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

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

XÂY DỰNG MÔ HÌNH DỊCH MÁY TỪ TIẾNG ANH SANG TIẾNG VIỆT

(Building English-Vietnamese machine translation model)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– TS. Ngô Huy Biên (Khoa Công nghệ Thông tin)

Nhóm Sinh viên thực hiện:

1. Trương Phạm Nhật Tiến (MSSV: 1612689)
2. Nguyễn Minh Trí (MSSV: 1612726)

Loại đề tài: Ứng dụng

Thời gian thực hiện: Từ *11/2019* đến *6/2020*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

- Xây dựng và thu thập dữ liệu đào tạo mô hình dịch một văn bản từ tiếng Anh sang tiếng Việt.
- Vai trò sinh viên: Data Scientist, Data Collector, Developer, Tester, Project Manager/Scrum Master.
- Kỹ năng yêu cầu: Python programming, Machine learning algorithms.
- Ngữ cảnh: Bài toán dịch máy đã được đặt ra từ hơn nửa thế kỷ qua nhưng vẫn đang thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu bởi ý nghĩa thực tiễn to lớn của nó trong sự phát triển của mạng thông tin. Các cách tiếp cận khác nhau đã ra đời và đều đạt được những thành công nhất định. Trong đó, cách tiếp cận thống kê đang được cộng đồng nghiên cứu quan tâm hơn cả bởi tính linh hoạt, mềm dẻo của nó trong việc tự động học các tri thức dịch dựa trên dữ liệu. Bên cạnh đó, mỗi cặp ngôn ngữ đều có những đặc trưng riêng và thông tin ngôn ngữ là yếu tố không thể thiếu góp phần nâng cao chất lượng dịch cho một cặp ngôn ngữ cụ thể.

2.2 Mục tiêu đề tài

- Trình bày lý thuyết nền tảng và giải pháp để xử lý việc dịch một văn bản từ tiếng Anh sang tiếng Việt.
- Xây dựng, thu thập dữ liệu và đào tạo mô hình để dịch một văn bản từ tiếng Anh sang tiếng Việt.
- Xây dựng một trang web demo việc sử dụng mô hình để dịch một văn bản từ tiếng Anh sang tiếng Việt.
- Viết 120 trang luận văn theo đúng chuẩn yêu cầu và trích dẫn các tài liệu tham khảo đầy đủ.

2.3 Phạm vi của đề tài

- Mô hình chỉ dịch từ tiếng Anh sang tiếng Việt.
- Mô hình chỉ sử dụng được khi có kết nối internet.

2.4 Cách tiếp cận dự kiến

2.4.1 Phương pháp dịch trực tiếp

Dịch trực tiếp sẽ thực hiện dịch ngôn ngữ bằng cách thay thế những từ trong ngôn ngữ nguồn với những từ trong ngôn ngữ đích một cách máy móc.

2.4.2 Phương pháp dịch chuyển đổi

- Dịch chuyển đổi cú pháp: Dịch chuyển đổi cú pháp thực hiện phân tích cú pháp câu được nhập vào và sau đó áp dụng những luật ngôn ngữ và từ vựng (hay còn được gọi là những luật chuyển đổi) để ánh xạ thông tin văn phạm từ ngôn ngữ này sang ngôn ngữ khác.
- Dịch chuyển đổi cú pháp cộng phân giải ngữ nghĩa: Dung hoà giữa mức độ phân tích cú pháp và phân giải ngữ nghĩa. Hệ dịch chủ yếu dựa vào phân tích cú pháp, và chỉ phân giải ngữ nghĩa ở mức cần thiết để khử nhập nhằng nghĩa.

2.4.3 Phương pháp dịch máy dựa trên thống kê

Thay vì xây dựng các từ điển, các quy luật chuyển đổi bằng tay, hệ dịch này tự động xây dựng các từ điển, các quy luật dựa trên thống kê. Cách tiếp cận này không đòi hỏi sự phân tích sâu về ngôn ngữ, chúng thực hiện hoàn toàn tự động các quá trình phân tích, chuyển đổi, tạo câu dựa trên kết quả thống kê có được từ kho ngữ liệu.

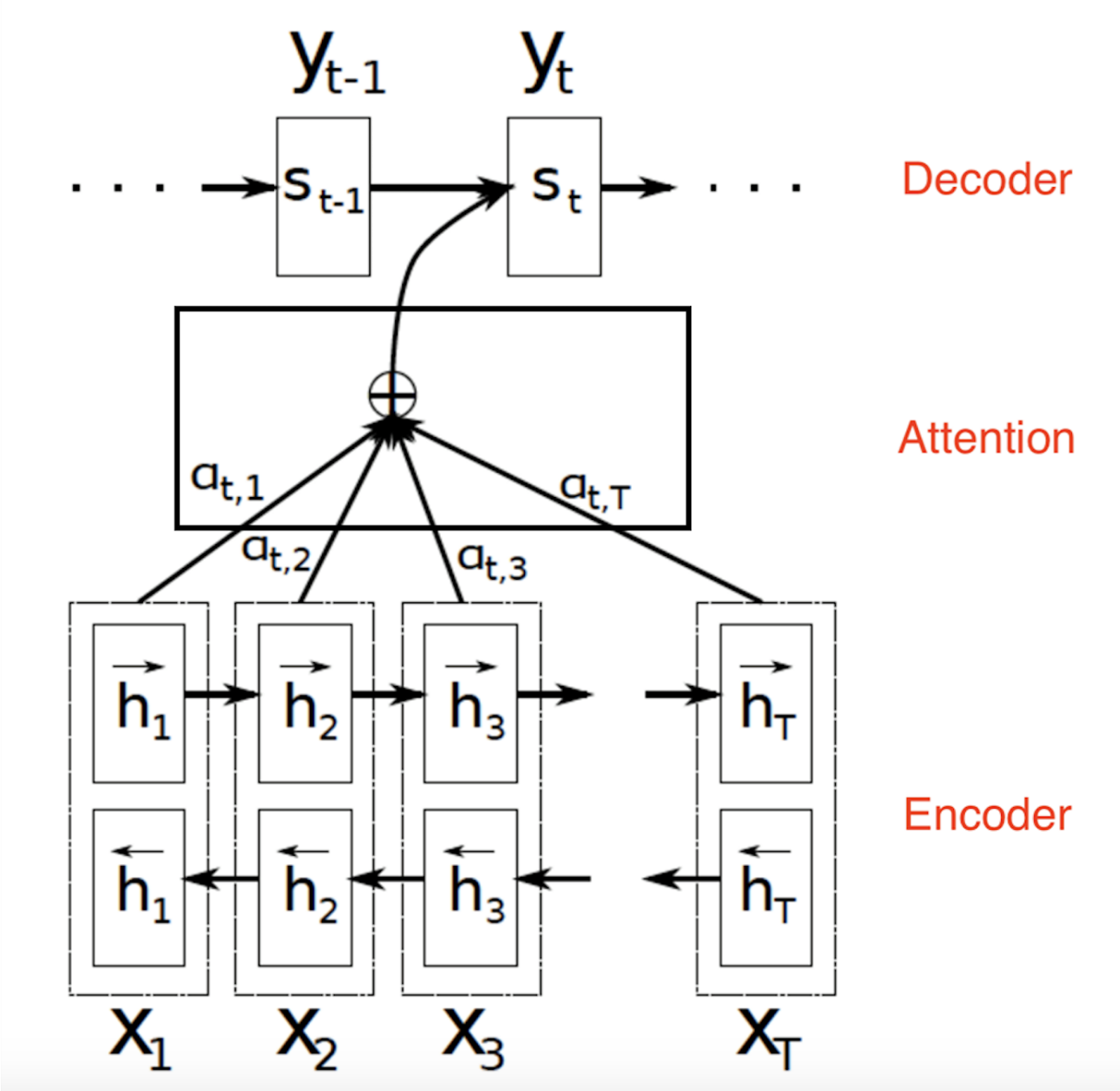
2.4.4 Phương pháp dịch máy dựa trên mẫu ví dụ

Để dịch một câu chúng ta có thể sử dụng kết quả dịch của một câu khác gần giống như vậy sửa đổi đi đôi chút.

2.4.5 Mô hình dự kiến thực hiện trong đề tài

Kiến trúc này dùng 2 mạng RNN(Recurrent neural network với các khối LSTM - Long short-term memory) và cơ chế Attention:

- Mạng Encoder(Bidirectional LSTM) đóng vai trò là mã hoá câu đầu vào thành các vector.
- Mạng Decoder(LSTM) đóng vai trò giải mã và sinh ra câu đầu ra.



- RNN Encoder cho phép chúng ta mã hoá câu đầu vào (ngôn ngữ gốc) thành một dãy các vector, trong đó mỗi vector ứng với một mã hoá của tại mỗi bước(time-step).
- Vùng hình chữ nhật đại diện cho cơ chế Attention là một mạng Neural Network đơn giản. Các trọng số trong mô hình này được tính bằng một mạng neural truyền thẳng. RNN encoder, RNN decoder và các tham số trong kỹ thuật attention được huấn luyện đồng thời từ dữ liệu.
- RNN Decoder sẽ lần lượt sinh từng từ trong chuỗi đầu ra dựa trên vector c(context vector) và những từ được dự đoán trước đó cho tới khi gặp từ kết thúc câu. Nó áp dụng kỹ thuật soft attention bằng cách lấy tổng có trọng số của dãy các vector mã hóa.

2.5 Kết quả dự kiến của đề tài

- Dịch vụ web API của mô hình dịch máy từ tiếng Anh sang tiếng Việt cho phép tải lên một từ, câu, hoặc đoạn văn bản và trả về dạng chữ tiếng Việt của từ, câu, đoạn văn đó.
- Website mẫu việc sử dụng API của mô hình dịch máy từ tiếng Anh sang tiếng Việt đã xây dựng.

2.6 Một số nguồn dữ liệu có sẵn có thể được sử dụng

Bộ dữ liệu IWSLT (The International Workshop on Spoken Language Translation) English-Vietnamese[1] được ra đời năm 2015. Đây là bộ dữ liệu khá nhỏ chỉ nhằm mục đích training và kiểm thử mô hình, từ đó đưa ra các đánh giá và ảnh hưởng của các tham số tới mô hình. Bộ dữ liệu dữ liệu có khoảng 133 ngàn cặp câu và hơn 50 ngàn từ vựng thường xuyên sử dụng nhất.

[1] https://github.com/tensorflow/nmt

Bộ dữ liệu Binhvq News Corpus[2] của tác giả Vương Quốc Bình công tác tại Thường Đại học công nghiệp Hà Nội được trích xuất từ khoảng 14.896.998 bài báo trên internet. Bộ dữ liệu gồm nhiều phần khác nhau như: Only Title có số lượng title là 10.787.976, Full TXT(title + description + body) V1 - số lượng câu lên tới 111.274.300 và được trải qua các xử lí cơ bản. Còn có các phần dữ liệu khác như Full TXT V2, CSV V2 và TXT Categorys.

EVBCorpus[3] - Một Corpus song ngữ Anh-Việt nhiều lớp cho các nhiệm vụ học tập trong ngôn ngữ học so sánh và dịch máy. EVBCopus chứa hơn 20.000.000 từ (20 triệu) từ 15 cuốn sách song ngữ, 100 văn bản song song Anh-Việt / Việt-Anh, 250 văn bản luật và pháp lệnh song song, 5.000 bài báo và 2.000 phụ đề phim. Thành phần, chú thích, mã hóa và tính sẵn có của kho văn bản nhằm tạo điều kiện cho sự phát triển của công nghệ ngôn ngữ và nghiên cứu về trích xuất thuật ngữ song ngữ, chủ yếu cho cặp ngôn ngữ Anh-Việt-Anh.

2.7 Kế hoạch thực hiện

Thời gian thực hiện	Công việc thực hiện	Người thực hiện
20/11/2019 - 25/11/2019	• Nhận đề tài. • Xây dựng bản kế hoạch sơ bộ cho các công việc cần thực hiện.	Tiến, Trí
26/11/2019 - 02/12/2019	• Tìm hiểu và phân tích các yêu cầu về kiến thức cho đề tài. • Khảo sát và dùng thử các hệ thống cung cấp dịch vụ mẫu có sẵn trên thị trường.	Tiến, Trí

[2]https://github.com/binhvq/news-corpus
[3]https://github.com/qhungngo/EVBCorpus

03/12/2019 - 05/12/2019	• Thống nhất nội dung chính của ứng dụng demo việc sử dụng API.	Tiến, Trí
05/12/2019 – 15/02/2020	• Tìm hiểu lý thuyết nền tảng trong máy học. • Tìm hiểu lý thuyết nền tảng trong việc dịch máy.	Tiến, Trí
20/12/2019 - 26/12/2019	• Executive Summary. • Project Vision.	Tiến, Trí
27/12/2019 - 02/01/2020	• Tạo EC2. • Trello.	Tiến, Trí
03/01/2020 - 10/01/2020	• Viết Release Plan. • Product backlog. • Risk Management.	Tiến, Trí
11/01/2020 - 01/02/2020	• Tìm hiểu về các thư viện Scikit-Learn, Tensorflow, Keras.	Tiến, Trí
02/02/2020 – 15/02/2020	• Tìm hiểu các model và kiến trúc, chạy thử các ví dụ để đánh giá.	Tiến, Trí
16/02/2020 - 22/02/2020	• Chạy thử mô hình dịch máy từ tiếng Anh sang các ngôn ngữ khác.	Tiến, Trí
23/02/2020 - 29/02/2020	• Thu thập dữ liệu ngôn ngữ. • Viết chương 1 luận văn.	Tiến, Trí
01/03/2020 - 06/03/2020	• Chỉnh sửa dữ liệu âm thanh. • Tìm hiểu và xây dựng mô hình dịch máy từ tiếng Anh sang tiếng Việt. • Chỉnh sửa chương 1 luận văn.	Tiến, Trí

07/03/2020 - 15/03/2020	• Huấn luyện mô hình. • Viết chương 2 luận văn.	Tiến, Trí
16/03/2020 - 21/03/2020	• Cải tiến mô hình. • Chỉnh sửa chương 2 luận văn.	Tiến, Trí
22/03/2020 - 30/03/2020	• Viết chương 3 luận văn. • Chỉnh sửa chương 3 luận văn.	Tiến, Trí
01/04/2020 - 07/04/2020	• Xây dựng và triển khai hệ thống cung cấp dịch vụ web (API). • Viết chương 4 luận văn.	Tiến, Trí
08/04/2020 - 15/04/2020	• Xây dựng ứng dụng demo việc sử dụng API trên nền tảng web. • Chỉnh sửa chương 4 luận văn.	Tiến, Trí
16/04/2020 - 21/04/2020	• Viết chương 5 luận văn. • Chỉnh sửa chương 5 luận văn.	Tiến, Trí
21/04/2020 - 30/04/2020	• Hoàn thành luận văn. • Chỉnh sửa và cải thiện hiệu năng ứng dụng demo.	Tiến, Trí
01/05/2020 - 30/05/2020	• Nâng cấp mô hình hoàn thiện hơn. • Cải thiện hiệu năng hệ thống cung cấp dịch vụ web(API).	Tiến, Trí
03/06/2019 - 19/06/2019	• Hoàn chỉnh cuốn luận văn.	Tiến, Trí
20/06/2019 - 30/06/2019	• Hoàn chỉnh slide trình bày. • Hoàn chỉnh sản phẩm khoá luận.	Tiến, Trí

Tài liệu

[1] A. Geron, *Hand-On Machine Learning with Scikit-Learn and TensorFlow*. Published by O'Reilly Media, 2017.

[2] A. Geitgey, "Machine learning is fun part 5: Language translation with deep learning and the magic of sequences." `https://tinyurl.com/mgltqjq`.

[3] I. Pestov, "A history of machine translation from the cold war to deep learning." `https://tinyurl.com/yxfwtsdg`.

[4] "Machine translation." https://paperswithcode.com/task/machine-translation.

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

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