****

**BỘ THÔNG TIN VÀ TRUYỀN THÔNG**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO KẾT QUẢ ĐỀ TÀI:**

**“Nghiên cứu Phát triển**

**Hệ thống dịch câu tiếng Việt sang câu tiếng Anh”**

**Nhóm 7:**

Vũ Viết Công – B18DCCN063

Trần Ngọc Đắc – B18DCCN149

Nguyễn Văn Tuấn – B18DCCN570

# MỞ ĐẦU

Dịch thuật là chiếc cầu nối giúp cho những người không cùng ngôn ngữ có thể hiểu được điều mà người khác muốn diễn đạt. Nói cách khác, dịch thuật là đem thông tin được diễn đạt bằng ngôn ngữ này (gọi là ngôn ngữ nguồn) diễn đạt lại bằng ngôn ngữ khác (gọi là ngôn ngữ đích). Tuy nhiên, chuyển tải ý nghĩa giữa các ngôn ngữ chỉ là bề nổi của tảng băng trôi. Trên thực tế, dịch thuật còn đi xa hơn những gì mà chúng ta thấy trên bề mặt nổi.

Dịch thuật từ xưa đến nay luôn đóng vai trò quan trọng và thiết yếu trong lịch sử dân tộc, bất kể phương đông hay phương tây. Không đơn thuần chỉ là chuyển tải từ ngôn ngữ gốc sang một hay nhiều ngôn ngữ khác, mà song song với việc chuyển dịch ngôn ngữ, dịch thuật còn chuyển tải cả một nền văn hóa.

Trong nền kinh tế hội nhập, thì dịch thuật càng đóng vai trò quan trọng. Hầu hết các lĩnh vực như xây dựng, y tế, tài chính, du lịch, … đều sử dụng những tài liệu chuyên ngành cả tiếng Việt cũng như ngoại ngữ. Với xu thế hội nhập kinh tế quốc tế thì dịch vụ dịch thuật đối với việc kinh doanh và trao đổi văn hóa là vô cùng thiết yếu.

Để giải quyết những vấn đề gặp phải khi giao tiếp với bạn bè quốc tế, cũng như các hồ sơ hay hợp đồng của đối tác thì cần phải có dịch vụ dịch thuật đi kèm. Dịch thuật là yếu tố quan trọng giúp hai bên hiểu rõ nhau hơn vì vậy các doanh nghiệp đều phải có sự hợp tác với một dịch vụ dịch thuật để hỗ trợ họ khi cần thiết. Nó vừa có tính thống nhất vừa nâng tầm thương hiệu của doanh nghiệp trong mắt đối tác. Không chỉ mang yêu tố kinh tế việc dịch thuật cũng là một điều không thể thiếu để kết nối cũng như thấu hiểu giữa các quốc gia trên thế giới. Việc dịch thuật như một cầu nối để hiểu thêm về những quốc gia khác từ nên văn hoá, chính trị và lịch sử. Từ đó xây dựng được liên kết, giao dịch, trao đổi giữa những lãnh thổ trên khắp thế giới.

Có thể thấy rõ rằng việc dịch thuật đóng một vai trò vô cùng to lớn với mỗi quốc gia từ việc kết nối văn hoá, xã hội, chính trị, kinh tế. Với vai trò quan trọng trong việc thúc đẩy hội nhập kinh tế, dịch thuật chính là cầu nối quan trọng để các nền kinh tế có thể rút ngắn khoảng cách tuy nhiên việc dịch thuật được thực hiện bởi chủ thể chính là con người đã có những ưu thế riêng của nó nhưng ngược lại với xu thế toàn cầu hoá thêm vào đó là sự phát triển không ngừng của khoa học công nghệ thì con người dường như chưa thể đáp ứng các yêu cầu có thể đến như: thời gian, đa dạng lĩnh vực ngành nghề.

Ngày nay các hệ thống dịch tự động ra đời và đã thực sự phát triển với sự thông minh và chính xác đến mức độ có thể nói là tương đối gần với khả năng của con người. Như các bạn cũng biết, trí tuệ nhân tạo (AI) đang ngày càng phát triển, thời gian gần đây nhiều ứng dụng sử dụng công nghệ AI đã được đưa vào cuộc sống thực tiễn.

Trong đề tài này chúng tôi xin đề xuất nghiên cứu phát triển hệ thống dịch câu tiếng Việt sang câu tiếng Anh.

Nội dung nghiên cứu của đề tài bao gồm những nội dung chính sau đây:

**Nội dung 1:** Nghiên cứu khảo sát tổng quan các hệ thống dịch thuật ở Việt Nam và trên thế giới.

**Nội dung 2:** Xây dựng và thiết kế hệ thống dịch câu tiếng Việt sang câu tiếng Anh

- Khảo sát các bộ dữ liệu sẵn có.

- Xây dựng thuật toán.

**Nội dung 3:** Nghiên cứu, thiết kế phần mềm quản lý, lưu trữ, hiển thị.

**Nội dung 4:** Nghiên cứu, xây dựng hệ thống.

**Nội dung 5:** Thử nghiệm đánh giá chất lượng hệ thống.

**Nội dung 6:**Viết báo cáo và xây dựng tài liệu hướng dẫn.

**MỤC LỤC**

[MỞ ĐẦU 2](#_Toc118957276)

[Chương 1 Nghiên cứu khảo sát tổng quan các hệ thống dịch thuật ở Việt Nam và trên thế giới 5](#_Toc118957277)

[1.1 Ngoài nước 5](#_Toc118957278)

[1.2 Trong nước 6](#_Toc118957279)

[Chương 2 Xây dựng và thiết kế hệ thống dịch câu tiếng Việt sang câu tiếng Anh 10](#_Toc118957280)

[2.1 Tổng quan hệ thống 10](#_Toc118957281)

[2.1.1 Hệ thống dịch máy thống kê (Statistical Machine Translation) 10](#_Toc118957282)

[**Quá trình Lựa chọn** 19](#_Toc118957283)

[**Quá trình Lai tạo** 19](#_Toc118957284)

[**Quá trình Thay thế** 20](#_Toc118957285)

[Chương 3 Kịch bản vận hành hệ thống 22](#_Toc118957286)

[Chương 4 Nghiên cứu, xây dựng hệ thống 25](#_Toc118957287)

[Chương 5 Thử nghiệm đánh giá chất lượng hệ thống 28](#_Toc118957288)

[5.1 Tiêu chuẩn đánh giá 28](#_Toc118957289)

[5.1.1 Đánh giá độ chính xác của hệ thống 28](#_Toc118957290)

[5.1.2 Đánh giá sự ổn định của hệ thống 29](#_Toc118957291)

[5.2 Kết quả chạy thử nghiệm trên thực tế 29](#_Toc118957292)

[5.2.1 Đánh giá độ chính xác của hệ thống 29](#_Toc118957293)

[5.2.2 Đánh giá độ ổn định của hệ thống 29](#_Toc118957294)

[Kết luận và Định hướng nghiên cứu trong tương lai 31](#_Toc118957295)

[TÀI LIỆU THAM KHẢO 31](#_Toc118957296)

# Nghiên cứu khảo sát tổng quan các hệ thống dịch thuật ở Việt Nam và trên thế giới

## Ngoài nước

Hiện nay việc giao thương giữa các nền văn hoá trở nên khá phổ biến khi mạng xã hội phát triển, công nghệ khoa học tiên tiến đã dẫn đến việc đưa con người, văn hoá, danh lam thắng cảnh của Việt Nam đến bạn bè quốc tế. Cũng chính bởi lý do này việc tiếp thu và cảm nhận nền văn hoá mang đậm Á châu này là một nhu cầu thiết yếu. Các nước trên thế giới cũng như chính phủ của nước Việt Nam luôn thúc đẩy du lịch bằng các chính sách, hiệp định, các gói hỗ trợ để thu hút nhiều du khách đến.

Cụ thể có thể kể đến như: Nhật Bản, Hàn Quốc, Trung Quốc và các nước Châu Âu đã và đang thực hiện nhiều chính sách khác nhau nhằm tăng lượng khách đến với Việt Nam. Khi người nước ngoài đến với Việt Nam càng nhiều thì nhu cầu dịch thuật càng cao. Như vậy các hệ thống dịch tự động ra đời nhằm giải quyết vấn đề về ngôn ngữ. Thêm vào đó tiếng Anh được coi là ngôn ngữ chung của thế giới khi có đến 1 tỷ người, hơn 50 quốc gia nói tiếng Anh, biến tiếng Anh là ngôn ngữ chính phổ biến thứ 3 trên thế giới, sau tiếng Trung và tiếng Tây Ban Nha. Các hệ thống dịch tự động được nhiều người dùng có thể kể đến như: Google dịch, Từ điển Vdict, TFLAT, Microsoft Translator, Laban Dictionary, …

**Phân tích một vài hệ thống dịch tự động được sử dụng phổ biến hiện nay**

1. Google Translate

Chắc chắn những người đã từng sử dụng những công cụ để dịch thuật không thể không biết đến cái tên Google Translate này.

Công cụ Google Translate hỗ trợ dịch thuật nhanh với hàng chục loại ngôn ngữ khác nhau trên khắp thế giới. Chất lượng bản dịch khá tốt và làm hài lòng phần lớn người sử dụng hiện nay.

Đặc biệt, bạn có thể sử dụng ngay chỉ với một chiếc smartphone và hoàn toàn miễn phí.

2. Linguee

Linguee là một trong những công cụ được đánh giá cao về chức năng dịch thuật. Có thể nói, công cụ này là một quyển từ điển đầy giá trị cho những người đam mê học hỏi.

Mặc dù nó mạnh về mặt dịch ngôn ngữ như Google Translator, tuy nhiên, điểm mạnh của công cụ này là dịch những từ, cụm từ khó theo từng ngữ cảnh khác nhau một cách chính xác. Đó cũng là lý do mà người dùng thường sử dụng kết hợp giữa hai công cụ: Linguee và Google Translator.

3. Bing Translator

Bing Translator cũng là một công cụ về dịch thuật rất mạnh và tiện ích hiện nay. Đây là một sản phẩm ấn tượng của hãng Microsoft và có thể dịch hàng chục ngôn ngữ trên thế giới.

Đặc biệt, Bing Translator có thể dịch trang web và tài liệu tải lên một cách chính xác đến ngạc nhiên.

**Đánh giá tiện ích công cụ dịch thuật với dịch thuật viên**

Với những tiện ích mà những công cụ giới thiệu ở trên đem lại cho người dùng, chúng ta thấy được những lợi ích to lớn của các ứng dụng này.

Tuy nhiên, khi làm phép so sánh các công cụ dịch thuật và những dịch thuật viên sẽ là một sự khập khiễng.

Những công cụ dịch thuật có thể hỗ trợ dịch tức thời và miễn phí. Nội dung bản dịch chính xác và giúp người dùng hiểu được một cách nhanh chóng.

Tuy nhiên, với những bản dịch có độ khó cao về nhiều lĩnh vực chuyên ngành, các công cụ này sẽ lộ ra điểm yếu. Nó không thể dịch một cách chính xác tất cả các từ chuyên ngành và trong mọi ngữ cảnh khác nhau được.

Chính vì vậy, những công cụ này không thể thay thế được công việc của những dịch thuật viên chuyên nghiệp.

## Trong nước

Trong quá trình phát triển của đất nước có sự giao lưu văn hóa, kinh tế, chính trị và xã hội với rất nhiều quốc gia khác trên thế giới. Tuy nhiên, rào cản ngôn ngữ giữa các quốc gia ảnh hưởng to lớn tới các hoạt động như: Thương mại, du lịch, hợp tác đầu tư, nghiên cứu văn hóa các nước, nghiên cứu vận hành kinh tế của các nền kinh tế khác trên thế giới…

Ngành dịch thuật đóng vai trò là cầu nối, là phương thức truyền tải thông tin chính xác nhất giữa ngôn ngữ của các quốc gia. Chính lĩnh vực dịch thuật giúp cho các hoạt động trên diễn ra nhanh chóng và thuận lợi.

Ngoài ra, ngành dịch thuật trong nước cũng biên dịch hàng loạt tài liệu, sách báo, hồ sơ tiếng nước ngoài sang tiếng Việt, giúp cho các cá nhân, tổ chức, doanh nghiệp và quốc gia nghiên cứu và áp dụng vào thực tiễn. Hầu hết lĩnh vực đều cần sự có mặt của biên dịch viên và phiên dịch viên.

Các hoạt động offline, giao thương, hội họp giữa người Việt Nam và người nước ngoài cũng có sự tham gia của các phiên dịch viên nhằm mục đích tăng cường truyền tải thông tin nhanh chóng, chính xác, giúp cho hai bên hiểu rõ nội dung vấn đề để đạt được sự thống nhất nào đó.

Có thể nói rằng, cùng với sự phát triển của hệ thống máy tính, dịch tự động từ tiếng Việt sang tiếng Anh bắt đầu nghiên cứu ở Việt Nam từ cuối những năm 80 thế kỉ XX. Trong những năm gần đây, việc sử dụng được ngoại ngữ trong giao lưu quốc tế đóng vai trò ngày càng quan trọng. Nhiều loại từ điển điện tử, các sách song ngữ đã và đang được hình thành nhằm phục vụ cho việc dạy và học ngoại ngữ. Một số nhóm gồm các lập trình viên đã bắt đầu nghiên cứu tổ chức hệ thống dịch tự động từ tiếng Việt sang tiếng Anh. Việc xây dựng hệ thông dịch tự động từ tiếng Việt sang tiếng Anh là rất cần thiết để đáp ứng nhu cầu to lớn của sự phát triển kinh tế và xã hội mang tính chất toàn cầu.

Dịch thuật có nhiều vai trò khác nhau. Nó có thể giúp các sản phẩm, công nghệ, thông tin, văn hóa giữa các quốc gia được phổ biến rộng rãi trên toàn cầu.

Theo báo cáo của McKinsey, việc dịch thuạt đã giúp nhiều nước tăng trưởng mạnh mẽ:

Luồng dữ liệu xuyên biên giới đang tăng với tốc độ gấp 50 lần so với thập kỷ trước.

Khoảng 1 tỷ người dùng mạng xã hội có ít nhất một kết nối nước ngoài

Hơn 350 triệu người mua sắm xuyên biên giới, thương mại điện tử – mở rộng cơ hội cho các doanh nghiệp vừa và nhỏ trở thành “công ty đa quốc gia siêu nhỏ”.

Bất cứ thứ gì được tạo ra để được phổ biến đến các nơi khác trên thế giới với nhiều mục đích khác nhau. Để đạt được mục tiêu, chúng đều phải thực hiện bằng cách dịch thuật.

Mỗi khi nhắc đến dịch thuật thì người ta lại đi kèm với việc toàn cầu hóa. Đây là hai khái niệm có mối quan hệ đôi bên cùng có lợi. Nhiều công ty cho rằng, việc vươn ra toàn cầu mang lại cho họ cơ hội mở rộng phạm vi tiếp cận đến những khách hàng mới. Tuy nhiên, để một doanh nghiệp thành công vươn tầm quốc tế, thì dịch thuật đóng vai trò rất quan trọng.

Để minh họa sự cần thiết của các dịch vụ dịch thuật ngôn ngữ chuyên nghiệp, hãy cùng theo dõi Noel, Giám đốc điều hành của Everet & Elbert Sản Xuất, khi ông chuẩn bị đưa sản phẩm hàng đầu của mình (Bàn Ghế Clover) sang châu Âu – cụ thể là Pháp, Đức, Tây Ban Nha và Vương quốc Anh.

Thoạt nhìn, đưa sản phẩm vào các quốc gia này một cách liên tiếp có vẻ quá sức – và đúng như vậy. Tuy nhiên, Noel vẫn quyết tâm phát triển và tích cực ra quốc tế. Để có được bước đầu thuận lợi, đội ngũ nội bộ của anh ấy bắt đầu một công việc khó khăn đó là dịch trang web, sản phẩm và tài liệu tiếp thị từ tiếng Anh sang tiếng Pháp, Đức và Tây Ban Nha.

Nhưng những khách hàng ở nước Anh thì sao? Mặc dù chúng tôi sử dụng cùng một ngôn ngữ cơ bản, nhưng có sự khác biệt giữa tiếng Anh-Mỹ và tiếng Anh-Anh cho nên chúng tôi không nên bỏ qua vấn đề này.

Nói về các vấn đề bị bỏ qua trước đây, Noel đã thất bại trong kế hoạch dịch các tài liệu pháp lý như hợp đồng và nhãn hiệu, bằng cấp sáng chế và bản quyền, cấu trúc công ty và tài liệu quy trình, cũng như đăng ký tên miền dành riêng cho quốc gia.

Mặc dù việc dịch các tài liệu và trang web là cần thiết nhưng việc toàn cầu hóa, quốc tế hóa và bản địa hóa cũng không kém phần quan trọng. Ba hoạt động tất yếu này nên được thực hiện trước quá trình dịch ngôn ngữ.

Hãy cùng xem xét toàn cầu hóa, quốc tế hóa, địa phương hóa ngôn ngữ và dịch thuật (Globalization, International, Localization and Translation) – chúng là gì và tại sao chúng phải diễn ra theo một trình tự cụ thể.

Sử dụng công cụ dịch thuật đem lại nhiều tiện ích người dùng hiện nay. Bạn có thể hiểu được nội dung tài liệu để học tập, nghiên cứu hay giải trí một cách nhanh chóng và hoàn toàn miễn phí. Vậy, chất lượng dịch thuật của những công cụ này như thế nào? Những công cụ nào đang được người dùng đánh giá cao?

Hệ thống dịch câu tiếng việt sang câu tiếng anh hỗ trợ chuyển đổi ngôn ngữ nhanh chóng, đơn giản, tiện lợi và đặc biệt là miễn phí. Hỗ trợ đặc lực cho học tập, nghiên cứu, giao tiếp hay đi du lịch, …

* Hoàn toàn miễn phí

Một ưu điểm lớn khi sử dụng các công cụ này là phần lớn chức năng đều hoàn toàn miễn phí. Như vậy, bạn có thể đọc hiểu những tài liệu nước ngoài một cách nhanh chóng mà không phải trả bất kỳ khoản tiền nào.

Điều này đã hỗ trợ rất tốt cho người dùng trong mọi ngành nghề, đặc biệt là học sinh, sinh viên.

* Thời gian dịch nhanh

Khi đưa nội dung vào công cụ, bản dịch sẽ xuất hiện ngay tức thời sau vài giây. Điều này giúp rất nhiều cho người sử dụng khi cần đọc hiểu nhanh tài liệu, hướng dẫn. Đặc biệt, đối tượng nghiên cứu, học tập hay đi du lịch rất cần đến các công cụ này.

* Nội dung dịch chính xác

Mặc dù thời gian dịch nhanh và hoàn toàn miễn phí, tuy nhiên nội dung dịch vẫn đảm bảo chất lượng và độ chính xác. Đây cũng là lý do mà ngày càng có nhiều người tham gia và sử dụng các ứng dụng này.

Bên cạnh những lợi ích tuyệt vời mà hệ thống dịch thuật đem lại cũng không thể phủ nhận một vài rủi ro khi dịch sai hay dịch thiếu.

# Xây dựng và thiết kế hệ thống dịch câu tiếng Việt sang câu tiếng Anh

## 2.1 Tổng quan hệ thống

## 2.1.1 Hệ thống dịch máy thống kê (Statistical Machine Translation)

Với yêu cầu dịch thuật từ tiếng việt sang tiếng anh thì cần đòi hỏi 2 kiến thức chính trong quá trình dịch thuật của dịch thuật viên

* Lựa chọn từ cho phù hợp với nghĩa của từ ngôn ngữ gốc
* Đúng ngữ pháp (Cấu trúc câu)

Từ hai kiến thức chính trong quá trình dịch thuật của dịch thuật viên, người ta đã nghiên cứu và xây dựng ra nhiều hệ thống dịch thuật sử dụng các công nghệ khác nhau. Sau một thời gian tìm hiểu và tham khảo, nhóm em đã quyết định sử dụng Statistical Machine Translation (SMT) để giải quyết bài toán phiên dịch câu tiếng việt sang tiếng anh.

“Dịch máy, đôi khi được gọi bằng chữ viết tắt MT, là một lĩnh vực phụ của Ngôn ngữ học tính toán điều tra việc sử dụng phần mềm máy tính để dịch văn bản hoặc lời nói từ ngôn ngữ tự nhiên này sang ngôn ngữ tự nhiên khác.” 1 Có một số phần mềm và dịch vụ Dịch máy gần như tốt chẳng hạn như Google Translate và Microsoft Translator có thể dịch giữa các ngôn ngữ khác nhau. Hiện tại các hệ thống Dịch máy đang tạo ra bản dịch có thể chấp nhận được nhưng còn quá xa để tạo ra bản dịch tốt. Trong hầu hết các trường hợp, nó có thể chỉ được sử dụng cho bản dịch nháp và cần thêm ấn bản cho con người. Chúng ta có thể nói rằng Dịch máy, với tư cách là một lĩnh vực nghiên cứu đang hoạt động, còn lâu mới có chất lượng dịch lý tưởng. Mục tiêu của báo cáo này là đề xuất một thuật toán Di truyền để xây dựng Bộ giải mã như một phần chính của Dịch máy thống kê. Bộ giải mã trong bản dịch máy thống kê (SMT) như một thuật toán tìm kiếm cần tìm kiếm trong số nhiều bản dịch có thể có cho một câu nhất định. Không gian tìm kiếm này lớn vì các ý nghĩa có thể có cho mỗi từ trong ngôn ngữ nguồn và cũng vì rất nhiều hoán vị của các từ đích trong câu đích. Để tìm một giải pháp tối ưu đối với vấn đề này, chúng em cần tìm kiếm trong không gian tìm kiếm khổng lồ này và không thể kiểm tra tất cả các khả năng. Để giải quyết vấn đề Decoder (tìm kiếm), các nhà nghiên cứu đang tìm kiếm các thuật toán để tìm ra giải pháp tối ưu thay vì giải pháp chính xác: Giải thuật di truyền, một thuật toán tiến hóa, có thể giúp chúng ta giải quyết vấn đề. Có các thuật toán khác nhau cho Bộ giải mã trong SMT như Beam Search hoặc Greedy Decoder. Cả hai thuật toán này đều đưa ra quyết định dựa trên một phần câu trái ngược với Thuật toán di truyền mà chúng ta có quyền truy cập vào câu hoàn chỉnh. Có thể đưa ra quyết định về bản dịch dựa trên câu hoàn chỉnh. Do đó, ví dụ, phân tích ngữ pháp hoặc ngữ nghĩa chính thức có thể được tích hợp vào Thuật toán di truyền.

Mục tiêu cụ thể: Xây dựng hệ thống dịch câu tiếng việt sang câu tiếng anh với chức năng chính:

* Xây dựng hệt thống đưa ra một câu tiếng anh khi người dùng nhập vào một câu tiếng việt

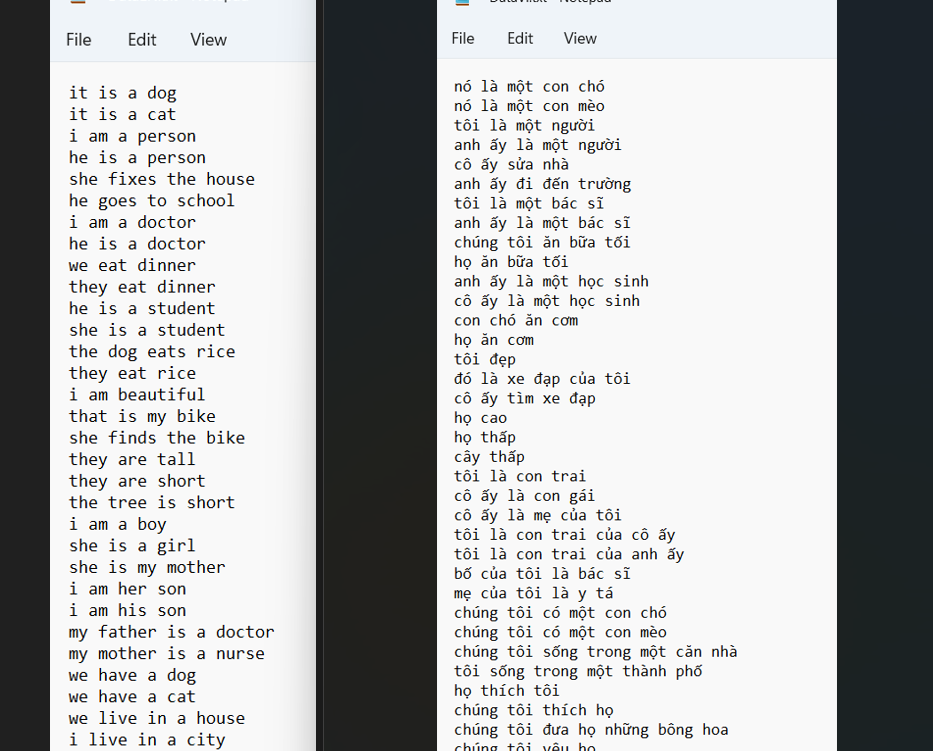
Kiến trúc tổng quan của hệ thống dịch câu tiếng anh sang tiếng việt được mô tả trong hình 1.

Diagram

Description automatically generated

**Hình 1 Hệ thống dịch câu tiếng việt sang tiếng anh**

* + 1. ***Khảo sát các bộ dữ liệu có trong hệ thống dịch câu tiếng Việt sang tiếng Anh***



**Hình 2.** **Hệ thống dịch câu tiếng việt sang tiếng anh**

Hình 2 minh hoạ các câu trong tập dữ liệu sử dụng cho hệ thống. Tập dữ liệu đầu vào của hệ thống là các câu đơn với định dạng text, mỗi câu được trình bày trên một dòng. Mỗi câu sẽ bao gồm chủ ngữ + động từ (+ tân ngữ).

**2.2 Thuật toán di truyền (Genetic Algorithm)**

Decoder là thành phần chính trong hệ thống SMT. Bản chất của Decoder là một thuật toán tìm kiếm. Vai trò của nó là tìm kiếm bản dịch tốt nhất trong tất cả các bản dịch có thể có. Tuy nhiên, việc sinh ra tất cả các bản dịch và tìm kiếm trong đó là bất khả thi, vì vậy bài toán tìm kiếm chỉ có thể đưa ra lời giải cực trị địa phương. Ở đây, ta sẽ sử dụng giải thuật di truyền làm thuật toán cho Decoder (GADecoder).

**2.2.1 Các nghiên cứu liên quan**

Về bản chất của hệ thống dịch câu là đi tìm kiếm bản ghi tốt nhất trong một tập hợp gồm rất nhiều bản ghi, đã có rất nhiều phương pháp để giải quyết việc này như Giải thuật **Beam Search Decoder, Greedy Decoder, …**

Về cơ bản, Beam Search là một thuật toán tìm kiếm ưu tiên tốt nhất được tối ưu hóa với số lượng nút hạn chế được lưu giữ trong mỗi bước tìm kiếm. Bằng cách giới hạn số lượng nút được giữ trong mỗi bước, Beam Search cần ít bộ nhớ hơn là thuật toán tìm kiếm ưu tiên nhất. Vì thuật toán ưu tiên nhất không thể được sử dụng cho vấn đề tìm kiếm không gian lớn do sử dụng bộ nhớ phức tạp, Beam Search được sử dụng để duy trì khả năng theo dõi trong vấn đề không gian tìm kiếm lớn với việc sử dụng bộ nhớ hạn chế. Trong hệ thống dịch máy thống kê, nó được sử dụng để tìm các bản dịch tốt nhất trong số một số bản dịch tốt trong từng bước. Beam Search đã được sử dụng hầu hết trong các nghiên cứu hiện tại về Dịch máy thống kê vì hiệu suất tốt.

Greedy Decoder giải quyết vấn đề về không gian tìm kiếm bằng cách sử dụng kết quả tối ưu cục bộ trong mỗi bước với hy vọng tìm được giá trị tối ưu toàn cục. Nó nhanh nhưng nó không thành công đối với một số vấn đề để tìm ra mức tối ưu toàn cầu. Nó sẽ Dự án. bắt đầu với bản dịch từ sang chữ tốt nhất cho câu nguồn nhất định và sau đó cố gắng cải thiện chất lượng bản dịch bằng cách áp dụng một hoặc nhiều thao tác xác định trước như TranslateOneOrTwoWords, TranslateAndInsert, RemoveWordOfFertility,

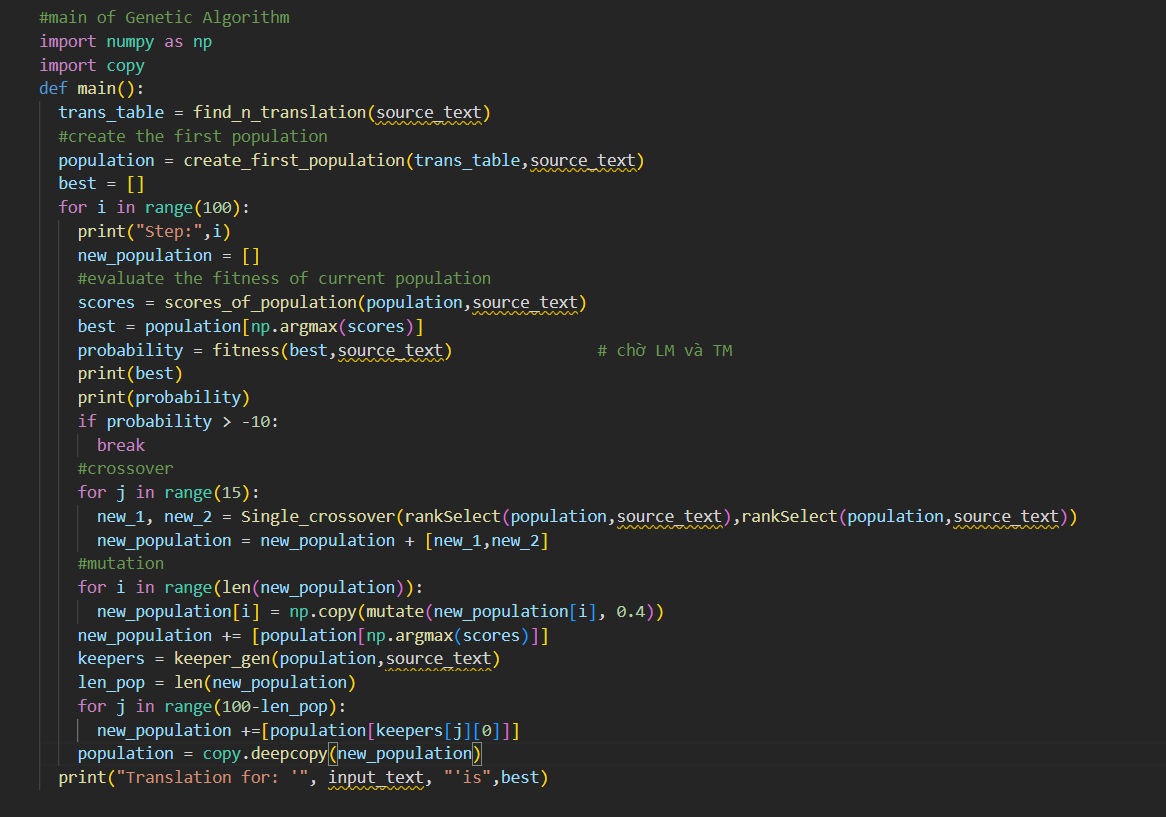
Giải thuật di truyền (GA) hoạt động với cơ chế tương tự như cơ chế hành vi của gen người. Để khởi tạo quần thể ban đầu, GA tạo ra các NST (nhiễm sắc thể) - lời giải - một cách ngẫu nhiên, trong đó mỗi NST là 1 chuỗi các gen -thành phần tạo nên lời giải. Sau đó, các NST trong quần thể sẽ trải qua quá trình Lai tạo bao gồm 2 bước: Lai ghép và Đột biến. Thông thường, để có thể tham gia quá trình Lai tạo, các NST sẽ phải trải qua bước Lựa chọn, trong đó GA sẽ dựa vào 1 thuật toán nào đó để chọn NST được Lai tạo. Kết quả sau quá trình Lai tạo là một quần thể NST con được sử dụng cho thế hệ NST tiếp theo. GA sẽ dựa vào thuật toán Thay thế để chọn các NST con và các NST từ quần thể hiện tại để đưa vào quần thể thế hệ mới. Bằng cách lặp lại quá trình này, GA có thể tìm được nhiều NST khác nhau với hi vọng là NST tốt sẽ xuất hiện ở quần thể thế hệ cuối. Thuật toán sẽ dừng lại sau khi số thế hệ vượt quá số lượng cho phép hoặc hàm fitness vượt qua ngưỡng mong muốn của bài toán.

Luồng hoạt động của GA

Diagram

Description automatically generated

Hình ảnh mô tả giải thuật di truyền



Các thuật toán được sử dụng cho quá trình Lựa chọn, Lai tạo, Thay thế sẽ được trình bày ở phần tiếp theo.

**2.2.2 Phương pháp đề xuất**

**2.2.2.1 Tổng quát về hệ thống**

Ý tưởng chính đằng sau Hệ thống dịch máy dựa trên xác suất (SMT) đó là SMT sẽ tạo ra Bản dịch (T) từ Câu cần dịch (S), sao cho xác suất là tốt nhất có thể. Xác suất này có thể được tính dựa theo công thức Bayes như sau

Định lý Bayes

Trong đó: : Xác suất câu T được dịch thành câu S

: Xác suất xuất hiện của câu T

Để tối đa hóa xác suất , ta sẽ cần tối đa hóa giá trị của . Giá trị của thể hiện xác suất mà ta nhận được câu S – thứ mà ta không thể tính toán được vì nó hoàn toàn phụ thuộc vào dữ liệu nhập vào của người dùng – vì vậy sẽ không tham gia vào quá trình tối đa hóa xác suất của ta.

Phương trình được sử dụng bởi SMT

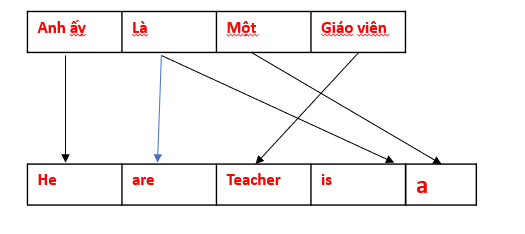
P(S|T) là xác suất nhận được câu nguồn S khi có Bản dịch T. Xác suất này là để giải quyết vấn đề về việc **lựa chọn các từ để dịch sao cho phù hợp**. P(S|T) sẽ được tính toán bởi Mô hình dịch thuật (Translation Model). P(T) là xác suất mà Bản dịch T là một bản dịch (câu) có nghĩa, hay nói cách khác là Bản dịch T có đúng **cấu trúc ngữ pháp** hay không. Xác suất này sẽ được tính bởi Mô hình Ngôn ngữ (Language Model). Ngoài 2 model kể trên, hệ thống SMT cần thêm thành phần thứ 3 được gọi là Decoder. Thành phần này đảm nhận việc tối đa hóa giá trị , và từ đó tìm được bản dịch với xác suất là lớn nhất. Với hệ thống của mình, chúng em lựa chọn Genetic Algorithm làm Decoder

**2.2.2.2 Bài toán dịch câu tiếng việt sang tiếng anh**

**a. Translation Model (Mô hình Dịch thuật)**

Mô hình dịch thuật dựa trên mối quan hệ song ngữ của hai câu từ ngữ liệu song song (parallel corpora).

Biểu thức P(S|T) có thể được hiểu là xác suất tạo ra câu Nguồn (S) từ Bản dịch (T), giá trị của nó có thể được ước tính từ một ngữ liệu song song (parallel corpora). Mỗi câu trong ngôn ngữ nguồn có một câu tương ứng trong ngôn ngữ đích. Ánh xạ (Alignment) là một tập hợp các mối quan hệ giữa các từ trong câu nguồn đến các từ trong câu đích với xác suất ước tính.



Hình 1

Để định nghĩa ánh xạ (Alignment), ta xét một câu nguồn dưới dạng chuỗi S1I = s1,…,si,…,sI và câu đích T1J = t1,…tj,…,tJ cần được ánh xạ. Ta định nghĩa ánh xạ giữa hai chuỗi từ là tập hợp con của tích Descartes của vị trí từ; đó là một ánh xạ A được định nghĩa:



Vì một từ trong câu nguồn có thể có nhiều từ trong câu đích nên ta cần chú ý tới ánh xạ nhiều nhiều. Translation Model có thể sử dụng xác suất ánh xạ để dịch từng từ trong câu nguồn thành một hoặc nhiều từ trong câu đích như hình 1. Xác suất P(S|T) có thể được tính qua phương trình:

Shape

Description automatically generated with medium confidence

S = S1 S2 … Ss, T = t1 t2 … tt, aij: Si tương ứng với tj

Trong đó Si là một từ trong câu nguồn () và tj là một bản dịch của Si trong câu đích (Shape

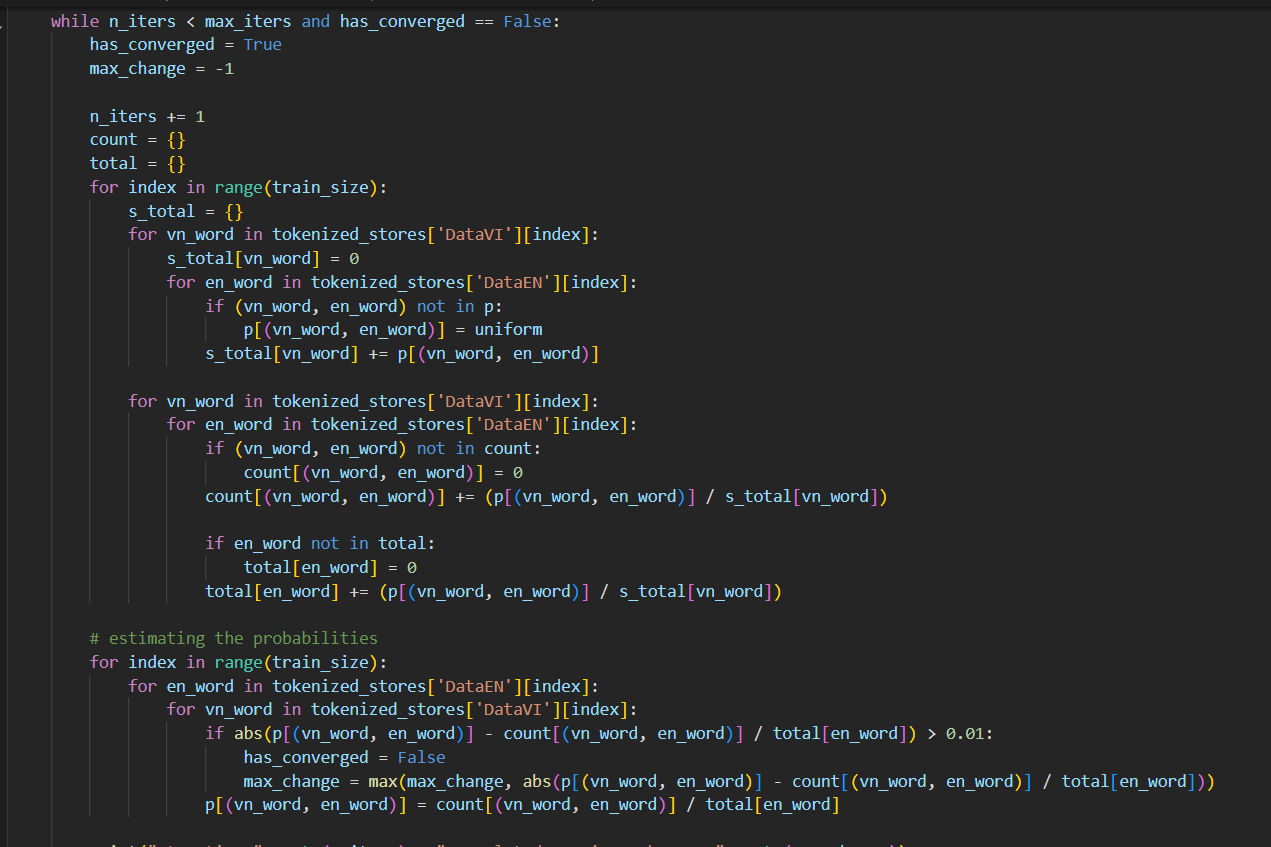
Description automatically generated with medium confidence) và aij là ánh xạ giữa si và tj.

Nếu chúng ta có một ngữ liệu song song (parallel corpora) lý tưởng, xác suất P(S|T) có thể dễ dàng tính theo công thức:

Count(si,tj): tần số của 2 từ gốc và nguồn cùng xuất hiện trong cùng 2 câu song song

Count(tj): tần số tj trong kho ngữ liệu

Hình ảnh dưới đây mô tả thuật toán xử lý Translation Model



**b. Language Model (Mô hình Ngôn ngữ)**

Mô hình ngôn ngữ thống kê là một phân phối xác suất trên các chuỗi từ. Với một câu có độ dài m, nó gán một xác suất P(w1,…,wm) cho toàn bộ câu.

Mô hình ngôn ngữ được sử dụng trong các ứng dụng xử lý ngôn ngữ tự nhiên khác nhau như Dịch máy, Nhận dạng giọng nói, Phân tích và truy xuất thông tin,… Language Model được sử dụng để tìm đầu ra tốt nhất trong ngôn ngữ đích. Language Model sử dụng phương trình sau để ước tính khả năng của một từ sẽ xuất hiện sau một chuỗi từ nhất định. Xác suất chuỗi từ đó được kí hiệu là P(w1 … wm); dựa vào chain rule, ta có thể tính xác suất đó dựa theo công thức



trong đó m là số từ trong chuỗi (câu) và n là số lượng các từ đứng trước từ Wi

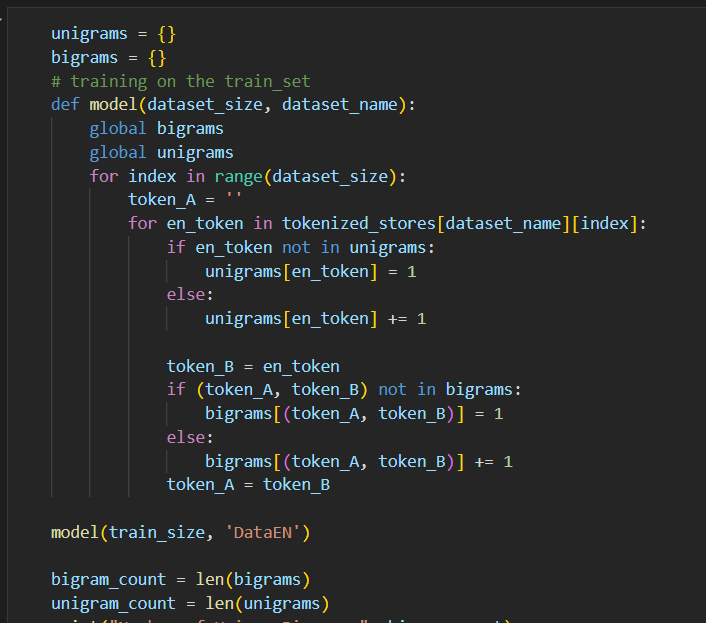
Shape

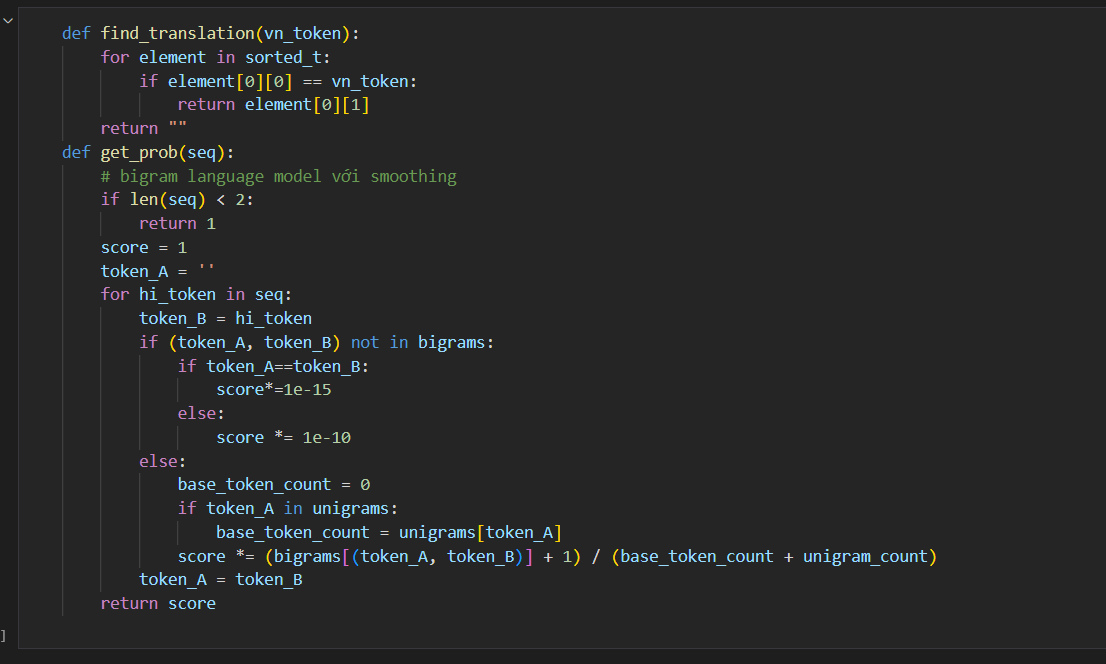
Description automatically generated with medium confidence

Một vấn đề thường xảy ra của mô hình ngôn ngữ n-gram là sự thưa thớt dữ liệu, có những trình tự từ có nghĩa nhưng không xuất hiện trong ngữ liệu huấn luyện. Những trình tự từ đó cần được gán một xác suất nhỏ thông qua kỹ thuật làm mịn. Kỹ thuật làm mịn được sử dụng trong hệ thống này là kỹ thuật làm mịn add-1 dựa theo công thức sau

với V là số các từ độc nhất trong ngữ liệu huấn luyện

Hình ảnh dưới đây mô tả thuật toán xử lý Language Model





**2.2.2.3 Hàm fitness**

Để GA có thể được sử dụng làm Decoder cho SMT, hàm fitness của GA cần nhận vào 2 tham số, đó chính là các xác suất được tính bởi Language Model và Translation Model. Giá trị hàm fitness được tính dựa theo công thức sau:

Trong đó: a: Tham số cho Translation Model

b: Tham số cho Language Model

### **Quá trình Lựa chọn**

Chúng em sẽ áp dụng thuật toán Roulette wheel cho quá trình Lựa chọn. Theo phương thức này, xác suất mỗi NST sẽ được lựa chọn dựa trên giá trị fitness. Các NST với giá trị fitness cao sẽ có xác suất được lựa chọn cao hơn.

1. Tính tổng giá trị fitness của tất cả các NST trong quần thể - tổng ***S.***
2. Sinh một số ngẫu nhiên nằm trong khoảng (0, ***S***) – số ***r.***
3. Lần lượt lấy từng NST trong quần thể và cộng dồn giá trị fitness của chúng từ 0 – tổng ***s.*** Khi ***s***>***r***, dừng thuật toán và chọn NST vừa được lấy ra, đưa NST đó vào danh sách được Lai tạo.

Thuật toán Roulette wheel

### **Quá trình Lai tạo**

Đối với bước Lai ghép, thuật toán Lai ghép đơn được sử dụng

1. Sinh ra NST con 1 có cùng số bộ gen với NST cha.
2. Sinh ra NST con 2 có cùng số bộ gen với NST mẹ.
3. for j=0 to length\_of\_selected\_parent (Chọn NST có độ dài ngắn nhất)
   1. Chọn ngẫu nhiên NST cha hoặc mẹ
   2. Thêm từ ở vị trí j trong NST được chọn vào NST con 1
   3. Thêm từ ở vị trí j trong NST không được chọn vào NST con 2
4. for k=j to length\_of\_remaining\_parent (NST không được chọn ở bước trước)
   1. Thêm từ ở vị trí k trong NST không được chọn ở bước trước vào NST con 1/2 (chọn NST con có độ dài lớn nhất)
5. Trả về 2 NST con.

Thuật toán Lai ghép đơn

Một số NST con sẽ được Đột biến. Hệ thống sẽ sinh ra một giá trị ngẫu nhiên ***p.*** Nếu ***p*** < ***P*** (***P*** là một giá trị được cố định), quá trình Đột biến sẽ được thực hiện.

### **Quá trình Thay thế**

Những NST cha mẹ sẽ được sắp xếp dựa trên giá trị hàm fitness. Những NST có giá trị fitness cao sẽ được giữ lại và đưa vào quần thể thế hệ tiếp theo.

**2.2.3 Thực nghiệm và đánh giá kết quả**

a. Mô tả tập dữ liệu

Tập dữ liệu bao gồm các câu đơn tiếng anh và tiếng việt. Có tất cả 500 câu bao gồm 100 câu tiếng việt và 100 tiếng anh. Mỗi câu gồm ít nhất 2 thành phần chính là chủ ngữ và vị ngữ, độ dài các câu là khác nhau.

b. Mô tả thực nghiệm

Bộ dữ liệu được chia thành hai bộ dữ liệu huấn luyện – kiểm tra với tỉ lệ 80%-20% để tiến hành quá trình thực nghiệm:

* + Bộ dữ liệu thứ nhất là bộ dữ liệu huấn luyện (training dataset): bao gồm 200 câu tiếng việt và 200 câu tiếng anh dùng để huấn luyện mô hình.
  + Bộ dữ liệu thứ hai là bộ dữ liệu kiểm tra (test dataset): bao gồm 5 tiếng việt và 50 câu tiếng anh dùng để kiểm tra mô hình.

c. Kết quả thực nghiệm

Với những câu có trong dữ liệu huấn luyện, hệ thống đưa ra kết quả chính xác khoảng 98%.

Với những câu nằm ngoài dữ liệu huấn luyện, độ chính xác của hệ thống đạt được khoảng 80%

**2.2.4 Kết luận**

Dịch ngôn ngữ là một trong những yếu tố quan trọng cho sự phát triển của xã hội ngày nay. Đề tài đã đưa ra phương pháp dịch câu tiếng việt sang tiếng anh có hiệu quả cao, bằng việc sử dụng giải thuật di truyền để tìm ra câu dịch tốt nhất so với ngôn ngữ gốc. Các kết quả thực nghiệm cũng cho thấy phương pháp của chúng em đủ nhanh, độ chính xác cao. Trong tương lai, chúng em sẽ cố gắng xây dựng một tập dữ liệu lớn hơn để hệ thống có thể dịch một cách chính xác hơn nữa. Nhược điểm của phương pháp hiện tại là độ chính xác khi dịch các từ không phổ biến chưa được cao. Chúng em sẽ tiếp tục cải tiến phương pháp và cung cấp giải pháp cho các vấn đề còn lại.

# Kịch bản vận hành hệ thống

Kịch bản 1: Các từ nằm trong câu S đều tồn tại trong corpus

1. Người dùng nhập vào hệ thống câu tiếng Việt “Tôi là một bác sĩ” (S).
2. Hệ thống biến câu S thành một danh sách các từ có nghĩa.
3. Hệ thống gửi danh sách từ có nghĩa đó cho GA Decoder.
4. GA Decoder sinh ra các bản dịch tiếng Anh ngẫu nhiên (T) cho câu S dựa vào Translation Model. Các bản dịch đó là NST cho quần thể đầu tiên.
5. Language Model tính xác suất mà các câu T đó là câu đúng ngữ nghĩa dựa trên xác suất bi-gram/tri-gram của mỗi cặp từ trong câu.
6. Translation Model tính xác suất mà các câu T đó là bản dịch đúng dựa trên xác suất mà từ trong câu S được ánh xạ sang từ tại vị trí tương ứng trong câu T.
7. GA Decoder tính hàm fitness của các câu T dựa trên 2 xác suất vừa được tính. Nếu giá trị hàm fitness của 1 câu t trong quần thể bằng hoặc lớn hơn giá trị ngưỡng thì đi tới Bước 11.
8. GA Decoder lựa chọn các câu T để thực hiện Lai ghép chéo dựa trên thuật toán Roulette wheel.
9. GA Decoder quyết định dựa trên xác suất liệu quần thể NST con vừa được tạo có được Đột biến hay không. Nếu có, NST con thực hiện Đột biến dựa trên thuật toán Đột biến trao đổi.
10. GA Decoder tính hàm fitness của các NST con dựa trên xác suất được tính từ Language Model và Translation Model. Nếu giá trị hàm fitness của 1 câu T trong quần thể bằng hoặc lớn hơn giá trị ngưỡng thì đi tới Bước 11. Nếu không, quay lại Bước 7.
11. GA Decoder đưa ra NST con có giá trị hàm fitness tốt nhất (câu “I am a doctor”) ra màn hình.

Kịch bản 2: Trong câu S tồn tại từ không nằm trong corpora

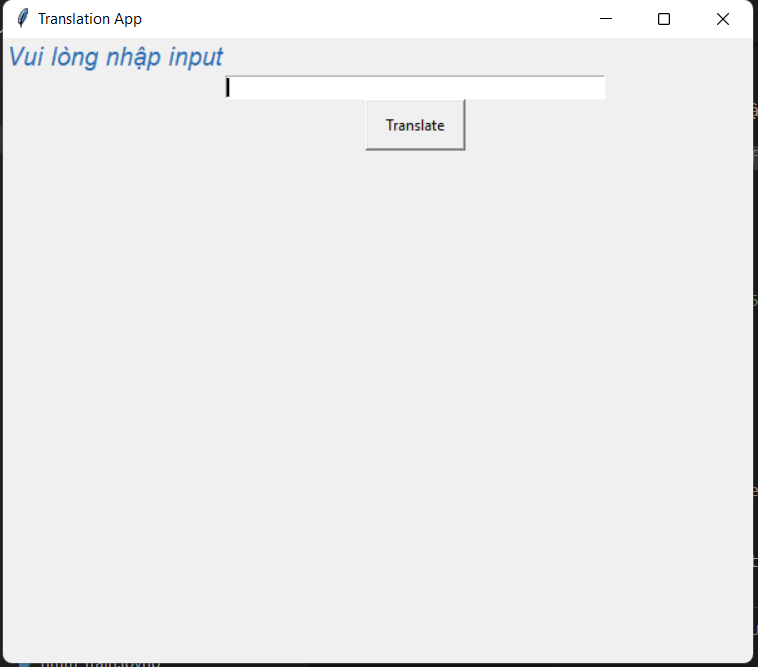
1. Người dùng nhập vào hệ thống câu tiếng Việt “Đó là một con kiến” (S). Từ “con\_kiến” không tồn tại trong corpora
2. Hệ thống biến câu S thành một danh sách các từ có nghĩa.
3. Hệ thống gửi danh sách từ có nghĩa đó cho GA Decoder.
4. GA Decoder sinh ra các bản dịch tiếng Anh ngẫu nhiên (T) cho câu S dựa vào Translation Model. Các bản dịch đó là NST cho quần thể đầu tiên.
5. Language Model tính xác suất mà các câu T đó là câu đúng ngữ nghĩa dựa trên xác suất bi-gram/tri-gram của mỗi cặp từ trong câu.
6. Translation Model tính xác suất mà các câu T đó là bản dịch đúng dựa trên xác suất mà từ trong câu S được ánh xạ sang từ tại vị trí tương ứng trong câu T. Từ “con\_kiến” không tồn tại trong corpora nên xác suất ánh xạ từ “con\_kiến” sang một từ tiếng Anh bất kì được gán là một giá trị rất nhỏ.
7. GA Decoder tính hàm fitness của các câu T dựa trên 2 xác suất vừa được tính. Nếu giá trị hàm fitness của 1 câu *t* trong quần thể bằng hoặc lớn hơn giá trị ngưỡng thì đi tới Bước 11.
8. GA Decoder lựa chọn các câu T để thực hiện Lai ghép chéo dựa trên thuật toán Roulette wheel.
9. GA Decoder quyết định dựa trên xác suất liệu quần thể NST con vừa được tạo có được Đột biến hay không. Nếu có, NST con thực hiện Đột biến dựa trên thuật toán Đột biến trao đổi.
10. GA Decoder tính hàm fitness của các NST con dựa trên xác suất được tính từ Language Model và Translation Model. Nếu giá trị hàm fitness của 1 câu T trong quần thể bằng hoặc lớn hơn giá trị ngưỡng thì đi tới Bước 11. Nếu không, quay lại Bước 7.
11. GA Decoder đưa ra NST con có giá trị hàm fitness tốt nhất ra màn hình.

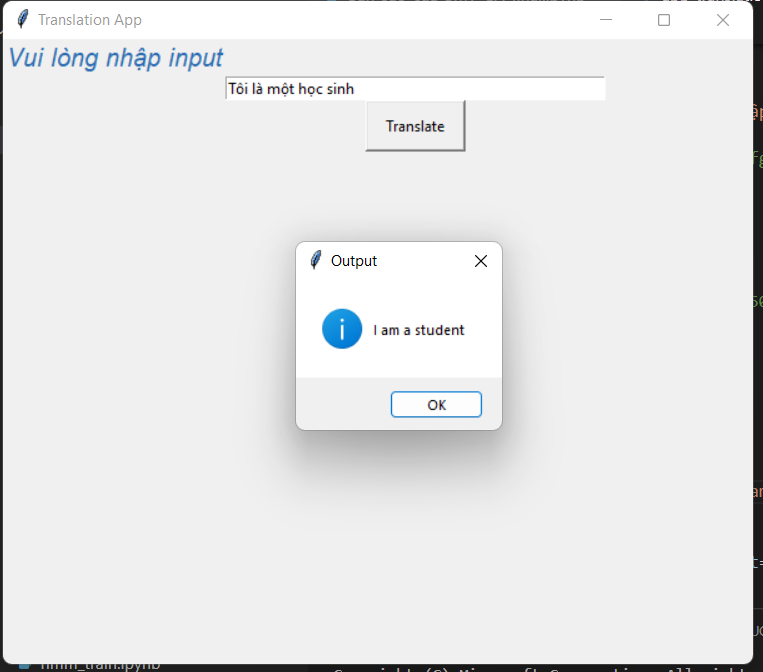
Kịch bản 3: Trong câu S tồn tại chữ số

1. Người dùng nhập vào hệ thống câu tiếng Việt “Tôi là 1 bác\_sĩ” (S).
2. Hệ thống biến câu S thành một danh sách các từ có nghĩa.
3. Hệ thống gửi danh sách từ có nghĩa đó cho GA Decoder.
4. GA Decoder sinh ra các bản dịch tiếng Anh ngẫu nhiên (T) cho câu S dựa vào Translation Model. Trong mỗi bản dịch đều chứa ký tự “1” (do câu người dùng nhập vào chứa chữ số “1”). Các bản dịch đó là NST cho quần thể đầu tiên.
5. Language Model tính xác suất mà các câu T đó là câu đúng ngữ nghĩa dựa trên xác suất bi-gram/tri-gram của mỗi cặp từ trong câu.
6. Translation Model tính xác suất mà các câu T đó là bản dịch đúng dựa trên xác suất mà từ trong câu S được ánh xạ sang từ tại vị trí tương ứng trong câu T.
7. GA Decoder tính hàm fitness của các câu T dựa trên 2 xác suất vừa được tính. Nếu giá trị hàm fitness của 1 câu t trong quần thể bằng hoặc lớn hơn giá trị ngưỡng thì đi tới Bước 11.
8. GA Decoder lựa chọn các câu T để thực hiện Lai ghép chéo dựa trên thuật toán Roulette wheel.
9. GA Decoder quyết định dựa trên xác suất liệu quần thể NST con vừa được tạo có được Đột biến hay không. Nếu có, NST con thực hiện Đột biến dựa trên thuật toán Đột biến trao đổi.
10. GA Decoder tính hàm fitness của các NST con dựa trên xác suất được tính từ Language Model và Translation Model. Nếu giá trị hàm fitness của 1 câu T trong quần thể bằng hoặc lớn hơn giá trị ngưỡng thì đi tới Bước 11. Nếu không, quay lại Bước 7.
11. GA Decoder đưa ra NST con có giá trị hàm fitness tốt nhất (câu “I am a doctor”) ra màn hình.

# Nghiên cứu, xây dựng hệ thống

Đối với chức năng dịch câu từ tiếng Việt sang tiếng Anh được chúng tôi thiết kế sao cho thuận tiện nhất cho người sử dụng. Dữ liệu được lưu trữ dưới định dạng text. Hình ảnh dưới đây mô tả giao diện dịch câu được chúng tôi xây dựng.

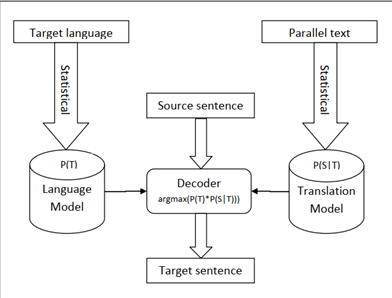




* Giao diện chính bao gồm ba thành phần cơ bản sau:

1. Màn hình hiển thị: cho phép chúng ta nhập dữ liệu để dịch câu
2. Nút bấm: khi chúng ta click vào sẽ trả ra kết quả dịch tương ứng với ngôn ngữ gốc
3. Cửa sổ hiển thị: trả ra kết quả dịch tương ứng với ngôn ngữ gốc

Sơ đồ hoạt động dịch của hệ thống:



# Thử nghiệm đánh giá chất lượng hệ thống

## 5.1 Tiêu chuẩn đánh giá

Hiện nay trong nước chưa ban hành đầy đủ các tiêu chuẩn, cũng như phương thức đánh giá chính quy nào về việc đánh giá tiêu chuẩn chất lượng một hệ thống dịch thuật AI. Vì vậy chúng tôi tham khảo các tiêu chuẩn cũng như phương pháp đánh giá của nước ngoài để áp dụng cho việc đánh giá hệ thống.

## 5.1.1 Đánh giá độ chính xác của hệ thống

Phương pháp đánh giá là sử dụng phép đo BLEU trong dịch máy:

- BLUE là viết tắt của **Bilingual Evaluation Understudy**, là phương pháp đánh giá một bản dịch dựa trên các bản dịch tham khảo. BLEU được thiết kế để sử dụng trong dịch máy (Machine Translation), nhưng thực tế, phép đo này cũng được sử dụng trong các nhiệm vụ như tóm tắt văn bản, nhận dạng giọng nói, sinh nhãn ảnh v... v...

- Điều kiện tiên quyết để có thể sử dụng BLEU là bạn phải có một (hoặc nhiều) câu mẫu. Đối với bài toán dịch máy, câu mẫu chính là câu đầu ra của cặp câu trong tập dữ liệu. **Cách tính điểm:** BLUE đánh giá một câu thông qua việc so khớp câu đó với các câu mẫu và cho thang điểm từ 0 (sai lệch tuyệt đối) đến 1 (khớp tuyệt đối).

- BLEU được biết đến như một phương pháp đơn giản, dễ hiểu, chi phí tính toán thấp và tương đồng với cách đánh giá của con người. Mặc dù vậy, yếu tố con người trong việc sinh câu mẫu làm cho BLEU không hoàn toàn khách quan. Ví dụ, cùng một câu có thể có nhiều bản dịch tốt và việc viết tất cả các bản dịch đó vào tập câu mẫu đôi khi bất khả thi.

- Cách tính của BLEU là đếm số n-gram khớp nhau giữa câu mẫu (R) và câu được đánh giá (C) sau đó chia cho số token của C. Việc chọn n phụ thuộc vào ngôn ngữ, nhiệm vụ và mục tiêu cụ thể. Đơn giản nhất ta có thể sử dụng unigram là n-gram chứa 1 token (n=1). Một cách trực quan, n càng lớn, câu văn càng mượt.

- Việc so khớp này không phụ thuộc vào vị trí, do đó BLEU không thể đánh giá được thứ tự của từ. Đây vừa là ưu điểm vừa là hạn chế của BLEU. Trong ngôn ngữ, một câu có thể được biểu diễn bởi các thứ tự từ khác nhau nhưng vẫn phải tuân theo những quy tắc nhất định.

- Ngoài ra, để tránh việc một bản dịch lặp đi lặp lại một từ mà vẫn được “chấm điểm” cao (ví dụ như **“this this this this”** được so với **“this is a cat”**), BLEU tính đến số lần xuất hiện lớn nhất của mỗi n-gram trong toàn bộ các câu mẫu để giới hạn số lần khớp tối đa. Trong ví dụ trên, **this** sẽ chỉ được tính 1 lần”



Số lần khớp tối đa được giới hạn bởi số lần xuất hiện tối đa của n-gram trong các câu mẫu.

- Trong Python, để tính điểm BLEU chúng ta có thể thực hiện dễ dàng với sự hỗ trợ của thư viện NLTK (Natural Language Toolkit). Ví dụ như chúng ta có hai bản dịch cần đánh giá của câu **“Có con mèo nằm trên chiếu”**. Ta sẽ có điểm BLEU của hai câu như sau:

**import**  nltk.translate.bleu\_score **as** bleu

reference\_translation=['The cat is on the mat.'.split(),

                      'There is a cat on the mat.'.split()

                     ]

candidate\_translation\_1='the the the mat on the the.'.split()

candidate\_translation\_2='The cat is on the mat.'.split()

print("BLEU Score candidate 1: ",bleu.sentence\_bleu(reference\_translation, candidate\_translation\_1))

print("BLEU Score candidate 2: ",bleu.sentence\_bleu(reference\_translation, candidate\_translation\_2))

Với kết quả 0.47 và 1, [BLEU](https://trituenhantao.io/tu-dien-thuat-ngu/bleu/) đánh giá câu dịch thứ 2 tốt hơn câu dịch thứ 1 vì nó khớp tuyệt đối với câu trong tập ví dụ.

Hệ thống đạt chất lượng nếu điểm BLEU lớn hơn hoặc bằng 0.55.

## 5.1.2 Đánh giá sự ổn định của hệ thống

Phương pháp đánh giá:

- Yêu cầu hệ thống có thể chạy liên tục trong thời gian dài.

- Dịch câu và đưa ra kết quả trong thời gian ngắn.

- Trong quá trình chạy hệ thống bị lỗi thì phải tiến hành đánh giá lại từ đầu.

## 5.2 Kết quả chạy thử nghiệm trên thực tế

Cấu Hình Phần cứng của hệ thống:

• RAM: 32GB

• OS: Windows 11

• GPU: GTX 1060 6G

• CPU: intel core i7-8750H

## 5.2.1 Đánh giá độ chính xác của hệ thống

Để đánh giá kết quả thực nghiệm, nhóm dự án tiến hành dịch 100 câu để lấy mẫu thử nghiệm, kết quả cụ thể như sau:

Precision = Độ chính xác của toàn hệ thống đạt 94% đạt yêu cầu.

## 5.2.2 Đánh giá độ ổn định của hệ thống

Kết quả đánh giá độ ổn định cũng cho thấy hệ thống có thể chạy liên tục, nhanh mà không xảy ra bất kỳ sự cố nào.

# Kết luận và Định hướng nghiên cứu trong tương lai

Hệ thống tự động dịch thuật là một hệ thống quan trọng và cần thiết giúp nâng cao năng lực ngoại ngữ, khả năng từ vựng góp phần xây dựng đô thị thông minh.

Kết quả thực nghiệm cũng cho thấy, thuật toán Giải thuật di truyền do nhóm nghiên cứu phát triển đạt độ chính xác cao, với thời gian xử lý nhanh hoàn toàn phù hợp với ứng dụng thời gian thực. Ngoài ra phần mềm do nhóm nghiên cứu phát triển cũng đạt độ ổn định cao đủ tiêu chuẩn để đưa vào hoạt động trên thực tế. Hệ thống của nhóm phát triển vẫn còn những khuyết điểm cần phải cải thiện thêm.

Kết quả thực nghiệm cũng cho thấy, thuật toán phát hiện bạo lực do nhóm nghiên cứu phát triển đạt độ chính xác cao, với thời gian xử lý nhanh hoàn toàn phù hợp với ứng dụng thời gian thực. Ngoài ra phần mềm do nhóm nghiên cứu phát triển cũng đạt độ ổn đinh cao đủ tiêu chuẩn để đưa vào hoạt động trên thực tế.

Hệ thống của nhóm phát triển vẫn còn những khuyết điểm cần phải cải thiện thêm. Trước hết cơ sở dữ liệu vẫn chưa thực sự đầy đủ cần phải thu thập thêm để cải thiện độ chính xác huấn luyện mô hình. Tiếp theo nhóm nghiên cứu cũng có kế hoạch cải thiện kiến trúc mô hình nhằm tăng cường độ chính xác của mô hình hơn nữa.

Ngoài ra hệ thống mới chỉ được thực nghiệm trong phạm vi của Học viện điều này chưa đảm bảo được hệ thống sẽ chạy tốt trong môi trường khác. Do vậy trong thời gian tới nhóm nghiên cứu sẽ tiến hành thực nghiệm đối với nhiều địa điểm môi trường khác nhau, để đảm bảo hệ thống có thể chạy tốt trong mọi điều kiện hoàn cảnh khác nhau.

# TÀI LIỆU THAM KHẢO

[1] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T.-K. Kim, “Fast fight detection,” PLoS ONE, vol. 10, no. 4, Apr. 2015, Art. no. e0120448.

[2] P. C. Ribeiro, R. Audigier, and Q. C. Pham, “RIMOC, a feature to discriminate unstructured motions: Application to violence detection for videosurveillance,” Comput. Vis. Image Understand., vol. 144, pp. 121–143, Mar. 2016.

[3] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, “Automatic fight detection in surveillance videos,” in Proc. 14th Int. Conf. Adv. Mobile Comput. Multi Media, Nov. 2016, pp. 225–234.

[4] R. Nar, A. Singal, and P. Kumar, “Abnormal activity detection for bank ATM surveillance,” in Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI), Sep. 2016, pp. 2042–2046.

[5] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, ‘‘Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation,’’ IEEE Trans. Inf. Forensics Security, vol. 12, no. 12, pp. 2945–2956, Dec. 2017.

[6] V. E. M. Arceda, K. M. F. Fabian, P. C. L. Laura, J. J. R. Tito, and J. C. G. Caceres, ‘‘Fast face detection in violent video scenes,’’ Electron. Notes Theor. Comput. Sci., vol. 329, pp. 5–26, Dec. 2016.

[7] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, ‘‘A new method for violence detection in surveillance scenes,’’ Multimedia Tools Appl., vol. 75, no. 12, pp. 7327–7349, 2016.

[8] J. Xie, W. Yan, C. Mu, T. Liu, P. Li, and S. Yan, ‘‘Recognizing violent activity without decoding video streams,’’ Optik, vol. 127, no. 2, pp. 795–801, Jan. 2016.

[9] E. Y. Fu, M. X. Huang, H. Va Leong, and G. Ngai, ‘‘Cross-species learning: A low-cost approach to learning human fight from animal fight,’’ in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 320–327

[10] M. Coletto, C. Lucchese, and S. Orlando, ‘‘Do violent people smile: Social media analysis of their profile pictures,’’ in Proc. Companion Proc. Web Conf., Apr. 2018, pp. 1465–1468.

[11] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEng-Technol.2017.8308186.

[12] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[13] Alex Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network”. arXiv, 2018.

[14] Ralf C. Staudemeyer and Eric Rothstein Morris, “Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks”. arXiv, 2019.

[15] Seymanur Akti, Gozde Ayse Tataroglu and Hazim Kemal Ekenel, “Vision-based Fight Detection from Surveillance Cameras”. IEEE, 2019.

[16] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin, “Attention Is All You Need”. arXiv, 2017.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”. arXiv, 2016.

[19] Nicolai Wojke, Alex Bewley and Dietrich Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric”. arXiv, 2017.

[20] Alexey Bochkovskiy, Chien-Yao Wang and Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection”. arXiv, 2020.

[21] Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q. Weinberger, “Densely Connected Convolutional Networks”. arXiv, 2016.

[22] Mingxing Tan and Quoc V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. arXiv, 2019.

[23] Joseph Redmon and Ali Farhadi, “YOLOv3: An Incremental Improvement”. arXiv, 2018.

[24] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos and Ben Upcroft, “Simple Online and Realtime Tracking”. arXiv, 2016.

[25] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[26] Alex Krizhevsky, Alex Krizhevsky and Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”. NIPS, 2012.

[27] Matthew D Zeiler and Rob Fergus, “Visualizing and Understanding Convolutional Networks”. arXiv, 2013.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich, “Going Deeper with Convolutions”. arXiv, 2014.

[29] Karen Simonyan and Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”. arXiv, 2015.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition”. arXiv, 2015.