ỦY BAN NHÂN DÂN TỈNH BÌNH DƯƠNG
TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT

NGUYỄN THỊ THANH HƯƠNG

XÂY DỰNG HỆ THỐNG TRẢ LỜI TỰ ĐỘNG CHATBOT BẰNG TIẾNG VIỆT SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 8480104

LUẬN VĂN THẠC SỸ

GIẢNG VIÊN HƯỚNG DẪN
TS. Bùi Thanh Hùng

BÌNH DƯƠNG - 2019

LỜI CAM ĐOAN

Tôi là Nguyễn Thị Thanh Hương, học viên lớp CH16HT01, ngành Hệ thống thông tin, trường Đại học Thủ Dầu Một. Tôi xin cam đoan luận văn "Xây dựng hệ thống trả lời tự động chatbot bằng tiếng việt sử dựng phương pháp học sâu" là do tôi nghiên cứu, tìm hiểu và phát triển dưới sự hướng dẫn của TS. Bùi Thanh Hùng, không phải sự sao chép từ các tài liệu, công trình nghiên cứu của người khác mà không ghi rõ trong tài liệu tham khảo. Tôi xin chịu trách nhiệm về lời cam đoan này.

Bình Dương, ngày 31 tháng 1 năm 2019

Tác giả

Nguyễn Thị Thanh Hương

LỜI CẢM ƠN

Để hoàn thành luận văn này, tôi xin gửi lời cảm ơn đến tất cả các Thầy cô trường Đại học Thủ Dầu Một đã tận tình giảng dạy và truyền đạt cho tôi những kiến thức hữu ích trong suốt quá trình học tập tại trường. Tôi cũng xin cảm ơn các anh chị, các bạn, các em trong lớp cao học đã chia sẻ, hỗ trợ cho tôi để tôi có thể thực hiện tốt luận văn của mình.

Hơn hết, tôi xin chân thành cảm ơn thầy hướng dẫn TS. Bùi Thanh Hùng, người đã tận tình truyền đạt, chỉ dạy cho tôi những kiến thức bổ ích về máy học và học tập sâu, cảm ơn thầy đã nhiệt tình hướng dẫn, chỉ bảo cho tôi trong suốt quá trình tôi nghiên cứu, xây dựng và hoàn thiện luận văn này.

Xin gửi lời cảm ơn sâu sắc tới gia đình, các anh chị em đồng nghiệp và bạn bè đã luôn động viên, chia sẻ kinh nghiệm, cung cấp các tài liệu hữu ích cho tôi để tôi thực hiện tốt luận văn của mình.

TÓM TẮT LUẬN VĂN

Cùng với sự phát triển của công nghệ, con người ngoài giao tiếp trực tiếp với nhau bằng ngôn ngữ tự nhiên, chúng ta còn thường xuyên liên lạc và kết nối với nhau mọi lúc mọi nơi thông qua mạng xã hội. Một hệ thống trả lời tự động thông minh có thể giúp con người trò chuyện, nhắc nhở hay làm trợ lý công việc và có thể theo dõi tình trạng sức khỏe bản thân,… đã được nhiều nhà nghiên cứu bằng cách sử dụng các kỹ thuật trong học máy để xây dựng và phát triển. Đối thoại thông minh là một nhiệm vụ quan trọng trong bài toán hiểu và xử lý ngôn ngữ tự nhiên. Những phương pháp tiếp cận trước đây chỉ giới hạn trong vài lĩnh vực nhưng hiệu quả đạt được chưa cao, khó mở rộng mô hình và ứng dụng rộng rãi. Hướng nghiên cứu dựa trên Deep Learning kết hợp với trí tuệ nhân tạo là một trong những xu hướng phát triển hiện nay.

Đề tài luận văn dựa trên những nghiên cứu trước đây để đề xuất nghiên cứu và phát triển một hệ thống trả lời tự động dựa trên hai hướng tiếp cận theo hướng dịch máy và trích xuất thông tin, hai hướng này đều sử dụng mạng học sâu LSTM bằng cách áp dụng phương pháp học chuỗi liên tiếp (sequence-to-sequence) và kĩ thuật attention để sinh ra câu trả lời tự động từ một chuỗi đầu vào tương ứng. Mô hình dịch máy bằng mạng Nơ-ron nhân tạo của Google (Google’s Neural Machine Translation) và mô hình phân loại câu hỏi theo hướng mạng Nơ-ron sâu được áp dụng để huấn luyện trên bộ dữ liệu chuẩn và bộ dữ liệu tiếng Việt được thu thập, sau đó so sánh kết quả thực nghiệm trên hai bộ dữ liệu này.

Bộ dữ liệu thu thập sẽ phân tách thành hai bộ câu hỏi và câu trả lời tương ứng, sau đó tiến hành tách từ để tiến hành thiết lập biểu diễn các từ dưới dạng các véc-tơ và các bộ từ vựng để tiến hành huấn luyện và kết hợp với các phương pháp đánh giá để cho ra mô hình dự đoán đưa ra các câu trả lời tối ưu.

Luận văn cũng đề xuất xây dựng một ứng dụng web hỗ trợ tư vấn trả lời tự động các câu hỏi của sinh viên liên quan đến chương trình đào tạo, tuyển sinh, thông tin về giảng viên và các văn bản thường gặp của trường Đại học Thủ Dầu Một.

Kết quả thực nghiệm cho thấy mô hình đề xuất đạt được kết quả tốt trên hai độ đo: Độ chính xác (Accuracy) và đánh giá BLEU dựa trên các bộ dữ liệu được sử dụng để huấn luyện và đánh giá.

MỤC LỤC

DANH MỤC THUẬT NGỮ VÀ CÁC TỪ VIẾT TẮT

Từ viết tắt	Từ chuẩn	Diễn giải
Chatbot	Chatbot	Hệ thống trả lời tự động
NLP	Natural Languague Processing	Xử lý ngôn ngữ tự nhiên
RNN	Recurrent Neural Network	Mạng nơ ron tái phát
LSTM	Long short-term memory	Mạng cải tiến để giải quyết vấn đề phụ thuộc quá dài
Seq2Seq	sequence to sequence	Phương pháp học chuỗi liên tiếp trong DeepLearning
GNMT	Google’s Neural Machine Translation	Trình dịch máy của Google
BLEU	BiLingual Evaluation Understudy	Độ đo BLEU

DANH MỤC CÁC BẢNG

DANH MỤC HÌNH VẼ, ĐỒ THỊ

CHƯƠNG 1
TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU

1.1. Lí do chọn đề tài

Trí tuệ nhân tạo (AI) và học máy (machine learning - ML) là thành phần chính trong Cuộc cách mạng công nghiệp 4.0 đang bùng nổ và phát triển mạnh mẽ. Xử lí ngôn ngữ tự nhiên Natural Language Processing (NLP) là một trong số những bài toán cơ bản của Trí tuệ nhân tạo với nhiều chủ đề như: Tìm kiếm, Trả lời tự động, Tóm tắt văn bản, Phân loại văn bản, Truy xuất thông tin, ...Chatbot (hay là một hệ thống trả lời tự động) được biết đến là một chương trình máy tính tương tác với người dùng bằng ngôn ngữ tự nhiên dưới một giao diện đơn giản, âm thanh hoặc dưới dạng tin nhắn. Chatbot được ứng dụng rất rộng rãi trong nhiều lĩnh vực như Tài chính ngân hàng, Kinh doanh – Sản xuất, Y tế, Giáo dục,... với mục đích làm Trợ lý cá nhân, chăm sóc khách hàng, đặt chỗ, mua hàng, bán hàng tự động, hỗ trợ dạy và học, tư vấn dịch vụ công…

Rất nhiều công ty lớn đã phát triển chatbot của mình nhằm trả lời các hỏi đáp trực tuyến, bằng việc sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên và các kỹ thuật học sâu Deep Learning làm tăng chất lượng và hiệu quả của hệ thống chatbot, giúp tiết kiệm chi phí, giúp tư vấn khách hàng liên tục ngay cả khi không có nhân viên tư vấn trực tiếp.

Với một số lượng lớn học sinh, sinh viên đang theo học ở các trường như hiện nay thì việc các em có nhiều thắc mắc, nhiều câu hỏi, nhiều vấn đề cần giải đáp liên quan đến việc học của mình, các em cần có người có thể giải đáp các câu hỏi đó một cách chính xác và nhanh nhất. Việc phải trả lời các câu hỏi giống nhau của nhiều em sẽ tạo áp lực lớn cho đội ngũ tư vấn. Vì thế cho nên tôi đã chọn đề tài "Xây dựng hệ thống trả lời tự động chatbot bằng Tiếng Việt sử dụng phương pháp học sâu", nhằm tạo ra một cổng thông tin hỏi đáp trực tuyến cho các em.

1.2. Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của luận văn là vận dụng kiến thức đã học để xây dựng một hệ thống trả lời tự động, sử dụng mạng học sâu Deep Neural Networks, dựa trên khung làm việc sequence-to-sequence và cơ chế attention để sinh ra câu trả lời tự

động từ một chuỗi đầu vào tương ứng. Mô hình được huấn luyện end-to-end GNMT (Google's Neural Machine Translation) trên tập dữ liệu miền mở có sẵn. Từ đó xây dựng, cài đặt và thử nghiệm một mô hình trả lời các câu hỏi theo hướng tự động thông qua việc sử dụng mạng nơ-ron để huấn luyện kho dữ liệu Tiếng Việt dựa trên bộ dữ liệu là các thông tin liên quan đến chương trình đào tạo, hỗ trợ tuyển sinh, thông tin về giảng viên và các văn bản thường gặp của trường Đại học Thủ Dầu Một.

1.3. Đối tượng, phạm vi nghiên cứu

Nghiên cứu các Mô hình huấn luyện dựa trên nền tảng học sâu Deep Neural Networks để xây dựng hệ thống trả lời tự động.

Lĩnh vực nghiên cứu: xây dựng mô hình trả lời tự động các câu hỏi của sinh viên liên quan đến chương trình đào tạo của Trường Đại học Thủ Dầu Một thông qua một hệ thống câu hỏi và trả lời được xây dựng từ trước. Qua cơ chế huấn luyện từ các phương pháp của DeepLearning như: RNN, CNN, LSTM, sau đó tiến hành dự đoán để trả lời các câu hỏi của sinh viên.

Phạm vi nghiên cứu: Tập trung nghiên cứu các câu hỏi mà sinh viên đặt ra liên quan đến chương trình đào tạo, tuyển sinh, thông tin về giảng viên và các văn bản thường gặp của trường Đại học Thủ Dầu Một để xây dựng hệ thống trả lời các câu hỏi theo hướng tự động.

1.4. Phương pháp nghiên cứu

Vận dụng các lý thuyết đã học, các bài báo khoa học và các nghiên cứu trước đây của các tác giả, cùng với sự hướng dẫn của thầy để tiến hành phân tích thu thập dữ liệu, đề xuất mô hình xây dựng chatbot.

1.5. Bố cục luận văn

Luận văn được chia thành 5 chương với các nội dung như sau:

- ✓ Chương 1 – Tổng quan về lĩnh vực nghiên cứu

  Sơ lược tổng quan về vấn đề nghiên cứu trên phương diện tổng quan nhất, nêu ra mục tiêu, phương pháp nghiên cứu và bố cục luận văn.

- ✓ Chương 2 – Cơ sở lý thuyết và các nghiên cứu liên quan

  Giới thiệu tổng quan về xử lí ngôn ngữ tự nhiên, về Word2Vector; giới thiệu về mạng nơ-ron nhân tạo, các mô hình mạng nơ-ron cải tiến là cơ sở của mạng học sâu. Nghiên cứu các mô hình phát sinh văn bản trong hệ thống đối

thoại, giới thiệu về mô hình đối thoại seq2seq, kĩ thuật attention và các vấn đề chung có thể gặp phải khi xây dựng mô hình đối thoại; Trình bày cơ bản về hệ thống trả lời tự động, cùng với tình hình nghiên cứu trong nước và ngoài nước

- ✓ Chương 3 – Mô hình đề xuất
- ✓ Chương 4 – Thực nghiệm

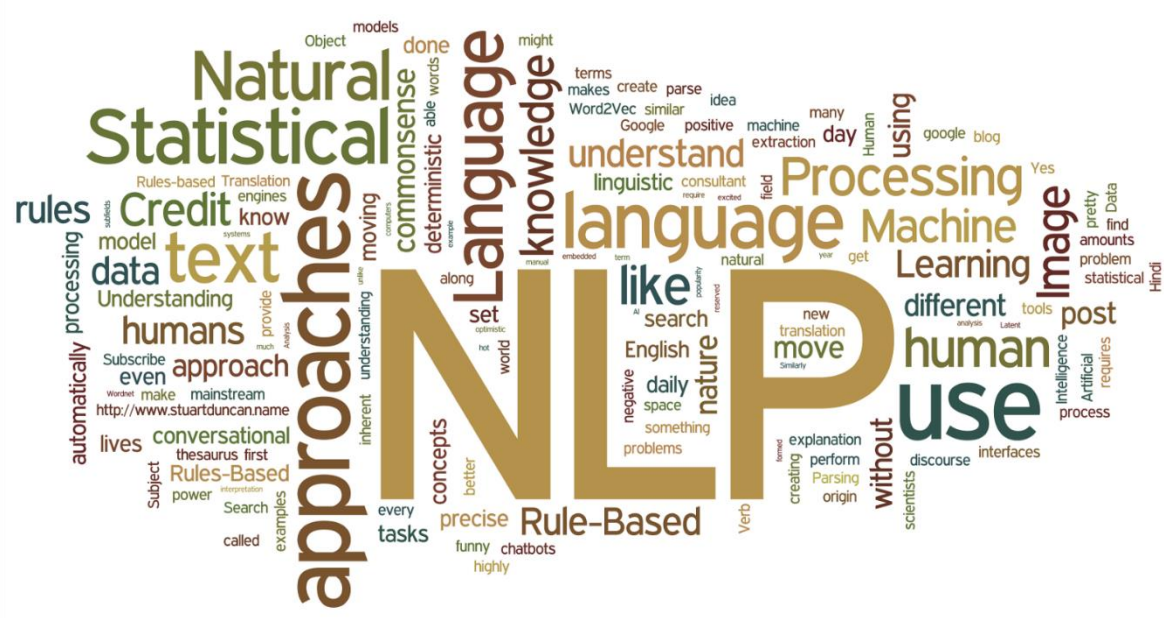
Giới thiệu bộ dữ liệu, quá trình xử lí dữ liệu, phần thực nghiệm và đánh giá thực nghiệm theo hai hướng dựa trên bộ dữ liệu có sẵn và bộ dữ liệu thu thập được.

- ✓ Chương 5 – Kết luận và hướng phát triển

CHƯƠNG 2
CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1. Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (gọi tắt NLP - Natural Language Processing) là các kĩ thuật, phương pháp nhằm hỗ trợ cho các hệ thống máy tính hiểu và xử lý, nhận dạng ngôn ngữ tự nhiên như tiếng Việt, tiếng Anh. Dịch máy, rút trích thông tin và truy hồi thông tin là một trong những nghiên cứu của xử lý ngôn ngữ tự nhiên.



Hình 2.1: Mô hình xử lý ngôn ngữ tự nhiên

Dịch máy tự động là một trong những nghiên cứu đã được phát triển nhiều năm qua và đã đạt được những kết quả tốt trong thời gian gần đây nhờ áp dụng các phương pháp học sâu thông minh. Có nhiều phương pháp tiếp cận giải quyết bài toán này như: dịch máy trên cơ sở luật, dịch máy thống kê và dịch máy trên cơ sở ví dụ,…..

+ Phương pháp dịch máy trên cơ sở luật được xây dựng dựa trên hệ thống luật cú pháp, ngữ nghĩa và phải có một từ điển khá đầy đủ thông tin cho các mục từ như ngữ nghĩa, ngữ dụng,….

+ Dịch máy bằng phương pháp thống kê (SMT - Statistical Machine Translation – SMT) được xây dựng dựa trên các kết quả thống kê từ kho ngữ liệu song ngữ. Kết quả trung gian của phương pháp dịch máy này là các bảng thống kê về các từ, ngữ và các qui luật chuyển đổi mà không cần đến tri thức ngôn ngữ. Với

phương pháp này, ngữ liệu càng lớn và có chất lượng tốt thì hệ dịch sẽ càng hiệu quả. Phương pháp dịch thống kê hiện tại đang cải thiện được chất lượng dịch bằng các mô hình huấn luyện không chỉ dựa trên cơ sở các từ đơn mà còn dựa trên các cụm từ.

+ Phương pháp dịch máy trên cơ sở ví dụ truyền thống sử dụng các câu mẫu hay còn gọi là câu ví dụ. Các câu này được lưu trữ trên cơ sở dữ liệu với đầy đủ các thông tin như câu chú giải, các liên kết giữa các thành phần của hai câu thuộc hai ngôn ngữ. Phương pháp này cũng cần tập luật cú pháp của các câu ngôn ngữ nguồn để xây dựng cơ sở dữ liệu cho mẫu câu ví dụ. Sự khác biệt từ sẽ được xác định thông qua từ điển phân lớp, câu nhập sẽ được phân tích bằng tập luật cú pháp và xác định cặp câu cú pháp của câu nguồn và câu đích. Một tiếp cận khác với phương pháp dịch máy trên cơ sở ví dụ là xây dựng ngân hàng mẫu câu ví dụ. Câu nguồn chỉ cần so trùng từng phần với mẫu câu ví dụ bằng các giải thuật phù hợp (có sử dụng từ đồng nghĩa trong từ điển phân lớp).

+ Dịch máy dựa trên ngữ liệu hiện nay cũng đang được áp dụng vào nhiều hệ thống dịch tự động, việc lấy đúng được cặp ánh xạ đích và nguồn một cách tự động là một yêu cầu thiết yếu cho các phương pháp dịch dựa trên ngữ liệu.

Rút trích thông tin (IE - Information extraction) là một nhánh nghiên cứu khác thiên về rút trích thông tin ngữ nghĩa có cấu trúc một cách tự động từ các nguồn dữ liệu không có cấu trúc hay bán cấu trúc (unstructured/semi-structure) ví dụ như các tài liệu văn bản hay các trang web. Có nhiều hướng tiếp cận cơ bản trong việc rút trích thông tin như sau:

+ Hướng tiếp cận dựa trên Rule-based: sử dụng các pattern khớp nối các thông tin trong văn bản, trong một vài lĩnh vực cụ thể thì cách tiếp cận này cho hiệu quả tương đối cao nhưng cần phải mất nhiều thời gian và quan trọng là phải có kiến thức nghiệp vụ, chuyên gia mới xây dựng được.

+ Hướng tiếp cận dựa trên máy học thống kê (statistical machine learning): sử dụng phương pháp tách nhỏ các bài toán thành các bài toán nhỏ hơn để xử lý.

+ Hướng tiếp cận đang sử dụng hiện nay đó là việc cố gắng rút trích tất cả các quan hệ thực thể được cho là hữu ích đã được thu thập. Khi đó đầu ra của hệ thống sẽ bao gồm tên của quan hệ và mô tả chi tiết của quan hệ thực thể đó.

Truy hồi thông tin (Information Retrieval - IR) là cách tổ chức trình bày, lưu trữ và truy cập các mục thông tin. Truy hồi thông tin là hoạt động thu thập tài nguyên

hệ thống thông tin có liên quan đến nhu cầu thông tin từ tập hợp các nguồn thông tin tin cậy. Các tìm kiếm có thể dựa trên tìm kiếm toàn văn bản hoặc các chỉ mục. Truy hồi thông tin là nhánh nghiên cứu nhằm tìm kiếm thông tin trong các tài liệu, siêu dữ liệu mô tả dữ liệu và cơ sở dữ liệu văn bản, hình ảnh hoặc âm thanh. Một tính năng khác của truy xuất thông tin là nó không thực sự nạp tài liệu, mà có thể chỉ thông báo cho người dùng về sự tồn tại và nơi lưu trữ các tài liệu liên quan đến câu truy vấn. Có hai hướng tiếp cận cơ bản trong truy hồi thông tin:

+ Hướng tiếp cận dựa trên chỉ mục theo cặp từ: Trong hướng tiếp cận này, xem mỗi cặp từ liên tiếp nhau trong nhóm tài liệu, văn bản là một cặp từ. Khi đó, mỗi cặp từ được xem là một chỉ mục. Hướng tiếp cận này không phải là một giải pháp chuẩn, tuy nhiên hướng tiếp cận này có thể kết hợp với các hướng tiếp cận khác.

+ Hướng tiếp cận dựa trên chỉ mục theo vị trí:

- Ứng với mỗi từ chỉ mục, lưu lại vị trí mà nó lưu trữ theo cách thức sau:

  <từ chỉ mục: số tài liệu chứa từ chỉ mục;

  văn bản 1: vị trí 1, vị trí 2, ……;

  văn bản 2: vị trí 1’, vị trí 2’, ……;

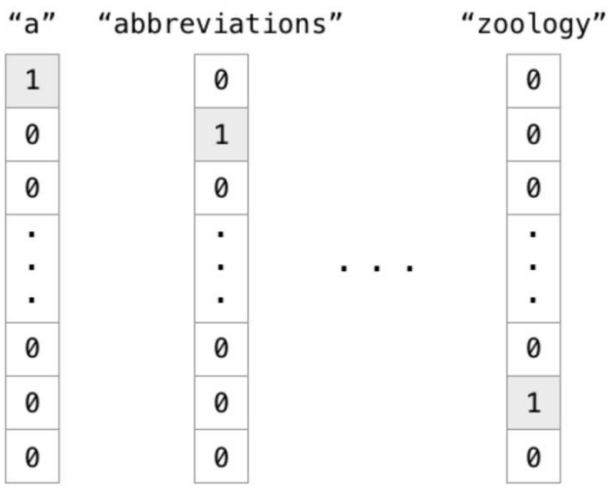
  ……>

- Quá trình xử lý truy vấn được thực hiện bằng cách trộn tất cả các danh sách <Văn bản: vị trí> để liệt kê tất cả các vị trí chứa nhóm từ.

Các ứng dụng cơ bản của NLP: Chế tạo các hệ thống Máy dịch (Google translation, xử lý văn bản và ngôn ngữ, tìm kiếm thông tin, chiết suất thông tin, tóm tắt văn bản, phân loại văn bản, data mining, web mining.

2.2. Word2vector

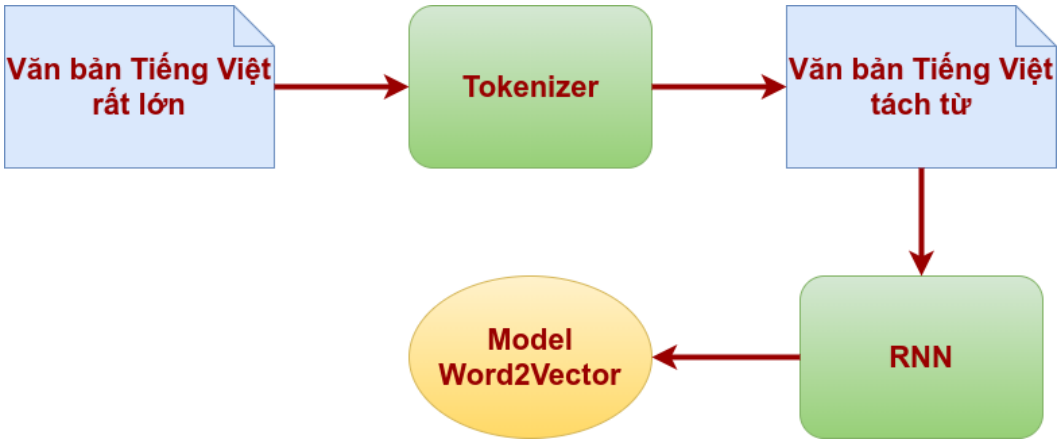
Thông thường, cách truyền thống để biểu diễn một từ là dùng one-hot vector, khi đó độ lớn véc-tơ sẽ đúng bằng số lượng từ vựng có trong văn bản.



Hình 2.2: Biểu diễn véc-tơ one-hot

Vấn đề ở đây là làm thế nào để thể hiện mối quan hệ giữa các từ và tính tương đồng giữa chúng trong văn bản. Do đó, Word2Vector là giải pháp để giải quyết vấn đề này.

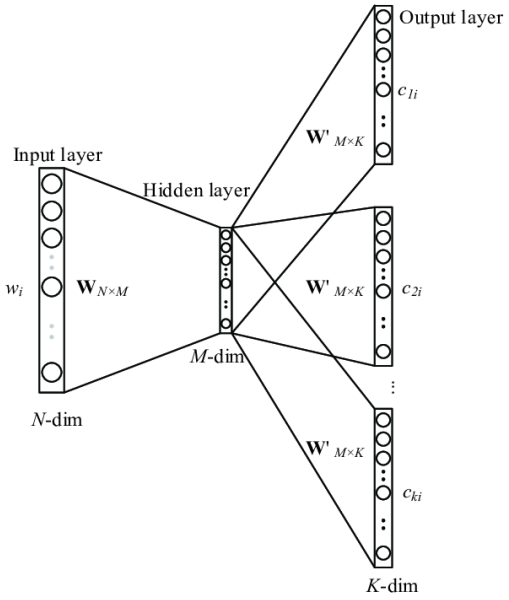
Word2Vector là cách chúng ta biểu diễn 1 từ trong từ điển thành một vector trọng số, có số chiều cụ thể. Word2Vector được giới thiệu bởi một nhóm các nhà nghiên cứu tại Google vào năm 2013. Word2Vector sử dụng các kỹ thuật dựa trên mạng thần kinh và học tập sâu để chuyển đổi các từ thành các vectơ tương ứng theo về mặt ngữ nghĩa gần nhau trong không gian N chiều.



Hình 2.3: Mô hình Word2vector

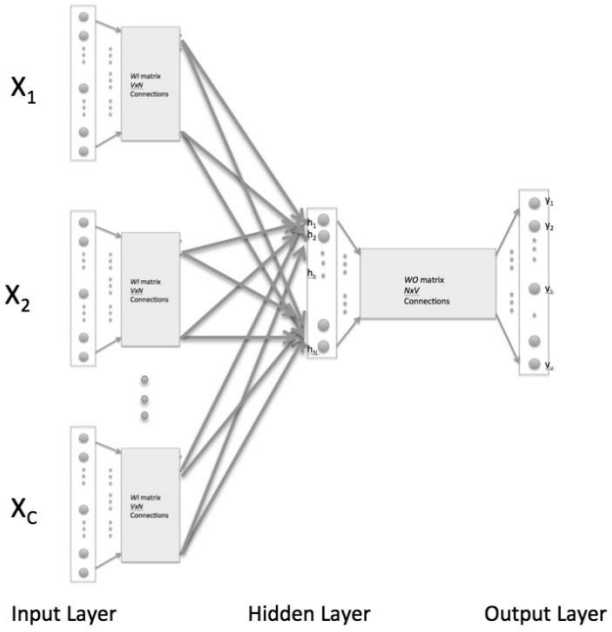
Hai trong số mô hình word2vector được áp dụng để biểu diễn các từ là Skip-gram và Continuous Bag of Words (CBOW):

+ Mô hình Skip-gram[16]: Đầu vào (input) là từ cần tìm mối quan hệ, đầu ra (output) là các từ có quan hệ gần nhất với từ được đưa ở đầu vào.



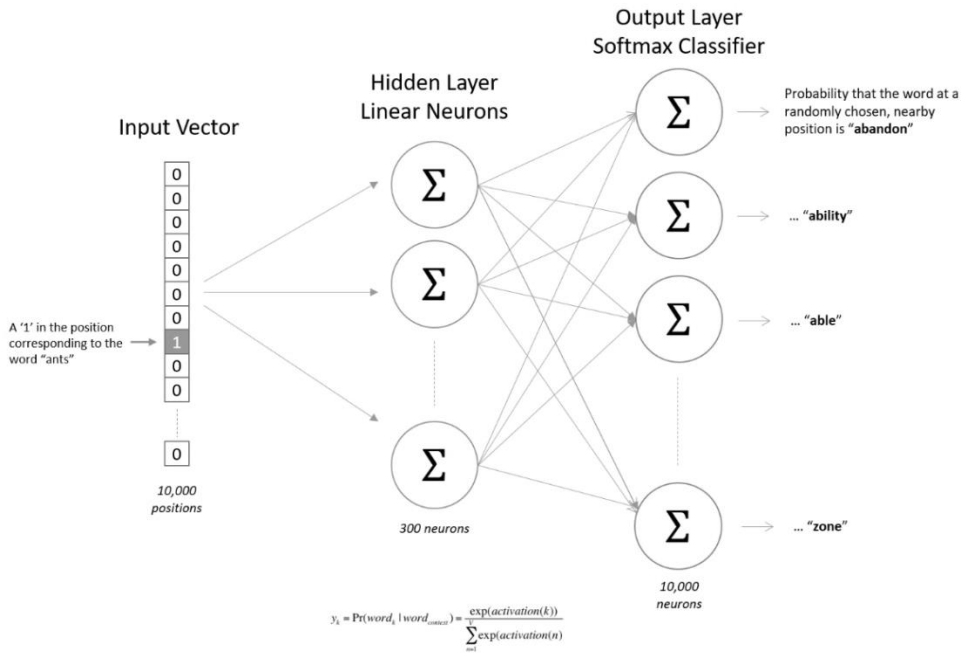
Hình 2.4: Mô hình Skip-gram

+ Mô hình Continuous Bag of Words (CBOW)[15]: Mô hình này ngược lại với mô hình Skip-gram, có nghĩa đầu ra (output) là từ gần với nội dung đầu vào



Hình 2.5: Mô hình Continuous Bag of Words

Mô hình huấn luyện dựa trên word2vector được thực hiện với 2 cách tính toán cơ bản là: sigmoid probability hoặc dự đoán probability của từ.



Hình 2.6: Mô hình huấn luyện dựa trên word2vector

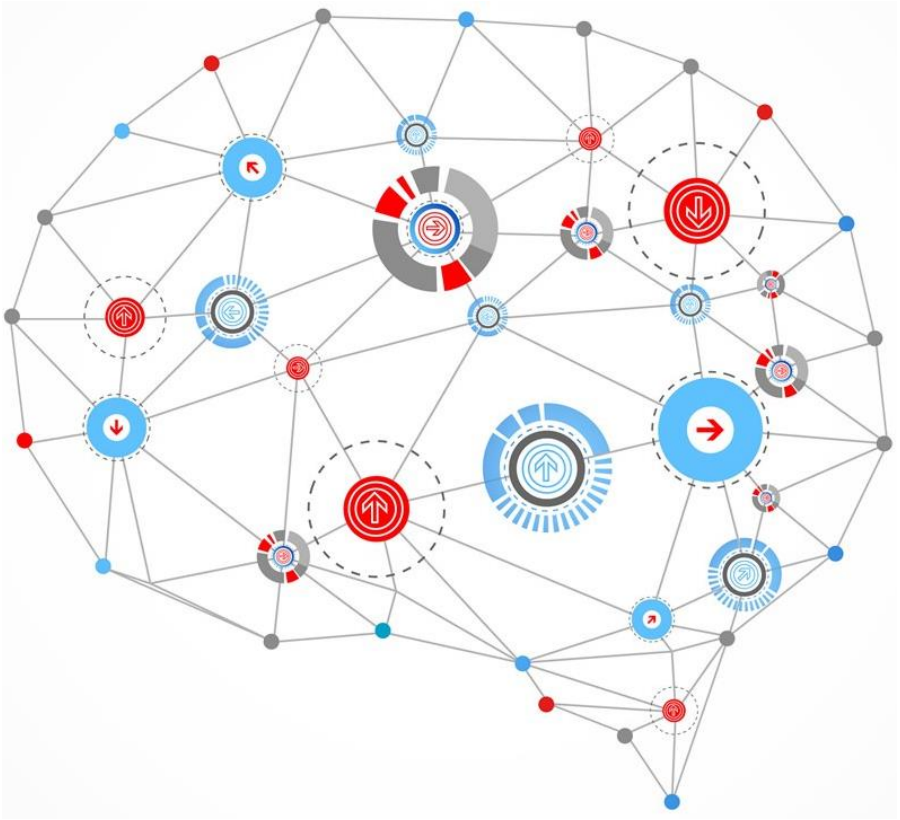
2.3. Học sâu - Deep Learning

Học máy (Machine Learning) là một lĩnh vực của trí tuệ nhân tạo (Artificial Intelligence - AI). Các thuật toán học máy cho phép máy tính đào tạo đầu vào dữ liệu và sử dụng phân tích thống kê để đưa ra các giá trị nằm trong một phạm vi cụ thể.

Ngày nay, những người sử dụng công nghệ đều được hưởng lợi từ việc học máy. Công nghệ nhận diện khuôn mặt giúp người dùng gắn thẻ và chia sẻ ảnh của bạn bè. Công nghệ nhận dạng ký tự quang học (OCR) chuyển đổi hình ảnh văn bản sang dạng di chuyển.

Khi mà khả năng tính toán của máy tính được nâng lên một tầm cao mới cùng với lượng dữ liệu khổng lồ được thu thập, Machine Learning đã tiến thêm một bước dài và Deep Learning (DL) một lĩnh vực mới được ra đời.

Deep Learning được lấy cảm hứng từ mạng nơ-ron sinh học và bao gồm nhiều lớp trong mạng nơ-ron nhân tạo được tạo thành từ phần cứng và GPU. Deep Learning sử dụng một tầng các lớp đơn vị xử lý phi tuyến để trích xuất hoặc chuyển đổi các tính năng (hoặc biểu diễn) của dữ liệu. Đầu ra của một lớp phục vụ như là đầu vào của lớp kế tiếp. Deep learning tập trung giải quyết các vấn đề liên quan đến mạng thần kinh nhân tạo nhằm nâng cấp các công nghệ như nhận diện giọng nói, dịch tự động (machine translation), xử lý ngôn ngữ tự nhiên…



Hình 2.7: Mô hình Deep Learning[1]

Trong số các thuật toán học máy hiện đang được sử dụng và phát triển, học sâu thu hút được nhiều dữ liệu nhất và có thể đánh bại con người trong một số nhiệm vụ nhận thức. Do những thuộc tính này, học tập sâu đã trở thành phương pháp tiếp cận có tiềm năng đáng kể trong lĩnh vực trí tuệ nhân tạo.

2.3.1. Mạng nơ-ron hồi quy RNN (Recurrent Neural Network)

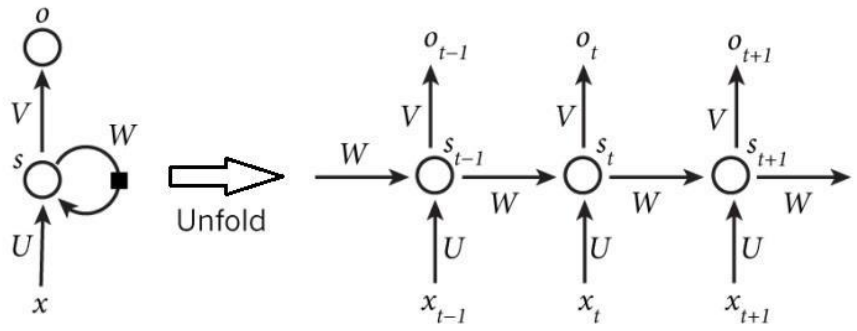
Mạng nơ-ron hồi quy RNN (Recurrent Neural Network) được giới thiệu bởi John Hopfield năm 1982 [2], là một trong những mô hình học sâu - Deep learning**.** Recurrent có nghĩa là thực hiện lặp lại cùng một tác vụ cho mỗi thành phần trong chuỗi. Trong đó, kết quả đầu ra tại thời điểm hiện tại phụ thuộc vào kết quả tính toán của các thành phần ở những thời điểm trước đó.

RNN là một mô hình có trí nhớ **(**memory**)**, có khả năng nhớ được thông tin đã tính toán trước đó. Không như các mô hình Neural Network truyền thống trước đó là thông tin đầu vào **(**input**)** hoàn toàn độc lập với thông tin đầu ra **(**output**).**

Hầu hết RNN được thiết kế như là một chuỗi các module được lặp đi lặp lại, các môdun này thường có cấu trúc đơn giản chỉ có một lớp mạng *tanh*. Huấn luyện

[1] *Nguồn từ learning-systems.org*

RNN tương tự như huấn luyện ANN truyền thống. Giá trị tại mỗi output không chỉ phụ thuộc vào kết quả tính toán của bước hiện tại mà còn phụ thuộc vào kết quả tính toán của các bước trước đó.



Hình 2.8 Quá trình xử lý thông tin trong mạng RNN

RNN có khả năng biểu diễn mối quan hệ phụ thuộc giữa các thành phần trong chuỗi (nếu chuỗi đầu vào có 6 từ thì RNN sẽ dàn ra thành 6 layer, mỗi layer ứng với mỗi từ, chỉ số mỗi từ được đánh từ 0 đến 5. Trong hình 2.8 ở trên, x_t là input tại thời điểm thứ t, s_t là hidden state (trạng thái ẩn) tại thời điểm thứ t, được tính dựa trên các hidden state trước đó kết hợp với input của thời điểm hiện tại với công thức:

$$s_t = \tanh(U_{x_t} + W_{s_{t-1}})$$

s_{t-1} là hidden state được khởi tạo là một vector 0.

Hàm tanh là hàm phi tuyến tính có dạng hàm tang hyperbolic được xác định bởi công thức:

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

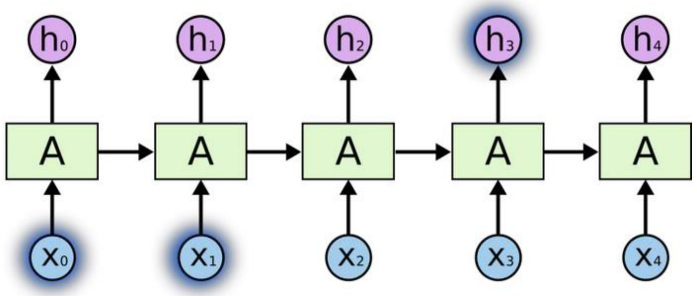
O_t là output tại thời điểm thứ t, là một vector chứa xác xuất của toàn bộ các từ trong từ điển. softmax gọi là hàm trung bình mũ nhằm khái quát hóa của hàm lôgic biến không gian K-chiều véctơ với các giá trị thực đến bất kỳ không gian K-chiều véctơ mang giá trị trong phạm vi (0,1].

$$O_t = \text{softmax}(V_{s_t})$$

Không như ANN truyền thống, tại mỗi layer cần phải sử dụng một tham số khác, RNNs chỉ sử dụng một bộ parameters (U,V,W) cho toàn bộ các bước.

Ý tưởng ban đầu của RNN là kết nối những thông tin trước đó nhằm hỗ trợ cho các xử lý hiện tại. Nhưng đôi khi, chỉ cần dựa vào một số thông tin gần nhất để thực hiện tác vụ hiện tại. Ví dụ, chúng ta dự đoán từ cuối cùng trong câu "chuồn_chuồn bay thấp thì mưa", thì chúng ta không cần truy tìm quá nhiều từ trước

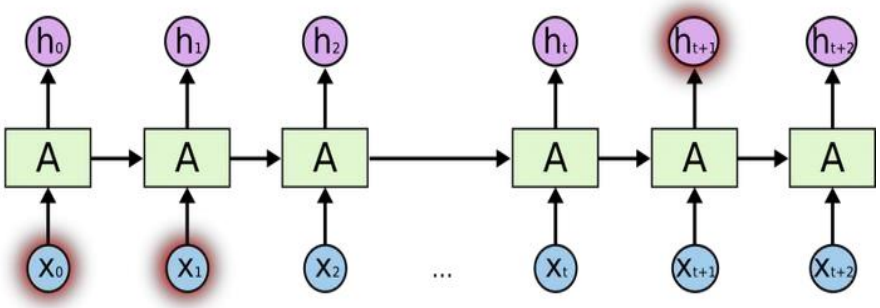
đó, ta có thể đoán ngay từ tiếp theo sẽ là "mưa". Trong trường hợp này, khoảng cách tới thông tin liên quan được rút ngắn lại, mạng RNN có thể học và sử dụng các thông tin quá khứ.



Hình 2.9: RNN phụ thuộc short-term

Trường hợp có nhiều thông tin hơn trong một câu, nghĩa là phụ thuộc vào ngữ cảnh. Ví dụ nhưng khi dự đoán từ cuối cùng trong đoạn văn bản **"Tôi sinh ra và lớn lên ở Việt_Nam … Tôi có_thể nói thuần_thục Tiếng_Việt."** Từ thông tin gần nhất cho thấy rằng từ tiếp theo là tên một ngôn ngữ, nhưng khi chúng ta muốn biết cụ thể ngôn ngữ nào, thì cần quay về quá khứ xa hơn, để tìm được ngữ cảnh **Việt_Nam**. Và như vậy, RNN có thể phải tìm những thông tin có liên quan và số lượng các điểm đó trở nên rất lớn.

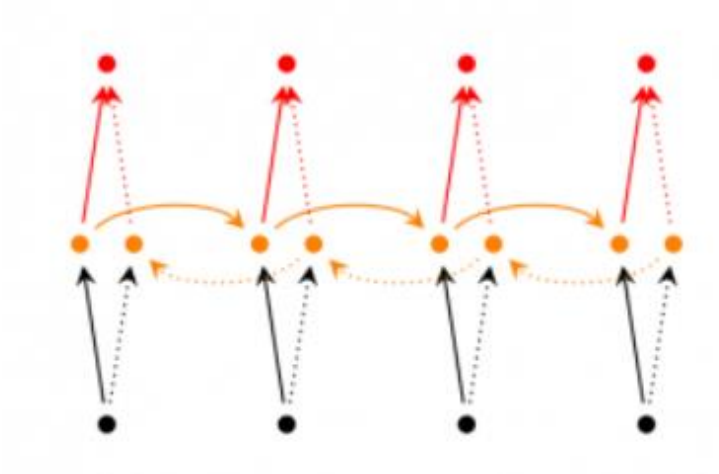
Không được như mong đợi, RNN không thể học để kết nối các thông tin lại với nhau.



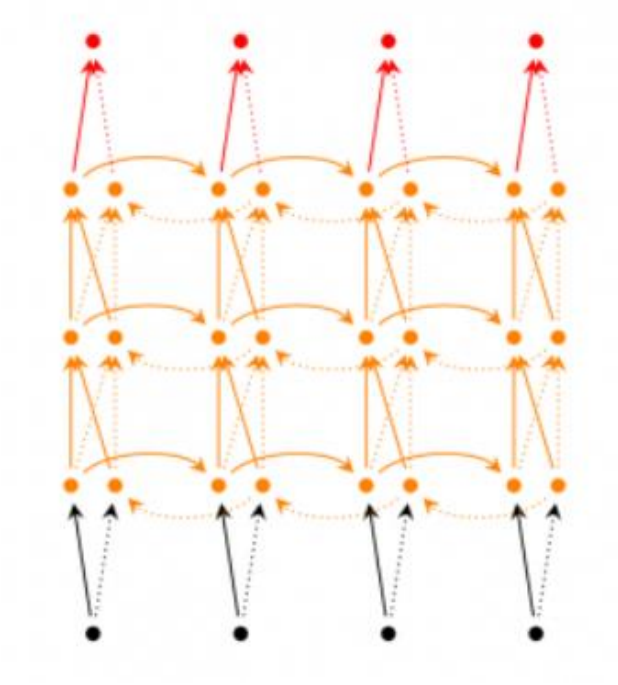
Hình 2.10: RNN phụ thuộc long-term

Về lý thuyết, RNN có thể nhớ được thông tin của chuỗi có chiều dài bất kì, nhưng trong thực tế mô hình này chỉ nhớ được thông tin ở vài bước trước đó[13].

RNN có các phiên bản mở rộng như: Bidirectional RNN **(**RNN hai chiều**),** Deep (Bidirectional) RNN**,** Long short-term memory networks (LSTM) [4] [10] [11]**.**



Hình 2.11 Bidirectional RNN



Hình 2.12. Deep (Bidirectional) RNN

2.3.2. Bộ nhớ dài-ngắn LSTM (Long-short term memory)

Là một dạng đặc biệt của mạng Noron hồi quy, một kĩ thuật dựa trên Gradient. Trong quá trình hoạt động nó cho phép cắt bỏ những Gradient dư thừa. Trong quá trình học LSTM có thể thu hẹp thời gian trễ dư thừa của các bước thực hiện thông qua tập hằng số lỗi (theo Hochreiter & Schmidhuber- 1997[4]).

LSTM có các thành phần cơ bản sau:

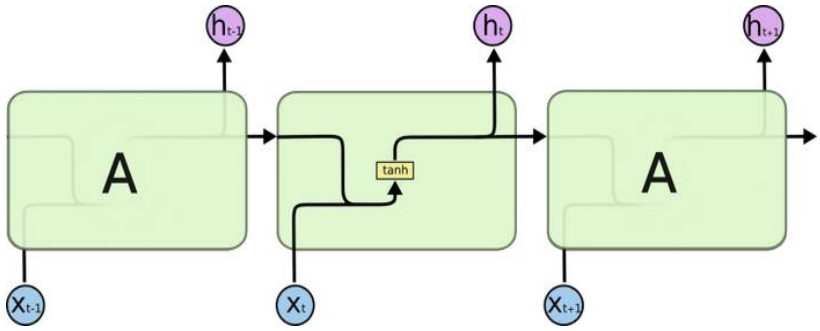
+ Tế bào trạng thái (cell state)

+ Cổng (gates)

+ Sigmoid

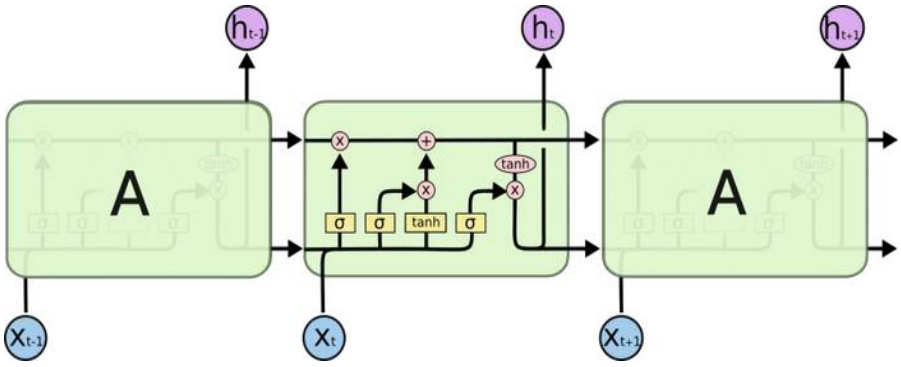
+ Tanh

LSTM được thiết kế nhằm loại bỏ vấn đề phụ thuộc quá dài. Ta quan sát lại mô hình RNN bên dưới, các layer đều mắc nối với nhau thành các module neural network. Trong RNN chuẩn, module repeating này có cấu trúc rất đơn giản chỉ gồm một lớp đơn giản **tanh layer**.



Hình 2.13: Các mô-đun lặp của mạng RNN chứa một layer

Về kiến trúc mạng LSTM**:** giống như RNN, nó là một chuỗi các module được lặp đi lặp lại. Tuy nhiên, xét về cấu trúc thì LSTM có 4 tầng mạng noron tương tác với nhau, gọi đó là các tầng ẩn (hiden layer). Một số biến thể của LSTM được thực hiện dựa trên việc thay đổi vị trí kết nối giữa các tầng và cổng.



Hình 2.14:Các mô-đun lặp của mạng LSTM chứa bốn layer

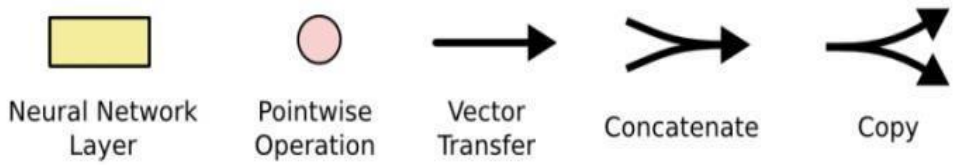
Trong đó, các ký hiệu sử dụng trong mạng LSTM được giải nghĩa như hình 2.15 sau đây:

• Hình chữ nhật là các lớp ẩn của mạng nơ-ron

• Hình tròn biểu diễn toán tử *Pointwise*

• Đường kẻ gộp lại với nhau biểu thị phép nối các toán hạng

• Và đường rẽ nhánh biểu thị cho sự sao chép từ vị trí này sang vị trí khác



Hình 2.15: Các kí hiệu sử dụng trong mạng *LSTM*

Cho một chuỗi các vectơ (x1, x2, …,xn), σ là hàm sigmoid logistic, trạng thái ẩn h_t, của LSTM tại thời điểm t được tính như sau:

$$h_t = o_t * tanh\ (c_t) \quad (1)$$

$$o_t = tanh\ (Wx_0x_t + Wh_0h_{t-1} + Wc_0c_t + b_o) \quad (2)$$

$$c_t = f_t * c_{t-1} + i_t * tanh(Wx_cx_t + Wh_ch_{t-1} + b_c) \quad (3)$$

$$f_t = \sigma\ (Wx_fx_t + Wh_fh_{t-1} + Wc_fc_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma\ (Wx_ix_t + Wh_ih_{t-1} + Wc_ic_{t-1} + b_i) \quad (5)$$

Trong đó:

+ tanh là tang hyperbolic

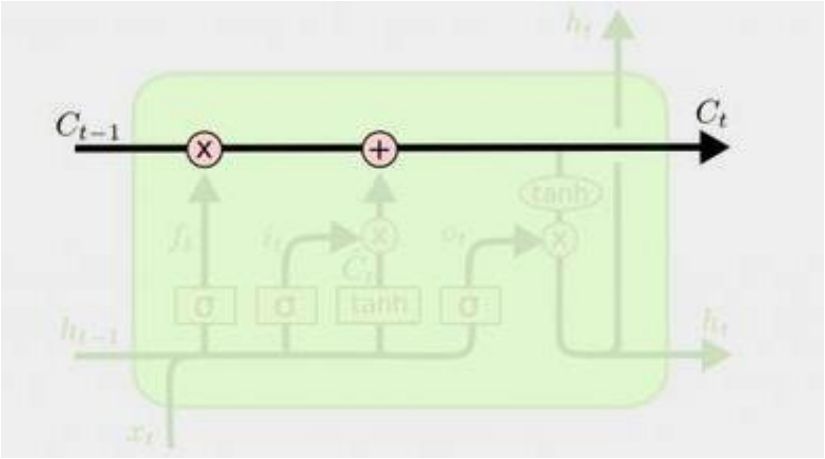
+ σ là hàm signmoid hay còn gọi là hàm lôgit chuẩn là nghiệm của phương trình sai phân phi tuyến bậc 1:

$$\frac{dP}{dt} = P(1-P)\ trong\ \text{đó}\ P(t) = \frac{1}{1+\ e^t}$$

Phân tích mô hình *LSTM*:

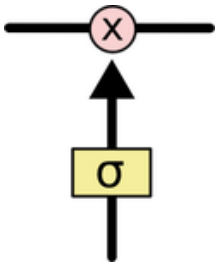
+ LSTM được thiết kế là một bảng mạch số, gồm các mạch logic và các phép toán logic. Thông tin di chuyển trong mạch sẽ được lưu trữ, lan truyền theo cách mà chúng ta thiết kế mạch.

+ Mấu chốt của *LSTM* là *cell state* (tế bào trạng thái), là đường kẻ ngang chạy dọc ở trên *top diagram*. *Nó* giống như băng chuyền, chạy xuyên thẳng toàn bộ mắc xích, chỉ một vài tương tác nhỏ tuyến tính (*minor linear interaction*) được thực hiện, giúp thông tin trong quá trình lan truyền ít bị thay đổi.



Hình 2.16: Tế bào trạng thái LSTM giống như một băng truyền

+ Trong LSTM cấu trúc cổng (*gate*) có khả năng thêm hoặc bớt thông tin vào *cell state*. Các cổng này được tạo bởi hàm *sigmoid* và một toán tử nhân(*pointwise*)

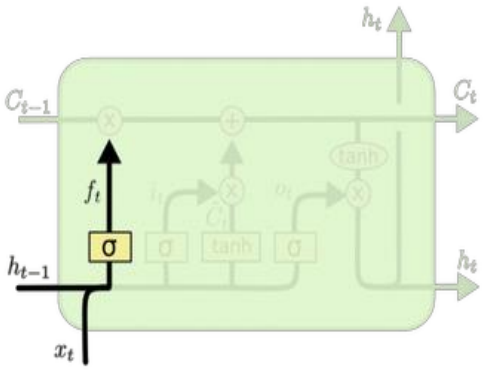


Hình 2.17: Cổng trạng thái LSTM

+ Hàm kích hoạt *Sigmoid* mô tả độ lớn thông tin được phép truyền qua tại mỗi lớp mạng, có giá trị từ 0 – 1. Thu giá trị 0 có nghĩa là "không cho bất kỳ cái gì đi qua", ngược lại nếu thu được giá trị là 1 có nghĩa là "cho phép mọi thứ đi qua". Một *LSTM* có ba cổng như vậy để bảo vệ và điều khiển *cell state*.

Quá trình hoạt động của LSTM được thông qua các bước cơ bản sau:

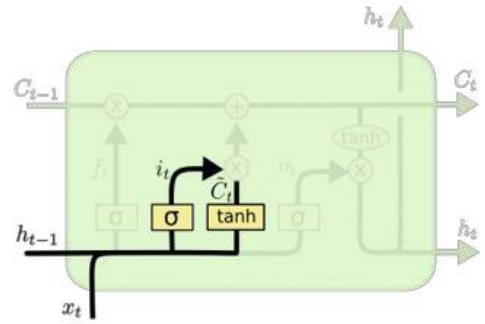
+ Bước đầu tiên của mô hình *LSTM* là quyết định xem thông tin nào chúng ta cần loại bỏ khỏi *cell state*. Một *sigmoid layer* gọi là "*forget gate layer*" – cổng chặn sẽ thực hiện tiến trình này. Đầu vào là h_{t-1} *và* x_t , đầu ra là một giá trị nằm trong khoảng [0, 1] cho *cell state* C_{t-1} (1 tương đương với "giữ lại thông tin", 0 tương đương với "loại bỏ thông tin").



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Hình 2.18: LSTM focus f

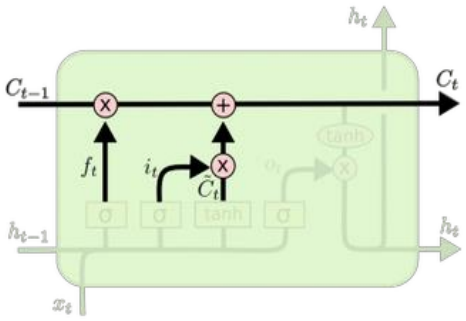
+ Bước tiếp theo, quyết định thông tin nào cần được lưu lại tại *cell state*, được thực hiện bởi *single sigmoid layer* được gọi là "*input gate layer*" quyết định các giá trị chúng ta sẽ cập nhật, và một *tanh layer* tạo ra một *vector* ứng viên mới, được thêm $\tilde{C}_t$ vào trong ô trạng thái.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Hình 2.19: LSTM focus i

+ Kế tiếp, kết hợp hai thành phần này lại để cập nhật vào cell state. Lúc cập nhật vào cell state cũ C_{t-1} vào cell state mới C_t. Ta sẽ đưa *state* cũ vào hàm f_t , để quên đi những gì trước đó. Sau đó, ta sẽ thêm i_t*C_t. Đây là giá trị ứng viên mới, co giãn (*scale*) số lượng giá trị mà ta muốn cập nhật cho mỗi *state*.



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Hình 2.20: LSTM focus c

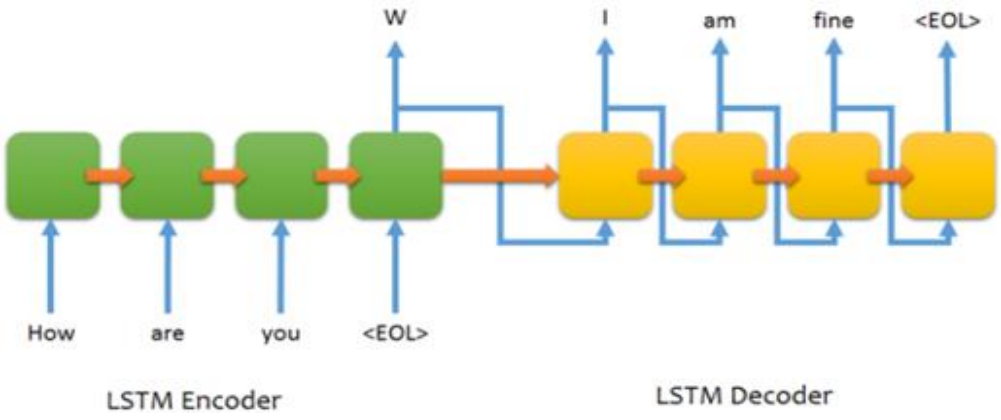
+ Cuối cùng, ta cần quyết định xem thông tin *output* là gì. *Output* này cần dựa trên *cell state* của chúng ta, nhưng sẽ được lọc bớt thông tin. Đầu tiên, ta sẽ áp dụng *single sigmoid layer* để quyết định xem phần nào của *cell state* chúng ta dự định sẽ *output*. Sau đó, ta sẽ đẩy *cell state* qua *tanh* (đẩy giá trị vào khoảng - 1 và 1) và nhân với một *output sigmoid gate*, để giữ lại những phần ta muốn *output* ra ngoài [13].

Mô hình LSTM là một bước đột phá đạt được từ mô hình RNN. Mô hình này giải quyết triệt để vấn đề không xử lý được câu hỏi dài mà những mô hình chatbot Skype đang gặp phải.

2.3.3. Mô hình sequence to sequence

Mô hình chuỗi sang chuỗi sequence to sequence (seq2seq)[1], được giới thiệu trong bài báo "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". Mô hình bao gồm hai mạng *nơ-ron***,** một đóng vai trò là mạng mã hóa (encoder) câu đầu vào thành một vector có độ dài cố định được gọi là vectơ ngữ cảnh và một mạng nơ-ron đóng vai trò giải mã (decoder) để sinh ra chuỗi đầu ra. Mô hình được áp dụng cho các bài toán khi cần sinh ra chuỗi đầu ra từ câu đầu vào cho trước. Mô hình còn có tên gọi khác là Encoder–Decoder framework. Mô hình này được trình bày trong Hình 2.21

Mạng *nơ-ron encoder* mã hóa chuỗi đầu vào thành một vector *c* có độ dài cố định. Mạng *nơ-ron decoder* sẽ lần lượt sinh từng từ trong chuỗi đầu ra dựa trên vector *c* và những từ được dự đoán trước đó cho tới khi gặp từ kết thúc câu. Chúng ta có thể sử dụng các kiến trúc mạng khác nhau cho thành phần encoder decoder trong mô hình sequence-to-sequence này.



Hình 2.21: Mô hình sequence-to-sequence sử dụng 2 mạng nơ-ron LSTM [14]

Thành phần RNN encoder sinh ra vector c với độ dài cố định từ một chuỗi các vector đầu vào: $c=RNN^{enc}(x_{1:n})$. Sau đó, thành phần RNN decoder sẽ sinh lần lượt từng từ cho chuỗi đầu ra theo mô hình sinh có điều kiện (conditional generation).

$$p(t_{j+1} = k|\hat{t}_{1:j}, c) = f(0(s_{j+1}))$$

$$s_{j+1} = R(s_j, [\hat{t}_j; c])$$

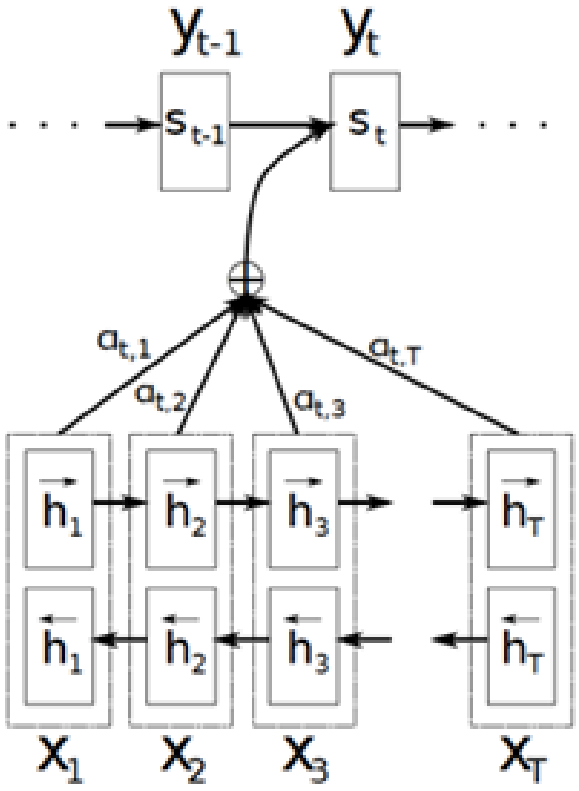
$$\hat{t}_j \sim p(t_j|\hat{t}_{1:j-1}, c)$$

ở đây t_j là bước thời gian thứ j trong chuỗi output, $\hat{t}_j$ là giá trị dự đoán tại các bước thứ j và s_j là các trạng thái ẩn.

2.3.4. Kỹ thuật attention

Mô hình sequence-to-sequence cơ bản có 2 nhược điểm đó là yêu cầu RNN decoder sử dụng toàn bộ thông tin mã hóa từ chuỗi đầu vào cho dù chuỗi đó dài hay ngắn và RNN encoder cần phải mã hóa chuỗi đầu vào thành một vec-tơ duy nhất, có độ dài cố định. Tuy nhiên, việc sinh ra từ tại một thời điểm trong chuỗi đầu ra còn phụ thuộc nhiều hơn vào một số những thành phần nhất định trong chuỗi đầu vào (chẳng hạn như ngữ cảnh xung quanh từ hiện tại so với các từ khác trong câu). Kỹ thuật attention được đưa ra để giải quyết vấn đề này.

Attention được đưa ra lần đầu vào năm 2014 bởi Bahdanau và cộng sự [5] trong công trình nghiên cứu về dịch máy và kỹ thuật này được sử dụng rộng rãi trong nhiều ứng dụng của xử lý ngôn ngữ tự nhiên [7] [8] [9]. Thay vì giải mã chuỗi đầu vào thành một vectơ bối cảnh cố định, với kỹ thuật attention này, các từ trong chuỗi đầu vào sẽ được RNN encoder mã hóa thành một dãy các vector. Tiếp đến, RNN decoder áp dụng kỹ thuật attention mềm dẻo (soft attention) bằng cách lấy tổng trọng số của dãy các vector mã hóa (các trọng số được tính bằng một mạng nơ-ron truyền thẳng).



Hình 2.22: Mô hình sequence-to-sequence dùng soft attention trong dịch máy[14]

RNN encoder, RNN decoder và cơ chế attention được huấn luyện đồng thời từ dữ liệu. Attention sử dụng các đơn vị GRU thay vì các tế bào bộ nhớ LSTM. Trong trường hợp này, một đầu vào hai chiều được sử dụng trong đó các chuỗi đầu vào được cung cấp cả về phía trước và phía sau, sau đó được ghép nối trước khi được truyền vào bộ giải mã.

Kỹ thuật attention được thực hiện từ việc nhận chuỗi đầu vào $x_{1:n}$ đã được mã hóa bằng biRNN (bidirectional RNN) để sinh ra n vector $c_{1:n}$ có dạng như sau:

$$c_{1:n} = \text{ENC}(x_{1:n}) = \text{biRNN*}(x_{1:n}) \quad (6)$$

Tại mỗi bước j, decoder sẽ chọn những phần nào trong $c_{1:n}$ để sinh ra véctơ ngữ cảnh c^j thông qua kỹ thuật attention để dự đoán bước thứ j (c^j = attend $(c_{1:n}, \hat{t}_{1:j})$)

$$p(t_{j+1} = k|\hat{t}_{1:j}, x_{1:n}) = f(O(s_{j+1})) \quad (7)$$

$$s_{j+1} = R(s_j, [\widehat{t_j}\,;\, c^j]) \quad (8)$$

$$c^j = \text{attend}\,(c_{1:n}, \hat{t}_{1:j}) \quad (9)$$

$$\hat{t}_j \sim p(\,t_j\,|\hat{t}_{1:j-1}, x_{1:n}) \quad (10)$$

Tại mỗi bước thực hiện, vector ngữ cảnh c^j được sinh ra thông qua kỹ thuật attention chính là tổng trọng số của các vector $c_{1:n}$ có dạng như sau:

$$c^j = \sum_{i=1}^{n} \alpha_{[i]}^{j}.\, c_i \quad (11)$$

Các giá trị $\alpha_{[i]}^{j}$ được tính qua mạng nơ-ron MLP (muti layer perception) sẽ được chuẩn hóa bằng hàm softmax.

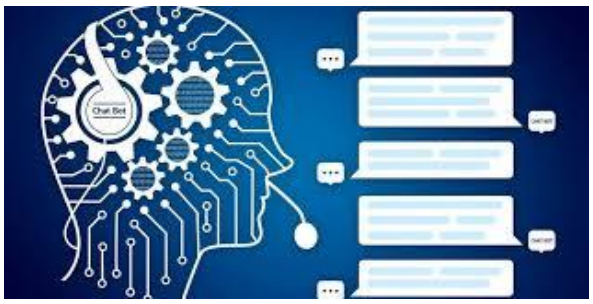
Ngoài phương pháp soft attention, kỹ thuật attention còn có các kiểu attention khác khác như: additive attention multiplicative attention, self-attention, key-value attention [7] [8] [9].

2.4. Hệ thống trả lời tự động Chatbot

2.4.1. Tổng quan

Chatbot là một hệ thống trả lời tự động thông minh, hay là một chương trình mô phỏng cuộc trò chuyện của con người thông qua văn bản hoặc bằng giọng nói với máy. Người dùng có thể yêu cầu một câu hỏi hoặc thực hiện một lệnh và chatbot sẽ trả lời hoặc thực hiện các hành động được yêu cầu. Mức độ chuẩn xác và tự nhiên của câu trả lời phụ thuộc vào khả năng xử lý dữ liệu đầu vào cũng như độ phức tạp của thuật toán lựa chọn đầu ra của hệ thống.

Nhiều nhà nghiên cứu đã sử dụng các kỹ thuật học máy để xây dựng Chatbot có khả năng hỗ trợ con người trò chuyện, nhắc nhở hay làm trợ lý công việc và có thể theo dõi tình trạng sức khỏe cá nhân mọi lúc, mọi nơi. Rất nhiều công ty lớn đã phát triển các trợ lý ảo có thể hiểu được ngôn ngữ tự nhiên của con người và tương tác được với con người một cách tự nhiên hơn, nhằm làm tăng chất lượng và hiệu quả trong việc chăm sóc khách hàng, giúp khách hàng có những trải nghiệm tốt nhất về sản phẩm và các dịch vụ mà họ được cung cấp.



Hình 2.23: Tổng quan Chatbot

2.4.2. Tình hình sử dụng các ứng dụng nhắn tin

Theo thống kê của The Statistics Portal, trong 2017 có khoảng 28,2 tỷ tin nhắn di động đã được gửi với tỷ lệ 98% đã được đọc trong khi e-mail tỷ lệ này chiếm khoảng 22%. Riêng ở Việt Nam, số lượng tin nhắn được luân chuyển khoảng 800 triệu tin nhắn.

Cũng theo trang này, số lượng người dùng hàng tháng của các ứng dụng nhắn tin từ tháng 4/2014 cho đến tháng 9/2017 tăng khá cao, cụ thể:

- WhatsApp: 1,5 tỷ[2]
- Facebook messenger: 1,2 tỷ[3]
- Wechat: 963 triệu[4]
- Kik: 300 triệu[5]
- LINE: 217 triệu[6]
- Zalo: 80 triệu[7]

Cùng với việc bùng nổ cách mạng công nghiệp 4.0, Chatbot đã trở thành cơ hội kinh doanh của các công ty thông qua sự soi sáng của cuộc cách mạng 4.0 này. Việc xử lý và hiểu ngôn ngữ tự nhiên thông qua các kỹ thuật, công nghệ mới trở nên đơn giản và chính xác hơn rất nhiều. Từ đó, các công ty lớn như Google, Yahoo, Facebook,…. đã đều xây dựng các platform của mình để phục vụ cho việc hoàn thiện hệ thống Chatbot.

Với công nghệ Facts, mỗi năm công ty tiêu tốn 1,3 nghìn tỷ USD cho 265 triệu cuộc gọi và khoảng 91% khách hàng khi không được thoả mãn sẽ không sử dụng lại dịch vụ. Tuy nhiên khi ứng dụng Chatbot, các công ty đã giảm hơn 30% chi phí chăm sóc khách hàng[8] như ở IBM với ứng dụng Watson conversation đã giảm 80% các cuộc gọi đơn giản không cần nhân viên tư vấn, các công ty ở Nga vẫn có thể bán từng đó sản phẩm chỉ mất 1-2 tiếng so với việc bán sản phẩm trong 8 tiếng so với trước đây.

Chính vì sự phát triển, bùng nổ của Chatbot mà nhiều nhà nghiên cứu, khoa học đã nghiên cứu để hỗ trợ cho việc xây dựng hệ thống Chatbot ngày càng hoàn thiện hơn.

2.4.3. Các hướng tiếp cận

Có 2 hướng tiếp cận chính trong bài toán xây dựng chatbot là: Hướng tiếp cận dựa trên luật (Rule-based) và Hướng tiếp cận dựa trên dữ liệu (Corpus-based)

[2] https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/
[3] https://www.statista.com/statistics/417295/facebook-messenger-monthly-active-users/
[4] https://www.statista.com/statistics/255778/number-of-active-wechat-messenger-accounts/
[5] https://www.statista.com/statistics/327312/number-of-registered-kik-messenger-users/
[6] https://www.statista.com/statistics/327292/number-of-monthly-active-line-app-users/
[7] https://news.zing.vn/zalo-da-co-80-trieu-nguoi-dung-post770245.html
[8] https://chatbotsmagazine.com/how-with-the-help-of-chatbots-customer-service-costs-could-be-reduced-up-to-30-b9266a369945

❖ Với hướng tiếp cận Rule-based có các nghiên cứu như:

Pattern-action rules (Eliza)

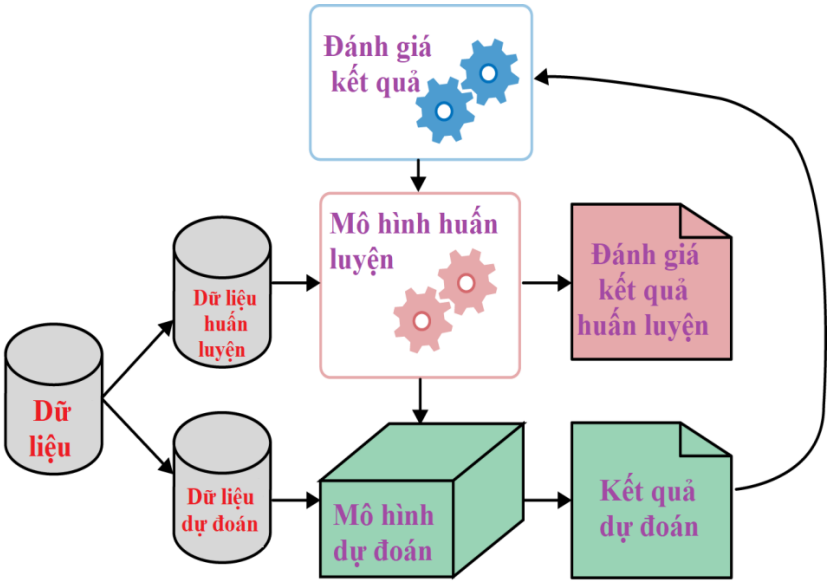
A mental model (Parry)

❖ Hướng tiếp cận Corpus-based có các nghiên cứu như:

Information Retrieval

Deep neural nets

Trong hai hướng tiếp cận này, hướng tiếp cận dựa trên dữ liệu được nghiên cứu, triển khai nhiều trong những năm gần đây và trở thành hướng nghiên cứu chính. Trong luận văn này, tôi nghiên cứu theo hướng tiếp cận thứ 2 dựa trên dữ liệu, sử dụng mạng học sâu LSTM, áp dụng phương pháp học chuỗi liên tiếp (sequence-to-sequence) và cơ chế attention để sinh ra câu trả lời tự động từ một chuỗi đầu vào tương ứng. Mô hình được huấn luyện end-to-end theo hướng GNMT (Google's Neural Machine Translation) và theo hướng phân loại câu hỏi, dựa trên bộ dữ liệu có sẵn CORNELL-MOVIE-DIALOGS và bộ dữ liệu câu hỏi thu thập được. Mô hình nghiên cứu được trình bày trong Hình 2.24.



Hình 2.24: Đề xuất mô hình huấn luyện dữ liệu và dự đoán kết quả

2.4.4. Tình hình nghiên cứu

2.4.4.1.Tình hình nghiên cứu ngoài nước

Tình hình nghiên cứu về lý thuyết

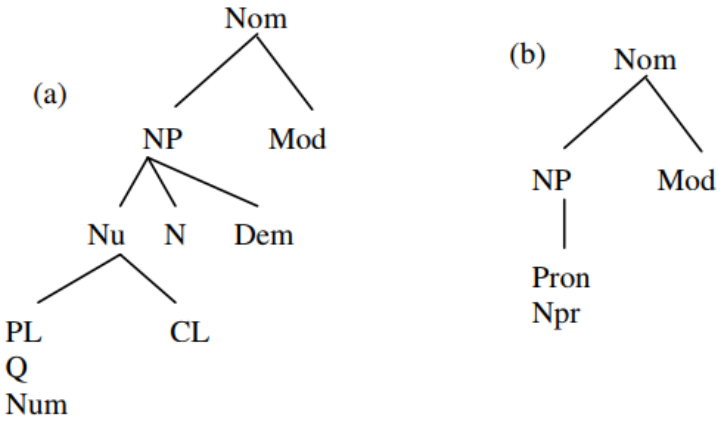
- Nền tảng cơ bản để phát triển chatbot là dựa trên nền tảng dịch máy và truy xuất thông tin. Trên nền tảng dịch máy và truy xuất thông tin, nhiều nghiên cứu đã phát triển hơn 50 năm qua, nhiều cải tiến lần lượt được đề xuất.

- Shum biểu diễn cụm danh từ cả ở dạng luật sinh và dạng cây như sau:

Nom → NP Mod
NP → Nu N Dem
NP → Pron
NP → Npr
Nu → PL CL
Nu → Q CL
Nu → Num CL
N → N' N"

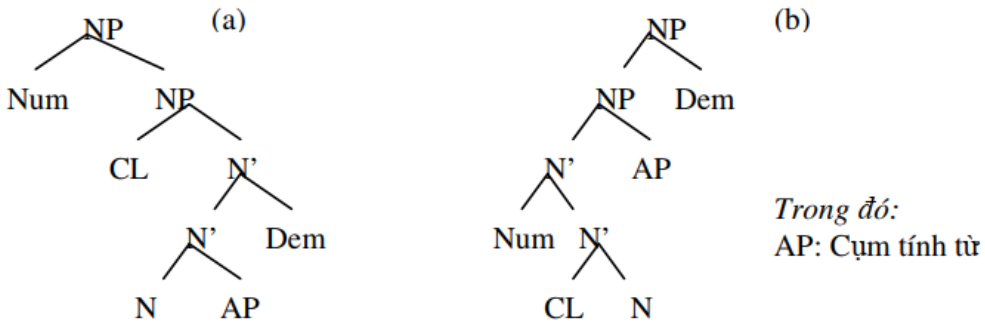
Trong đó:
Nom: Chủ ngữ
Mod: Bổ ngữ
Nu: Số đếm
Dem: Chỉ định từ
Pron: Đại từ

Npr: Danh từ riêng
N: Danh từ
N': danh từ phân loại
N": danh từ không phân loại
PL: Số nhiều
Q: Lượng từ



Hình 2.25: Mô hình nghiên cứu của Shum

- Beatty đưa ra hai khả năng có thể có của cụm danh từ được biểu diễn qua cấu trúc cây như sau:



Hình 2.26: Mô hình nghiên cứu của Beatty

Tình hình phát triển ứng dụng

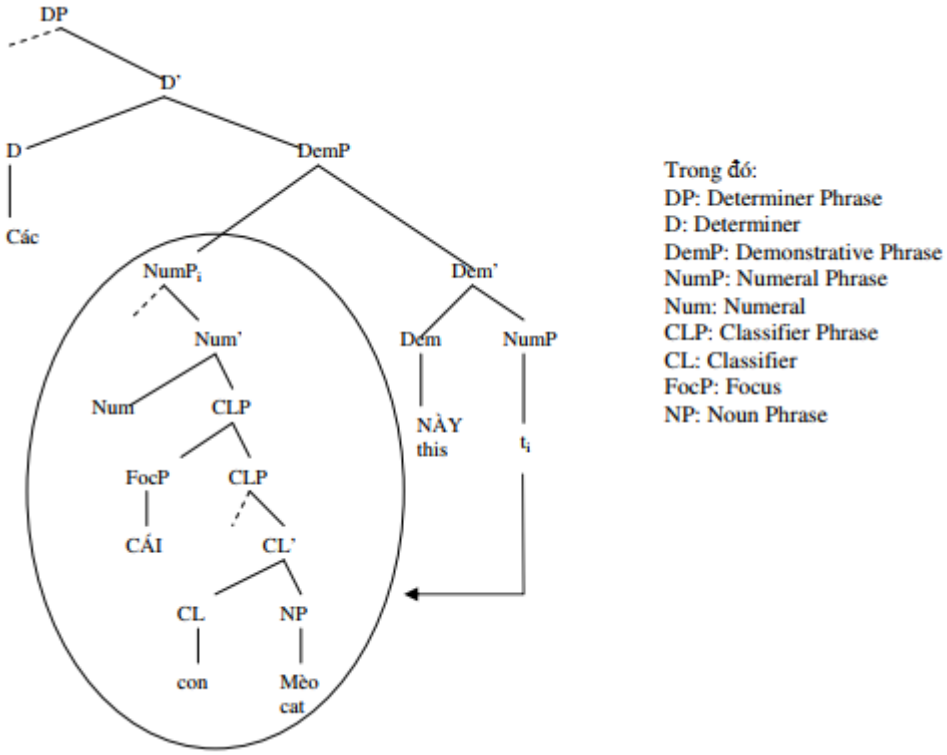
- Một số chatbot đã được giới thiệu đầu tiên như: ELIZA (1966); PARRY (1968) (The first system to pass the Turing test); ALICE; CLEVER hay Microsoft XiaoIce 小冰. Đến đầu năm 2016, các công ty lớn như Microsoft (Cortana); Google (Google Assistant); Facebook (M), Apple (Siri), Samsung (Viv), WeChat, Slack, …… cũng đã giới thiệu mô hình trợ lí ảo của mình. Hai trong số chatbot thông minh là Eugene Goostman và SmarterChild đã đẩy làn sóng chatbot lên cao.

- Gần đây nhất là hệ thống dịch máy Google's Neural Machine Translation (GNMT) cũng đang áp dụng mô hình sequence-to-sequence với cơ chế *attention* cho chất lượng dịch văn bản vượt trội. [3] [4] [5] [6].

2.4.4.2 Tình hình nghiên cứu trong nước

Ở Việt Nam, để hỗ trợ cho việc phát triển chatbot, cách đây hơn 20 năm, các đã nhà nghiên cứu, khoa học đã kế thừa những thành tựu của thế giới để đưa ra nhiều mô hình lý thuyết và cải tiến làm nền tảng cho việc phát triển các sản phẩm.

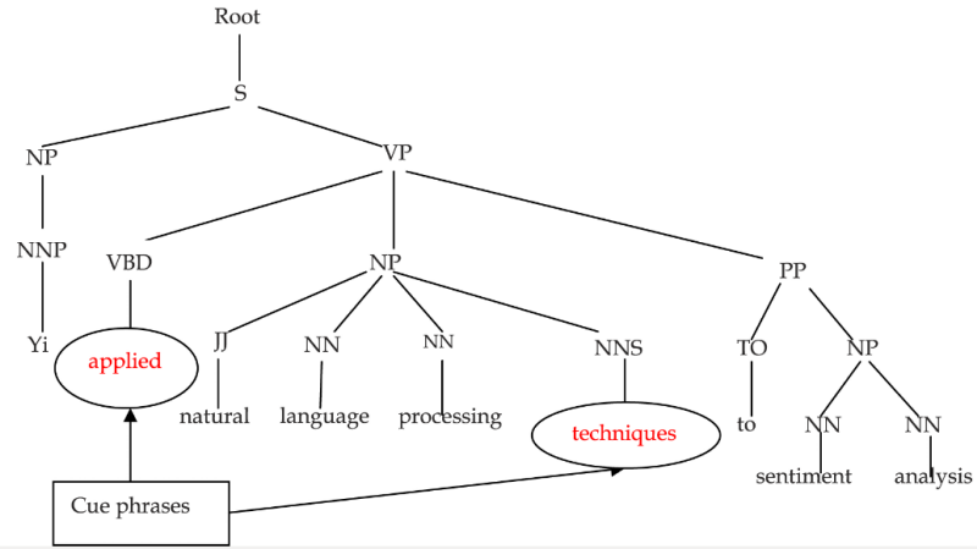
Tuong Hung Nguyen phát triển thêm những vấn đề mà Beatty chưa bàn đến và đưa ra cấu trúc tổng quát của cụm danh từ như hình 2.27 sau:



Hình 2.27: Mô hình của Tuong Hung Nguyen

Diệp Quang Ban đưa ra cấu tạo chung của cụm danh từ có ba phần là phần trung tâm, phần phụ trước và phần phụ sau. Phần trung tâm thường là một danh từ hoặc một ngữ danh từ. Trong phần phụ trước người ta đã xác định được ba vị trí khác nhau sắp xếp theo một trật tự nhất định. Ở phần phụ sau thường nhận được hai vị trí có trật tự ổn định. Phần phụ trước cụm danh từ thường dùng chỉ yếu tố số lượng của sự vật nêu ở trung tâm, phần phụ sau chủ yếu dùng chỉ yếu tố chất lượng của sự vật nêu ở thành phần trung tâm.

Hiện nay, Chatbot chỉ mới đang được nghiên cứu và phát triển nhằm giải quyết các vấn đề cơ bản của xử lí ngôn ngữ Tiếng Việt: gán nhãn từ; xây dựng cây phân tích cú pháp, ...)



Hình 2.28. Gán nhãn cây cú pháp (parse tree)

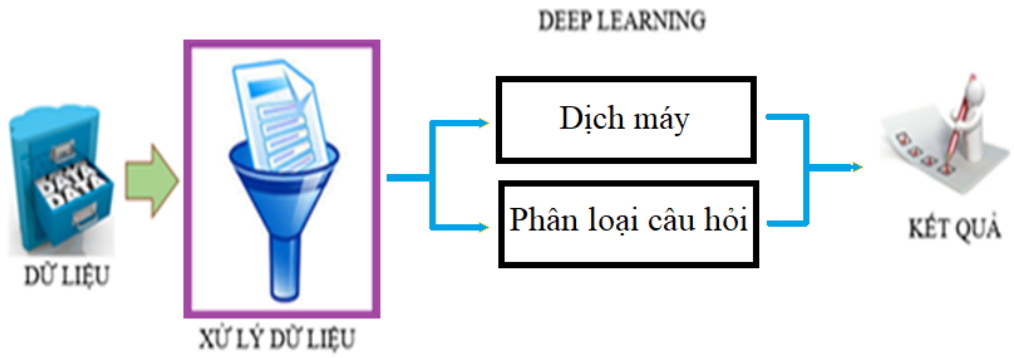
2.4.4.3 Hướng đề xuất nghiên cứu

Dựa trên những nghiên cứu trước đây, luận văn đã đưa ra hướng tiếp cận mới dựa trên nguyên tắc kế thừa từ các nghiên cứu của các tác giả về dịch máy và trích xuất thông để đề xuất nghiên cứu và phát triển một hệ thống trả lời tự động để trả lời các câu hỏi của sinh viên bằng ngôn ngữ Tiếng Việt. Hai hướng này đều sử dụng phương pháp học sâu bằng cách áp dụng phương pháp học chuỗi liên tiếp (sequence-to-sequence) và kĩ thuật attention để sinh ra câu trả lời tự động từ một chuỗi đầu vào tương ứng. Mô hình dịch máy bằng mạng Nơ-ron nhân tạo của Google (Google's Neural Machine Translation) và mô hình phân loại câu hỏi theo hướng mạng Nơ-ron sâu được đề xuất nhằm huấn luyện và đánh giá kết quả trên cả hai bộ dữ liệu gồm: bộ dữ liệu tiếng Việt được thu thập và bộ dữ liệu chuẩn. Luận văn cũng đưa ra phương pháp đánh giá kết quả huấn luyện và so sánh kết quả thực nghiệm trên hai bộ dữ liệu và đồng thời đề xuất mô hình ứng dụng trên nền tảng Web-based để hỗ trợ nhà trường trả lời các câu hỏi của sinh viên theo hai hướng dịch máy, trích xuất thông tin và theo từng chủ đề.

CHƯƠNG 3
MÔ HÌNH ĐỀ XUẤT

3.1. Tổng quan mô hình đề xuất

Mô hình trả lời tự động đề xuất trong luận văn được áp dụng dựa trên hai mô hình dịch máy bằng mạng Nơ-ron nhân tạo của Google (Google's Neural Machine Translation) [6] và mô hình phân loại câu hỏi theo hướng mạng Nơ-ron sâu để huấn luyện dữ liệu và kết hợp các phương pháp đánh giá dựa trên hai đánh giá độ chính xác (Accuracy) và đánh giá BLEU để đưa ra mô hình dự đoán tối ưu nhất nhằm mục đích trả lời các câu hỏi của sinh viên. Hai mô hình này đều sử dụng mạng học sâu LSTM bằng cách áp dụng mô hình học chuỗi liên tiếp (sequence-to-sequence) với bộ mã hóa, bộ giải mã (encoder and decoder) và kĩ thuật attention để sinh ra câu trả lời tự động từ một chuỗi đầu vào tương ứng để huấn luyện trên tập dữ liệu miền mở. Mô hình đề xuất được trình bày cụ thể ở hình 3.1 với 4 thành phần chính: dữ liệu đầu vào, xử lý dữ liệu, mô hình học sâu encode-decode (trên cơ sở LSTM 2 chiều, áp dụng kỹ thuật sequence-to-sequence và cơ chế attention) và phần kết quả đầu ra.



Hình 3.1: Đề xuất mô hình xây dựng chatbot

Với mô hình dịch máy bằng mạng Nơ-ron nhân tạo của Google:

+ Từ bộ dữ liệu đầu vào là các câu hỏi và câu trả lời sẽ tiến hành phân tách thành 4 bộ dữ liệu gồm:

(1) Một bộ dữ liệu huấn luyện chứa các câu hỏi

(2) Một bộ dữ liệu huấn luyện chứa các câu trả lời tương ứng với (1)

(3) Một bộ dữ liệu kiểm tra chứa các câu hỏi

(4) Một bộ dữ liệu kiểm tra chứa các câu trả lời tương ứng với (3)

+ Bộ dữ liệu huấn luyện và kiểm tra sẽ được tách từ và véc-tơ hóa bằng kỹ thuật word2vector, bên cạnh đó bộ dữ liệu huấn luyện sẽ được lưu trữ làm bộ từ điển để tiến hành huấn luyện dữ liệu. Cơ chế sequence-to-sequence được áp dụng trong quá trình huấn luyện thông qua việc thực hiện từng step, mỗi step sử dụng kỹ thuật attention để quyết định việc thu nhận hay loại bỏ thông tin và đánh giá độ mất mát thông tin.

+ Để đánh hiệu quả mô hình huấn luyện, đánh giá BLEU được áp dụng để quyết định mô hình dự đoán tối ưu.

Với mô hình theo hướng trích xuất thông tin, mô hình phân loại câu hỏi theo hướng mạng Nơ-ron sâu được áp dụng để thực hiện việc huấn luyện dữ liệu và đánh giá kết quả để đưa ra mô hình tối ưu để trả lời câu hỏi cho sinh viên. Từ bộ dữ liệu đầu vào là tập các câu hỏi và câu trả lời được phân hoạch theo một lớp cụ thể. Bộ dữ liệu câu hỏi sau khi loại bỏ các từ dừng sẽ được tách từ bằng phương pháp word2vector và sẽ được phân lớp cụ thể. Quá trình huấn luyện được huấn luyện dựa trên bộ câu hỏi đã được phân lớp, hàm softmax được sử dụng để thể hiện xác suất của lớp (kết quả dự đoán) thông qua các từ ở trong bộ câu hỏi đầu vào. Bên cạnh đó Entropy chéo được định nghĩa sẽ thể hiện cho xác suất mục tiêu của đầu ra. Việc đánh giá hiệu quả của mô hình huấn luyện, độ chính xác (accuracy) được áp dụng để quyết định mô hình dự đoán cuối cùng.

Căn cứ vào tình hình nghiên cứu trong nước cũng như ngoài nước, hướng tiếp cận dựa trên dữ liệu áp dụng mô hình dịch máy và trích rút thông tin đã thu được nhiều kết quả khả quan, bên cạnh đó ứng dụng trả lời tự động hiện đang là xu hướng của các công ty nhằm hỗ trợ khách hàng của họ trong chiến lược phát triển kinh doanh cũng như xây dựng các hệ thống chăm sóc con người,…Đây chính là lý do mà mô hình dịch máy bằng mạng Nơ-ron nhân tạo của google GNMT và mô hình phân loại câu hỏi theo hướng mạng Nơ-ron sâu được sử dụng là baseline trong xây dựng hệ thống trả lời tự động của luận văn.

3.1.1. Mô hình huấn luyện dữ liệu tổng quát

Như mọi các phương pháp ứng dụng của Machine Learning, việc đầu tiên là thu thập dữ liệu rồi sau đó tiến việc xử lý dữ liệu thô để tiến hành huấn luyện dữ liệu từ các phương pháp học máy như phân lớp dữ liệu, phân cụm, học sâu,… để sinh ra mô hình dự đoán kết quả nhằm trả lời các câu hỏi tương ứng của sinh viên.

Phương pháp huấn luyện được lựa chọn dựa trên các phương pháp học sâu để huấn luyện bộ dữ liệu, kết quả cuối cùng của việc huấn luyện là cho ra mô hình huấn luyện được lưu lại để thực hiện việc dự đoán kết quả. Quá trình huấn luyện sẽ được thực hiện liên tục nhằm mục đích tìm kiếm mô hình tối ưu nhất thông qua việc thay đổi bộ tham số huấn luyện đầu vào và phương pháp đánh giá kết quả dự đoán dựa trên bộ tham số đầu vào.

+ Bước 1: Thu thập dữ liệu

Thực hiện việc thu thập thông tin dữ liệu là các câu hỏi của các sinh viên liên quan đến trường Đại học Thủ Dầu Một như: thông tin ngành, nghề và chương trình đào tạo tại trường, thông tin về các giáo viên, thông tin về trường,…

+ Bước 2: Chuẩn hóa dữ liệu

Sau khi thu thập dữ liệu hoàn thành, việc chuẩn hóa dữ liệu sẽ được thực hiện nhằm loại bỏ những giá trị không phù hợp, ví dụ như các từ dừng (stop word), các ký hiệu dấu câu,….; khôi phục các từ lỗi,….

+Bước 3: Lựa chọn thuật toán huấn luyện

Tiến hành lựa chọn một trong các giải thuật của Machine Learning để tiến hành huấn luyện bộ dữ liệu huấn luyện.

+ Bước 4: Phân chia bộ dữ liệu huấn luyện và dự đoán

Phân chia bộ dữ liệu thu thập đã được chuẩn hóa theo tỷ lệ n:m, n phần dùng để thực hiện việc huấn luyện, m phần dùng để dự đoán

+ Bước 5: Huấn luyện bộ dữ liệu huấn luyện

Với phương pháp huấn luyện đã được chọn lựa, tiến hành huấn luyện dữ liệu huấn luyện với các tham số huấn luyện phù hợp

+ Bước 6: Lưu trữ mô hình dự đoán

Sau khi thực hiện việc huấn luyện dữ liệu hoàn tất, tiến hành lưu trữ mô hình huấn luyện để phục vụ cho việc dự đoán sau này.

3.1.2. Mô hình dự đoán kết quả

Để tiến hành việc dự đoán kết quả, từ mô hình huấn luyện đã lưu trữ trước đó, tiến hành dự đoán kết quả trên dữ liệu dự đoán, sau đó tiến hành đánh giá kết quả dự đoán.

3.1.3. Mô hình huấn luyện dữ liệu - dự đoán kết quả

Quá trình huấn luyện và dự đoán kết quả sẽ được lặp đi lặp lại cho đến khi mô hình dự đoán kết quả là tối ưu nhất thông qua việc đánh giá kết quả dự đoán. Mỗi quá trình huấn luyện dữ liệu, các tham số huấn luyện sẽ được thay đổi để tìm ra kết mô hình huấn luyện tối ưu nhất. Mô hình huấn luyện này sẽ được sử dụng để dự đoán kết quả cho các bộ dữ liệu mà người dùng nhập vào.

+ Bước 1: Thu thập dữ liệu

+ **Bước 2:** Chuẩn hóa dữ liệu

+ Bước 3: Lựa chọn thuật toán huấn luyện

+ Bước 4: Chia bộ dữ liệu theo tỷ lệ n:m

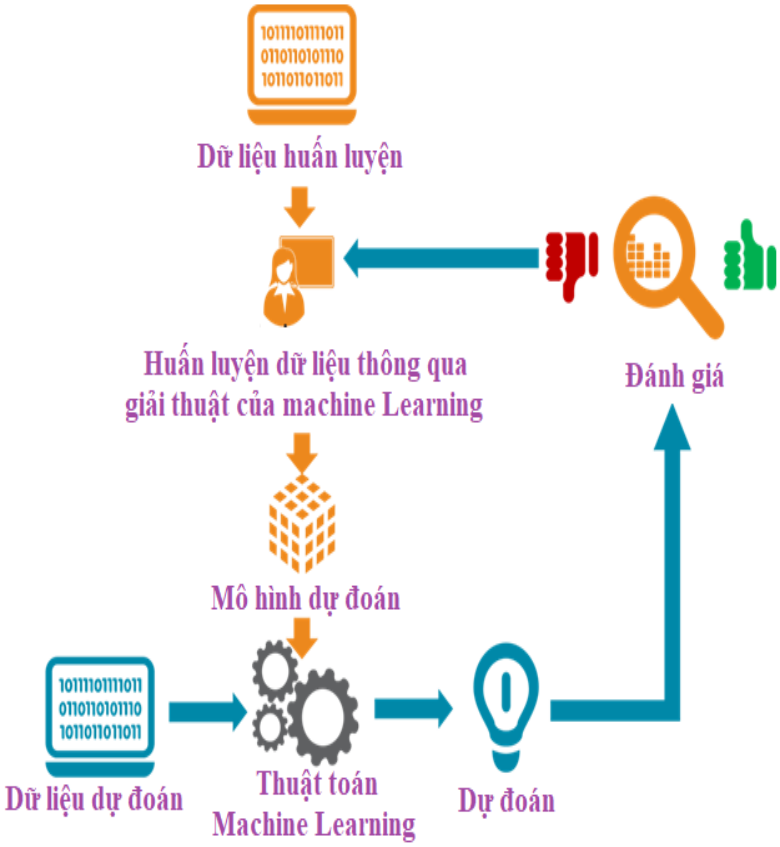
+ **Bước 5**: Xây dựng mô hình dự đoán bằng cách đưa vào các tham số huấn luyện phù hợp, sau đó tiến hành huấn luyện trên bộ dữ liệu huấn luyện

+ Bước 6: Lưu trữ mô hình dự đoán để phục vụ cho việc dự đoán kết quả sau này

+ Bước 7: Từ thuật toán huấn luyện được chọn lựa ở Bước 3, kết hợp với mô hình dự đoán đã lưu trữ ở Bước 6, tiến hành dự đoán kết quả với bộ dữ liệu dự đoán.

+ Bước 8: Đánh giá kết quả dự đoán ở Bước 7.

- Nếu kết quả dự đoán thỏa một tiêu chí đánh giá thì kết thúc, mô hình dự đoán sẽ được sử dụng để thực hiện việc dự đoán kết quả từ bộ dữ liệu do người dùng đưa vào.
- Nếu kết quả chưa đạt thì tiến hành quay lại Bước 4



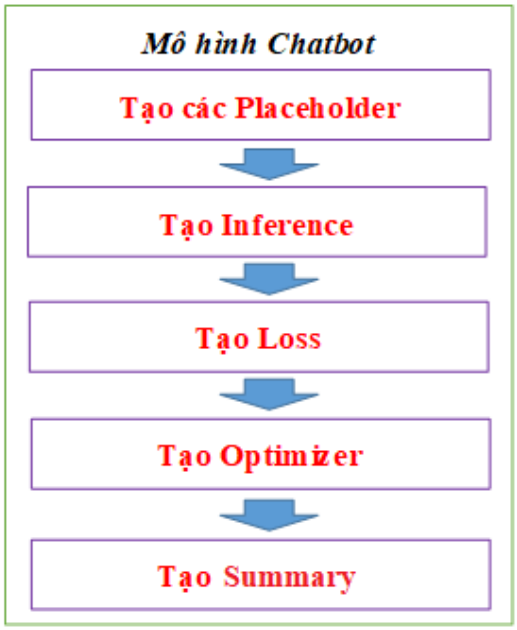
Hình 3.2. Quy trình huấn luyện dữ liệu - dự đoán kết quả

3.2. Mô hình dịch máy bằng mạng Nơ-ron nhân tạo của Google (GNMT)

3.2.1. Mô hình huấn luyện dữ liệu

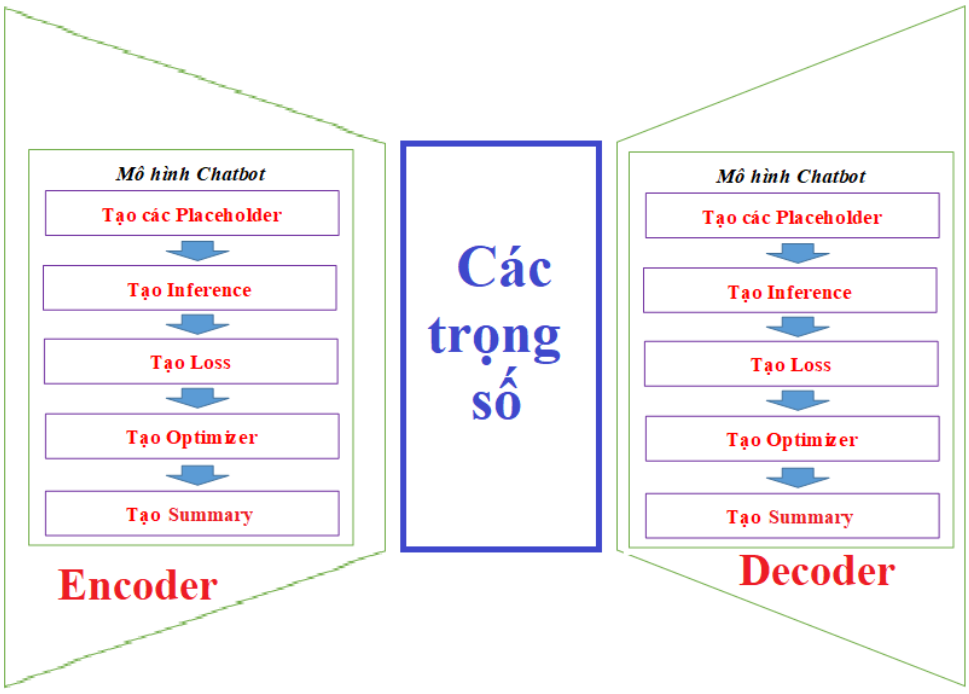
Hiện nay, Google Translate được tích hợp hệ thống dịch thuật mạnh mẽ mới có tên gọi là Google Neural Machine Translation (GNMT), mang lại kết quả dịch chính xác và tự nhiên hơn. Đề tài sử dụng Mô hình Chatbot dựa trên GNMT để huấn luyện dữ liệu. Mô hình Chatbot được xây dựng dựa trên 05 thành phần khởi tạo mà dữ liệu phải đi qua theo tuần tự:

+ Tạo các Placeholder
+ Tạo Inference
+ Tạo Loss
+ Tạo các Optimizer
+ Tạo Summary



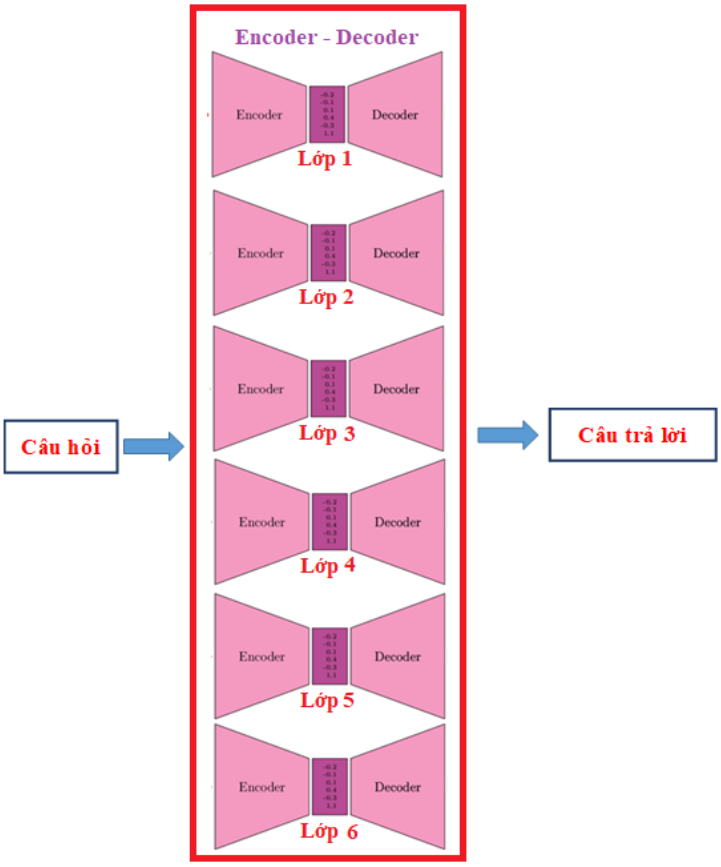
Hình 3.3:Mô hình Chatbot

Mỗi lớp trong mô hình dịch máy bằng mạng Nơ-ron nhân tạo được xây dựng dựa trên sự kết hợp giữa hai lớp Encoder và lớp Decoder. Mỗi lớp Encoder và Decoder đều được thiết kế dựa trên mô hình Chatbot theo 05 thành phần như đã xây dựng ở trên.



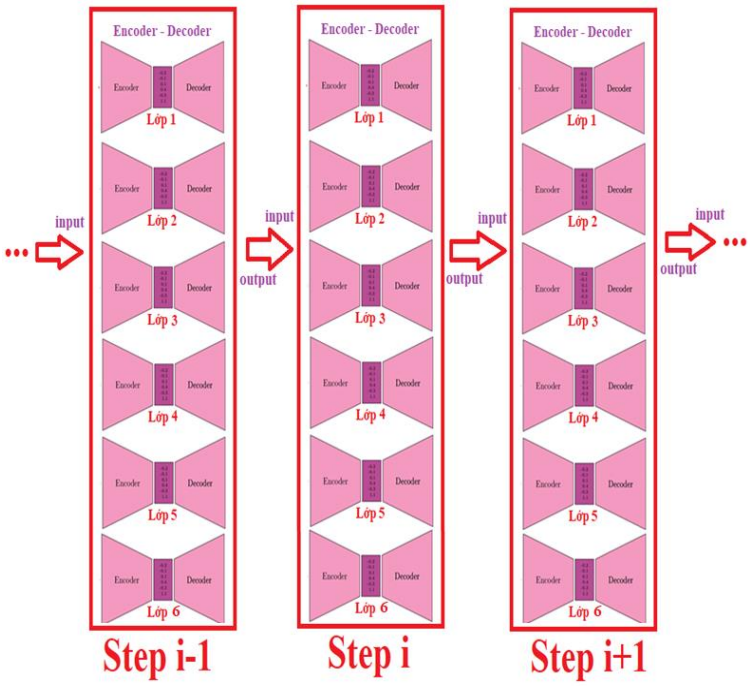
Hình 3.4: Sự kết hợp giữa hai lớp Encoder và lớp Decoder

Một step thực hiện huấn luyện một câu ở dữ liệu đầu vào trong mô hình sẽ đi qua 6 lớp Encoder-Decoder để cho đầu ra là câu trả lời tương ứng



Hình 3.5: Mô hình 6 lớp Encoder-Decoder

Mô hình dịch máy bằng mạng Nơ-ron nhân tạo dựa trên GNMT là sự kết hợp giữa các step, khi đó đầu ra của step trước đó là đầu vào của step hiện tại.



Hình 3.6: Mô hình các Step thực hiện

Trong quá trình huấn luyện mô hình dự đoán sẽ được lưu lại để phục vụ cho việc dự đoán kết quả.

3.2.2. Mô hình đánh giá quá trình huấn luyện

Trong quá trình thực hiện huấn luyện dữ liệu, tại mỗi step sau khi thực việc lưu trữ mô hình thì việc đánh giá sẽ được thực hiện trước khi chuyển sang step tiếp theo. Việc đánh giá sẽ được tính dựa trên Loss thông tin và thời gian thực hiện.

Tại mỗi step, quá trình huấn luyện sẽ được thực hiện theo 5 bước như sau:

+ **Bước 1:** Khởi tạo các lớp Encoder – Decoder

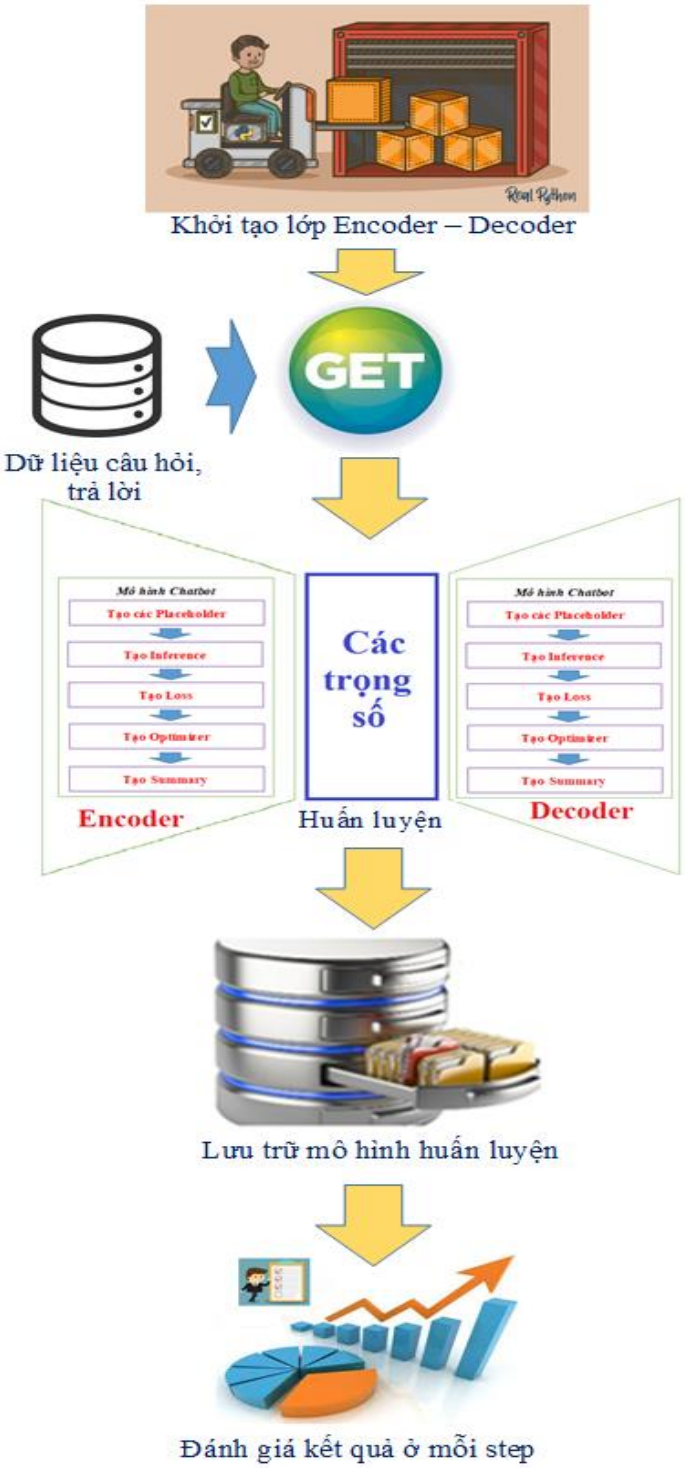
+ **Bước 2**: Nhận dữ liệu đầu vào

+ **Bước 3**: Thực hiện huấn luyện dữ liệu trên mô hình Chatbot (lớp Encoder – Decoder)

+ **Bước 4**: Thực hiện việc tính Loss giữa đầu ra với dữ liệu là các câu trả lời thực tế

+ **Bước 5**: Lưu trữ lại mô hình huấn luyện để thực hiện việc dự đoán sau này

+ **Bước 6**: Đánh giá kết quả thực hiện

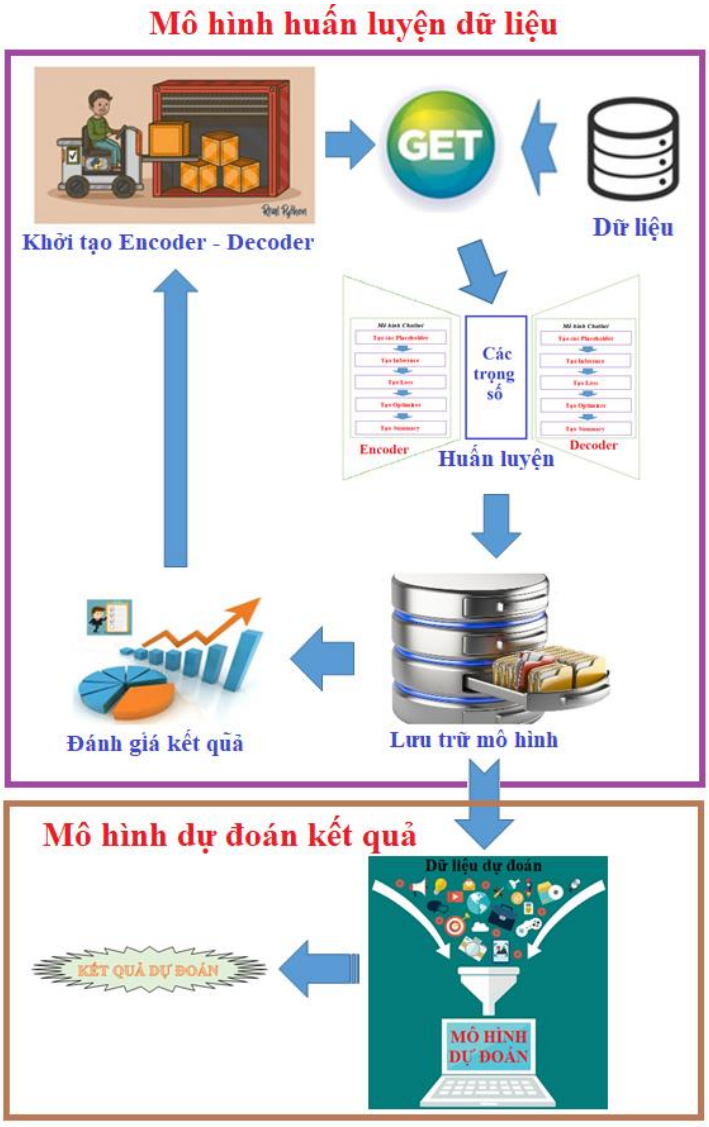


Hình 3.7: Quy trình đánh giá quá trình huấn luyện

3.2.3. Mô hình huấn luyện dữ liệu – dự đoán kết quả

Trước tiên, bộ dữ liệu sẽ được đưa vào huấn luyện, do mô hình được thiết kế theo từng step cho nên trong mỗi step sẽ đánh giá kết quả huấn luyện tại mỗi step đồng thời sẽ lưu trữ mô hình để thực hiện cho việc dự đoán các câu hỏi của sinh viên.

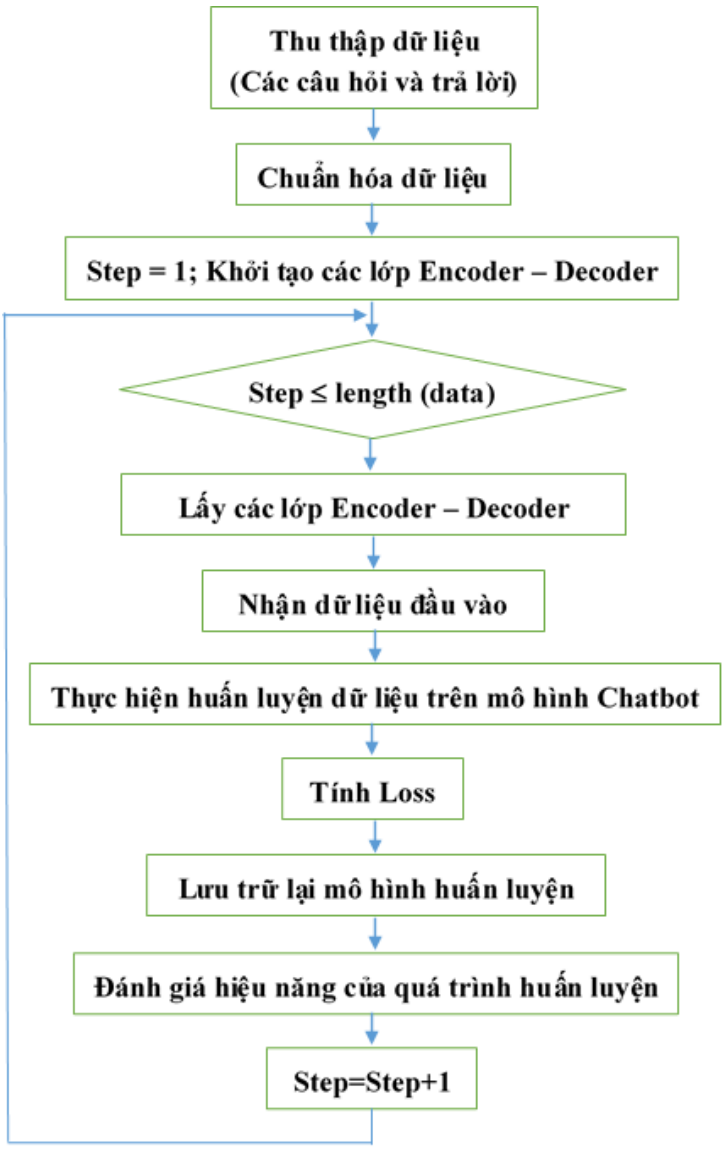
Sau quá trình huấn luyện, sử dụng mô hình đã lưu trữ để thực hiện việc trả lời các câu hỏi của sinh viên bằng cách dự đoán dựa trên mô hình đã lưu trữ trước đó.



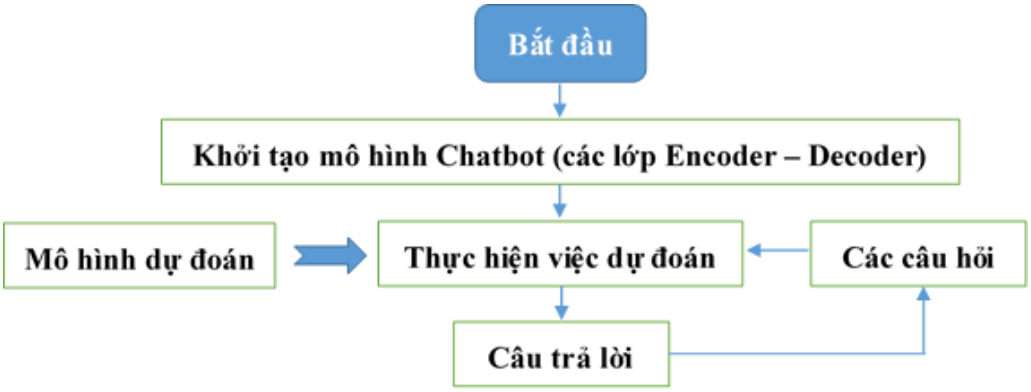
Hình 3.8: Huấn luyện dữ liệu – dự đoán kết quả

3.2.4. Giải thuật sử dụng mạng Nơ-ron nhân tạo của Google (GNMT)

3.2.4.1. Giải thuật huấn luyện dữ liệu



3.2.4.2. Giải thuật dự đoán kết quả



3.3. Mô hình phân loại câu hỏi bằng phương pháp học sâu

Mô hình trả lời các câu hỏi của sinh viên được xây dựng dựa trên mô hình phân loại câu hỏi bằng phương pháp học sâu. Để trả lời các câu hỏi của sinh viên, việc đầu tiên đó là quá trình huấn luyện dữ liệu đã được chuẩn hóa sau quá trình thu

thập. Tuy nhiên, trong quá trình đưa thông tin đầu vào là các câu hỏi, một số sinh viên có thể nhập các từ viết tắt không theo một quy luật nào, do đó việc huấn luyện dữ liệu là các từ viết tắc được thực hiện song song với việc huấn luyện dữ liệu thu thập là các câu hỏi và trả lời. Quá trình huấn luyện hoàn tất sẽ được lưu trữ thành để thực hiện cho việc dự đoán các câu hỏi của sinh viên.

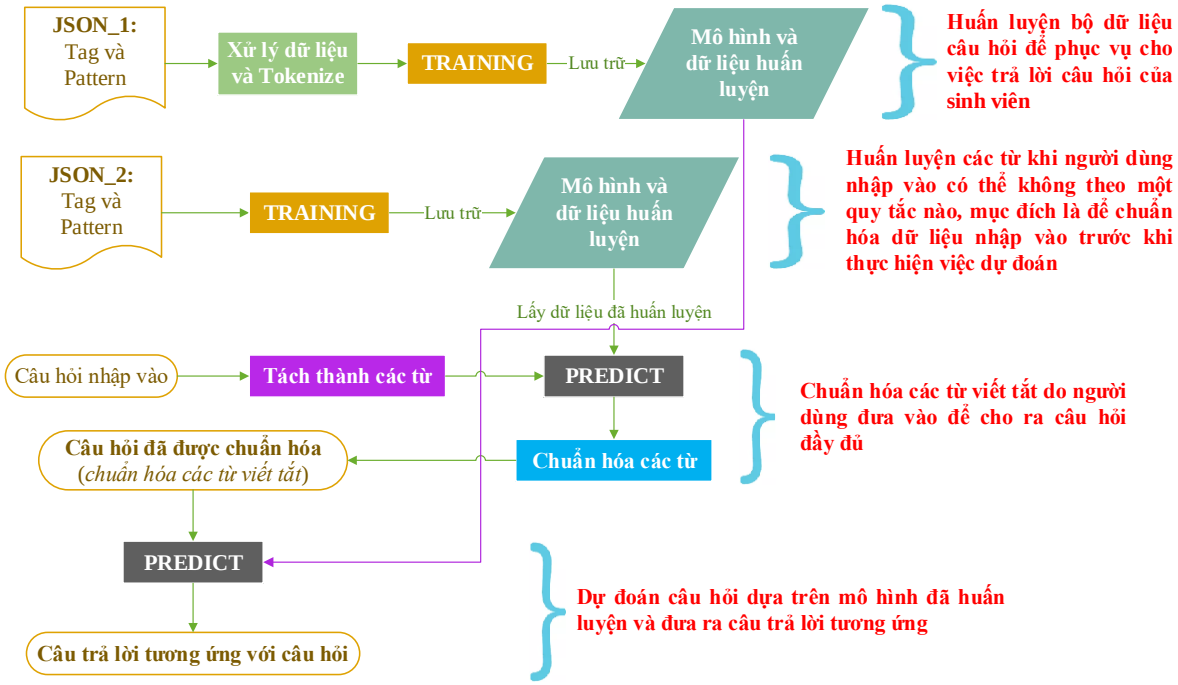
Quá trình trả lời các câu hỏi của sinh viên được thực hiện như sau:

+ Nhận dữ liệu là câu hỏi của sinh viên

+ Chuẩn hóa các từ viết tắt trong câu hỏi bằng cách tách các từ trong câu hỏi và tiến hành dự đoán từ đầy đủ của từ (nếu có)

+ Ghép các từ lại thành câu hoàn chỉnh theo trình tự

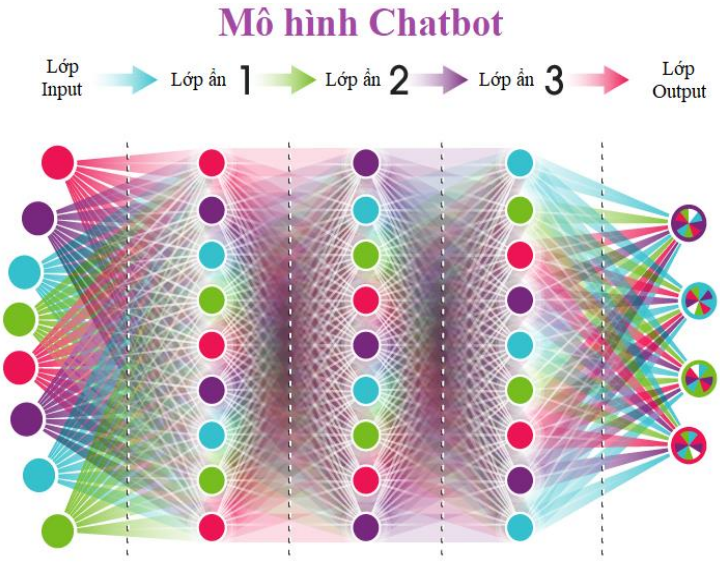
+ Tiến hành dự đoán câu trả lời trên câu hoàn chỉnh để trả lời cho sinh viên



Hình 3.9: Quy trình huấn luyện dữ liệu và dự đoán kết quả

3.3.1. Mô hình huấn luyện dữ liệu

Mô hình huấn luyện dữ liệu của Chatbot bao gồm: Một lớp đầu vào (Lớp input), lớp ẩn, một lớp đầu ra.



Hình 3.10: Mô hình Chatbot theo mạng Nơ-ron sâu

- Lớp đầu vào (lớp input): đây là lớp nhận dữ liệu đầu vào là các câu

- Lớp ẩn gồm có 03 lớp, mỗi lớp chứa các hàm, chức năng nhằm thực hiện việc tính toán các giá trị cho các câu được đưa vào từ lớp input được gọi là activation function, các activation function có các thành phần:

+ Step function

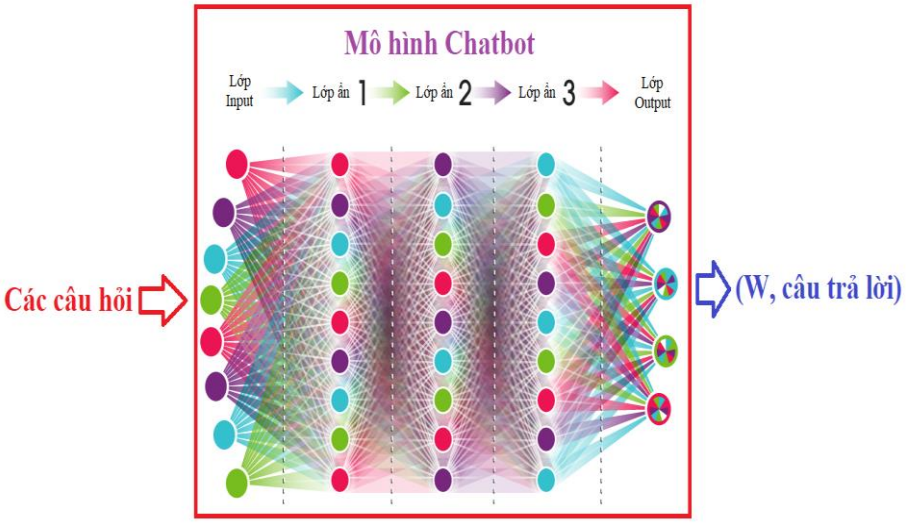
+ Linear function

+ Sigmoid function

+ Tanh function

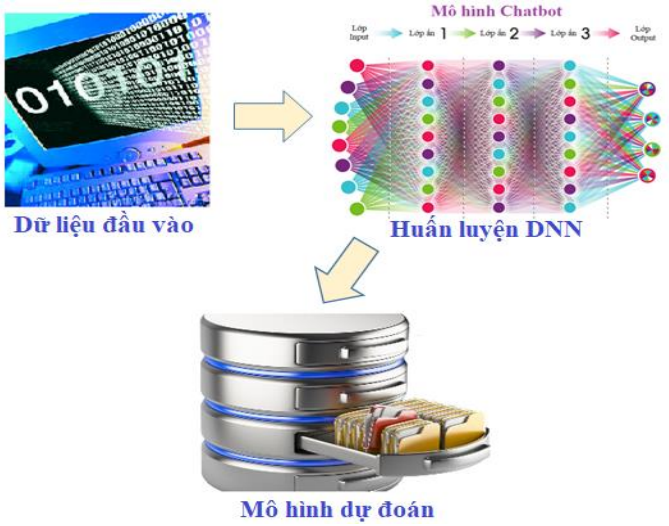
+ ReLu

- Lớp đầu ra (lớp output): là lớp chứa các câu trả lời với các trọng số đã được tính toán tương ứng.



Hình 3.11: Mô hình huấn luyện

Dữ liệu đầu vào là bộ các câu hỏi và các câu trả lời tương ứng, quá trình thực hiện với 1000 epochs và mạng đánh giá batch_size là 8. Sau khi thực hiện huấn luyện hoàn tất, tiến hành lưu lại mô hình dự đoán để thực hiện việc dự đoán kết quả sau này.



Hình 3.12: Quy trình huấn luyện dữ liệu

3.3.2. Mô hình đánh giá quá trình huấn luyện và dự đoán kết quả

Để đánh giá hiệu quả của mô hình huấn luyện, một tham số được đề xuất nhằm đánh giá hiệu năng của mô hình đề xuất đó là thông số Accuracy hoặc F1-Score. Quá trình huấn luyện được thực hiện theo tuần tự các bước như sau:

- **Bước 1**: Chuẩn hóa dữ liệu đầu vào

Với dữ liệu đầu vào là các câu hỏi, sử dụng kỹ thuật tokenize và loại bỏ các từ dừng (stop word).

- **Bước 2**: Khởi tạo các tham số huấn luyện

Khởi tạo các tham số huấn luyện: activation, optimizer, batch_size, loss,….

- **Bước 3**: Huấn luyện dữ liệu

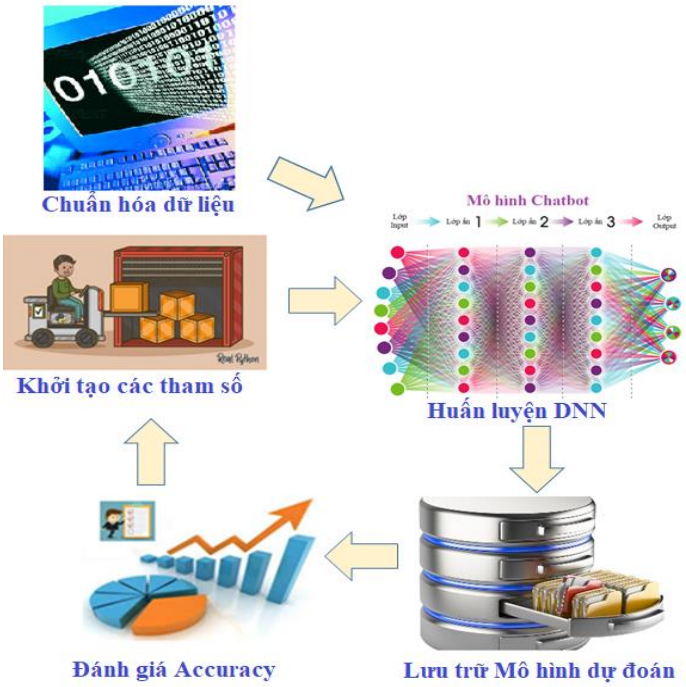
Sử dụng mô hình mạng Nơ-ron sâu để tiến hành huấn luyện với bộ tham số huấn luyện đã khởi tạo.

- **Bước 4**: Lưu trữ mô hình dự đoán

Tiến hành lưu trữ mô hình dự đoán với các tham số huấn luyện đã được khởi tạo trước đó để thực hiện cho việc dự đoán sau này

- **Bước 5**: Đánh giá hiệu quả của việc huấn luyện

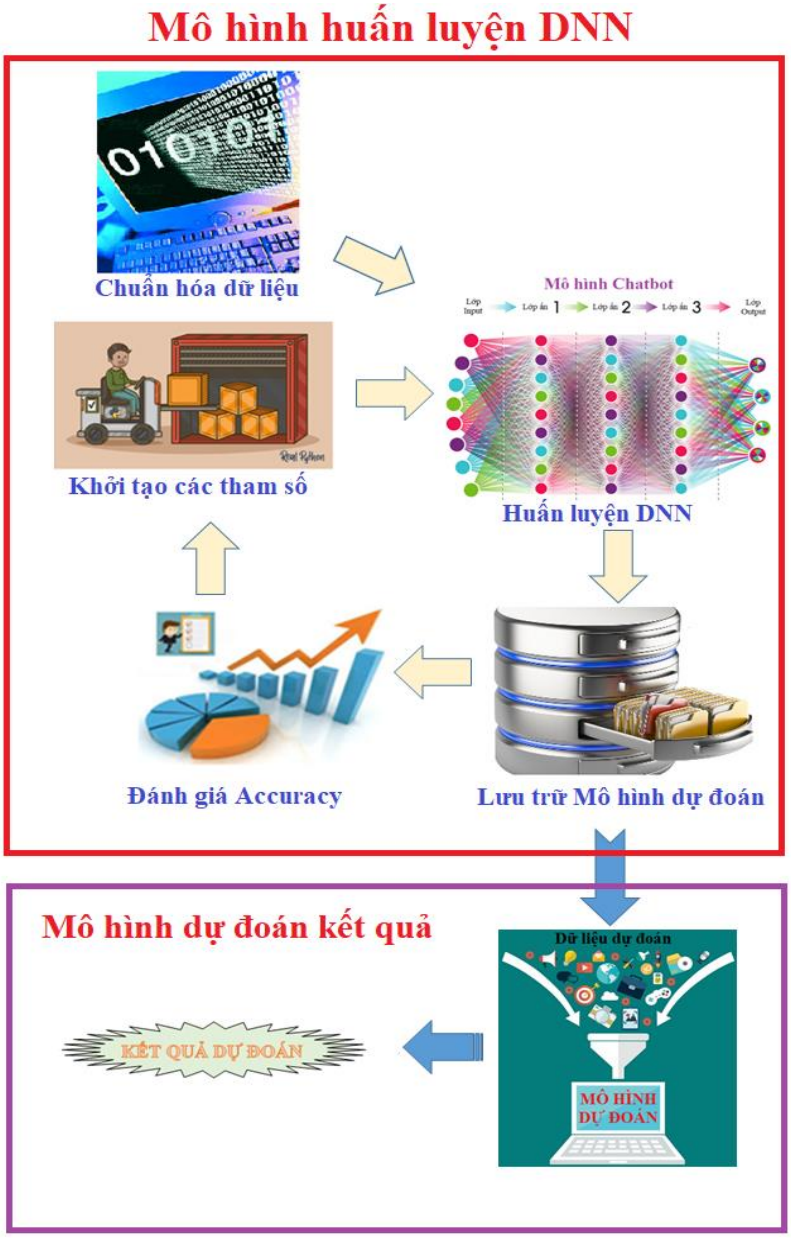
Sử dụng thông số accuracy để đánh giá mô hình huấn luyện, nếu tham số accuracy đạt trên 95% thì tiến hành kết thúc quá trình huấn luyện, nếu ngược lại thì quay lại Bước 2.



Hình 3.13: Quy trình đánh giá quá trình huấn luyện và dự đoán kết quả

3.3.3. Mô hình huấn luyện – dự đoán kết quả

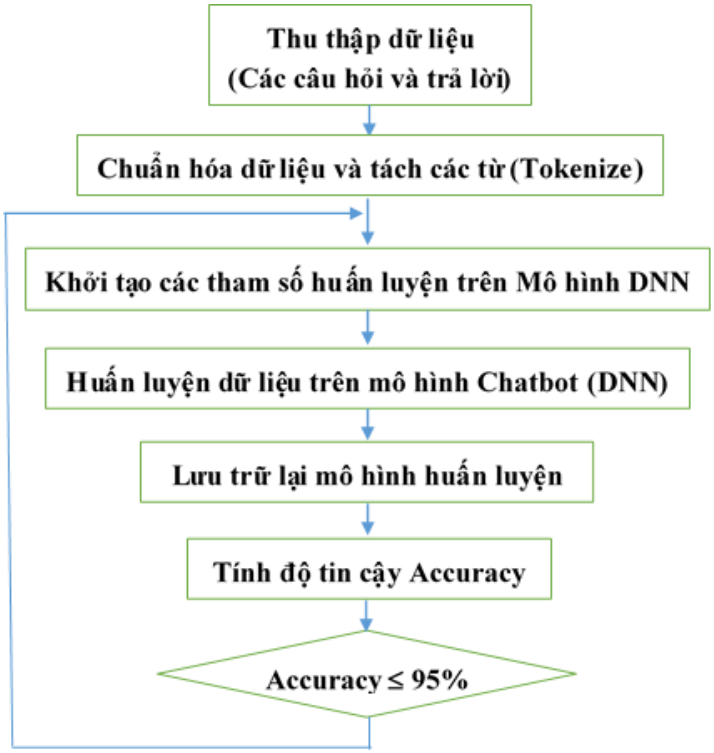
Quá trình huấn luyện sẽ được thực hiện lặp đi lặp lại cho đến khi tham số đánh giá hiệu quả của mô hình accuracy đạt ở mức cho phép thì kết thúc quá trình huấn luyện dữ liệu. Ở mỗi quá trình huấn luyện được thực hiện lại, các tham số huấn luyện sẽ được thay đổi. Khi quá trình huấn luyện kết thúc, mô hình dự đoán sẽ được lưu lại để thực hiện việc trả lời các câu hỏi của sinh viên bằng cách sử dụng mô hình đã lưu trữ trước đó để dự đoán kết quả.



Hình 3.14:Quy trình đánh giá quá trình huấn luyện và dự đoán kết quả

3.3.4. Giải thuật sử dụng mạng Nơ-ron sâu (DNN)

3.3.4.1. Giải thuật huấn luyện dữ liệu



3.3.4.2. Giải thuật dự đoán kết quả

CHƯƠNG 4
THỰC NGHIỆM

4.1. Theo hướng dịch máy bằng mạng Nơ-ron nhân tạo

4.1.1. Dữ liệu

4.1.1.1. Bộ dữ liệu Cornell Movie-Dialogs Corpus

Dữ liệu thử nghiệm là bộ dữ liệu the Cornell Movie-Dialogs Corpus. Chi tiết về bộ dữ liệu này được trình bày trong Bảng 4.1 sau đây.

Dữ liệu	**Tổng số hội thoại (Người 1 và Người 2)**	**Dung lượng (MB)**
Train	225000 + 225000	17 + 17
Test	5000 + 5000	
Validation	5000 + 5000	

Bảng 4.1:Bộ dữ liệu Cornell movie-dialogs

Dữ liệu đã được xử lý để đưa vào đường ống đầu vào mô hình dịch máy bằng mạng nơ ron nhân tạo của Google: GNMT (Google's Neural Machine Translation) [6]. Trong mô hình dịch máy GNMT ban đầu, có hai file dữ liệu đầu vào (file nguồn) và đầu ra (file đích hay file dịch). Hai file từ vựng được tạo ra từ các file đầu vào và đầu ra đó. Bộ từ vựng chứa các từ vựng được xử lý cho hai file dữ liệu đầu và đầu ra của hai ngôn ngữ khác nhau riêng biệt. Bộ dữ liệu được chia thành 3 bộ: Train, Test và Develop. Nếu file ngôn ngữ 1 là đầu vào của mô hình dịch máy và đầu ra file ngôn ngữ 2 và ngược lại file ngôn ngữ 2 là đầu vào và file ngôn ngữ 1 sẽ là đầu ra.

4.1.1.2. Bộ dữ liệu thu thập

Dữ liệu thu thập là bộ câu hỏi và câu trả lời liên quan đến các chương trình đào tạo, hỗ trợ tuyển sinh, thông tin về giảng viên và các văn bản hướng dẫn thường gặp cho sinh viên của trường Đại học Thủ Dầu Một. Bộ dữ liệu được biểu diễn theo mẫu sau:

stt	question	answer	sample
1	các ngành đào tạo Đại học chính quy năm học 2018 -2022 của trường	Kế toán; Quản trị Kinh doanh; Tài chính - Ngân hàng; Kỹ thuật Xây dựng; Kỹ thuật Điện; Kỹ thuật Phần mềm - Công nghệ Thông tin; Hệ thống Thông tin - Công nghệ Thông tin; Kiến trúc; Quy hoạch Vùng và Đô thị - Quản Lý Đô thị; Hóa học; Sinh học ứng dụng; Khoa học Môi trường; Vật lý học; Toán học; Quản lý Tài nguyên và Môi trường; Quản lí Nhà nước; Quản lý Công nghiệp; Sư phạm ngữ văn; Sư phạm Lịch sử; Giáo dục học; Luật; Ngôn ngữ Anh; Ngôn Ngữ Trung Quốc; Công tác Xã hội; Địa lý học; Quản lý Đất đai; Giáo dục Mầm non; Giáo dục Tiểu học; Chính trị học; Văn hóa học	Cho em hỏi các ngành đào tạo Đại học chính quy năm học 2018 -2022 của trường là những ngành nào?
			Các ngành đào tạo Đại học chính quy năm học 2018 -2022 của trường là những ngành
			Năm học 2018, trường đào tạo Đại học chính quy những ngành nào
			Ngành đào tạo Đại học chính quy năm 2018 của trường
2	Thời gian đào tạo bậc đại học	Chương trình đào tạo được thực hiện trong 8 học kỳ. Để tốt nghiệp, sinh viên phải hoàn thành đồ án tốt nghiệp vào học kỳ 8. Theo Quyết định số 1157/QĐ-ĐHTDM ngày 08 tháng 08 năm 2014 và bổ sung, sửa đổi theo Quyết định số 890/QĐ-ĐHTDM ngày 04 tháng 08 năm 2016 của Hiệu trưởng Trường Đại Thủ Dầu Một về Quy chế đào tạo đại học và cao đẳng chính quy theo hệ thống tín chỉ.	Em muốn biết thời gian đào tạo bậc đại học là bao lâu?
			Đại học được đào tạo thời gian bao lâu
			Đào tạo đại học trong thời gian bao lâu
			đào tạo bậc đại học tốn thời gian bao lâu?

Bảng 4.2: Bộ dữ liệu thu thập về thông tin của Trường ĐH Thủ Dầu Một

Cột question, sample thể hiện các câu hỏi thường gặp của sinh viên về thông tin của trường. Cột answer thể hiện câu trả lời tương ứng với các câu hỏi ở trong cột question, sample.

4.1.2. Xử lý dữ liệu

4.1.2.1. Xử lý bộ dữ liệu Cornell Movie-Dialogs Corpus

Dữ liệu Cornell Movie-Dialogs Corpus là dữ liệu hội thoại trong lĩnh vực phim ảnh, dữ liệu này gồm các dòng hội thoại có chứa các ID phim (Movie ID), ID nhân vật (Character ID),và ID dòng phim (Movie Line) được phân tách bằng dấu '++ +++++'.

Để tiền xử lý, dữ liệu hội thoại đã được làm sạch bằng cách xóa bớt các thẻ đánh dấu (ví dụ: ID phim, ID nhân vật, ID dòng phim). Ngoài ra, các phân tách dữ liệu (+++++) cũng đã bị loại bỏ. Một số ký tự trong dữ liệu chứa định dạng mã hóa không được hỗ trợ theo tiêu chuẩn UTF-8 sẽ bị xóa. Dữ liệu sau khi đã làm sạch sẽ được tách thành hai file khác nhau tương thích với định dạng của GNMT. Trong đó file đầu là đối thoại 1 và file thứ 2 là phản hồi của đối thoại 1 hay chính là file đầu ra (file dịch).

Sau khi tách thành hai files, dữ liệu trong cả hai file sẽ tiếp tục được làm sạch. Mọi ký tự ngoại trừ ký tự chữ cái và một số dấu câu (.,?!') đã bị xóa vì chúng không

có ý nghĩa gì trong cuộc trò chuyện. Tất cả các ký tự đã được chuyển đổi thành chữ thường. Với bước xử lý trên đã giảm quá tải các dấu câu. Tiếp theo, tất cả các dấu chấm trừ dấu (‘) được phân tách bằng một khoảng trắng trước và sau để có hiệu suất tốt hơn trong mô-đun GNMT. Cuối cùng, tất cả các khoảng trắng liên tiếp đã được giảm xuống thành một và mỗi đoạn hội thoại đã được cắt bớt loại bỏ các khoảng trắng ở trước và sau.

Dữ liệu tiếp tục được làm sạch bằng cách loại bỏ các đối thoại lặp lại. Nếu các đoạn đối thoại xuất hiện nhiều lần thì sẽ loại bỏ bớt chỉ lưu đối thoại xuất hiện cuối. Với các hội thoại có độ dài hơn 100 sẽ bị loại bỏ cho cả câu đối thoại và câu phản hồi điều đó có nghĩa là khi độ dài của văn bản tăng lên, mức độ liên quan đến ngữ cảnh bắt đầu giảm do sự đa dạng và dữ liệu hạn chế.

Sau khi đã làm sạch dữ liệu, dữ liệu đầu vào và đầu ra sẽ được chia ra thành 3 bộ: Train, Test và Develop tương ứng với mỗi bộ dữ liệu ở mỗi đầu. Để tạo từ vựng, module Google’s Sub-word Neural Machine Translation sẽ được sử dụng như theo đề xuất trong tài liệu của Google Tensorflow và Google’s Neural Machine Translation.

4.1.2.2. Xử lý bộ dữ liệu thu thập

Dữ liệu sau khi thu thập tiến hành xử lý theo các bước như sau:

- Bước 1: Phân tách thành hai bộ câu hỏi và câu trả lời

Dữ liệu từ file excel đã thu thập tiến hành tách và lưu trữ vào dữ liệu theo hai bộ câu hỏi và trả lời theo quy tắc:

+ Cột question, sample được đưa vào bộ dữ liệu câu hỏi, cột answer được đưa vào bộ dữ liệu trả lời

+ Các câu hỏi và câu trả lời phải tương ứng theo chỉ mục với nhau theo đúng cấu trúc trong file excel

- Bước 2: Phân chia tập huấn luyện (train) và kiểm tra (test)

Chia bộ dữ liệu thành hai tập huấn luyện và kiểm tra theo tỷ lệ 4:1, tức là 4 phần để huấn luyện và một phần để kiểm tra, quá trình phân chia được thực hiện theo nguyên tắc sau:

+ Dữ liệu được chọn vào tập huấn luyện, kiểm tra được chọn một cách ngẫu nhiên

+ Các câu hỏi và câu trả lời phải tương ứng theo chỉ mục với nhau theo đúng cấu trúc trong file excel

+ Dữ liệu huấn luyện và kiểm tra được lưu trữ vào 4 file:

- Bộ huấn luyện: 1 file chứa bộ câu hỏi, 1 file chứa bộ câu trả lời
- Bộ kiểm tra: 1 file chứa bộ câu hỏi, 1 file chứa bộ câu trả lời

```
Các thầy cô giảng viên Bộ môn Kĩ thuật phần mềm?
Bộ môn Kĩ thuật phần mềm gồm giảng viên nào?
Các thầy cô giảng viên Bộ môn HTTT?
Bộ môn HTTT gồm những thầy cô nào
Giảng viên Bộ môn Điện tử - Viễn thông
giảng viên Bộ môn Điện tử - Viễn thông
Bộ môn Điện tử - Viễn thông gồm giảng viên nào?
giảng viên Bộ môn Điện tử - Viễn thông gồm những thầy cô nào
Giảng viên Bộ môn Kỹ thuật Điện?
giảng viên Bộ môn  Bộ môn Kỹ thuật Điện
giảng viên Bộ môn Kỹ thuật Điện gồm những thầy cô nào
điện thoại thầy Trần Vĩnh Phước
Số điện thoại thầy Trần Vĩnh Phước
điện thoại thầy Phước
Cách liên lạc thầy Phước
liên lạc thầy Phước qua số điện thoại nào
Email  thầy Trần Vĩnh Phước
Email thầy Trần Vĩnh Phước
```

Hình 4.1: Bộ câu hỏi – training

```
PGS.TS Trần Vĩnh Phước Th.S Bùi Thanh Khiết TS Hoàng Mạnh Hà Th.S Vũ Văn Nam
PGS.TS Trần Vĩnh Phước Th.S Bùi Thanh Khiết TS Hoàng Mạnh Hà Th.S Vũ Văn Nam
PGS. TS Lê Tuấn Anh Th.S Nguyễn Thị Thủy Th.S Võ Thị Hồng Thắm Th.S Dương Th
PGS. TS Lê Tuấn Anh Th.S Nguyễn Thị Thủy Th.S Võ Thị Hồng Thắm Th.S Dương Th
Th.S Đỗ Đắc Thiềm Th.S Ngô Sỹ Th.S Lê Trường An Th.S Võ Thành Nhân TS Trần H
Th.S Đỗ Đắc Thiềm Th.S Ngô Sỹ Th.S Lê Trường An Th.S Võ Thành Nhân TS Trần H
Th.S Đỗ Đắc Thiềm Th.S Ngô Sỹ Th.S Lê Trường An Th.S Võ Thành Nhân TS Trần H
Th.S Đỗ Đắc Thiềm Th.S Ngô Sỹ Th.S Lê Trường An Th.S Võ Thành Nhân TS Trần H
Th.S Nguyễn Phương Trà Th.S Nguyễn Cao Trí Th.S Phạm Quang Minh T.S Trần Văn
Th.S Nguyễn Phương Trà Th.S Nguyễn Cao Trí Th.S Phạm Quang Minh T.S Trần Văn
Th.S Nguyễn Phương Trà Th.S Nguyễn Cao Trí Th.S Phạm Quang Minh T.S Trần Văn
913805972
913805972
913805972
913805972
913805972
phuoctv@tdmu.edu.vn
phuoctv@tdmu.edu.vn
phuoctv@tdmu.edu.vn
```

Hình 4.2: Bộ câu trả lời – training

- **Bước 3**: Tách từ (tokenize) và lưu bộ từ điển

Từ bộ dữ liệu huấn luyện, tiến hành tách từ vựng và lưu vào 2 file riêng biệt theo quy tắc sau:

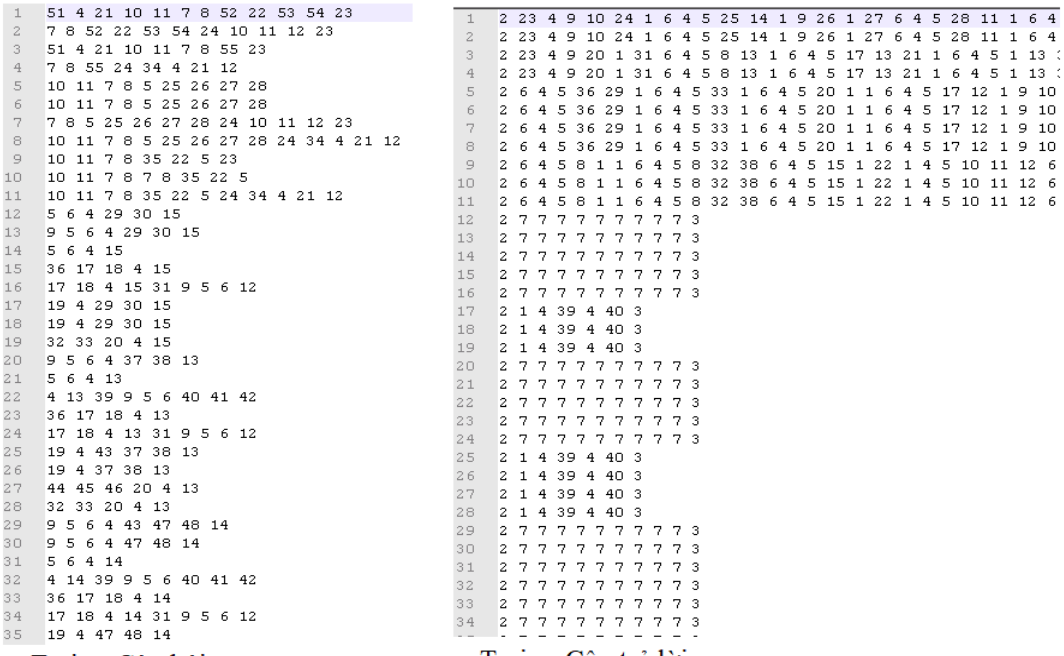
+ Từ thuộc bộ câu hỏi và trả lời được lưu vào hai file riêng biệt

+ Các từ lưu trong file được sắp xếp theo số lượng từ xuất hiện trong bộ câu hỏi và trả lời từ cao đến thấp.

Hình 4.3: Bộ từ điển Câu hỏi – Câu trả lời

- **Bước 4**: Word2vector: tiến hành chỉ mục hóa từ xuất hiện trong bộ từ điển

Chỉ mục hóa các file câu hỏi và trả lời theo các từ xuất hiện trong bộ từ điển



Train - Câu hỏi

Train - Câu trả lời

Hình 4.4: Chỉ mục từ

4.1.3. Huấn luyện dữ liệu:

Thực nghiệm trên bộ dữ liệu chuẩn và bộ dữ liệu thu thập, để huấn luyện tôi sử dụng mô-đun GNMT dựa trên các tế bào LSTM hai chiều. Cơ chế Attention đã được áp dụng để cải thiện hiệu suất. Beamsearch cũng được áp dụng vào đào tạo. Chúng tôi sử dụng 3200 lần lặp với 512 đơn vị ẩn trên mô hình huấn luyện 2 lớp. Chi tiết về mô hình huấn luyện và các tham số của mô hình huấn luyện được trình bày trong các Bảng 4.3 và Bảng 4.4.

Loại	Mô hình sử dụng
Mô hình áp dụng	Google's Neural Machine Translation (GNMT)
Mô hình Deep Learning	Deep Neural Network (DNN), Recurrent Neural Network (RNN)
Kỹ thuật	Sequence to Sequence (Sequence-to-sequence) modeling with encoder and decoder
Các kỹ thuật bổ trợ	Long Short Term Memory (LSTM) based RNN cell, Bidirectional LSTM, Neural Attention Model and Beam Search.

Bảng 4.3: Chi tiết về mô hình huấn luyện

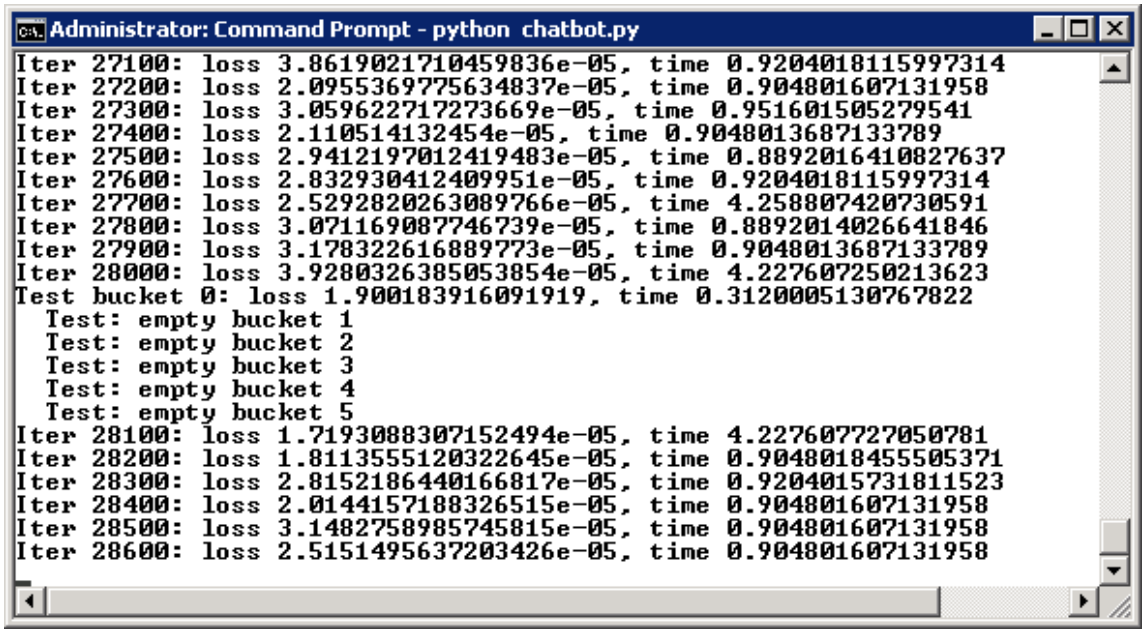
Tham số	Giá trị
Số bước huấn luyện	3200
Số bước lặp cho mỗi lần huấn luyện	100
Số đơn vị	512
Số tầng	2
Loại đơn vị	LSTM
Loại Encoder	Bidirectional
Kỹ thuật Neutral Attention	Scaled luong
Tối ưu hóa	Adam
Tốc độ học	0.0001
Số bước Decay	1
Số bước bắt đầu Decay	1

Chiều rộng của Beam	10
Tỉ lệ Dropout	0.2
Đơn vị đánh giá	BLEU

Bảng 4.4: Tham số của mô hình huấn luyện

Chúng tôi sử dụng máy tính với Bộ xử lý Intel Core i5 thế hệ thứ 7, đồ họa chuyên dụng NVIDIA GTX 1050 và 8GB Ram cho quá trình huấn luyện. Để huấn luyện mở rộng hơn, chúng tôi dự định sử dụng nền tảng điện toán đám mây Amazon AWS.

Quá trình trình huấn luyện được thực hiện như sau:



Hình 4.5: Quá trình huấn luyện

4.1.4. Kết quả

4.1.4.1. Trên bộ dữ Cornell Movie-Dialogs Corpus

Sau đây là một số phản hồi có được sau khi huấn luyện trên tập dữ liệu đầy đủ với dung lượng 34 MB (225000 + 225000). Kết quả thử nghiệm ban đầu tạo ra các câu phản hồi phù hợp. Các câu phản hồi sau đây được tạo ra trên cơ sở mô hình huấn luyện, mô hình huấn luyện sẽ tạo ra 10-30 câu phản hồi cho mỗi ứng viên đầu vào. Các ví dụ dưới đây là phù hợp nhất từ 10-30 phản hồi của mỗi ứng viên trong quá trình suy luận được trình bày trong Bảng 4.5.

Đầu vào (Người 1)	Đầu ra của GNMT (Người 2)
how are you	how am i ?
would you like some tea	would you like some help ?
you got something on your mind	Yeah
would you like to watch movie ?	No
where would you like to go for lunch today	you know where to go ?
where is the library located ?	is that where you were ?
did you like the new startrek movie sequel	i did
am i disturbing you ?	no , not at all . where are you ?
i’m heading out . how about you ?	i got to wait for a call .

Bảng 4.5: Một số kết quả minh họa

4.1.4.2. Trên bộ dữ liệu thu thập

Với bộ dữ liệu gồm 551 câu hỏi và 551 câu trả lời tương ứng. Kết quả cho dự đoán có khả quan.

Đầu vào (Người 1)	**Đầu ra của GNMT (Người 2)**
Đào tạo	Chương trình đào tạo được thực hiện trong 8 học kỳ. Để tốt nghiệp, sinh viên phải hoàn thành đồ án tốt nghiệp vào học kỳ 8. Theo Quyết định số 1157/QĐ-ĐHTDM ngày 08 tháng 08 năm 2014 và bổ sung, sửa đổi theo Quyết định số 890/QĐ-ĐHTDM ngày 04 tháng 08 năm 2016 của Hiệu trưởng Trường Đại Thủ Dầu Một về Quy chế đào tạo đại học và cao đẳng chính quy theo hệ thống tín chỉ.
Tốt nghiệp đại học ngành Hệ thống Thông tin	Theo Quyết định 1862/QĐ-ĐHTDM ngày 21/11/2016 của Hiệu trưởng Trường Đại Thủ Dầu Một về việc ban hành Quy định Chuẩn đầu ra cho học viên cao học, sinh viên đại học, cao đẳng hệ chính quy và thương xuyên đào tạo theo học chế tín chỉ.
Môn thi tốt nghiệp	Môn thi tốt nghiệp bao gồm: chính trị, lý thuyết tổng hợp nghề nghiệp, thực hành nghề nghiệp. Đối với hệ đào tạo mà đối tượng tuyển sinh là học sinh tốt nghiệp trung học cơ sở thì

	ngoài ba môn thi tốt nghiệp nêu trên, học sinh phải thi thêm ba môn văn hóa. Hiệu trưởng nhà trường quy định các môn thi, thời gian tổ chức thi các môn văn hóa và thông báo công khai từ đầu khóa học.
Không được chuyển trường	Sinh viên không được phép chuyển trường trong các trường hợp sau: a) Sinh viên năm thứ nhất và năm cuối khóa b) Sinh viên đang trong thời gian bị kỷ luật từ cảnh cáo trở lên
Thủ tục nhập học cao học khóa 5	https://tdmu.edu.vn/thong-bao/tuyen-sinh/huong-dan-lam-thu-tuc-nhap-hoc-cao-hoc-khoa-5-nam-2018
Kết quả thi tuyển sinh cao học khóa 5	https://tdmu.edu.vn/thong-bao/tuyen-sinh/ket-qua-thi-tuyen-sinh-cao-hoc-khoa-5-2018
Số điện thoại thầy Thanh Hùng	908542521

Bảng 4.6: Kết quả trên Bộ dữ liệu thu thập

4.1.5. Đánh giá

Để đánh giá kết quả chúng tôi sử dụng độ đo BLEU [12]. Đây là độ đo phổ biến được sử dụng trong dịch máy, được đề xuất bởi IBM tại hội nghị ACL ở Philadelphie. Độ đo BLEU sẽ đánh giá chất lượng văn bản trên cơ sở tính toán mức độ tương tự của văn bản đầu vào với các văn bản đầu ra. Kết quả được thể hiện trong Bảng 4.7.

Dữ liệu	BLEU	
	GNMT	NMT
Eval Dev	10.1	8.9
Eval Test	10.6	9.3

Bảng 4.7: Kết quả đánh giá Blue cho bộ dữ liệu Cornell movie-dialogs

4.2. Theo hướng phân loại câu hỏi bằng phương pháp học sâu

4.2.1. Quy trình thực hiện

Bước 1: Chuẩn bị dữ liệu

Với dữ liệu thu thập, các câu hỏi được chuyển vào JSON theo cấu trúc như sau: các câu hỏi được đưa vào trường patterns, các câu trả lời được đưa vào trường responses, lớp câu hỏi và trả lời được đưa vào trường tag.

Bước 2: Từ file JSON, tiến hành loại bỏ các từ dừng và tiến hành tokenize dữ liệu ở tag **Pattern** bằng cách sử dụng thư viện **nltk** véc-tơ hóa các từ.

Bước 3: Khởi tạo các tham số và tiến hành huấn luyện dữ liệu. Sau đó thực hiện lưu mô hình để thực hiện việc dự đoán sau này

Bước 4: Tiến hành đánh giá độ chính xác accuracy, nếu độ chính xác ≤ 95% thì tiếp tục thực hiện lại bước 3. Ngược lại thì chuyển sang bước 5.

Bước 5: Với câu được nhập vào từ người sử dụng, tiến hành loại bỏ các từ dừng và tách từ. Sau đó tiến hành dự đoán câu trả lời dựa trên mô hình đã lưu ở bước 4. Kết quả dự đoán là lớp tương ứng với câu hỏi được đưa vào, khi đó sẽ chọn lựa câu trả lời ngẫu nhiên ứng với lớp đã được dự đoán.

4.2.2. Dữ liệu

Dữ liệu thực nghiệm là dữ liệu (Câu hỏi, Câu trả lời) được thu thập là những câu hỏi thường gặp của sinh viên liên quan đến chương trình đào tạo, hỗ trợ tuyển sinh, thông tin về giảng viên và các văn bản thường gặp của trường Đại học Thủ Dầu Một. Thông tin thu thập được thể hiện như sau:

stt	question	answer	entity_faq_keyword	sample	entity_faq_subject	entity_faq_detail1	entity_faq_detail2
1	Cho em hỏi địa chỉ của Trường Đại Học Thủ Dầu Một?	Trường Đại học Thủ Dầu Một tọa lạc tại số 06 Trần Văn Ơn, Phú Hoà, Thủ Dầu Một, Bình	Địa chỉ, nơi tọa lạc	Trường Thủ Dầu Một ở đâu?	Địa chỉ Trường Đại học Thủ Dầu Một	địa chỉ; đ chỉ; đ/c; đ/chỉ	
				Vị trí trường Thủ Dầu Một		địa chỉ; đ chỉ; đ/c; đ/chỉ	
				Đại học Thủ Dầu Một ở chỗ nào?		địa chỉ; đ chỉ; đ/c; đ/chỉ	
2	Năm học 2018 trường nhận đào tạo bao nhiêu ngành? Đó là những ngành nào?	Năm học 2018 trường nhận đào tạo 30 ngành, đó là: Kế toán; Quản trị Kinh doanh; Tài chính - Ngân hàng; Kỹ thuật Xây dựng; Kỹ thuật Điện; Kỹ thuật Phần mềm - Công nghệ Thông tin; Hệ thống Thông tin - Công nghệ Thông tin; Kiến trúc; Quy hoạch Vùng và Đô thị - Quản Lý Đô thị; Hóa học; Sinh học ứng dụng; Khoa học Môi trường;	Ngành 2018; đào tạo 2018	Những ngành đào tạo năm 2018?	ngành đào tạo 2018	ngành đào tạo; ngành đ tạo; ngành đ/tao; ngành đ/t	
				Năm 2018, trường đào tạo những ngành nào?		ngành đào tạo; ngành đ tạo; ngành đ/tao; ngành đ/t	
				Những ngành trường đào tạo năm 2018 là ngành nào?		ngành đào tạo; ngành đ tạo; ngành đ/tao; ngành đ/t	

Bảng 4.8: Bộ dữ liệu thu thập theo hướng phân loại câu hỏi

Chuyển bộ dữ liệu thu thập sang file JSON để thực hiện việc chuẩn hóa dữ liệu. Đây là bộ dữ liệu phục vụ cho việc huấn luyện để trả lời các câu hỏi của sinh viên, dữ liệu trên file JSON_1 được tổ chức như sau:

+ Tag: lấy từ cột **entity_faq_subject** của file Excel

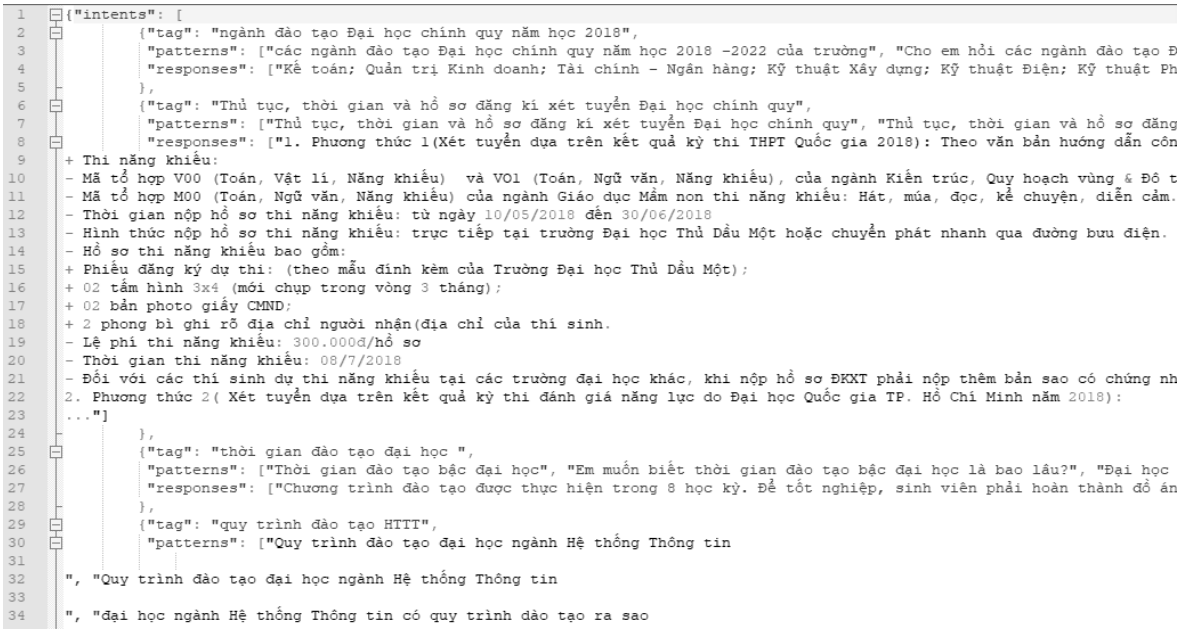
⇨ Phân lớp các câu hỏi khi đưa vào huấn luyện

+ Pattern: lấy từ cột **question** và cột **sample**

⇨ Dữ liệu dùng để training

Response: lấy từ cột **answer**

⇨ Dùng để trả lời các câu hỏi tương ứng của sinh viên



```
{"intents": [
        {"tag": "ngành đào tạo Đại học chính quy năm học 2018",
         "patterns": ["các ngành đào tạo Đại học chính quy năm học 2018 -2022 của trường", "Cho em hỏi các ngành đào tạo Đ
         "responses": ["Kế toán; Quản trị Kinh doanh; Tài chính - Ngân hàng; Kỹ thuật Xây dựng; Kỹ thuật Điện; Kỹ thuật Ph
        },
        {"tag": "Thủ tục, thời gian và hồ sơ đăng kí xét tuyển Đại học chính quy",
         "patterns": ["Thủ tục, thời gian và hồ sơ đăng kí xét tuyển Đại học chính quy", "Thủ tục, thời gian và hồ sơ đăng
         "responses": ["1. Phương thức 1(Xét tuyển dựa trên kết quả kỳ thi THPT Quốc gia 2018): Theo văn bản hướng dẫn côn
+ Thi năng khiếu:
- Mã tổ hợp V00 (Toán, Vật lí, Năng khiếu)  và V01 (Toán, Ngữ văn, Năng khiếu), của ngành Kiến trúc, Quy hoạch vùng & Đô t
- Mã tổ hợp M00 (Toán, Ngữ văn, Năng khiếu) của ngành Giáo dục Mầm non thi năng khiếu: Hát, múa, đọc, kể chuyện, diễn cảm.
- Thời gian nộp hồ sơ thi năng khiếu: từ ngày 10/05/2018 đến 30/06/2018
- Hình thức nộp hồ sơ thi năng khiếu: trực tiếp tại trường Đại học Thủ Dầu Một hoặc chuyển phát nhanh qua đường bưu điện.
- Hồ sơ thi năng khiếu bao gồm:
+ Phiếu đăng ký dự thi: (theo mẫu đính kèm của Trường Đại học Thủ Dầu Một);
+ 02 tấm hình 3x4 (mới chụp trong vòng 3 tháng);
+ 02 bản photo giấy CMND;
+ 2 phong bì ghi rõ địa chỉ người nhận(địa chỉ của thí sinh.
- Lệ phí thi năng khiếu: 300.000đ/hồ sơ
- Thời gian thi năng khiếu: 08/7/2018
- Đối với các thí sinh dự thi năng khiếu tại các trường đại học khác, khi nộp hồ sơ ĐKXT phải nộp thêm bản sao có chứng nh
2. Phương thức 2( Xét tuyển dựa trên kết quả kỳ thi đánh giá năng lực do Đại học Quốc gia TP. Hồ Chí Minh năm 2018):
..."]
        },
        {"tag": "thời gian đào tạo đại học ",
         "patterns": ["Thời gian đào tạo bậc đại học", "Em muốn biết thời gian đào tạo bậc đại học là bao lâu?", "Đại học
         "responses": ["Chương trình đào tạo được thực hiện trong 8 học kỳ. Để tốt nghiệp, sinh viên phải hoàn thành đồ án
        },
        {"tag": "quy trình đào tạo HTTT",
         "patterns": ["Quy trình đào tạo đại học ngành Hệ thống Thông tin

", "Quy trình đào tạo đại học ngành Hệ thống Thông tin

", "đại học ngành Hệ thống Thông tin có quy trình đào tạo ra sao
```

Bảng 4.9: Bộ dữ liệu thu thập - JSON

4.2.3. Xử lí dữ liệu

Để thực hiện việc huấn luyện dữ liệu, tiến hành xử lý bộ dữ liệu theo các bước sau:

- **Bước 1**: Đọc dữ liệu từ file Json và loại bỏ các từ dừng

- **Bước 2**: Tách từ

Tiến hành tách các trong bộ câu hỏi trên file json tương ứng với thẻ patterns

- **Bước 3**: Chỉ mục các từ

Tính chỉ số xuất hiện các từ trong bộ câu hỏi và sắp sếp theo thứ tự từ cao đến thấp theo số lần xuất hiện của các từ.

- **Bước 4**: Gán nhãn các từ theo lớp

Phân lớp các từ đã xuất hiện trong các bộ câu hỏi được quy định trong file json

4.2.4. Huấn luyện dữ liệu

Thực hiện việc huấn luyện dữ liệu trên mô hình với bộ dữ liệu đã được thu thập.

4.2.5. Đánh giá

Trên bộ dữ liệu thu thập:

Độ chính xác: Accurracy = 96.25%

Kết quả:

- Câu hỏi: Đào tạo
  - ⇨ Trả lời: Các hệ đào tạo của trường gồm: Đại học, Liên thông, Thạc sĩ, Tiến sĩ
- Câu hỏi: Hồ sơ nhập học
  - ⇨ Trả lời: 1. Bản chính giấy chứng nhận kết quả thi THPT quốc gia 2018 2.Học phí, lệ phí học kì 1 3. CMND(2 bản photo) 4. Thẻ BHYT (1 bản photo) nếu có
- Câu hỏi: miễn giảm học phí
  - ⇨ Trả lời:

    a) Người có công với cách mạng và thân nhân của người có công với cách mạng theo pháp lệnh ưu đãi người có công với cách mạng số 26/2005/PL-U BTVQH11 ngày 29/6/2005;

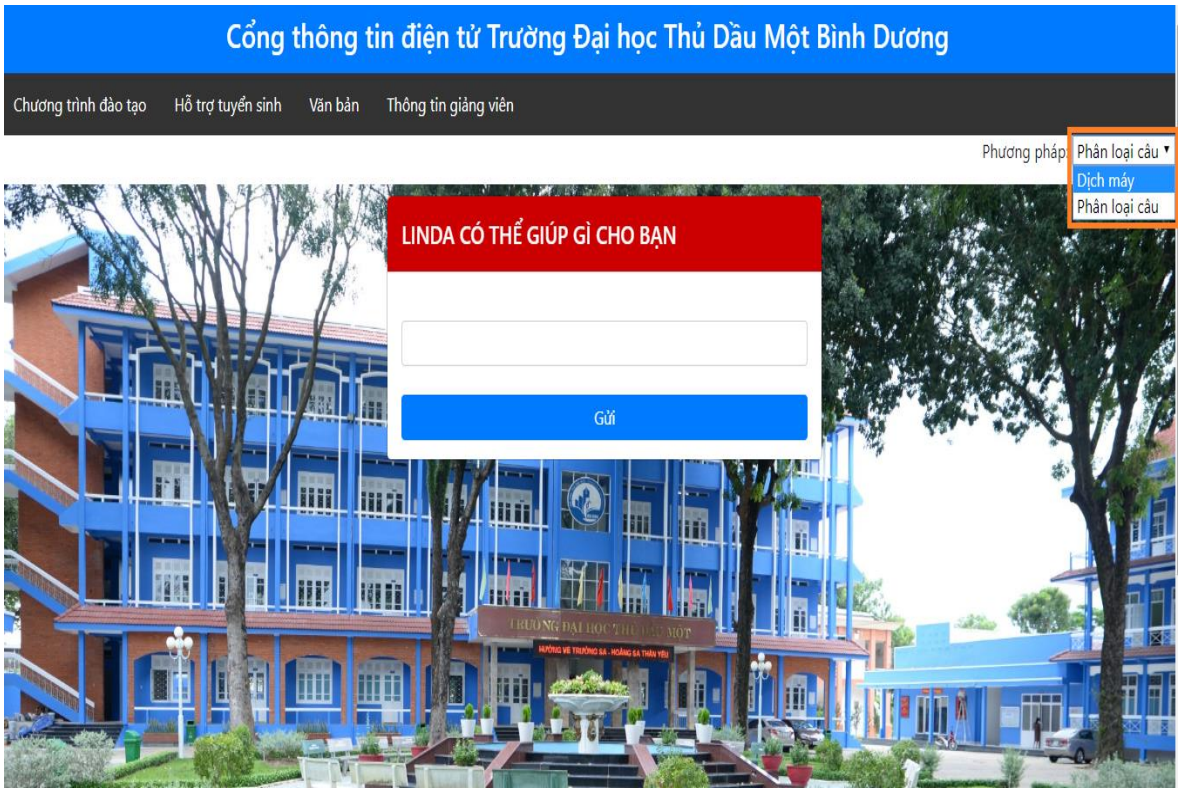
    b) Trẻ em học ….. thuộc hộ nghèo và hộ có thu nhập tối đa bằng 150% thu nhập của hộ nghèo theo tiêu chí của tỉnh.
- Câu hỏi: tín chỉ đại học
  - ⇨ Trả lời: Tín chỉ học tập (gọi tắt là tín chỉ) là đơn vị dùng để xác định khối lượng học tập của sinh viên, qua đó đánh giá kết quả học tập của sinh viên dựa trên số lượng tín chỉ học tập tích lũy được. Một tín chỉ được quy định bằng 15 tiết học lý thuyết (hoặc 30 tiết thực hành, thảo luận; hoặc 45 giờ thực tập tại cơ sở, 60 giờ làm bài tập lớn, khóa luận tốt nghiệp). Một tiết học được tính bằng 50 phút.

So sánh kết quả với phương pháp LSTM:

Phương pháp / Nội dung	Phương pháp học sâu	Mạng LSTM
Accurracy	96.25%	95.17%
Số câu trả lời chính xác theo mong đợi	14/15	12/15
Thời gian huấn luyện	Nhanh hơn	Mất thời gian nhiều hơn khi độ dài vectơ lớn

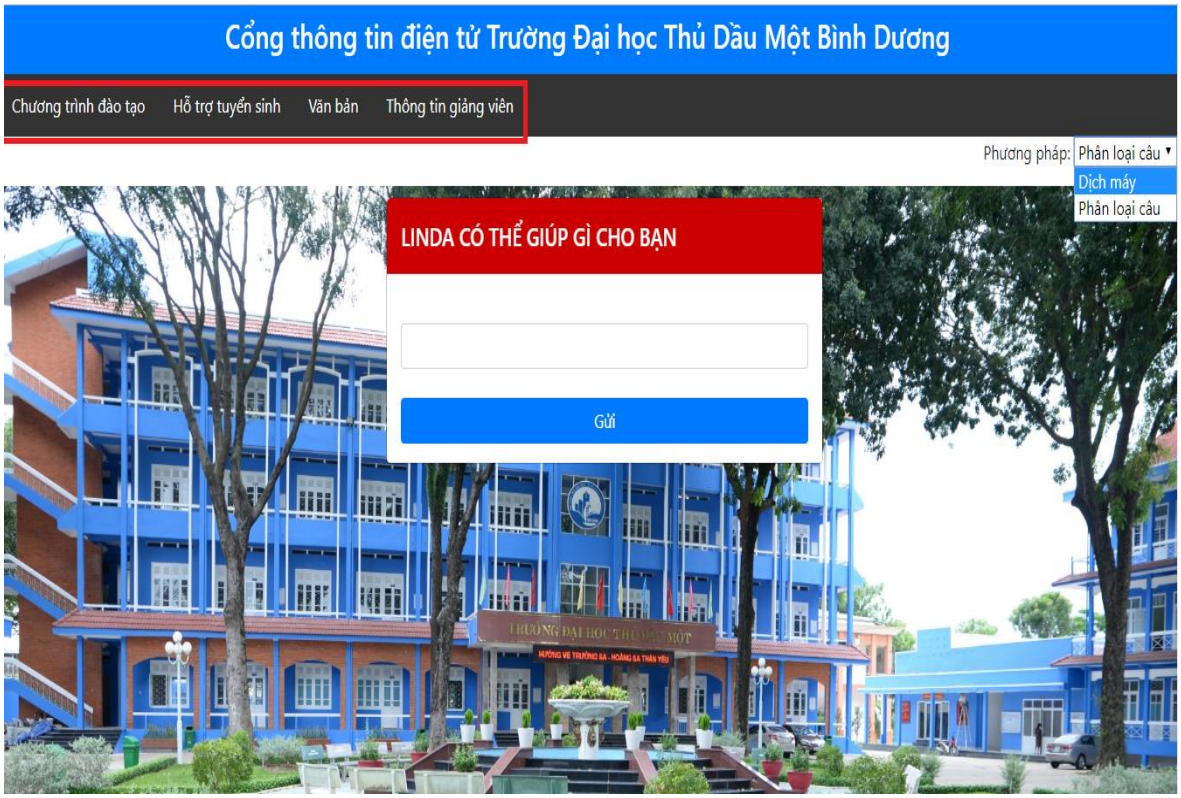
4.3. Mô phỏng ứng dụng Chatbot trên nền tảng web

- Người dùng có thể chọn lựa phương pháp huấn luyện dữ liệu để tra cứu câu hỏi: Phương pháp dịch máy hoặc phương pháp phân loại câu.



Hình 4.6: Giao diện Web - Chọn lựa phương pháp thực hiện

- Người dùng cỏ thể chọn lựa một trong các danh mục để hỏi như Chương trình đào tạo, Hồ sơ truyển sinh, văn bản hoặc thông tin giảng viên



Hình 4.7: Giao diện Web - Chọn lựa mục để hỏi

- Người dùng nhập các câu hỏi để nhận câu trả lời

Hình 4.8: Giao diện Web - Hỏi và trả lời tự động

CHƯƠNG 5
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả đạt được

Hệ thống trả lời tự động là một trong những hướng phát triển nhằm hỗ trợ con người trong việc giảm tải nguồn nhân lực để chăm sóc khách hàng hoặc tư vấn dịch vụ cụ thể nào đó. Trong bối cảnh hiện nay, máy học là một trong các hướng tiếp cận để xây dựng hiệu quả này. Do đó, luận văn đã kế thừa những tinh hoa từ các nghiên cứu của các tác giả để xây dựng hệ thống trả lời tự động nhằm giúp trường đại học Thủ Dầu Một trả lời các câu hỏi của sinh viên về các thông tin liên quan đến nhà trường. Kết quả đã đạt của luận văn gồm:

+ Xây dựng hệ thống trả lời tự động theo hướng dịch máy bằng mạng Nơ-ron của Google: Bộ dữ liệu đầu vào bao gồm các câu hỏi và câu trả lời được phân tách thành hai tập huấn luyện và kiểm tra và tiến hành tách từ và véc-tơ hóa bằng kỹ thuật word2vector để tiến hành huấn luyện dựa trên phương pháp học chuỗi liên tiếp (sequence-to-sequence) và kĩ thuật attention để sinh ra câu trả lời tự động với bộ từ vựng được tham chiếu từ tập huấn luyện. Phương pháp đánh giá BLEU được áp dụng để đánh giá vì kết quả huấn luyện.

+ Xây dựng hệ thống trả lời tự động dựa trên mô hình phân loại câu hỏi theo hướng mạng Nơ-ron sâu: Bộ dữ liệu đầu vào là các câu hỏi và câu trả lời được phân lớp sau khi loại bỏ các từ dừng sẽ được tách từ bằng phương pháp word2vector. Quá trình huấn luyện được tiến hành dựa trên kỹ thuật mạng Nơ-ron sâu thông qua hàm softmax để thể hiện xác suất của lớp và Emtropy chéo được định nghĩa để đánh giá mục tiêu của đầu ra để dự đoán các câu hỏi được đưa vào của sinh viên. Một phương pháp đánh giá dựa trên độ đo chính xác được sử dụng trong mô hình này nhằm đánh giá kết quả để đưa ra mô hình dự đoán tối ưu.

+ Xây dựng ứng dụng dựa trên nền tảng Web-based: Luận văn cũng đã xây dựng một giao diện dựa trên nền tảng Web-based nhằm trực quan kết quả và trả lời tự động các câu hỏi của sinh viên liên quan đến chương trình đào tạo, tuyển sinh, thông tin về giảng viên và các văn bản thường gặp của trường Đại học Thủ Dầu Một.

Với đánh giá BLEU và độ đo chính xác (Accuracy) đã giải quyết mặt hạn chế do dữ liệu thu thập đầu vào không được phong phú. Cả hai mô hình dịch máy và trích xuất thông tin đều có những mặt mạnh và hạn chế riêng nhưng nhìn chung đã đưa ra kết quả dự đoán chính xác. Mô hình dịch máy thời gian huấn luyện lâu hơn nhưng cho kết quả tốt hơn mô hình truy xuất thông tin.

5.2. Hướng phát triển

Tiếp tục kế thừa những nghiên cứu trước đây và phát triển mô hình chatbot mới có khả năng trả lời sát với ngữ cảnh, nhằm làm cho hệ thống trả lời tự động của tôi đạt chất lượng tốt hơn.

Mở rộng mô hình chatbot trong các lĩnh vực khác, thu thập bộ dữ liệu tối ưu nhất nhằm gia tăng tốc độ huấn luyện và tăng độ chính xác cho câu trả lời.

Xây dựng một hệ thống tự động thu thập câu hỏi từ người dùng và có khả năng tự động cập nhật thông tin vào bộ dữ liệu đã có.

TÀI LIỆU THAM KHẢO

[1] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, Sep 2014. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation".

[2] https://en.wikipedia.org/wiki/Hopfield_network

[3] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, "Sequence to Sequence Learning with Neural Networks", pp. 1–9, Dec 2014.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neural Computation, vol. 9, pp. 1735–1780, 1997.

[5] Bahdanau, D., Bengio, Y., & Cho, K., "Neural Machine Translation by Jointly Learning to Align and Translate", CoRR, abs/1409.0473. https://arxiv.org/abs/1409.0473, 2014.

[6] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al, "Google's neural machine translation system: Bridging the gap between human and machine translation", arXiv preprint arXiv:1609.08144, 2016.

[7] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification", In Proc ACL, 2016

[8] Sumit Chopra, Michael Auli, Alexander M Rush, and SEAS Harvard, "Abstractive sentence summarization with attentive recurrent neural networks", Proceedings of NAACL-HLT16 pages 93–98, 2016.

[9] Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria, "Recent Trends in Deep Learning Based Natural Language Processing, IEEE Computational Intelligence Magazine, 2018.

[10] Wang P, Qian Y, Soong F K, He L, Zhao H, "Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network", Cornell University, 2015

[11] Sundermeyer M, Ney H and Schluter R, "From Feedforward to Recurrent LSTM Neural Networks for Language Modelling", J. IEEE/ACM Trans, Audio Speech Lang Process, Issue 3, pp 517–29, 2015

[12] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J, "BLEU: a method for automatic evaluation of machine translation", ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

[13] Nhữ Bảo Vũ, "Xây dựng mô hình đối thoại cho tiếng việt trên miền mở dựa vào phương pháp học chuỗi liên tiếp", đại học quốc gia Hà Nội, trường Đại học Công Nghệ 2016.

[14] Ramamoorthy, S. Chatbots with Seq2Seq. Retrieved from http://suriyadeepan.github.io/2016-06-28-easy-seq2seq/

[15] https://cs224d.stanford.edu/lecture_notes/notes1.pdf

[16] L. Vergeest, “Using N-grams and Word Embeddings for Twitter Hashtag Suggestion”, 2014, Tilburg University (School of Humanities).