation of Machine Translation'.In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL*), Philadelphia, July 2002, pp. 311-318.

[25] Schwenk, H.; Chaudhary, V.; Sun S.; et al. (2019). WikiMatrix: Bitext extraction of 135 million Wikipedia sentences in 1,620 language pairs. In *https://arxiv.org/abs/1907.05791*.