BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN THỊ BÍCH ĐIỆP

NGHIÊN CỨU VÀ PHÁT TRIỂN PHƯƠNG PHÁP TIẾP CẬN DỰA TRÊN CẤU TRÚC VÀ THỐNG KÊ TRONG DỊCH TỰ ĐỘNG NGÔN NGỮ KÝ HIỆU VIỆT NAM

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Mã số: 9 48 01 01

Hà Nội – 2023

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:
1. Người hướng dẫn 1: TS. Vũ Tất Thắng, Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam
2. Người hướng dẫn 2: PGS.TS. Phùng Trung Nghĩa, Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi ……. giờ ……, ngày ….. tháng ….. năm ……..

Có thể tìm hiểu luận án tại:
1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

Ngôn ngữ ký hiệu (SL) là ngôn ngữ chính thức của cộng đồng người khiếm thính Việt Nam. Dịch ngôn ngữ ký hiệu là quá trình chuyển từ SL sang ngôn ngữ thông thường và ngược lại. Trong đó, bài toán dịch từ ngôn ngữ thông thường sang SL là bài toán có ý nghĩa quan trọng nhằm truyền đạt thông tin, mang lại tri thức xã hội cho người điếc.

Quá trình dịch ngôn ngữ thông thường thành SL bao gồm 3 bước:

Bước 1: Nhận dạng tiếng nói thành văn bản

Bước 2: Xử lý từ văn bản thông thường sang dạng đúng cú pháp trong ngôn ngữ ký hiệu

Bước 3: Mô phỏng từ dạng văn bản đúng cú pháp trong ngôn ngữ ký hiệu thành các biểu diễn trực quan

Bước 2 là bước quan trọng nhất trong quá trình này, vì nó quyết định thông điệp được truyền tải. Tuy nhiên, bước này cũng là thách thức lớn nhất, vì ngôn ngữ ký hiệu có vốn từ vựng hạn chế so với ngôn ngữ nói/viết. Nếu bản dịch máy được thực hiện không tốt, thông tin có thể không được truyền đạt thành công, hoặc trong một số trường hợp, thông điệp được truyền tải có ý nghĩa khác với nguyên bản.

Các nghiên cứu về dịch ngôn ngữ ký hiệu trên thế giới thường áp dụng cả phương pháp cổ điển và hiện đại. Phương pháp cổ điển sử dụng các quy tắc ngữ pháp để chuyển đổi từ ngôn ngữ thông thường sang SL. Phương pháp hiện đại dựa trên học sâu (Deep Learning) để tự động học các đặc trưng của ngôn ngữ ký hiệu từ dữ liệu đầu vào.

Luận án "Dịch ngôn ngữ ký hiệu Việt Nam" tập trung vào bài toán dịch ngôn ngữ thông thường dạng văn bản sang văn bản đúng cấu trúc cú pháp trong ngôn ngữ ký hiệu Việt Nam (text-to-text). Luận án đề xuất triển khai, đề xuất và cải tiến các mô hình dịch máy trong việc dịch ngôn ngữ thông thường dạng văn bản sang văn bản đúng cấu trúc cú pháp trong ngôn ngữ ký hiệu Việt Nam. Đồng thời, luận án cũng xây dựng các bộ tự liệu từ điển VSL

và các bộ dữ liệu song ngữ từ việc thử nghiệm và phát triển phương pháp làm giàu dữ liệu cho bài toán.

Dịch ngôn ngữ ký hiệu là một bài toán khó, nhưng có ý nghĩa quan trọng đối với cộng đồng người khiếm thính. Luận án "Dịch ngôn ngữ ký hiệu Việt Nam" là một nghiên cứu quan trọng, góp phần nâng cao chất lượng dịch ngôn ngữ ký hiệu Việt Nam, giúp người khiếm thính có thể tiếp cận được thông tin và tri thức xã hội một cách đầy đủ và chính xác.

2. Mục tiêu của luận án

Mục tiêu chính của luận án là giải quyết một bài toán có ý nghĩa về cả mặt khoa học và thực tiễn. Trong đó việc đề xuất các mô hình và phương pháp tốt trong lĩnh vực dịch máy để giải quyết bài toán dịch tiếng Việt sang văn bản đúng cú pháp trong ngôn ngữ ký hiệu Việt Nam là trọng tâm. Tiếp đó là quá trình thực nghiệm, phân tích và đánh giá kết quả của bài toán so với các phương pháp đã đề xuất đối các ngôn ngữ ký hiệu trên thế giới và ngôn ngữ ký hiệu Việt Nam.

3. Đóng góp của luận án

Luận án nghiên cứu giải quyết vấn đề dịch máy từ câu tiếng Việt sang câu đúng cấu trúc cú pháp trong ngôn ngữ ký hiệu Việt Nam. Các đóng góp chính của luận án gồm:

1) Luận án đề xuất một phương án dịch đơn giản và hiệu quả cho bài toán sử dụng mô hình dịch dựa trên luật. Tuy là một phương pháp cổ điển nhưng phù hợp với bài toán đặt ra. Đóng góp này được công bố trong các công trình số [CT1], [CT2], [CT3].

2) Đề xuất một phương pháp làm giàu dữ liệu dựa trên mạng từ cho dữ liệu song ngữ câu tiếng Việt – câu đúng cú pháp trong VSL. Đóng góp này được công bố trong các công trình số [CT5].

3) Cải tiến một mô hình dịch thống kê cơ bản và một số mô hình dịch hiện đại dựa trên mạng Noron cho bài toán. Đóng góp này được công bố trong các công trình số [CT4], [CT6].

Đồng thời luận án đã xây dựng các bộ dữ liệu: từ điển ngôn ngữ ký hiệu Việt Nam **VSL-Lexicon**; dữ liệu "song ngữ" **Vie-VSL10k**, **Vie-VSL60k** công bố cho cộng đồng nghiên cứu sử dụng.

4. Phạm vi của luận án

Phạm vi của luận án tập trung vào các phương pháp dịch máy cho bài toán dịch câu tiếng Việt sang câu đúng cú pháp trong ngôn ngữ ký hiệu Việt Nam. Các mô hình diễn hoạ 3D hay các đầu ra cuối cùng khác của ngôn ngữ ký hiệu Việt Nam không được đề cập đến trong luận án này.

5. Cấu trúc của luận án

Nội dung chính của luận án luận án được tổ chức thành phần mở đầu và bốn chương có bố cục như sau:

- Phần Mở đầu: giới thiệu về bài toán dịch ngôn ngữ ký hiệu trong đó trọng tâm của luận án đề cập đến các phương pháp dịch máy cho việc dịch từ văn bản tiếng Việt thông thường sang dạng văn bản đúng cú pháp trong ngôn ngữ ký hiệu. Nội dung này đề cập ý nghĩa và tính cấp thiết của luận án, tổng quan về bối cảnh nghiên cứu.
- Chương 1 giới thiệu tổng quan vấn đề nghiên cứu trong luận án, trình bày và phân tích các vấn đề còn tồn tại trong các nghiên cứu trong nước và thế giới liên dến bài toán dịch ngôn ngữ ký hiệu.
- Chương 2: Giới thiệu một số kiến thức cơ sở liên quan đến nội dung nghiên cứu của luận án.
- Chương 3: Nghiên cứu phương pháp tiếp cận dựa trên cấu trúc trong dịch tự động ngôn ngữ ký hiệu Việt Nam, thực nghiệm và đánh giá các kết quả trên phương pháp này.
- Chương 4: Trình bày một phương pháp làm giàu dữ liệu dựa trên mạng từ cho bài toán.
- Chương 5: Nghiên cứu một số mô hình dịch máy thống kê cổ điển và dịch máy hiện đại dựa trên mạng noron trong dịch tự động ngôn ngữ ký hiệu Việt Nam, thực nghiệm và đánh giá các kết quả trên các phương pháp này.
- Cuối cùng là phần kết luận về những kết quả đạt được của luận án; nêu ưu nhược điểm và định hướng phát triển.

CHƯƠNG 1
TỔNG QUAN VỀ BÀI TOÁN DỊCH NGÔN NGỮ KÝ HIỆU VIỆT NAM

1.1. Tổng quan về ngôn ngữ ký hiệu

Ngôn ngữ ký hiệu được hình thành từ rất sớm gắn với sự phát triển ngôn ngữ thông thường. Cộng đồng người khiếm thính tạo ra ngôn ngữ ký là một loại ngôn ngữ riêng biệt để giao tiếp và thu nhận kiến thức của nhân loại. Thay vì ngôn ngữ thông thường diễn đạt bằng âm thanh, lời nói thì ngôn ngữ ký hiệu có thể là sự kết hợp giữ sự chuyển động của bàn tay, cả cánh tay kết hợp nét biểu cảm trên khuôn mặt. Vì vậy trong ngôn ngữ học nó cũng thuộc một dạng ngôn ngữ tự nhiên. Tuy nhiên nó không phải là ngôn ngữ cơ thể - một loại giao tiếp phi ngôn ngữ. Ngôn ngữ ký hiệu có những đặc trưng về cú pháp: tính rút gọn, nhấn mạnh trọng tâm và thay đổi trật tự cú pháp so với ngôn ngữ thông thường.

1.2. Các nghiên cứu liên quan

Vấn đề về dịch ngôn ngữ ký hiệu trên thế giới được chia thành 2 lớp bài toán. Một là dịch từ ngôn ngữ thông thường sang ngôn ngữ ký hiệu. Hai là dịch theo chiều ngược lại tức là từ ngôn ngữ ký hiệu sang dạng chữ viết hoặc giọng nói trong ngôn ngữ thông thường.

Tuy nhiên, ***luận án này chỉ xem xét các bài nghiên cứu liên quan đến dịch văn bản/giọng nói sang ngôn ngữ ký hiệu***. Bởi vì, đây là một bài toán có ý nghĩa quan trọng nhằm truyền đạt thông tin, mang lại tri thức xã hội cho người điếc.Trong bài toán này, nhiều nghiên cứu chú trọng đến vấn đề dịch văn bản thông thường sang dạng văn bản đúng cú pháp SL.

Những năm gần đây, dịch dựa trên cấu trúc vẫn được ứng dụng trong một số bài toán dịch ngôn ngữ ký hiệu. Phương pháp thống kê cũng thường xuyên được áp dụng cho mục tiêu dịch text-to-text trong bài toán dịch ngôn ngữ ký hiệu với một lượng dữ liệu nhỏ. Các nghiên cứu gần đây đang tận dụng tối đa những tiến bộ kỹ thuật trong các lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), Mạng thần kinh sâu (DNN) và Dịch máy (MT), với mục đích phát triển các hệ thống có khả năng dịch giữa ngôn ngữ ký hiệu và ngôn

ngữ nói nhằm lấp đầy khoảng cách giao tiếp giữa cộng đồng nói tiếng SL và cộng đồng sử dụng ngôn ngữ nói.

Tóm lại, một trong những nhược điểm lớn nhất của nhiều dự án kể trên là ít chú trọng đến cú pháp ngôn ngữ ký hiệu với những đặc điểm riêng của từng ngôn ngữ độc lập này dẫn đến các vấn đề về hiểu ngôn ngữ. Ngoài ra còn là vấn đề với cơ sở dữ liệu không đủ lớn. Đặc biệt là những đánh giá trong cộng đồng người điếc ít được xem xét đến.

1.3. Bài toán dịch ngôn ngữ ký hiệu Việt Nam

Đối với bài toán dịch VSL, hiện chưa có một cơ sở dữ liệu nào được công bố cho cộng đồng nghiên cứu. Bởi vậy, luận án này cũng tập trung vào một mục tiêu quan trọng là xây dựng được bộ cơ sở dữ liệu cho dịch máy VSL. Với kỳ vọng ban đầu là xây dựng đầy đủ bộ từ vựng VSL (VSL-lexicon) với các chú giải là mỗi từ vựng gắn mới một mô hình diễn hoạ 3D. Đồng thời xây dựng bộ "dữ liệu song ngữ" bao gồm các cặp câu tiếng Việt – câu đúng cú pháp trong VSL.

Bởi vậy, trong luận án này vấn đề về dịch máy VSL được chú trọng tới các vấn đề cụ thể là các phương pháp dịch máy cổ điển và hiện đại (dịch máy dựa trên cấu trúc, dịch thống kê và dựa trên mạng noron) và xây dựng dữ liệu cho bài toán.

1.4. Kết luận chương

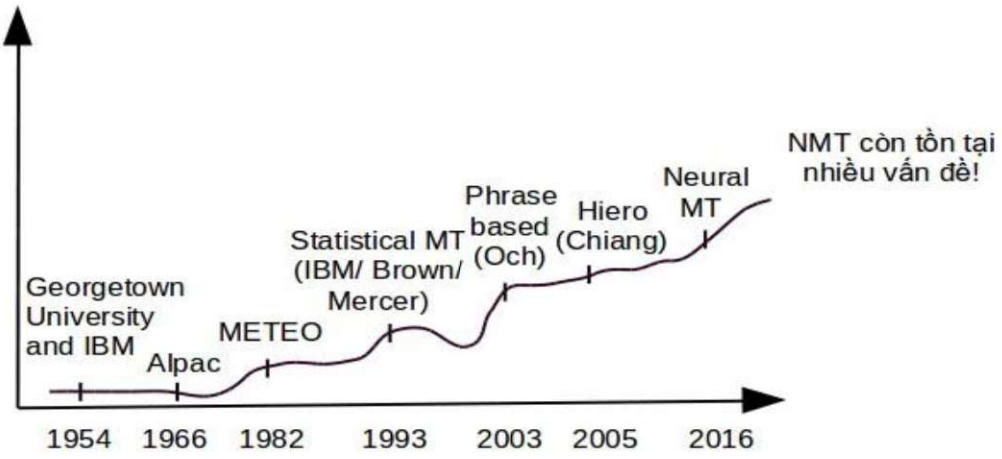
Trong chương này luận án đã trình bày những vấn đề tổng quan về ngôn ngữ ký hiệu nói chung và những đặc điểm cú pháp đặc trưng của ngôn ngữ Việt Nam nói riêng. Tính cấp thiết của bài toán dịch ngôn ngữ ký hiệu được thể hiện qua phân tích và đánh giá một số công trình nghiên cứu về dịch ngôn ngữ ký hiệu trên thế giới. Từ những nhận định đó, đặt ra 3 vấn đề chính cho bài toán dịch máy VSL. Một là việc áp dụng những phương pháp dịch máy được cho là cổ điển, tuy nhiên chúng được đánh giá là hiệu quả và phù hợp với bài toán dịch VSL. Hai là triển khai phương pháp làm giàu dữ liệu – một trong những nội dung trọng tâm cho việc đánh giá, thử nghiệm các mô hình dịch. Ba là đề xuất mô hình dịch máy thống kê hiện đại phù hợp

với bài toán dịch VSL. Ngoài ra, phạm vi của bài toán cũng thu hẹp lại với việc dịch văn bản thông thường sang dạng văn bản đúng cú pháp của VSL trong bài toán tổng thể (bước thứ 2). Vì luận án nhận định rằng trong 3 bước của quy trình dịch ngôn ngữ ký hiệu đã phân tích thì bước thứ 2 là trọng tâm và có ý nghĩa nhất với bài toán này vì thực sự chú ý tới cú pháp ngôn ngữ riêng của người khiếm thính trong thực tế.

CHƯƠNG 2
CÁC KIẾN THỨC CƠ SỞ

2.1. Kiến thức cơ sở về dịch máy

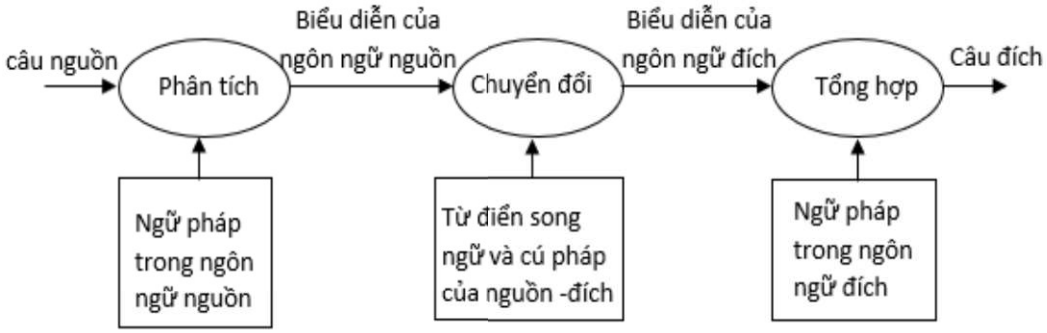
Dịch máy (Machine translation) gọi tắt là MT hay còn gọi là dịch tự động, là quá trình phần mềm máy tính dịch các văn bản từ một ngôn ngữ nguồn sang một văn bản thuộc một ngôn ngữ đích.



Hình 2.1. Quá trình phát triển của MT

2.2. Dịch dựa trên luật

Kỹ thuật dịch dựa vào luật (Rules based machine translation - RBMT) sử dụng một tập các luật về hình thái, cú pháp, ngữ nghĩa giữa các cặp ngôn ngữ nguồn và đích [33]. Tiếng Việt và VSL có liên quan chặt chẽ về cú pháp. Bởi vậy, việc dịch có thể được thực hiện bằng phân tích cú pháp và một số kỹ thuật. Hình 2.1 mô tả một hệ thống dịch theo luật.



Hình 2.2. Sơ đồ dịch máy dựa trên luật.

2.2. Dịch máy thống kê

Phương pháp dịch máy thống kê lần đầu tiên được Brown đề xuất năm 1993 với phương pháp sử dụng là mô hình kênh nhiễu. Bài toán được phát biểu như sau:

Cho một câu f thuộc ngôn ngữ nguồn $f \in f^J = \{f_1, f_2, ..., f_J\}$, hệ thống cần dịch sang câu e thuộc ngôn ngữ đích $e \in e^I = \{e_1, e_2, ..., e_I\}$. Hệ thống dịch sẽ chọn một câu e có xác suất cao nhất trong rất nhiều khả năng dịch được đưa ra.

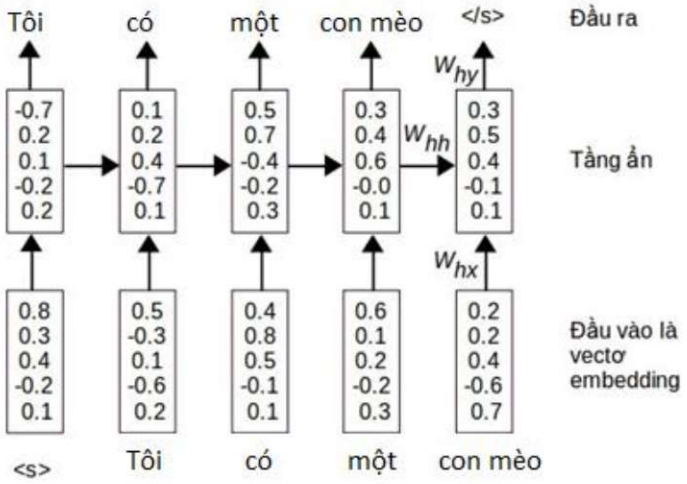
$$e* = argmax_e p(e)p(f|e) \quad (2.3)$$

Với công thức 2.3 mô hình SMT được mô hình hóa thành hai mô hình con là mô hình ngôn ngữ $p(e)$ và mô hình dịch $p(f|e)$.

Mô hình dịch là bài toán trung tâm của SMT. Trong mô hình dịch, vấn đề trọng tâm của việc mô hình hoá xác suất dịch $p(f|e)$ là việc xác định sự tương ứng giữa các từ của câu nguồn với các từ của câu đích.

2.3. Dịch máy dựa trên mạng rơron

Mạng nơron hồi quy (RNN) được đề xuất bởi Elman năm 1990 là một kiến trúc cho phép nhận một trình tự dữ liệu đầu vào và tính toán đầu ra thông qua các trạng thái ẩn bên trong. Các mạng RNN được áp dụng thành công cho mô hình ngôn ngữ trong các nghiên cứu gần đây của Mikolov và các cộng sự [38]. Trong dịch máy, các mạng RNN nhận một trình tự các vectơ đầu vào, ứng với mỗi vectơ tại thời điểm t các RNN cập nhật bộ nhớ của nó để sinh ra các trạng thái ẩn thông qua một biểu thức hồi quy.



Hình 2.5. Mô hình ngôn ngữ sử dụng mạng RNN

2.3.1. Mô hình Sequence to Sequence

Với việc ứng dụng thành công mạng RNN cho mô hình ngôn ngữ, các nhà nghiên cứu đã đề xuất mô hình sequence to sequence (gọi tắt là seq2seq) dựa trên kiến trúc encoder-decoder với các mạng RNN là thành phần trung tâm.

2.3.2. Mô hình Transformer

Mô hình Transformer là một kiến trúc mạng nơ-ron nhân tạo được giới thiệu trong bài báo "Attention Is All You Need" của Vaswani và các cộng sự vào năm 2017. Mô hình này đã trở thành một trong những kiến trúc quan trọng nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Mô hình Transformer sử dụng kiến trúc mã hóa-giải mã để thực hiện các tác vụ xử lý ngôn ngữ tự nhiên. Mỗi lớp của mô hình Transformer bao gồm hai phần chính: một lớp tự chú ý (self-attention) và một lớp mạng truyền thẳng (feed-forward). Lớp tự chú ý cho phép mô hình chú ý đến các phần khác nhau của câu đầu vào trong quá trình mã hóa và giải mã. Lớp mạng truyền thẳng sau đó áp dụng một hàm phi tuyến tính để tính toán đầu ra. Kiến trúc tổng thể của mô hình Transformer được tạo thành bởi sự kết nối của nhiều lớp mã hóa và giải mã. Trong quá trình mã hóa, mỗi lớp của bộ mã hóa sẽ nhận đầu vào là một chuỗi từ và sản xuất một chuỗi trạng thái ẩn. Trong quá trình giải mã, mỗi lớp của bộ giải mã sẽ nhận đầu vào là một chuỗi đích và trạng thái ẩn được tính toán bởi lớp trước đó của bộ giải mã.

2.4. Điểm đánh giá chất lượng bản dịch máy

Phương pháp đánh giá tự động phổ biến nhất là BLEU (Bilingual Evaluation Understudy). BLEU tính toán độ tương đồng giữa bản dịch máy và bản gốc bằng cách so sánh các n-gram (các cụm từ có độ dài n) trong hai văn bản. Độ tương đồng được đánh giá bằng cách tính tỷ lệ của số lượng các n-gram giống nhau trong bản dịch và bản gốc.

Công thức để tính điểm đánh giá của IBM là như sau:

$$score = exp\left\{\sum_{i=1}^{N} w_i \, log(p_i) - \max\left(\frac{L_{\text{ref}}}{L_{\text{tra}}} - 1{,}0\right)\right\}$$

- $P_i = \frac{\sum_j NR_j}{\sum_j NT_j}$
- NR_j: là số lượng các n-grams trong phân đoạn *j* của bản dịch dùng để tham khảo.
- NT_j: là số lượng các n-grams trong phân đoạn *j* của bản dịch bằng máy.
- $w_i = N^{-1}$
- L_{ref}: là số lượng các từ trong bản dịch tham khảo, độ dài của nó thường là gần bằng độ dài của bản dịch bằng máy.
- L_{tra}: là số lượng các từ trong bản dịch bằng máy

Một ưu điểm của BLEU là phương pháp này đơn giản và tính toán nhanh chóng, cho phép đánh giá nhanh chóng chất lượng của một hệ thống dịch máy. Do vậy luận án chọn phương pháp đánh giá bản dịch BLEU cho bài toán dịch.

2.5. Kết luận chương

Chương 2 trình bày các kiến thức nền tảng cơ sở được sử dụng trong luận án này. Nội dung bao gồm: một số khái niệm cơ bản về dịch máy; các mô hình dịch máy cổ điển và hiện đại cho bài toán dịch NGÔN NGỮ KÝ HIỆU bao gồm: dịch máy dựa trên luật, mô hình dịch máy thống kê IBM và mô hình dịch dựa trên mạng noron (Seq2Seq và Transformer);. kiến thức cơ sở về điểm đánh giá các bản dịch máy và trình bày cụ thể về công thức tính toán điểm BLEU – điểm đánh giá chất lượng bản dịch máy dùng trong luận án này.

CHƯƠNG 3
PHƯƠNG PHÁP TIẾP CẬN DỰA TRÊN CẤU TRÚC TRONG DỊCH TỰ ĐỘNG NGÔN NGỮ KÝ HIỆU VIỆT NAM

3.1. Xây dựng cơ sở dữ liệu cho bài toán

3.1.1. Tập từ điển VSL-Lexicon

Trong dữ liệu **VSL-Lexicon** lưu trữ các đơn vị từ vựng với các thông tin đi kèm như: từ loại, mã số chú thích, từ đồng nghĩa và mô hình diễn hoạ tương ứng. Bảng 3.1 mô tả cấu trúc của dữ liệu VSL-lexicon.

Bảng 3.1. Bảng mô tả từ điển VSL-Lexicon

STT	Đơn vị từ vựng	Từ loại	Đồng nghĩa	Mã số chú thích	Mô hình diễn hoạ 3D tương ứng
1	a	Alphabet		VSL0001	M3D0001.FBX
2	ă	Alphabet		VSL0002	M3D0002.FBX
153	tôi	Đại từ (P)	tao, tớ	VSL0153	M3D0153.FBX
154	họ	Đại từ (P)		VSL0154	M3D0154.FBX
296	chết	Động từ (V)	hi sinh,..	VSL0296	M3D0296.FBX
3035	trường học	Danh từ (N)		VSL3035	M3D3035.FBX
3036	nhà	Danh từ (N)		VSL3036	M3D3036.FBX
6176	xương rồng	Danh từ (N)		VSL6176	Chưa có CSDL

3.1.2. Bộ dữ liệu song ngữ Vie-VSL10k

Bộ dữ liệu xây dựng được đặt tên là ***Vie-VSL10k***. Bộ dữ liệu này được xây dựng bán thủ công với 10.000 cặp câu trong miền giao tiếp thông thường. Các dữ liệu thu thập được xử lý bán tự động một phần qua một số thuật toán rút gọn văn bản và chuyển đổi cú pháp sơ khai. Sau đó được đánh giá lại bởi một số chuyên gia ngôn ngữ. Dữ liệu cuối cùng luận án thu thập được 10.000 cặp câu song ngữ Vie – VSL cho phần xây dựng hệ thống dịch dựa trên luật với 4626 đơn vị từ vựng. Các số liệu thống kê về cơ sở dữ liệu Vie-VSL-10k được thể hiện trong bảng.

Bảng 3.2. Các số liệu thống kê về dữ liệu câu tiếng Việt trong Vie-VSL-10k

STT	Loại từ	Ký hiệu	Số lượng trong câu tiếng Việt	Số lượng trong câu VSL
1	Danh từ	N	16182	16182
2	Danh từ riêng	Np	7030	7030
3	Danh từ chỉ loại	Nc	1069	1069

4	Danh từ đơn vị	Nu	172	172
5	Động từ	V	15528	13559
6	Tính từ	A	4241	4241
7	Đại từ	P	3424	3424
8	Định từ	L	537	0
9	Số từ	M	1560	1560
10	Phụ từ	R	8477	4689
11	Giới từ	E	4471	2910
12	Liên từ	C	1480	0
13	Thán từ	I	175	0
14	Trợ từ, tiểu từ, từ tình thái	T	878	0
15	Yếu tố cấu tạo từ	S	10	0
16	Các từ không phân loại được	X	322	0

3.2. Vấn đề tổng hợp luật

Với các đặc điểm đặc trưng về cú pháp trong VSL đã trình bày, có một số đặc điểm rút gọn và chuyển đổi cú pháp của câu trong VSL được tổng hợp lại cho bài toán dịch dựa trên cấu trúc.Luận áp trình bày việc áp dụng các công cụ phân tách cú pháp vào dữ liệu 10000 cặp câu song ngữ Vie-VSL để trích rút luật. Từ đó xây dựng được 8025 luật từ dữ liệu song ngữ. Bảng dưới đây mô tả một số luật được trích rút.

Bảng 3.13. Một số luật trích rút cho hệ thống dịch Rule-based

STT	Câu tiếng việt được phân tích cú pháp	Quy tắc ngữ pháp	Câu ngôn ngữ ký hiệu được phân tích cú pháp	Luật trích rút
1	SQ (NP (N Bạn) (N tên)) (VP (V là) (WHNP (P gì))) (? ?)	1	SQ (NP (N Bạn) (N tên) (P gì)) (? ?)	SQ (NP (N) (N)) (VP (V) (WHNP (P)) (? ?) → SQ (NP (N) (N) (P)) (? ?)
2	S (NP (P Tôi)) (NP (N tên)) (VP (V là) (NP (Np Hiếu))) (..)	1	S (NP (P Tôi)) (NP (N tên) (Np Hiếu)) (..)	S (NP (P)) (NP (N)) (VP (V) (NP (Np)) (..) → S (NP (P)) (NP (N) (Np)) (..)
3	S (NP (N Khế)) (C thì) (AP (A chua)) (..)	1	S (NP (N Khế)) (AP (A chua)) (..)	S (NP (N)) (C) (AP (A)) (..) → S (NP (N)) (AP (A)) (..)
4	S (NP (N Mít)) (C thì) (AP (A ngọt)) (..)	1	S (NP (N Mít)) (AP (A ngọt)) (..)	
5	S (NP (P Tôi)) (NP (M 19) (N tuổi)) (..)	2	S (NP (P Tôi)) (NP (N tuổi) (M 19)) (..)	S (NP (P)) (NP (M) (N))) (..) → S (NP (P)) (NP (N) (M)) (..)
..	…	..	…	…

Từ 8025 luật được trích rút từ kho dữ liệu 10000 cặp câu song ngữ, ta tiến hành xây dựng hệ thống dịch máy dựa trên luật. Hiệu quả của phương pháp dịch này được phân tích và đánh giá ở phần sau. Tham khảo 8025 luật tại https://github.com/BichDiep/rules-VSL.git.

3.3. Xây dựng hệ thống dịch dựa trên luật

Thuật toán khối tổng hợp luật và hệ thống dịch máy trên luật được miêu tả dựa trên mã giả như sau:

Algorithm: Rule-based-MT-VSL

Input: Sentence S in Vietnamese,
Output: Sentence S' in the syntax of VSL.

```
1. R is set of syntax conversion rules
2. WD = ∅; (WD: Waiting Dataset)
3. SYN is Synonyms files with n line: SYN[n,1] in VSL dictiary;
   SYN[n,i] is a synonym of SYN[n,1]; (i=1:m).
4. Si ⇐ Tokenization(S)
5. While ∃ Si in SYN:
        Si = SYN[n,1]
6. (TS, PS)⇐ Parsing (S)
7.  If (Find Ps in R)
        ST= Transform(TS)
   Else
        Add S to WD
8. S'= Shorten(ST)
9. Return S'
```

3.4. Các thực nghiệm và đánh giá hệ thống dịch dựa trên luật

Đánh giá hiệu quả của phương pháp dịch dựa trên luật cho bài toán dịch Vie-VSL trên các tập kiểm tra được chuẩn bị. Điểm BLEU đáng giá bản dịch trong dịch tự động Vie-VSL đối với các tập dữ liệu trong bảng 3.16. Nhìn chung điểm BLEU trên các tập test đều vượt trội so với điểm BLEU của một số ngôn ngữ khác như bởi vì trong bài toán của luận án, mô hình dịch gần như không thay đổi với hầu hết các đơn vị ngôn ngữ là giống nhau ở 2 ngôn ngữ. Chỉ một số từ không có trong ngôn ngữ ký hiệu được thay thế bằng từ đồng

nghĩa. Với thứ tự trong câu thì VSL hầu hết là các mẫu câu đơn giản, chúng kém đa dạng hơn rất nhiều so với các cặp ngôn ngữ khác.

Bảng 3.16. Tổng hợp điểm BLEU hệ thống dịch dựa trên luật với một số tập kiểm tra

Tập dữ liệu	**BLEU Score**
Data set 1: Miền các câu trong giao tiếp	81.15
Data set 2: Miền các câu trong văn học	48.68
Data set 3: Miền các câu trong kỹ thuật	64.13
Data set 4: Miền các câu trong y học	55.72
Trung bình	62.55

Do vậy mô hình ngôn ngữ đơn giản hơn so với máy vì mô hình xác xuất là hội tụ. Tuy nhiên chúng có sự khác biệt giữa các tập test khác nhau. Sự khác biệt này chủ yếu phụ thuộc vào độ dài của câu, sự phức tạp và từ vựng trong từng miền. Đối với miền giao tiếp, các câu chủ yếu là ngắn gọn, đơn giản và tỉ lệ từ vựng thuộc tập từ điển VSL cao hơn so với các miền dữ liệu khác.

3.5. Kết luận chương

Trong chương này, luận án trình bày một phương pháp giải quyết bài toán với mô hình dịch dựa trên luật. Để giải quyết bài toán với mô hình này cần có các dữ liệu tài nguyên về từ điển VSL và các quy tắc ngữ pháp. Mô hình dịch máy sẽ sử dụng các quy tắc ngữ pháp để phân tích và dịch câu. Các quy tắc này được tổng hợp xác định trước và được cấu trúc theo các luật chuyển đổi Vie-VSL. Kết quả đạt được của phần này bao gồm: cơ sở dữ liệu từ điển **VSL-Lexicon** với các thành phần và đặc trưng khác với các loại từ điển thông thường, cơ sở dữ liệu song ngữ **Vie-VSL-10k** bao gồm 10.000 cặp câu tiếng Việt – câu đúng quy tắc cú pháp VSL phục vụ cho việc xây dựng các quy tắc cú pháp của mô hình dịch luật; cuối cùng là một **mô hình dịch luật** đơn giản và hiệu quả cho bài toán. Điểm đánh giá chất lượng bản dịch BLEU đạt 62.55 với những đặc điểm đã phân tích. Các công trình công bố liên quan đến phần này bao gồm [CT1], [CT2], [CT3].

CHƯƠNG 4
LÀM GIÀU DỮ LIỆU CHO BÀI TOÁN DỊCH TỰ ĐỘNG NGÔN NGỮ KÝ HIỆU VIỆT NAM

4.1. Cơ sở của phương pháp đề xuất

Với những đặc điểm của Wordnet về quan hệ ngữ nghĩa giữa các từ, ý tưởng của việc làm giàu dữ liệu là thực hiện thay thế từ trong câu để sinh ra một dữ liệu mới. Câu mới được sinh ra về mặt cú pháp không thay đổi và ngữ nghĩa hợp logic, vì vậy để dịch nó sang VSL ta vẫn giữ nguyên luật chuyển đổi. Như vậy việc dịch thực hiện đúng và đảm bảo về ngữ nghĩa với các đánh giá độ tương đồng ở phần thực nghiệm.

Ở đây sử dụng 3 tiêu chuẩn:

Tiêu chuẩn anh em:

Tức là: $SV = \{S_i^{jk} / S_i^{jk} \in S_i^j (\forall j: 0 \le j \le n_i^j) : \exists S_p : (S_P\ is\_hyper\ S_i^{jk})\}$

Tiêu chuẩn cha con:

$SV = \{S_i^{jk} / \exists S_p \in S_i^h\ (h \in [1 \dots n_i^j])\ S_i^{jk} \in S_i^h (\forall j: 0 \le j \le n_i^j, \quad j \ne h) : (S_P\ is\_hyper\ S_i^{jk})\}$

Tiêu chuẩn ông cháu:

$SV = \{S_i^{jk} / \exists S_g \in S_i^h\ (h \in [1 \dots n_i^j]), S_i^{jk} \in S_i^j (\forall j: 0 \le j \le n_i^j, j \ne h)(S_g\ is\_dist\_hyper\ S_i^{jk})\}$

Như vậy, với một từ W trong câu, ta có thể thay W bằng W' với điều kiện W và W' đảm bảo các tiêu chuẩn anh em, tiêu chuẩn cha con, tiêu chuẩn ông cháu.Thuật toán làm giàu dữ liệu được mô tả bởi mã giả như sau:

Algorithm: Data-Augment-VSL

Input: Sentences S
Output: Set of sentences S' are generated based on S.

```
1: Split W word ∈ S
2: X ⇐ W. hypernyms()
   n = len(X);
3: For i=1,n do
     Xi⇐X.hyponyms()
     Add Xi to set T
4:  While !∃ Xi.hyponyms:
          Yi ⇐ Xi.hyponyms()
          Add Yi to set T
5: S'= Replace(W,Ti)
```

4.3. Kết quả thực nghiệm và đánh giá

Để đánh giá các thực nghiệm của mình, luận án căn cứ vào các tiêu chí mức độ làm giàu dữ liệu và độ tương đồng về dữ liệu để phân tích.

Đầu tiên về mức độ làm giàu dữ liệu, luận án đề cập đến vấn đề số lượng tập T xây dựng được từ thuật toán làm giàu dữ liệu và xem xét một số khía cạnh về ngữ nghĩa của câu mới sinh ra từ dữ liệu gốc. Với nhóm các từ vựng là động từ thì các thử nghiệm này cho kết quả không hợp lý về ngữ nghĩa trong câu tiếng Việt.

Sau quá trình thực nghiệm với một số dữ liệu, từ loại động từ khi sử dụng phương pháp tìm kiếm từ có hạ danh với các tiêu chuẩn anh em, cha-con và ông cháu không phù hợp về mặt ngữ nghĩa. Do vậy chỉ xét đến các nhóm đơn vị từ bao gồm đại từ, danh từ và tính từ. Bảng 4.1. trình bày một số tập T và tổng kết số câu được làm giàu dữ liệu từ thuật toán đề xuất (trong đó T là tập các từ có cùng hạ danh với các tiêu chuẩn đã áp dụng với từng nhóm từ loại, W_S là số câu dữ liệu gốc có chứa 1 từ thuộc nhóm từ loại đang xét, W'_S) là số câu được làm giàu từ tất cả các câu gốc có chứa 1 từ thuộc nhóm từ loại đang xét)

Bảng 4.1. Kết quả của thuật toán làm giàu dữ liệu từ Vie-VSL10k

Từ loại	Nhóm	Ví dụ	T	W_S	W'_S
Danh từ	Thực vật 1 (trái cây)	Bưởi, cam, nho, táo,..	92	35	3220
	Thực vật 2 (hoa)	Hoa cúc, hoa hồng, hoa ly,…	183	5	915
	Thực vật 3 (chung)	Cây, hoa, cỏ, lá, rau	438	10	2628
	Thực phẩm	Bánh, kẹo, bia, thịt, rau…	471	3	1413
	Động vật 1 (vật nuôi)	chó, chó con, chó xù, gà, mèo,…	25	5	125
	Động vật 2 (khác)	Báo, hổ, hươu sao, kỳ đà	708	3	2124
	Đồ vật 1 (gia dụng)	Bàn, ghế, tủ,..	257	11	2827
	Đồ vật 2 (đồ tạo tác)	Buá, kéo, máy,..	1564	4	5056
	Đồ vật 3 (phương tiện)	Xe máy, ô tô, xe chở hàng, ..	78	7	546
	Thời tiết	Nắng, mưa, gió,..	63	5	315

	Nghề nghiệp	Giáo viên, công nhân,	21	8	168
	Cơ thể	Chân, tay, tóc, má, môi,…	231	4	924
	Hình khối	Tam giác, hình tròn,…	134	3	402
Tính từ	Màu sắc	Đỏ, xanh, vàng, tím,…	12	36	432
	Tính chất vật chất	Nặng, nhẹ, Cứng, mềm,…	45	2	90
	Độ lớn nhỏ	To, rộng, dài, ngắn…	15	4	60
	Cảm xúc	vui, buồn, lo lắng	279	7	1953
	Tính cách	hài hước, cục cằn, dễ thương…	23	4	92
Đại từ		Tôi, họ, chúng ta, ..	12	3424	41088
				Tổng:	64378

Trong dữ liệu 10000 câu ban đầu, với miền được chọn là câu giao tiếp nên đại từ chiếm số lượng từ vựng lớn trong kho ngữ liệu.

Độ tương đồng của kho ngữ liệu trước và sau khi làm giàu có thể được đánh giá dựa trên độ hỗn loạn mô hình ngôn ngữ (perplexity) của mỗi loại. Perplexity là một độ đo được sử dụng trong xác suất và thống kê để đánh giá hiệu quả của mô hình ngôn ngữ. Trong mô hình ngôn ngữ n-gram, perplexity đo lường khả năng dự đoán của mô hình trên một đoạn văn bản mới dựa trên xác suất của chuỗi n-gram trong mô hình. Bảng 4.3. trình bày chỉ số Perplexity đối với các kho ngữ liệu đã xây dựng với mô hình ngôn ngữ 3-gram:

Bảng 4.3. Chỉ số Perplexity đối với các kho ngữ liệu đã xây dựng

Kho ngữ liệu	**Chỉ số Perplexity trung bình khoảng**
Vie10k	P_1= 420
VSL10k	P_2= 300
Vie60K	P_1'= 520
VSL60K	P_2'= 450

Như vậy kích thước lớn hơn gấp 6 lần so với dữ liệu gốc, nhưng điểm Perplexity cao không quá 1,5 lần. Điều đó cho thấy kho ngữ liệu với mô hình 3-gram có hiệu suất tốt. Với sự tương đồng cao giữa các câu gốc và câu mới sinh vì giữ nguyên cấu trúc cú pháp. Về mặt ngữ nghĩa, sự tương đồng được đảm bảo bởi tính chất của các từ cùng hạ danh với các tiêu chuẩn đã áp dụng.Điểm BLEU cũng là một tiêu chí để so sánh hiệu quả các mô hình dịch sẽ được trình bày trong chương 5.

4.4. Kết luận chương

Trong chương này, luận án trình bày việc xây dựng 2 bộ dữ liệu **Vie-VSL-10k** và **Vie-VSL-60k** gồm các cặp câu song ngữ tiếng Việt – câu đúng cú pháp trong VSL. Trong đó bộ dữ liệu Vie-VSL-60k là kết quả của một phương pháp làm giàu dữ liệu từ bộ dữ liệu cơ sở Vie-VSL-10k. Ý tưởng đề xuất của phương pháp làm giàu dữ liệu là dựa trên cơ sở về cấu trúc thượng danh và hạ danh của mạng từ (Wordnet) và sử dụng cơ sở dữ liệu Wordnet tiếng Việt. Thuật toán làm giàu dữ liệu giúp sinh ra các cặp câu song ngữ Vie-VSL dựa trên dữ liệu gốc bao gồm 10.000 câu song ngữ Vie-VSL. Từ những phân tích đánh giá quá trình thực nghiệm thuật toán làm giàu dữ liệu ta thấy rằng bộ dữ liệu sau làm giàu có tính tương đồng cao với dữ liệu gốc vì vẫn giữ nguyên được cấu trúc cú pháp câu ban đầu. Đồng thời câu mới sinh từ việc thay thế từ đảm bảo các tiêu chuẩn dựa trên các tính chất của mạng từ đảm bảo sự phù hợp về ngữ nghĩa. Bộ dữ liệu Vie-VSL-60K được sử dụng cho các thực nghiệm đánh giá bài toán của luận án với một số mô hình dịch máy thống kê và dịch máy dựa trên mạng noron ở chương tiếp theo.

CHƯƠNG 5
PHƯƠNG PHÁP TIẾP CẬN DỰA TRÊN THỐNG KÊ VÀ MẠNG NORON TRONG DỊCH TỰ ĐỘNG NGÔN NGỮ KÝ HIỆU VIỆT NAM

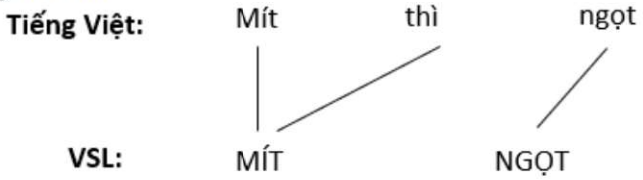
5.1. Cải tiến mô hình dịch IBM cho bài toán dịch Vie-VSL

Trong phần này, luận án trình bày một mô hình dịch thống kê đơn giản cho dịch máy ngôn ngữ ký hiệu dựa trên dịch từ vựng, dịch từ. Phương pháp này yêu cầu một từ điển ánh xạ các từ từ ngôn ngữ nguồn sang ngôn ngữ đích. Trong bài toán dịch Vie-VSL, từ điển ánh xạ này đơn giản hơn rất nhiều các bài toán dịch giữa các ngôn ngữ khác như dịch Anh – Việt, Việt – Trung hay Việt- Nhật. Bởi dầu hết các từ đều ánh xạ 1-1. Luận án đề cập đến việc sử dụng số liệu thống kê dựa trên số lượng từ trong kho văn bản hoặc văn bản song ngữ. Ta cần ước tính về phân phối xác suất dịch từ vựng. Hàm này sẽ trả về

một xác suất, đối với mỗi lựa chọn bản dịch VSL , cho biết khả năng bản dịch đó như thế nào.

$$P_f = e \rightarrow P_f(e)$$

Nhờ phân phối xác suất cho dịch từ vựng, ta có thể thực hiện bước nhảy sang mô hình đầu tiên chỉ sử dụng xác suất dịch từ vựng. Chúng ta biểu thị xác suất dịch một từ tiếng Việt f sang một từ VSL e bằng hàm xác suất có điều kiện $t(e|f)$. Sự liên kết giữa các từ đầu vào và các từ đầu ra có thể được minh họa bằng sơ đồ:

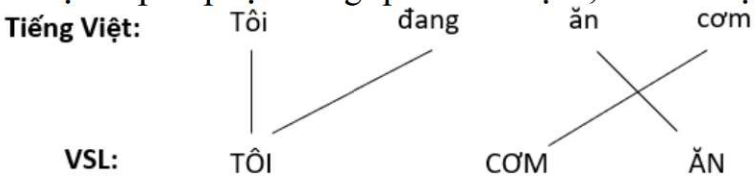


W: (1 →1; 1 →2; 2 →3)

Hình 5.1. Liên kết giữa các từ đầu vào và các từ đầu ra trong dịch câu Vie-VSL

Hàm ánh xạ trong ví dụ trên, mỗi từ đầu ra VSL ở vị trí i thành một từ đầu vào tiếng Việt ở vị trí j: w: $(j \rightarrow i)$

Đây là cách sắp xếp rất đơn giản, vì các từ tiếng Việt và các từ VSL tương ứng không theo một thứ tự hoàn toàn giống nhau. Điều này có nghĩa là các từ phải được sắp xếp lại trong quá trình dịch, như ví dụ sau minh họa:



W: (1 →1; 1 →2; 2 →4; 3 →3)

Hình 5.2. Ví dụ minh hoạ về sắp xếp lại từ trong dịch câu Vie-VSL

Với mô hình căn chỉnh dựa trên các từ, mỗi đầu ra có thể được liên kết với một hoặc nhiều từ đầu vào, như được xác định bởi chức năng căn chỉnh. Mô hình IBM có thể triển khai để căn chỉnh từ dựa trên xác suất dịch từ vựng. Có 3

mô hình IBM để ánh xạ các từ từ ngôn ngữ nguồn Vie và ngôn ngữ đích VSL với thuật toán cải tiến dựa trên khớp chuỗi cho bài toán dịch Vie-VSL.

Mô hình IBM1 xác định xác suất dịch cho một câu tiếng Việt $f = (f_1, \dots, f_{l_f})$ có độ dài l_f sang một câu VSL $e = (e_1, \dots, e_{l_e})$ có độ dài l_e với sự liên kết của từng từ VSL e_j sang một từ tiếng Anh từ f_i theo hàm căn chỉnh w: $(j \rightarrow i)$ như sau:

$$p(e, w|f) = \frac{\varepsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{w(j)})$$

Xét thuật toán trên 1 phần ngữ liệu nhỏ của kho dữ liệu Vie-VSL-10k với 3 từ tiếng Việt là đầu vào: "*tôi*", "*ăn*", "*cơm*", và 3 từ trong VSL là đầu ra: "*TÔI*", "*ĂN*", "*CƠM*".

Tối ưu hoá EM trong mô hình IBM-1

Từ công thức:

$$p(w|e, f) = \frac{p(e, w|f)}{p(e|f)}$$

Ta có thể cải thiện kết quả bằng cách thêm d_w giữa e và f, ta có:

$$p(w|e, f) = \frac{\alpha . p(e, w|f) + (1 - \alpha) . d_w(e, f)}{p(e|f)}$$

Trong đó α là hệ số tương đồng giữa hai từ e và f. Giá trị tiêu chuẩn của α được sử dụng cho các thí nghiệm là 0.5. Bảng 5.4 trình bày các kết quả so sánh được áp dụng cho một ngữ liệu nhỏ bao gồm hai cặp câu tương ứng.

Bảng 5.4. Kết quả xác suất dịch với mô hình IBM 1 có tối ưu hoá

e	f	Ban đầu	Lần lặp 3	Lần lặp 3 và so khớp chuỗi
TÔI	tôi	0.33	0.75	0.96
TÔI	ăn	0.33	0.21	0.03
TÔI	cơm	0.33	0.21	0.02
ĂN	tôi	0.33	0.04	0.01
ĂN	ăn	0.33	0.42	0.77
ĂN	cơm	0.33	0.42	0.32

CƠM	tôi	0.33	0.04	0.01
CƠM	ăn	0.33	0.42	0.28
CƠM	cơm	0.33	0.77	0.95

Giống như mô hình IBM-1, thì mô hình IBM-2 thêm một hệ số α để so khớp chuỗi vào quy trình căn chỉnh. Kết quả cho thấy chúng chỉ hội tụ về 1 sau 2 lần lặp. Như vậy, đối với mô hình thống kê cho bài toán, kỹ thuật so khớp chuỗi và dịch dựa trên cụm từ được sử dụng là phù hợp. Luận án sử dụng các kho ngữ liệu Vie-VSL-10K và Vie-VSL-60K cho việc thực nghiệm mô hình dịch dựa trên thống kê đã đề xuất. Các phân tích đánh giá kết quả được trình bày cụ thể ở mục 5.4.

5.2. Mô hình Sequence to Sequence cho bài toán

5.2.1. Mô hình bộ mã hóa và giải mã

Sơ đồ tổng quan về mô hình được minh hoạ trong hình 5.3. Các hằng số cho mô hình: embedding_dim = 256; units = 1024. Bắt đầu bằng cách xây dựng bộ mã hóa.

Bộ mã hóa:

1. Lấy danh sách các mã ID token (từ input_text_processor).
2. Tìm kiếm một vecto Embeding cho mỗi mã thông báo (Sử dụng một kỹ thuật nhúng).
3. Xử lý “embeddings” thành một chuỗi mới.
4. Kết quả:
   - Trình tự được xử lý - sẽ được chuyển đến đầu chú ý.
   - Trạng thái bên trong - sẽ được sử dụng để khởi tạo bộ giải mã

Bộ mã hóa trả về trạng thái bên trong của nó để trạng thái có thể được sử dụng để khởi tạo bộ giải mã. RNN cũng thường trả về trạng thái của nó để nó có thể xử lý một chuỗi qua nhiều lần gọi.

Cơ chế chú ý

Bộ giải mã sử dụng cơ chế “chú ý” để tập trung có chọn lọc vào các phần của chuỗi đầu vào. Cơ chế chú ý lấy một chuỗi các vectơ làm đầu vào cho mỗi ví dụ và trả về một vectơ "chú ý" cho mỗi ví dụ. Lớp “chú ý” này

cũng tương tự như một lớp tổng hợp trung bình nhưng lớp chú ý thực hiện một mức trung bình có trọng số - (*a weighted average).*

Ta sử dụng các dữ liệu Vie-VSL-10K và Vie-VSL-60k cho mô hình dịch với các thông số thiết lập và quá trình này. Sau đó các số liệu đánh giá thực nghiệm được phân tích và so sánh ở phần 4.4. Mô hình Seq2Seq cho bài toán dịch VSL được công bố trên Github tại địa chỉ https://github.com/BichDiep/Seq2seq-VSL.

5.3. Mô hình Transformer cho bài toán dịch

- Cài đặt siêu tham số: Mô hình cơ sở được được sử dụng là:

num_layers = 6, d_model = 512, dff = 2048.

- Trình tối ưu hóa: Sử dụng trình tối ưu hóa Adam với công cụ lập lịch tốc độ học tập tùy chỉnh (Thuật toán tối ưu hóa Adam là một phần mở rộng cho quá trình giảm độ dốc ngẫu nhiên mà gần đây đã được áp dụng rộng rãi hơn cho các ứng dụng học sâu trong thị giác máy tính và xử lý ngôn ngữ tự nhiên).
- Huấn luyện và kiểm tra:

Sau mỗi bước huấn luyện việc lưu các checkpoint được thực hiện bằng cách tạo đường dẫn checkpoint và trình quản lý checkpoint sử dụng. Đầu vào của bài toán là câu tiếng Việt thông thường và câu đúng cú pháp trong ngôn ngữ ký hiệu Việt Nam là đầu ra.

- Để suy luận, ta cần thực hiện các bước sau đây:

- Bước 1: Bộ mã hoá thực hiện cho các câu tiếng Việt đầu vào và sử dụng bằng trình mã hóa tiếng Việt (tokenizers.Vie). Bộ mã hoá sự dụng các thông tin này để làm đầu vào.
- Bước 2: Sau đó các giá trị này sẽ được khởi tạo thành mã thông báo (START).
- Bước 3 là quá trình tính toán mặt nạ đệm (padding masks) và mặt nạ nhìn trước (look ahead masks).
- Bước 4: Bộ giải mã sẽ đưa ra các dự đoán dựa trên sự xem xét đầu ra của bộ mã hóa và đầu ra của chính nó (cơ chế tự chú ý - self-attention).

- Nối mã thông báo được dự đoán với đầu vào của bộ giải mã và chuyển nó đến bộ giải mã. Trong cách tiếp cận này, bộ giải mã dự đoán mã thông báo tiếp theo dựa trên các mã thông báo trước đó nó đã dự đoán.

• Hiển thị Attention: Lớp Translator trả về từ điển bản đồ Attention cho ta cái nhìn trực quan để hiểu được mô hình hoạt động bên trong như thế nào.

5.4. Các thực nghiệm và đánh giá kết quả

Đánh giá các thực nghiệm của các phương án đề xuất căn cứ vào điểm BLEU đánh giá kho dữ liệu mới làm giàu so sánh với tập dữ liệu gốc trên một số mô hình dịch máy. Trong các thực nghiệm này đánh giá hiệu suất dịch bằng điểm BLEU.

Bảng 5.5. So sánh điểm BLEU trên một số mô hình dịch giữa dữ liệu gốc và dữ liệu làm giàu

	Mô hình dịch	**Dữ liệu gốc**	**Dữ liệu làm giàu**
1	Dựa trên luật	68.02	68.02
2	Dịch trên mô hình IBM	42.31	60.32
3	Dịch thống kê trên mô hình IBM cải tiến	48.75	76.25
4	Seq2Seq	58.53	81.44
4	Transformer	65.22	89.23

Như vậy qua quá trình thực nghiệm với một số mô hình như trên cho chúng ta thấy với dữ liệu huấn luyện ở 10.000 cặp câu thì dịch dựa trên luật cho kết quả dịch dựa trên điểm BLEU cao hơn các mô hình thống kê. Còn với dữ liệu lớn hơn thì các mô hình thống kê cho kết quả vượt trội và tăng dần. Trong các mô hình thống kê được sử dụng thì hiện tại trong nghiên cứu của chúng tôi, mô hình Transformer là cho kết quả tốt hơn cả.

Tham chiếu kết quả đạt được của luận án với một số nghiên cứu dịch ngôn ngữ ký hiệu của một số ngôn ngữ khác ta thấy rằng điểm BLEU trong các mô hình dịch áp dụng với bài toán Vie-VSL cao vượt trội hơn so với các mô hình dịch máy các cặp ngôn ngữ khác. Như vậy, ta thấy rằng mô hình

Transformer mang lại kết quả dịch tốt trong việc dịch ngôn ngữ ký hiệu Việt Nam trong phạm vi đặt ra của bài toán này. Điểm BLEU đánh giá chất lượng bản dịch rất cao với những lý do đã được phân tích. Cụ thể đó là do tính hội tụ của mô hình ngôn ngữ, mô hình dịch gần như không thay đổi với hầu hết các đơn vị ngôn ngữ là giống nhau ở 2 ngôn ngữ.

5.4. Kết luận chương

Chương 5 đã trình bày một số mô hình thống kê và những cải tiến áp dụng cho bài toán dịch. Cụ thể là mô hình dịch IBM với cải tiến dịch dựa trên cụm từ và thêm một hệ số căn chỉnh cùng với kỹ thuật so khớp chuỗi. Với các thử nghiệm từ một phần dữ liệu nhỏ cho đến toàn bộ kho dữ liệu cho thấy mô hình dịch đề xuất có những cải tiến đáng kể so với cơ sở. Đồng thời, nguồn dữ diệu sau khi làm giàu với thuật toán trình bày ở chương 3 được sử dụng làm dữ liệu thử nghiệm một số mô hình dịch máy hiện đại dựa trên mạng noron: Seq2Seq và Transformer. Cuối cùng là các phân tích và đánh giá các bộ dữ liệu với các mô hình dịch đề xuất. Với các mô hình đề xuất cho bài toán, ta thấy rằng mô hình Transformer mang lại kết quả dịch tốt nhất trong việc dịch ngôn ngữ ký hiệu Việt Nam trong phạm vi đặt ra của bài toán này.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN NGHIÊN CỨU

Dịch tự động ngôn ngữ ký hiệu Việt Nam (VSL) là một bài toán khó và thách thức đối với các nhà nghiên cứu và nhà phát triển trong lĩnh vực xử lý ngôn ngữ tự nhiên. Ngôn ngữ ký hiệu Việt Nam là một ngôn ngữ đặc biệt, có cấu trúc cú pháp riêng biệt so với ngôn ngữ nói/viết.

Luận án này tập trung vào bài toán dịch theo chiều từ tiếng Việt sang VSL. Quá trình dịch văn bản tiếng Việt sang câu đúng cú pháp trong VSL là một trong những quá trình quan trọng nhất của bài toán này.

Luận án đã đạt được một số kết quả chính sau:

- Đề xuất một phương án dịch đơn giản và hiệu quả cho bài toán sử dụng mô hình dịch dựa trên luật.
- Đề xuất một phương pháp làm giàu dữ liệu dựa trên mạng từ cho dữ liệu song ngữ câu tiếng Việt – câu đúng cú pháp trong VSL. Từ đó Xây dựng các bộ dữ liệu: từ điển VSL-Lexicon; dữ liệu "song ngữ" Vie-VSL10k, Vie-VSL60k.
- Cải tiến một mô hình dịch thống kê cơ bản và một số mô hình dịch hiện đại dựa trên mạng Noron cho bài toán.

Các kết quả này đã góp phần nâng cao chất lượng dịch tự động ngôn ngữ ký hiệu Việt Nam, giúp người khiếm thính có thể tiếp cận được thông tin và tri thức xã hội một cách đầy đủ và chính xác hơn. Trong tương lai, nghiên cứu tiếp theo sẽ tập trung vào việc đề xuất các mô hình và phương pháp mới để tiếp tục cải thiện dịch tự động ngôn ngữ ký hiệu. Đồng thời, cần phát triển các mô hình tối ưu hơn cho các bài toán dịch máy, đặc biệt là đối với các ngôn ngữ ít tài nguyên. Những mục tiêu này sẽ đóng góp cho việc xây dựng các hệ thống dịch hoàn chỉnh hơn, giúp người khiếm thính tương tác và hòa nhập một cách hiệu quả trong cộng đồng xã hội.

DANH MỤC CÁC CÔNG TRÌNH CỦA TÁC GIẢ

[CT1]. Diep Nguyen Thi Bich, Trung-Nghia Phung, Thang Vu Tat and Lam Phi Tung, "*Special Characters of Vietnamese Sign Language Recognition System Based on Virtual Reality Glove*", the International Conference on Advances in Information and Communication Technology – ICTA, 2016.

[CT2]. Thi Bich Diep Nguyen and Trung-Nghia Phung, "*Some issues on syntax transformation in Vietnamese sign language translation". Sign Language Studies*. IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.5, pp 292-297, 2017.

[CT3]. Thi Bich Diep Nguyen, Trung-Nghia Phung, Tat-Thang Vu , "*A rule-based method for text shortening in Vietnamese sign language translation".* Springer AISC, Vol. 672, Proc. of INDIA-2017, Vietnam, 2017.

[CT4]. Nguyễn Thị Bích Điệp, "*Ứng dụng mô hình dịch máy Transformer trong bài toán dịch tự động ngôn ngữ ký hiệu Việt Nam*", Kỷ yếu hội thảo quốc gia VNICT, 2021.

[CT5]. Diep Nguyen Thi Bich, Tuyen Ho Thi, "*Data Augmentation Techniques in Automatic Translation of Vietnamese Sign Language for the Deaf*", International Conference on the Development of Biomedical Engineering -BME9, 2022.

[CT6]. Thi Bich Diep Nguyen, Trung-Nghia Phung, Tat-Thang Vu, *A Study of Data Augmentation and Accuracy Improvement in Machine translation for Vietnamese sign language,* Journal of Computer Science and Cybernetics, Vol 39, N2, pp 143-158, 2023.