TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỀ CƯƠNG CHI TIẾT LUẬN VĂN THẠC SĨ**

**CẢI THIỆN DỊCH MÁY NƠ-RON**

**VIỆT-ANH BẰNG GIÓNG HÀNG TỪ HƯỚNG DẪN**

*Người hướng dẫn*: **TS. Nguyễn Chí Thiện**

*Người thực hiện*: **Nguyễn Duy Hàn Lâm (MSHV: 196005004)**

Khoá  **: 2019**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỀ CƯƠNG CHI TIẾT LUẬN VĂN THẠC SĨ**

**CẢI THIỆN DỊCH MÁY NƠ-RON**

**VIỆT-ANH BẰNG GIÓNG HÀNG TỪ HƯỚNG DẪN**

*Người hướng dẫn*: **TS. Nguyễn Chí Thiện**

*Người thực hiện*: **Nguyễn Duy Hàn Lâm (MSHV: 196005004)**

Khoá  **: 2019**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

**ĐỀ CƯƠNG CHI TIẾT ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đề cương chi tiết của riêng tôi và được sự hướng dẫn của TS. Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 17 tháng 11 năm 2020*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Duy Hàn Lâm*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN ii](#_Toc56767063)

[MỤC LỤC 1](#_Toc56767064)

[TÓM TẮT NỘI DUNG 3](#_Toc56767065)

[Chương 1: QR CODE 4](#_Toc56767066)

[1.1 Barcode 4](#_Toc56767067)

[1.2 QR code 7](#_Toc56767068)

[1.3 Cấu tạo của QR code 7](#_Toc56767069)

[1.3.1 Phân vùng cấu hình 9](#_Toc56767070)

[1.3.2 Hiệu chỉnh sai số 11](#_Toc56767071)

[1.3.3 Mã hóa 12](#_Toc56767072)

[1.4 Tổng kết chương 1 14](#_Toc56767073)

[CHƯƠNG 2: QR CODE VÀ ỨNG DỤNG 15](#_Toc56767074)

[2.1 Ứng dụng của QR code 15](#_Toc56767075)

[2.2 Xây dựng ứng dụng đăng nhập bằng QR code 15](#_Toc56767076)

[2.3 Một số rủi ro của QR code 19](#_Toc56767077)

[2.4 Tổng kết chương 2 19](#_Toc56767078)

[TÀI LIỆU THAM KHẢO 20](#_Toc56767079)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1.1. Minh họa về barcode 4](#_Toc56767080)

[Hình 1.2. Kết quả barcode sau khi thực hiện 6](#_Toc56767081)

[Hình 1.3. Cấu tạo của 1 QR code [1] 7](#_Toc56767082)

[Hình 1.4. Lưu trữ giá trị khả năng phục hồi dữ liệu 8](#_Toc56767083)

[Hình 1.5. Các mức phục hồi dữ liệu [1] 9](#_Toc56767084)

[Hình 1.6. Mô tả phân vùng cấu hình của QR code [1] 10](#_Toc56767085)

[Hình 1.7. Các loại phân phối dữ liệu của QR code [1] 10](#_Toc56767086)

[Hình 1.8. Chiều mã hóa của thông tin trong QR code 11](#_Toc56767087)

[Hình 1.9. Thứ tự ghi các bit mã hóa của dữ liệu trong các trường hợp đặc biệt [1] 12](#_Toc56767088)

[Hình 1.10. Thứ tự ghi các bit mã hóa của dữ liệu trong QR code 13](#_Toc56767089)

[Hình 2.1. Giao diện trang web đăng nhập 15](#_Toc56767090)

[Hình 2.2. Tiến hành đăng nhập vào hệ thống 16](#_Toc56767091)

[Hình 2.3. Đăng nhập thành công để lấy QR code 17](#_Toc56767092)

[Hình 2.4. Đăng nhập bằng QR code 18](#_Toc56767093)

[Hình 2.5. Đăng nhập thành công bằng QR code 18](#_Toc56767094)

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

TÓM TẮT

Hiện nay việc áp dụng các giải thuật, mô hình vào quá trình dịch đang là một lĩnh vực nghiên cứu nhằm giúp cải thiện và tăng tính tự động của quá trình dịch. Quá trình áp dụng các giải thuật hay mô hình vào quá trình dịch được gọi tắt là dịch máy.

Tài liệu này xin trình bày mô hình dịch máy nơ-ron transformer làm chủ đạo trong quá trình thực hiện. Mô hình gióng hàng từ (word alignment) sẽ được dùng trong quá trình xây dựng mô hình transformer.

Ngoài ra, trong cơ chế attention của mô hình encoder-decoder ta sẽ sử dụng thêm các yếu tố về phân loại.

Bằng việc kết hợp này, ta hy vọng có thể cải thiện được chất lượng của quá trình dịch máy từ tiếng Việt sang tiếng Anh.

Quá trình thực nghiệm sẽ được tiến hành trên bộ dữ liệu Việt – Anh của Stefan Shcweter tại github: <https://github.com/stefan-it/nmt-en-vi/tree/master/data>.

Sau quá trình thực nghiệm, ta sẽ nghiệm thu, đánh giá mô hình dựa trên kết quả đạt được.

Chương 1: TỔNG QUAN ĐỀ TÀI

* 1. Tính cấp thiết của đề tài
  2. Mục tiêu nghiên cứu
  3. Tổng quan nghiên cứu
  4. Đối tượng nghiên cứu
  5. Phạm vi nghiên cứu
  6. Phương pháp nghiên cứu
  7. Cấu trúc của đề cương

Chương 2: MÔ HÌNH GIÓNG HÀNG TỪ

2.1 Giới thiệu về dịch máy thống kê

2.2 Kĩ thuật gióng hàng từ

2.3 Các mô hình dịch máy sử dụng kĩ thuật gióng hàng từ của IBM

2.4 Kĩ thuật gióng hàng từ nhanh dựa trên mô hình 2 của IBM

Chương 3: MÔ HÌNH DỊCH MÁY NƠ-RON

3.1 Cấu trúc của 1 mạng nơ-ron cơ bản

3.2 Mô hình dịch máy nơ-ron sequence-to-sequence (encoder-decoder) cơ bản

3.3 Encoder 2 chiều (Bidirectional RNN Encoder)

3.4 Kĩ thuật soft alignment trong decoder

3.5

Chương 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

TÀI LIỆU THAM KHẢO

**Tiếng Anh:**

1. <https://en.wikipedia.org/wiki/QR_code>.