BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

----------⁂----------

MÔN HỌC HỌC MÁY (MACHINE LEARNING) TRONG PHÂN TÍCH KINH DOANH

Giảng viên hướng dẫn: Tiến sĩ TRẦN DUY THANH

Mã lớp học phần: 251BIM401407

CHỦ ĐỀ: PHÂN TÍCH VÀ MINH HỌA MACHINE TRANSLATION CỦA GOOGLE

SINH VIÊN THỰC HIỆN: LÊ PHƯỚC THỊNH

MÃ SỐ SINH VIÊN: K234161856

Thành phố Hồ Chí Minh, tháng 9 năm 2025

MỤC LỤC

I. BỐI CẢNH LỊCH SỬ HÌNH THÀNH CỦA MACHINE TRANSLATION.

a. Bối cảnh ra đời và khát vọng ban đầu

Sự đa dạng ngôn ngữ vừa là biểu hiện của nền văn hóa phong phú, vừa là rào cản to lớn trong giao tiếp và trao đổi tri thức giữa các cộng đồng trên thế giới. Khát vọng phá bỏ rào cản này đã manh nha từ nhiều thế kỷ, với các giả thuyết về "ngôn ngữ phổ quát" hay các từ điển song ngữ thủ công. Tuy nhiên, phải đến sự xuất hiện của máy tính điện tử vào giữa thế kỷ XX, ý tưởng về một cỗ máy có thể tự động dịch ngôn ngữ mới thực sự hình thành.

Bối cảnh lịch sử sau Chiến tranh Thế giới thứ Hai và trong thời kỳ Chiến tranh Lạnh đã tạo ra một nhu cầu cấp thiết: cần nhanh chóng giải mã và hiểu được một khối lượng khổng lồ các tài liệu khoa học, kỹ thuật và tình báo từ các ngôn ngữ khác nhau, đặc biệt là từ tiếng Nga. Việc dịch thủ công không thể đáp ứng kịp thời độ phức tạp và khối lượng thông tin này. Chính trong hoàn cảnh đó, Machine Translation (MT) - Dịch Máy - đã ra đời như một lời giải cho bài toán hóc búa, đánh dấu sự khởi đầu của một lĩnh vực nghiên cứu đầy tham vọng.

b. Kỷ nguyên Rule-Based Machine Translation (RBMT) và những thập niên đầu tiên.

Giai đoạn khai sinh của MT được định hình bởi mô hình Dịch Máy dựa trên Quy tắc (Rule-Based Machine Translation - RBMT). Người tiên phong cho ý tưởng này là Warren Weaver, một nhà khoa học người Mỹ, người vào năm 1949 đã gửi một bản ghi nhớ mang tính bước ngoặt có tựa đề "Dịch thuật", trong đó ông đề xuất việc áp dụng các khái niệm từ lý thuyết mã hóa và thống kê vào dịch ngôn ngữ. Ý tưởng của Weaver đã truyền cảm hứng cho các nghiên cứu thực tiễn đầu tiên.

Dự án GEORGETOWN-IBM (1954): Thường được coi là sự kiện đánh dấu sự ra đời chính thức của MT. Trong một buổi trình diễn lịch sử, hệ thống này đã dịch thành công hơn 60 câu tiếng Nga sang tiếng Anh. Mặc dù rất đơn giản, với một từ vựng chỉ khoảng 250 từ và 6 quy tắc ngữ pháp, nó đã chứng minh tính khả thi của việc dịch tự động bằng máy tính.

Nguyên lý hoạt động của RBMT: Các hệ thống RBMT hoạt động dựa trên một loạt các quy tắc ngôn ngữ học phức tạp được các chuyên gia con người xây dựng thủ công. Chúng bao gồm:

- *Từ điển song ngữ:* Chứa các cặp từ tương đương giữa ngôn ngữ nguồn và ngôn ngữ đích.
- *Quy tắc ngữ pháp:* Mô tả cấu trúc cú pháp của cả hai ngôn ngữ.
- *Quy tắc chuyển đổi:* Quy định cách biến đổi cấu trúc câu từ ngôn ngữ nguồn sang ngôn ngữ đích.
- *Quá trình dịch thường trải qua các bước:* phân tích hình thái (xác định dạng từ), phân tích cú pháp (xác định chức năng ngữ pháp), chuyển đổi cấu trúc, sinh từ vựng và sinh câu đích.

Các hệ thống RBMT như SYSTRAN (ra đời những năm 1970 và được sử dụng rộng rãi trong nhiều thập kỷ) có ưu điểm là cho kết quả dịch ổn định, có cấu trúc ngữ pháp rõ ràng vì tuân theo quy tắc. Tuy nhiên, chúng bộc lộ những hạn chế nghiêm trọng: chi phí xây dựng và bảo trì bộ quy tắc cực kỳ cao, kém linh hoạt trước sự đa dạng và biến đổi của ngôn ngữ tự nhiên, và đặc biệt không thể xử lý tốt các hiện tượng như thành ngữ, nghĩa bóng hay ngữ cảnh. Sự thất vọng với tiến độ của RBMT đã được phản ánh trong Báo cáo ALPAC nổi tiếng năm 1966 tại Mỹ, kết luận rằng dịch máy kém hiệu quả và không kinh tế so với dịch giả con người, dẫn đến việc cắt giảm mạnh tài trợ cho nghiên cứu MT trong một thời gian dài.

c. Sự bùng nổ của Statistical Machine Translation (SMT) và sự tham gia của các "gã khổng lồ" công nghệ.

Sự trỗi dậy của Internet và sự sẵn có của các kho ngữ liệu song ngữ khổng lồ (corpus) vào cuối thế kỷ 20 đã mở đường cho một cuộc cách mạng trong MT: sự ra đời của mô hình Dịch Máy Thống kê (Statistical Machine Translation - SMT). Thay vì dựa vào các quy tắc ngôn ngữ học do con người định nghĩa, SMT xem việc dịch thuật như một bài toán thống kê. Ý tưởng cốt lõi, được đề xuất bởi các nhà nghiên cứu tại IBM trong dự án Candide vào những năm 1990, rất đơn giản: tìm câu dịch trong ngôn ngữ đích có xác suất cao nhất khi cho trước một câu trong ngôn ngữ nguồn.

Nguyên lý hoạt động của SMT: Mô hình này dựa vào việc "học" từ dữ liệu. Nó phân tích hàng triệu cặp câu đã được dịch sẵn (ngữ liệu song song) để rút ra các mô hình thống kê về:

- Mô hình dịch (Translation Model): Xác suất một từ/cụm từ trong ngôn ngữ nguồn tương ứng với một từ/cụm từ trong ngôn ngữ đích.
- Mô hình ngôn ngữ (Language Model): Xác suất xuất hiện của một chuỗi từ trong ngôn ngữ đích, đảm bảo câu dịch ra trôi chảy, tự nhiên.

Vai trò của Google Translate: Sự thành công vượt bậc của SMT gắn liền với sự ra mắt của Google Translate vào năm 2006. Bằng cách tận dụng sức mạnh tính toán khổng lồ và kho ngữ liệu song ngữ khổng lồ có được từ việc crawl (thu thập dữ liệu) web, Google Translate đã đưa dịch máy đến với hàng trăm triệu người dùng toàn cầu. Dịch vụ này cung cấp khả năng dịch nhanh chóng, miễn phí giữa hàng chục ngôn ngữ, một điều không tưởng trong thời kỳ RBMT. Tuy nhiên, SMT vẫn tồn tại nhược điểm: lỗi dịch có thể rất lớn khi gặp cấu trúc câu phức tạp hoặc ngữ liệu huấn luyện ít, và việc dịch thường bị cục bộ, thiếu sự hiểu biết về ngữ cảnh tổng thể của văn bản.

d. Kỷ nguyên Neural Machine Translation (NMT) - Bước nhảy vọt về chất lượng.

Sự phát triển của mạng nơ-ron nhân tạo (Artificial Neural Networks) và Học sâu (Deep Learning) trong thập kỷ 2010 đã dẫn đến sự thay đổi mô hình một lần nữa, đưa MT bước vào kỷ nguyên của Dịch Máy Nơ-ron (Neural Machine Translation - NMT). Thay vì dịch từng phần của câu một cách độc lập như SMT, NMT sử dụng một mạng nơ-ron lớn để dịch toàn bộ câu đầu vào thành câu đầu ra trong một khung duy nhất, end-to-end (đầu cuối).

Sự khác biệt cốt lõi của Mô hình NMT (thường dựa trên kiến trúc Sequence-to-Sequence với cơ chế Attention) "mã hóa" toàn bộ ý nghĩa của câu nguồn thành một vector biểu diễn số (một dạng "ý tưởng"), sau đó "giải mã" vector này để sinh ra câu đích. Cơ chế Attention cho phép mô hình tập trung vào các phần khác nhau của câu nguồn khi sinh ra từng phần của câu đích, giúp xử lý tốt hơn các câu dài và phụ thuộc xa.

Kể từ khi được các tập đoàn lớn như Google (với Google Neural Machine Translation - GNMT, ra mắt 2016) và Microsoft giới thiệu, NMT đã cho thấy một

bước nhảy vọt rõ rệt về chất lượng dịch. Câu dịch trở nên trôi chảy, tự nhiên hơn, ít lỗi ngữ pháp và quan trọng nhất là có khả năng nắm bắt ngữ cảnh tốt hơn nhiều so với các mô hình trước đây. Sự ra đời của các mô hình kiến trúc Transformer vào năm 2017 càng củng cố vị thế của NMT, trở thành kiến trúc tiêu chuẩn cho hầu hết các hệ thống dịch máy hiện đại, bao gồm cả các dịch vụ của Amazon.

Sự phát triển của Machine Translation từ RBMT đến SMT và hiện tại là NMT phản ánh một hành trình dài từ việc mô phỏng quy tắc ngôn ngữ của con người đến việc để cho máy móc tự học các mô hình từ dữ liệu. Cuộc cách mạng NMT đã tạo ra một bối cảnh mới, nơi chất lượng dịch thuật tiệm cận gần hơn với con người, mở ra cánh cửa cho sự tích hợp sâu rộng của dịch máy vào mọi mặt của đời sống và công nghệ, từ các ứng dụng di động đến các nền tảng thương mại điện tử toàn cầu, và đặt nền móng cho sự xuất hiện của các giải pháp dịch máy chuyên biệt, hiệu năng cao như Amazon Translate.

II. MACHINE TRANSLATION LÀ GÌ?

a. Định nghĩa, Mục tiêu và Phạm vi ứng dụng

Về bản chất, Dịch máy (Machine Translation - MT) là một nhánh nghiên cứu của Trí tuệ nhân tạo (AI) và Xử lý ngôn ngữ tự nhiên (NLP), liên quan đến việc sử dụng phần mềm máy tính để tự động dịch văn bản hoặc lời nói từ một ngôn ngữ tự nhiên này (ngôn ngữ nguồn) sang một ngôn ngữ tự nhiên khác (ngôn ngữ đích) mà không có sự can thiệp trực tiếp của con người trong quá trình dịch. Mục tiêu lý tưởng của MT là đạt được chất lượng dịch thuật ngang bằng hoặc gần bằng với chất lượng của một biên dịch viên chuyên nghiệp — tức là bản dịch không chỉ chính xác về mặt ngữ nghĩa mà còn phải trôi chảy, tự nhiên, phù hợp với ngữ cảnh và văn hóa của ngôn ngữ đích.

Tuy nhiên, trên thực tế, MT thường được đánh giá dựa trên sự cân bằng giữa ba yếu tố chính: Chất lượng, Tốc độ và Chi phí. Trong khi con người có ưu thế về chất lượng, MT vượt trội về tốc độ và khả năng mở rộng. Một hệ thống MT có thể xử lý hàng triệu từ trong vài phút, một điều không tưởng đối với đội ngũ dịch giả. Nhờ đó, phạm vi ứng dụng của MT trong thế giới hiện đại là vô cùng rộng lớn. Nó không còn là một công cụ học thuật mà đã trở thành một yếu tố then chốt trong toàn cầu hóa, thúc đẩy giao tiếp xuyên biên giới trong các lĩnh vực như: dịch tin tức, tài liệu kỹ thuật, hướng dẫn sử dụng sản phẩm; hỗ trợ dịch phụ đề video; tích hợp vào các nền tảng thương mại điện tử để dịch mô tả sản phẩm; cung cấp bản dịch tức thời trong trò chuyện và hội nghị trực tuyến; và là công cụ hỗ trợ đắc lực cho các dịch giả chuyên nghiệp (trong quy trình gọi là Dịch máy có sự hỗ trợ của con người - MTPE).

b. Các thành phần cốt lõi và quy trình hoạt động cơ bản

Về mặt kỹ thuật, một hệ thống MT hiện đại, đặc biệt là các hệ thống dựa trên Học sâu (Deep Learning), là một kiến trúc phức tạp. Tuy nhiên, quy trình hoạt động của nó có thể được mô tả một cách khái quát thông qua ba giai đoạn chính: Phân tích đầu vào, Dịch thuật thực sự, và Tạo đầu ra.

Phân tích đầu vào (Input Analysis): Ở giai đoạn này, văn bản nguồn được hệ thống "đọc" và xử lý sơ bộ. Công việc bao gồm:

- Tokenization: Chia câu thành các đơn vị nhỏ hơn như từ, cụm từ hoặc các phần của từ (subwords). Ví dụ, câu "I love machine translation" có thể được tách thành các token: ["I", "love", "machine", "translation"].
- Chuẩn hóa (Normalization): Chuyển đổi văn bản về dạng chuẩn, như viết thường tất cả các chữ cái, loại bỏ các ký tự đặc biệt không cần thiết.
- Phân tích ngôn ngữ học cơ bản: Một số hệ thống có thể thực hiện nhận diện từ loại (danh từ, động từ, tính từ) hoặc phân tích cú pháp sơ bộ để hiểu cấu trúc câu.

Dịch thuật (Translation - Core Process): Đây là bước quan trọng nhất, nơi diễn ra sự chuyển đổi ngôn ngữ. Dựa trên mô hình đã được huấn luyện (ví dụ: mô hình Nơ-ron), hệ thống sẽ ánh xạ chuỗi token của ngôn ngữ nguồn sang một biểu diễn trung gian (thường là một vector số học đa chiều, còn gọi là "không gian đặc trưng" - feature space) nắm bắt ý nghĩa của câu. Sau đó, từ biểu diễn trung gian này, hệ thống sẽ "tạo sinh" (generate) ra chuỗi token tương đương trong ngôn ngữ đích. Trong các mô hình NMT tiên tiến, cơ chế "Attention" cho phép mô hình tập trung vào các phần khác nhau của câu nguồn khi sinh ra từng phần của câu đích, giúp xử lý tốt sự khác biệt về trật tự từ và ngữ pháp giữa các ngôn ngữ.

Tạo đầu ra (Output Generation): Các token trong ngôn ngữ đích sau khi được sinh ra sẽ được kết hợp lại để hình thành câu hoàn chỉnh. Hệ thống cũng thực hiện các điều chỉnh nhỏ để đảm bảo tính tự nhiên, chẳng hạn như chính tả, dấu câu, và dạng từ (ví dụ: chia thì cho động từ).

c. Phân loại các phương pháp Machine Translation chính

Trải qua lịch sử phát triển, MT đã chứng kiến sự thống trị của ba phương pháp chính, mỗi phương pháp có triết lý và kỹ thuật cốt lõi khác biệt. Việc hiểu rõ sự khác biệt này là rất quan trọng để đánh giá ưu nhược điểm của các hệ thống.

Dịch máy dựa trên Quy tắc (Rule-Based Machine Translation - RBMT): Dựa hoàn toàn vào các quy tắc ngôn ngữ học (ngữ pháp, cú pháp, từ vựng) do các chuyên gia con người xây dựng thủ công. Hệ thống sẽ phân tích cú pháp câu nguồn, áp dụng các quy tắc chuyển đổi, rồi tổng hợp câu đích.

- Ưu điểm: Dịch ổn định, có cấu trúc ngữ pháp rõ ràng; không cần dữ liệu huấn luyện lớn.
- Nhược điểm: Chi phí xây dựng và bảo trì rất cao; kém linh hoạt, không xử lý được thành ngữ, nghĩa bóng; khó mở rộng sang ngôn ngữ mới.

Dịch máy dựa trên Thống kê (Statistical Machine Translation - SMT): Bỏ qua các quy tắc ngôn ngữ học, thay vào đó sử dụng các mô hình thống kê học được từ một kho ngữ liệu song song khổng lồ (ví dụ: các văn bản đã được dịch song ngữ). Nó tính toán xác suất một cụm từ/câu trong ngôn ngữ nguồn sẽ được dịch thành một cụm từ/câu trong ngôn ngữ đích.

- Ưu điểm: Linh hoạt hơn RBMT, chất lượng tốt hơn khi có nhiều dữ liệu huấn luyện; giảm sự phụ thuộc vào kiến thức ngôn ngữ học thủ công.
- Nhược điểm: Chất lượng phụ thuộc hoàn toàn vào chất lượng và số lượng dữ liệu huấn luyện; bản dịch có thể thiếu tính tổng thể và mắc lỗi về độ trôi chảy.

Dịch máy Nơ-ron (Neural Machine Translation - NMT):Sử dụng mạng nơ-ron nhân tạo (đặc biệt là kiến trúc Transformer) để dịch toàn bộ câu đầu vào thành câu đầu ra trong một mô hình end-to-end (đầu cuối). Thay vì dịch từng phần, NMT "mã hóa" ý nghĩa của toàn bộ câu nguồn thành một biểu diễn số phong phú, sau đó "giải mã" biểu diễn đó để sinh ra câu đích một cách trôi chảy.

- Ưu điểm: Cho chất lượng dịch vượt trội so với SMT và RBMT; câu dịch trôi chảy, tự nhiên hơn nhờ nắm bắt được ngữ cảnh tổng thể; xử lý tốt hơn các cấu trúc phức tạp.
- Nhược điểm: Yêu cầu sức mạnh tính toán cực lớn và khối lượng dữ liệu huấn luyện khổng lồ; hoạt động như một "hộp đen", khó giải thích tại sao mô hình đưa ra một bản dịch cụ thể.

d. Các chỉ số đánh giá chất lượng Machine Translation

Để đo lường hiệu quả của các hệ thống MT, cộng đồng nghiên cứu sử dụng cả đánh giá tự động và đánh giá của con người.

Đánh giá tự động (Automatic Evaluation):

- *BLEU (Bilingual Evaluation Understudy):* Là chỉ số phổ biến nhất. BLEU so sánh bản dịch của máy với một hoặc nhiều bản dịch tham chiếu chất lượng cao do con người thực hiện, dựa trên sự trùng khớp của các cụm từ (n-gram). Điểm số càng cao (trên thang điểm 0 đến 1 hoặc 0-100%) cho thấy bản dịch càng gần với bản dịch của con người.
- *Các chỉ số khác:* TER (Translation Edit Rate) đo số lần sửa đổi cần thiết để biến bản dịch máy thành bản dịch tham chiếu; METEOR tập trung vào độ chính xác và độ hồi tưởng (recall).

Đánh giá của con người (Human Evaluation): Đây là tiêu chuẩn vàng vì nó đánh giá được các khía cạnh mà chỉ số tự động không nắm bắt được, như tính tự nhiên, sự phù hợp về văn phong và ngữ cảnh. Người đánh giá thường được yêu cầu chấm điểm theo thang điểm về Độ chính xác (Adequacy) - ý nghĩa có được bảo toàn không, và Độ trôi chảy (Fluency) - câu dịch có tự nhiên không.

Tóm lại, Machine Translation là một lĩnh vực năng động, nơi các kỹ thuật AI tiên tiến được ứng dụng để giải quyết bài toán chuyển đổi ngôn ngữ tự nhiên. Sự tiến hóa từ RBMT, SMT đến NMT đã đưa chất lượng dịch máy lên một tầm cao mới, tạo tiền đề cho các dịch vụ thương mại mạnh mẽ như Amazon Translate, nơi tận dụng sức mạnh của NMT để cung cấp các giải pháp dịch thuật có độ chính xác và khả năng mở rộng cao.

III. NGUYÊN LÝ MACHINE TRANSLATION CỦA GOOGLE

Google Translate (Dịch Google) đại diện cho sự tiến hóa của công nghệ dịch máy, từ các phương pháp dựa trên thống kê thuần túy sang các mô hình trí tuệ nhân tạo có khả năng hiểu ngữ cảnh một cách sâu sắc. Nguyên lý cốt lõi hiện nay của nó là Dịch máy Nơ-ron (Neural Machine Translation - NMT), một kiến trúc hoạt động dựa trên việc mô phỏng khả năng xử lý ngôn ngữ của não bộ con người thông qua các mạng nơ-ron nhân tạo. Để hiểu rõ nguyên lý này, cần phải phân tích cơ chế vận hành cụ thể của hệ thống.

a. Nguyên lý nền tảng: Từ Dịch máy Thống kê (SMT) đến Dịch máy Nơ-ron (NMT)

Sự phát triển công nghệ của Google Translate tuân theo một nguyên lý chung: chuyển từ việc dịch "cục bộ" sang dịch "toàn cục".

- Giai đoạn Dịch máy Thống kê (Statistical Machine Translation - SMT - Trước 2016): Nguyên lý hoạt động dựa trên việc phân tích xác suất. Hệ thống được "huấn luyện" trên một kho ngữ liệu song ngữ khổng lồ (ví dụ: các văn bản tiếng Anh và bản dịch tiếng Việt tương ứng). Cơ chế của SMT bao gồm:
    - Dịch từng cụm từ: Hệ thống chia nhỏ câu thành các cụm từ và tìm kiếm các bản dịch có xác suất cao nhất cho từng cụm đó dựa trên dữ liệu huấn luyện.
    - Ghép nối các cụm từ: Sau đó, một mô hình ngôn ngữ (được xây dựng từ các văn bản thuần túy bằng ngôn ngữ đích) sẽ giúp sắp xếp các cụm từ đã dịch thành một câu hoàn chỉnh và trôi chảy nhất có thể.
    - Hạn chế: Nguyên lý này khiến bản dịch thiếu tính mạch lạc tổng thể, vì nó không "hiểu" được ngữ cảnh của toàn bộ câu. Các lỗi thường gặp là dịch sai ngữ pháp, đặc biệt với các cấu trúc phức tạp hoặc giữa các ngôn ngữ có trật tự từ khác biệt.
- Giai đoạn Dịch máy Nơ-ron (Neural Machine Translation - NMT - Từ 2016 đến nay): Đây là một sự thay đổi mang tính nguyên lý cơ bản. Thay vì dịch từng phần, hệ thống NMT xem xét và dịch toàn bộ câu đầu vào cùng một lúc. Nguyên lý này mô phỏng cách con người dịch thuật: đọc hiểu ý nghĩa cả câu,

rồi diễn đạt lại ý nghĩa đó bằng ngôn ngữ đích một cách tự nhiên. Sự chuyển đổi này dẫn đến các bản dịch trôi chảy, chính xác về ngữ pháp và ngữ cảnh hơn hẳn.

b. Cơ chế hoạt động chi tiết của Mô hình Nơ-ron: Kiến trúc Encoder-Decoder với Attention

Cơ chế để hiện thực hóa nguyên lý "dịch toàn cục" của Google NMT là một kiến trúc mạng nơ-ron có tên Encoder-Decoder (Bộ Mã hóa - Giải mã) kết hợp với Cơ chế Tập trung (Attention Mechanism). Có thể hình dung cơ chế này qua một quy trình ba bước:

Bộ Mã hóa (Encoder): "Hiểu" câu nguồn

- Nhiệm vụ: Bộ mã hóa đảm nhận việc "đọc" và "thấu hiểu" câu văn đầu vào. Nó chuyển đổi câu thành một dạng biểu diễn số học mà máy tính có thể xử lý.
- Cách thức: Mỗi từ trong câu được chuyển thành một vector số (một dãy số biểu diễn ý nghĩa của từ). Các vector này sau đó được xử lý bởi một mạng nơ-ron (ban đầu là LSTM, nay chủ yếu là Transformer). Mạng này phân tích mối quan hệ giữa tất cả các từ với nhau, từ đó tạo ra một "vector ngữ cảnh" (context vector). Vector này là một dãy số phức tạp chứa đựng thông tin ngữ nghĩa và cấu trúc của toàn bộ câu gốc.

Cơ chế Tập trung (Attention Mechanism): "Điều phối viên" thông minh

- Vấn đề: Nếu chỉ sử dụng một vector ngữ cảnh duy nhất cho cả câu, thông tin có thể bị loãng, đặc biệt với các câu dài. Làm thế nào để khi dịch một từ cụ thể trong câu đích, hệ thống biết nên "chú ý" vào phần nào của câu nguồn?
- Giải pháp - Cơ chế Attention: Đây là trái tim của hệ thống NMT hiện đại. Với mỗi từ sắp được sinh ra trong bản dịch, cơ chế Attention tính toán một "trọng số chú ý" (attention weight) cho từng từ trong câu nguồn. Trọng số này cho biết mức độ quan trọng của từ nguồn tại thời điểm đó.
- Ví dụ: Khi dịch câu "I love cats" sang "Tôi yêu mèo", tại thời điểm dịch từ "mèo", cơ chế Attention sẽ gán trọng số rất cao cho từ "cats" trong câu gốc, đồng thời gán trọng số thấp cho các từ "I" và "love". Cơ chế này giúp hệ thống giải quyết hiệu quả vấn đề về sự khác biệt trong trật tự từ giữa các ngôn ngữ.

Bộ Giải mã (Decoder): "Sinh ra" bản dịch hoàn chỉnh

- Nhiệm vụ: Bộ giải mã hoạt động như một "người nói" thạo ngôn ngữ đích. Nó sử dụng thông tin từ vector ngữ cảnh và sự "hướng dẫn" của cơ chế Attention để lần lượt sinh ra từng từ trong bản dịch cuối cùng.
- Cách thức: Tại mỗi bước, bộ giải mã xem xét:
    - Các từ nó đã dịch trước đó.
    - Thông tin từ vector ngữ cảnh tổng thể.
    - Quan trọng nhất: Tín hiệu từ cơ chế Attention, cho biết nên tập trung vào đâu trong câu nguồn ở bước hiện tại.
- Dựa trên các thông tin này, nó dự đoán từ có xác suất cao nhất sẽ là từ tiếp theo. Quá trình này lặp lại cho đến khi sinh ra ký hiệu "kết thúc câu".

c. Sự tiến hóa về Cơ chế: Vai trò của Kiến trúc Transformer

Năm 2017, chính các nhà nghiên cứu tại Google đã công bố kiến trúc Transformer, một cải tiến vượt bậc so với các mạng nơ-ron trước đây. Transformer trở thành cơ chế xử lý nền tảng mới cho Google Translate nhờ các đặc điểm:

- Cơ chế Tự động Tập trung (Self-Attention): Cho phép mô hình đánh giá mối quan hệ giữa tất cả các từ trong câu một cách đồng thời, thay vì tuần tự như các mạng RNN/LSTM cũ. Điều này giúp mô hình hiểu ngữ cảnh tốt hơn và tốc độ xử lý nhanh hơn rất nhiều.
- Xử lý Song song: Kiến trúc Transformer cho phép xử lý toàn bộ câu cùng lúc, giúp tăng tốc độ đáng kể quá trình huấn luyện và dịch.
- Hiệu quả với Dữ liệu lớn: Transformer có khả năng tận dụng hiệu quả các bộ dữ liệu huấn luyện khổng lồ, dẫn đến các mô hình ngày càng mạnh mẽ và chính xác.

PHỤ LỤC

Hutchins, J. (2007). An introduction to machine translation. Encyclopedia of Language and Linguistics.

Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation.

Papineni, K., et al. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).

Lommel, A. R., & DePalma, D. A. (2016). Europe’s Leading Role in Machine Translation: A Report by the Association for Machine Translation in the Americas. (Nguồn về ứng dụng thực tế của MT trong công nghiệp).

Wu, Y., et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144.
Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS). (Đây là bài báo gốc giới thiệu kiến trúc Transformer).
Google AI Blog. (2016). A Neural Network for Machine Translation, at Production Scale. Truy cập từ: https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html
Brown, T. B., et al. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165. (Giới thiệu về mô hình GPT-3, minh họa cho sức mạnh của kiến trúc Transformer trong xử lý ngôn ngữ).
Koehn, P. (2009). Statistical Machine Translation. Cambridge University Press. (Tài liệu tham khảo về nguyên lý SMT để so sánh).