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

**CÁC KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

**CÁC CHỮ VIẾT TẮT**

MT Machine Translation

RBMT Rule-based Machine Translation

SMT Statistical Machine Translation

NMT Neural Machine Translation

LSTM Long-short Term Memory

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.

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

Dịch thuật là sự truyền đạt ý nghĩa của văn bản ngôn ngữ nguồn sang văn bản ngôn ngữ mong muốn tương đương [1].

Dịch thuật cần người có am hiểu ngôn ngữ của cả ngôn ngữ gốc (ngôn ngữ nguồn) của văn bản lẫn ngôn ngữ mong muốn (ngôn ngữ đích) mà nhu cầu đời sống cần sử dụng. Vì thế dịch thuật là một quá trình tốn rất nhiều thời gian và công sức. Vì thế với sự phát triển của thời buổi công nghệ hiện nay, ta có thể thực hiện việc dịch văn bản một cách tự động bằng các chương trình trên máy tính, với hy vọng ta có thể làm giảm chi phí thời gian, công sức, lẫn con người của quá trình dịch thuật.

Tuy nhiên, dịch thuật bằng máy (dịch máy, Machine translation (MT)) vẫn còn nhiều hạn chế. Các yếu tố này thường là do: một từ có nhiều nghĩa giữa cả 2 ngôn ngữ, sự khác biệt trong cấu trúc ngữ pháp của cả 2 ngôn ngữ, cách vận dụng, tình huống xảy ra trong văn bản của ngôn ngữ nguồn,… Đặc biệt là việc dịch từ tiếng Việt sang tiếng Anh. Vì tiếng Việt được mọi người trên thế giới đánh giá là một ngôn ngữ có mức độ khó và phức tạp về từ vựng, ngữ pháp, ý nghĩa.

Dựa trên cách thực hiện, dịch máy thường có thể được phân thành các hướng sau: dịch máy dựa trên tri thức hay kinh nghiệm của các chuyên gia trong lĩnh vực dịch thuật (Rule-based machine translation (RBMT)), dịch máy dựa trên văn bản bằng các tính toán lẫn phân tích về mặt thống kê (Statistical machine translation (SMT)), dịch máy kết hợp giữa rule-based machine translation và statistical machine translation được gọi là Hybrid machine translation, cuối cùng một phương hướng mới đó là sử dụng các mô hình mạng thần kinh (nơ-ron – nerual) nhân tạo (Neural machine translation – NMT).

Với các khó khăn hiện nay đã nêu ở trên, việc tìm ra những phương pháp giúp cải thiện quá trình dịch máy là một điều cần thiết. Đề tài luận văn xin được tập trung giới hạn nghiên cứu cải thiện dịch máy từ văn bản ngôn ngữ tiếng Việt sang ngôn ngữ tiếng Anh bằng mô hình transformer (một dạng của mô hình mạng neural) kết hợp với kỹ thuật gióng háng từ hướng dẫn (alignment). Vì thế đề tài có tên là: “Cải thiện dịch máy nơ-ron Việt-Anh bằng gióng hàng từ hướng dẫn.

* 1. Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của đề tài đó chính là sử dụng kỹ thuật gióng hàng từ vào mô hình mạng neural để cải thiện quá trình dịch máy từ văn bản tiếng Việt thành văn bản tiếng Anh. Mục tiêu cụ thể gồm:

* Tiền xử lý và tạo ra một tập dữ liệu phù hợp cho việc thực hiện kỹ thuật gióng hàng từ cũng như cho quá trình huấn luyện của mô hình transformer
* Chọn mô hình mạng neural và kỹ thuật gióng hàng từ cho quá trình dịch.
* Tiến hành thực nghiệm, đánh giá và so sánh giữa 2 phương pháp: phương pháp sử dụng mô hình mạng neural thuần với phương pháp sử dụng mô hình neural kết hợp với kỹ thuật alignment (cả 2 phương pháp sẽ đều dùng cùng một loại mô hình mạng neural).
  1. Tổng quan nghiên cứu

Dịch máy hiện nay thường được nghiên cứu dựa trên các mô hình mạng nơ-ron nhân tạo. Các mô hình nơ-ron này hoạt động dựa trên các cấu trúc hồi quy làm cốt lõi để trích xuất và nắm bắt cách phụ thuộc của các yếu tố trong cả 2 ngôn ngữ nguồn và đích được đem ra để dịch. Phương pháp chủ yếu của các giải thuật này đó chính là tối đa hóa các xác suất phụ thuộc trong quá trình huấn luyện, để từ đó mô hình có thể cho ra được kết quả dịch gần đúng và hợp lý.

Các mô hình mạng nơ-ron sơ khai ban đầu là các mạng nơ-ron với mỗi nhân trong mô hình là 1 hàm tuyến tính cơ bản, sau đó phát triển và được thay thế bằng các hàm phi tuyến tính. Theo thời gian, cấu trúc nhân của các mạng nơ-ron cũng được phát triển từ 1 hàm tính toán thành 1 tổ hợp các hàm tính toán nhằm làm giảm các hạn chế của việc chỉ sử dụng 1 hàm đơn thuần. Các cấu trúc tiêu biểu này thường là: Long-short Term Memory (LSTM) của Sepp Hochreiter; Jürgen Schmidhuber (1997), Gated Recurrent Unit (GRU) của Kyunghyun Cho và các cộng sự (2014),…

Bên cạnh sự phát triển nghiên cứu về nhân của nơ-ron, là sự phát triển trong nghiên cứu về cấu trúc của các mạng nơ-ron. Tiêu biểu là các mô hình Convolutional Neural Network (CNN) được Yann Lecun và cộng sự giới thiệu trong những thập niên 1980, mô hình Sequence-to-Sequence của Sutskever Ilya, Vinyals Oriol, Lê Quốc Việt (2014), Transformer của Ashish Vaswani và các cộng sự (2017),…

Thêm vào đó còn có các nghiên cứu về các kỹ thuật nhằm cải thiện chất lượng dịch. Tiêu biểu là các kỹ thuật alignment của Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio (2014), attention của Minh-Thang Luong, Hieu Pham, Christopher D. Manning (2015),…

Dựa trên các thành tựu nghiên cứu trên, ta sẽ tiến hành đi nghiên cứu và kết hợp các phương pháp nhằm cải thiện chất lượng dịch, nhất là quá trình dịch từ tiếng Việt sang tiếng Anh. Đó cũng chính là mục tiêu hướng tới của đề tài.

* 1. Đối tượng nghiên cứu

Các mô hình dịch máy phổ biến và cơ bản.

Dữ liệu văn bản tiếng Việt, tiếng Anh: <https://github.com/stefan-it/nmt-en-vi/tree/master/data>.

Kết quả giữa các mô hình dịch máy.

* 1. Phạm vi nghiên cứu

Nghiên cứu và ứng dụng các mô hình dịch máy phổ biến và cơ bản.

Sử dụng tập dữ liệu Việt-Anh phổ biến: <https://github.com/stefan-it/nmt-en-vi/tree/master/data> để thực nghiệm, đánh giá.

* 1. Phương pháp nghiên cứu

Tiến hành sử dụng dữ liệu để thực nghiệm và đánh giá kết quả của các mô hình dịch máy, sử dụng kết hợp với kỹ thuật gióng hàng từ (alignment). Từ đó rút ra kết luận cho đề tài.

* 1. Ý nghĩa của đề tài

Đóng góp vào việc cải thiện chất lượng của quá trình dịch máy nói chung cũng như quá trình dịch từ tiếng Việt sang tiếng Anh nói riêng nhằm góp phần đưa Việt Nam hội nhập hơn với thế giới.

* 1. Cấu trúc của đề cương

Đề tài được trình bày thành 4 chương:

Chương 1: Trình bày lý do nghiên cứu, mục tiêu nghiên cứu, tổng quan, phạm vi, đối tượng nghiên cứu cũng như ý nghĩa nghiên cứu của đề tài.

Chương 2: Trình bày cơ sở lý thuyết của kỹ thuật gióng hàng từ.

Chương 3: Trình bày cơ sở lý thuyết chung về mô hình no-ron cũng như mô hình nơ-ron cụ thể được thực nghiệm trong đề tài.

Chương 4: Trình bày quá trình thực nghiệm, kết quả sau khi thực nghiệm, tiến hành đánh giá, kết luận dựa trên kết quả thực nghiệm.

Chương 2: MÔ HÌNH GIÓNG HÀNG TỪ

2.1 Giới thiệu về dịch máy thống kê

Dịch máy thống kê (Statistical Machine Translation – SMT) là mô hình dịch tự động dựa trên các tham số thống kê được tính toán dựa trên các phân tích thống kê từ các tập dữ liệu văn bản song ngữ.

SMT lần đầu được giới thiệu bởi Warren Weaver vào năm 1947. Ông cho rằng các suy diễn logic của ngôn ngữ có thể được xem như các suy diễn logic của toán học. Kết quả dịch ở tập văn bản ngôn ngữ đích có thể được xác định dựa trên thống kê xuất hiện của các từ trên tập văn bản ngôn ngữ nguồn.

Quá trình dịch của SMT bắt đầu bằng các tập văn bản song ngữ đã được dịch một cách chính xác. Từ đây ta sẽ tiến hành đem các tập văn bản này đi thực hiện các phép thống kê để tạo ra 1 mô hình thống kê. Sau khi tạo được mô hình thống kê, ta sẽ tiến hành dịch các văn bản mong muốn dựa trên mô hình này. Kết quả dịch của văn chưa được dịch này sẽ tập các từ có xác suất phù hợp với tần số xuất hiện của các từ tương ứng trong văn bản nguồn.

Cho 1 cặp câu song ngữ với là câu của ngôn ngữ nguồn và là câu của ngôn ngữ đích được dịch từ tương ứng. Ta đi xây dựng mô hình dịch sao cho xác suất kết quả dịch của ngôn ngữ đích tương ứng có xác suất là cao nhất.

Ví dụ ta có câu trong văn bản tiếng Việt cần dịch là “Tôi ăn cơm” , ta xây dựng mô hình sao cho xác suất dịch xảy ra “I eat rice” lớn hơn xác suất dịch của “I eat dinner” , nói cách khác chính là .

Ta có kết quả dịch mong đợi:

Với là kết quả tính được của mô hình.

Điều ta mong muốn đó là có giá trị là cao nhất dẫn đến phải có giá trị cao nhất.

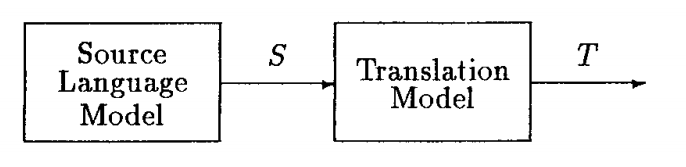
Ứng với mỗi của văn bản ngôn ngữ đích ta sẽ tiến hành đi tìm phù hợp ở văn bản ngôn ngữ nguồn có xác suất là cao nhất. Ta có xác suất Baye:

Với:

* là xác xuất của dựa vào .
* là xác suất số lần xuất hiện của trong tập dữ liệu.
* là xác suất số lần xuất hiện của khi biết .

Ta dễ dàng thấy không phụ thuộc vào , được tính dựa trên số lần xuất hiện của trong tập dữ liệu văn bản đích. Vì thế để đạt được cao nhất thì phải đạt được giá trị cao nhất.

Hệ thống dịch máy thống kê gồm 2 phần cơ bản: mô hình dịch (translation model), bộ giải mã (decoder).



1. Minh họa cho quá trình dịch máy thống kê [2]

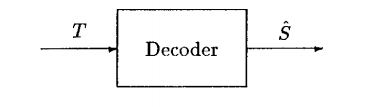
Trong hình 1.1 ta có thể thấy đầu vào của mô hình dịch (translation model) chính là các câu cần dịch và đầu ra là các kết quả dịch ứng với . Mô hình dịch sẽ tính toán các giá trị xác suất và từ đó tìm ra các giá trị khả thi sao cho giá trị đạt được là lớn nhất.

Ta đặt:

Nhiệm vụ của mô hình dịch chính là sử dụng giải thuật thống kê để tối ưu hóa .

Sau khi đã tính được từ mô hình dịch, ta sẽ đem đi thực hiện dịch bằng bộ giải mã (decoder).

Nhiệm vụ của decoder đó chính là ứng với mỗi câu T ở ngôn ngữ đích, decoder sẽ chọn các câu S sao cho giá trị là cực đại.



1. Minh họa cho quá trình decode [2]

Hình 1.2 ở trên là một minh họa về decoder. Đầu vào của decoder chính là các câu dịch ở ngôn ngữ đích, đầu ra là câu ngôn ngữ nguồn tương ứng, với:

Giá trị sẽ được tính dựa trên mô hình ngôn ngữ. Như ta đã biết là câu của ngôn ngữ nguồn, là một chuỗi gồm các từ .

2.1.1 Mô hình dịch (translation model) – kĩ thuật gióng hàng từ

Với cặp , (“Tôi uống trà”, “I drink tea”), ta có thể dễ dàng thấy rằng “Tôi” sẽ được dịch là “I”, “uống” sẽ được dịch là “drink”, “trà” sẽ được dịch là “tea”. Với cách dịch từ theo từ (word by word) như vậy, sẽ được gọi là gióng hàng từ, “Tôi” gióng thành “I”, “uống” gióng thành “drink”, “trà” gióng thành “tea”. Tuy nhiên, không phải lúc nào các cặp đều đơn giản như vậy. Ví dụ ta có cặp , (“Tôi đi ngủ”, “I go to sleep”), “Tôi” gióng thành “I”, “ngủ” gióng thành “sleep”, tuy nhiên “đi” gióng thành cho cả “go” và “to”. Đôi khi một từ trong tiếng Việt phải gióng với nhiều hoặc ít hơn một từ của tiếng Anh và ngược lại, đôi khi một từ của tiếng Anh phải gióng với nhiều hơn hoặc ít hơn một từ của tiếng Việt.

Ta đặt số lượng từ của được gióng với 1 từ của là fertility. Ta xét ví dụ: (“Tôi đi ngủ”, “I go to sleep”), , , . Xác suất của gióng hàng từ được tính bằng:

Với chính là xác suất gióng hàng từ của mỗi cặp gióng hàng từ.

Ví dụ:

Ta thấy rõ ràng trong cả 2 trường hợp (“Tôi uống trà”, “I drink tea”) và (“Tôi đi ngủ”, “I go to sleep”), vị trí của các từ Việt-Anh đều nằm ở gần nhau. Tuy nhiên, trong một vài trường hợp, một từ tiếng Việt khi dịch qua tiếng Anh, vị trí của 2 từ Việt-Anh này có thể nằm cách xa nhau. Những trường hợp này được gọi là “distortion” (dị biệt). Xác suất của distortion được tính bằng , với là vị trí của từ tiếng Anh tương ứng khi gióng, là vị trí của từ tiếng Việt, là số lượng từ gióng tương ứng.

2.1.2 Quá trình tìm kiếm

Quá trình tìm kiếm dựa trên kĩ thuật tìm kiếm ngăn xếp, ví dụ ta có câu: “Tôi uống trà”, quá trình tìm kiếm tuần tự sẽ là:

* .

Quá trình kết thúc khi kết quả dịch phù hợp với điều kiện yêu cầu.

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

Các mô hình được giới thiệu trong mục này được trích từ bài báo “The Mathematics of Statistical Machine Translation: Parameter Estimation” của nhóm tác giả ở trung tâm nghiên cứu T.J Watson của IBM (1993).

Các mô hình này chủ yếu giải quyết dựa trên xác suất đồng thời (joint distribution) , trong đó:

* là câu của ngôn ngữ nguồn.
* là câu của ngôn ngữ đích được dịch từ tương ứng.
* là biến ngẫu nhiên đại diện cho việc gióng hàng từ giữa và .

Ngoài ra còn có các biến và để biểu diễn cho độ dài của chuỗi và .

sẽ được viết lại thành:

là kết quả của phép tổng dựa trên tập các gióng hàng từ .

Việc gióng hàng từ của các mô hình cũng được giới hạn lại trong phạm vi 1 từ của sẽ chỉ gióng với 1 từ của hoặc không gióng với từ của . Nếu có từ và có từ, thì gióng hàng từ sẽ được biểu diễn thành cặp gióng hàng từ . Nếu 1 từ của ở vị trí gióng với 1 từ của ở vị trí , thì , còn nếu 1 từ của t ở vị trí j không gióng với bất kỳ từ nào của s thì .

Ta có:

Trong , ta có thể thấy rằng việc đầu tiên trước khi dịch đó là ta xem kết quả dịch sẽ gồm mấy từ. Ví dụ “Tôi đi ngủ” có thể dịch thành “I sleep” hoặc “I go to sleep” . Vì thế ta sẽ tính đầu tiên.

Tiếp theo ta sẽ tính tích xác suất của các cặp gióng hàng từ. Với một cặp gióng hàng từ ta sẽ tính tích của 2 giá trị xác suất. Giá trị thứ nhất chính là xác suất xảy ra từ thứ của sẽ được gióng với từ thứ của . Giá trị xác suất thứ hai chính là xác suất xảy ra sẽ được dùng để dịch.

2.2.1 Mô hình 1 của IBM

Trong mô hình 1, nhóm tác giả của IBM đã giả định rằng trong (Eq1), và là độc lập với nhau, vì thế sẽ chỉ còn phụ thuộc vào , chiều dài của chuỗi ngôn ngữ nguồn. Mỗi từ trong ngôn ngữ đích đều có khả năng không gióng với từ nào với ngôn ngữ nguồn hoặc có khả năng gióng từ 1 cho đến từ của ngôn ngữ nguồn, vì thế:

Còn xác suất xảy ra từ dịch sẽ chỉ còn phụ thuộc vào và .

Ta đặt:

Biểu thức sẽ trở thành:

Ứng với mỗi cũng đều có khả năng gióng từ 0 cho đến n từ của ngôn ngữ nguồn, vì thế:

Ta muốn kết quả dịch của ta là:

Vì thế ta sẽ vận dụng nhân tử Lagrange để đi tìm mục tiêu, sẽ trở thành:

Để tìm cực đại, ta sẽ tiến hành đạo hàm theo :

Rút gọn lại và giải phương trình, ta có điểm cực đại cần tìm:

Với và là hàm Kronecker :

Kết hợp , , , ta có thể rút gọn lại thành:

Ta có chính là tổng số lần mà gióng với .

Ta đặt số lần gióng mong đợi mà s gióng với t trong phép dịch chính là phép đếm (“”) của khi biết trong phép dịch, ta ký hiệu là .

Với:

Nếu ta chọn:

Thì biểu thức , trở thành:

Do phép dịch là tập hợp của phép dịch, nên:

Biểu thức trước khi áp dụng nhân tử Lagrange thì ta có thể rút gọn lại thành:

Ta thế biểu thức cho biểu thức và áp dụng Lagrange tương tự, ta sẽ được:

Ta sẽ kết hợp biểu thức và để tiến hành ước lượng cho các tham số theo trình tự sau:

1. Chọn các giá trị khởi tạo cho .
2. Ứng với mỗi cặp câu ta sẽ tính biểu thức . Các phép đếm trong khác không khi các và đều là các từ nằm trong . không phụ thuộc vào thứ tự sắp xếp của các từ, chỉ phụ thuộc vào số lần xuất hiện của các từ trong các cặp câu tương ứng.
3. Ứng với mỗi trong :

* Tính giá trị:
* Ứng với mỗi trong ta dùng để tính các giá trị .

1. Lặp lại bước 2 và 3 cho đến khi kết quả hội tụ đến giá trị mà ta mong muốn.

2.2.2 Mô hình 2 của IBM

Trong mô hình 1 của IBM, chúng ta không hề quan tâm đến vị trí của các từ trong các cặp câu dịch. Vì thế trong mô hình 2 chúng ta sẽ vừa kế thừa mô hình 1 vừa giả định rằng phụ thuộc vào và .

Ta đặt

Là xác suất gióng hàng từ, thỏa mãn:

Biểu thức được viết lại thành:

Áp dụng nhân tử Lagrange cho biểu thức , ta được:

Phép đếm mới sẽ là:

Áp dụng tương tự và , ta có:

Nếu cặp không có độ dài là hoặc không có độ dài là thì phép đếm tương ứng của sẽ bằng 0.

Ta có thể rút gọn lại thành:

Với biểu thức trên ta tạo được phép đếm:

Mô hình 1 là trường hợp đặc biệt của mô hình 2 vì trong mô hình 1 ta không quan tâm đến thứ tự gióng của phép dịch. Vì thế từ mô hình 1 mà ta có thể xây dựng nên mô hình tổng quát cho mô hình 2.

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

Trong mục 2.2.1 và 2.2.2 ta đã tìm hiểu các giải thuật dịch máy của IBM dựa trên các phép tính thống kê. Trong mục này ta sẽ tìm hiểu kỹ thuật gióng hàng từ nhanh dựa trên mô hình 2 của IBM.

Kĩ thuật này được phát triển bởi nhóm nghiên cứu Chris Dyer, Victor Chahuneau, Noah A. Smith [4].

Các giá trị được sử dụng trong mục 2.2.1 và 2.2.2 là hàm Kronecker :

Giá trị trong mục này sẽ được tính bằng:

Với là xác suất dịch của từ vựng.

Phân phối được tính dựa trên xác suất của gióng hàng từ khi không gióng với bất kỳ từ nào của ngôn ngữ nguồn và giá trị . Giá trị được dùng để kiểm soát mức độ gióng hàng từ giữa các từ. Trong trường hợp , phân phối sẽ càng tiến về dạng của mô hình 1. Trong trường hợp càng lớn thì mức độ gióng hàng từ giữa các từ sẽ càng giảm.

Tiếp theo ta sẽ đi nghiên cứu về các phép suy luận (inference) biên (marginal) và các phép đánh giá cải tiến.

2.3.1 Các phép suy luận biên

Ta có các phép tính:

2.3.2 Các phép đánh giá cải tiến

2.4 Tổng kết chương 2

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 Mô hình dịch máy transformer

3.4 Kĩ thuật soft alignment trong transformer

3.5 Tổng kết chương 3

Chương 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

TÀI LIỆU THAM KHẢO

1. Namit Bhatia, 1992, The Oxford Companion to the English Language, trang (pp.) 1051 – 1054.
2. Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, 1990, A STATISTICAL APPROACH TO MACHINE TRANSLATION.
3. Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer, 1993, The Mathematics of Statistical Machine Translation: Parameter Estimation.
4. Chris Dyer, Victor Chahuneau, Noah A. Smith, 2013, A Simple, Fast, and Effective Reparameterization of IBM Model 2.