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

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội – Năm 2023

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

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Mã số: 9 48 01 01

Xác nhận của Học viện Khoa học và Công nghệ	**Người hướng dẫn 1** *(Ký, ghi rõ họ tên)*	**Người hướng dẫn 2** *(Ký, ghi rõ họ tên)*
	TS. Vũ Tất Thắng	**PGS. TS. Phùng Trung Nghĩa**

Hà Nội – Năm 2023

LỜI CAM ĐOAN

Tôi xin cam đoan luận án: "***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***" là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của tập thể hướng dẫn. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ nguồn gốc. Các kết quả nghiên cứu của tôi được công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tôi làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hà Nội, tháng năm 2023

Tác giả

Nguyễn Thị Bích Điệp

LỜI CẢM ƠN

Lời đầu tiên, tôi xin được cảm ơn ***TS Vũ Tất Thắng*** và ***PGS. TS Phùng Trung Nghĩa***, các thầy đã tận tình hướng dẫn và định hướng trong quá trình nghiên cứu để tôi có thể hoàn thành luận án này.

Tôi xin cảm ơn ***TS Vũ Thị Hải Hà*** – Viện Ngôn ngữ học, Viện Hàn lâm Khoa học xã hội Việt Nam là người đã tận tình giúp đỡ trong quá trình xây dựng dữ liệu phục vụ cho bài toán.

Tôi cũng xin được bày tỏ lòng cảm ơn sâu sắc tới các Thầy, Cô của Viện Công nghệ thông tin, Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam và các nhà khoa học đầu ngành trong lĩnh vực nghiên cứu là cô ***PGS. TS Lương Chi Mai***, thầy ***TS Nguyễn Văn Vinh***, thầy ***TS Nguyễn Phương Thái*** đã có những góp ý khoa học rất xác đáng để tôi có thể bổ sung, chỉnh sửa và đánh giá kết quả của mình một các toàn diện hơn.

Cuối cùng tôi xin cảm ơn Ban giám hiệu trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên và các đồng nghiệp nơi tác giả công tác đã tạo điều kiện về công việc và ủng hộ để luận án được hoàn thành.

Hà Nội, tháng năm

Nguyễn Thị Bích Điệp

MỤC LỤC

Trang

DANH MỤC TỪ VIẾT TẮT

Ký hiệu	Tên đầy đủ	Tên tiếng Việt
ASL	American Sign Language	Ngôn ngữ ký hiệu Mỹ
BLEU	Bilingual Evaluation Understudy	Điểm đánh giá bản dịch song ngữ
BSL	British Sign Language	Ngôn ngữ ký hiệu Anh
DSG	Deutsche Gebärdensprache	Ngôn ngữ ký hiệu Đức
DRS	Discourse Representation Structure	Cấu trúc đại diện diễn đạt
KSL	Korean Sign Language	Ngôn ngữ ký hiệu Hàn Quốc
ISL	Indian Sign Language	Ngôn ngữ ký hiệu Ấn Độ
MT	Machine Translation	Dịch máy
NMT	Neural Machine Translation	Dịch máy dựa trên mạng Nơron
SMT	Statistical Machine Translation	Dịch máy thống kê
RBMT	Rules Based Machine Translation	Dịch máy dựa trên luật
PSL	Pakistani Sign Language	Ngôn ngữ ký hiệu Pakistan
SL	Sign Language	Ngôn ngữ ký hiệu
STAG	Synchronous Tree Adjoining Grammar	Cây đồng bộ ngữ pháp liền kề
VRML	Virtual Reality Modeling Language	Mô hình ngôn ngữ thực tế ảo
VSL	Vietnamese Sign Language	Ngôn ngữ ký hiệu Việt Nam
Vie-VSL	Vietnamese - Vietnamese Sign Language	Tiếng Việt - Ngôn ngữ ký hiệu Việt Nam
VLSP	Association for Vietnamese Language and Speech Processing	Cộng đồng xử lí văn bản và tiếng nói tiếng Việt

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

MỞ ĐẦU

1. Tính cấp thiết của luận án

Ngôn ngữ ký hiệu tiếng Việt và ngôn ngữ ký hiệu nói chung hiện có không nhiều các nghiên cứu cũng như các tài liệu chuyên môn chính thống như ngôn ngữ học, ngữ pháp, từ điển,… Vì thế đã giới hạn rất nhiều việc truyền tải thông tin tới những người khiếm thính và việc giao tiếp của người khiếm tính với cộng đồng. Tính cấp thiết của bài toán dịch ngôn ngữ kí hiệu Việt Nam được thể hiện qua bối cảnh xã hội và các nghiên cứu quốc tế đã đưa ra một số vấn đề nghiên cứu chính. Trên thế giới, đã có những nỗ lực đáng kể trong việc nghiên cứu và triển khai các hệ thống dịch ngôn ngữ kí hiệu:

- ASL-English Translation: Ở Hoa Kỳ, nơi sử dụng ngôn ngữ kí hiệu Mỹ (ASL), đã có các nghiên cứu về dịch từ ASL sang tiếng Anh và ngược lại. Điều này đã giúp cải thiện giao tiếp giữa người khiếm thính sử dụng ASL và người nói tiếng Anh, đặc biệt trong các tình huống quan trọng như chăm sóc sức khỏe và giáo dục.
- International Sign Language: Trong các sự kiện quốc tế như hội nghị, triển lãm, hay sự kiện thể thao, việc dịch ngôn ngữ kí hiệu trở nên quan trọng. Ngôn ngữ kí hiệu quốc tế (International Sign Language) đã được phát triển để giúp người khiếm thính từ các quốc gia khác nhau có thể giao tiếp. Tính cấp thiết của việc phát triển các công cụ dịch giữa các ngôn ngữ kí hiệu khác nhau đã được thể hiện rõ ràng.
- Educational Sign Language Translation: Trong lĩnh vực giáo dục, dịch ngôn ngữ kí hiệu đóng vai trò quan trọng để đảm bảo rằng người khiếm thính có quyền truy cập vào kiến thức. Nhiều quốc gia đã triển khai các hệ thống dịch máy trong lĩnh vực này để hỗ trợ việc giảng dạy và học tập.

Những ví dụ trên chỉ ra rằng bài toán dịch ngôn ngữ kí hiệu không chỉ là một vấn đề cấp thiết tại Việt Nam mà còn trên toàn thế giới. Việc cải thiện hiệu suất và chất lượng của các hệ thống dịch ngôn ngữ kí hiệu là cần thiết để tạo điều kiện giao tiếp tốt hơn cho cộng đồng người khiếm thính và giúp họ tham gia hoàn toàn vào xã hội và giáo dục.

Quá trình thông dịch ngôn ngữ ký hiệu bao gồm 2 bài toán là 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 khiếm thính.

Các dự án nghiên cứu ban đầu về bài toán này hầu hết đều là trên các mô hình dịch dựa trên cấu trúc. 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) 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.

Một vấn đề quan trọng được các nhà nghiên cứu về lĩnh vực dịch ngôn ngữ ký hiệu chỉ ra rằng, sự chấp nhận trong cộng đồng người khiếm thính là rất quan trọng đối với việc áp dụng các công nghệ tạo ngôn ngữ ký hiệu. Quan điểm của người dùng khiếm thính phải được phân tích chính xác và việc thực thi công nghệ đối với cộng đồng người khiếm thính sẽ không hiệu quả. Do vậy, việc chú trọng đến cấu trúc cú pháp trong ngôn ngữ ký hiệu là quan trọng, nhưng trong nhiều nghiên cứu lại xem nhẹ vấn đề này.

Các nghiên cứu về dịch Ngôn ngữ ký hiệu trên thế giới quan tâm đến bài toán dịch text-to-text thường áp dụng cả những phương pháp cổ điển và hiện đại. Một số phương pháp dịch máy dựa trên luật và thống kê cổ điển với cách tiếp cận đơn giản nhưng vẫn mang lại những hiệu quả nhất định bởi những đặc trưng trong cú pháp ngôn ngữ ký hiệu. Các phương pháp hiện đại dựa trên học sâu (Deep Learning) trong các nghiên cứu gần đây với các ưu điểm nổi bật như tự động học các đặc trưng của ngôn ngữ ký hiệu từ dữ liệu đầu vào mà không cần đến các quy tắc cụ thể; độ chính xác cao; có khả năng mở rộng với tốc độ xử lý nhanh. Tuy nhiên nhược điểm của hầu hết các nghiên cứu này là chưa có một bộ dữ liệu đủ lớn để đào tạo cũng như ít xem xét đến quan điểm và đánh giá của người khiếm thính.

Hầu hết các đặc điểm của bài toán dịch SL mang những tính chất tương đồng như vấn đề của các bài toán dịch ngôn ngữ khác. Chúng đều là quá trình sử dụng trí tuệ nhân tạo để tự động dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác mà không cần sự tham gia của con người. Tuy nhiên, trong bài toán dịch VSL là bài toán với

đầu vào là câu tiếng Việt thông thường, đầu ra cuối cùng là hình ảnh, video, các mô hình diễn họa 3D. Nhưng một bước trung gian quan trọng của quá trình dịch là từ câu tiếng Việt thông thường sang câu dạng đúng cú pháp trong VSL. Bởi lẽ, VSL có một số đặc trưng cơ bản như tính giản lược, nhấn mạnh trọng tâm và thay đổi trật tự từ so với ngôn ngữ tiếng Việt thông thường. Ngoài ra, việc biểu diễn từ câu đúng cú pháp trong VSL sang các dạng hình ảnh, mô hình 3D đã có những phương pháp kỹ thuật được đề xuất và có kết quả tốt. Điều này có nghĩa là các thành phần trong câu được tách ra và lưu trữ trong từ điển dưới dạng chú thích có 2 thành phần cơ bản là từ/cụm từ và mô hình 3D biểu diễn trực quan. Việc chuyển động liên kết mềm mại giữa các thành phần trong câu được xử lý bằng một số kỹ thuật nội suy. Do vậy trọng tâm của bài toán dịch ngôn ngữ ký hiệu Việt Nam vẫn là vấn đề 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).

Tóm lại, bài toán đặt ra trong luận án này bao gồm việc 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. Song song với đó là việc 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.

2. Mục tiêu của luận án

Mục tiêu của bài toán dịch tự động ngôn ngữ ký hiệu Việt Nam là tạo ra một hệ thống dịch máy hiệu quả và đáng tin cậy để dịch từ ngôn ngữ thông thường (văn bản tiếng Việt) sang ngôn ngữ kí hiệu Việt Nam (VSL). Mục tiêu này có thể được đảm bảo thông qua hai hướng tiếp cận chính:

- Mô hình dịch máy dựa trên mạng neuron hiện đại (Neural Network-Based Machine Translation): Mô hình này sử dụng tiến bộ trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên (NLP) để xây dựng một hệ thống dịch ngôn ngữ kí hiệu hiệu quả. Mô hình dựa trên mạng neuron có khả năng tự động học và hiểu các đặc trưng của ngôn ngữ kí hiệu từ dữ liệu đầu vào. Điều này bao gồm cả việc học cú pháp, ngữ nghĩa, và ngữ vựng của VSL. Mô hình dựa trên mạng neuron có thể

tối ưu hóa để cải thiện hiệu suất dịch và đảm bảo tính chính xác trong việc truyền đạt thông điệp giữa ngôn ngữ thông thường và VSL.

- Phương pháp dịch dựa trên luật (Rule-Based Translation): Mặc dù mô hình dựa trên mạng neuron là tiến bộ và mạnh mẽ, nhưng cũng có thể kết hợp phương pháp dịch dựa trên luật. Điều này đặc biệt hữu ích khi chúng ta cần đảm bảo tính chính xác và đúng cú pháp trong ngôn ngữ kí hiệu. Phương pháp dịch dựa trên luật có thể xác định các quy tắc cụ thể để ánh xạ từ ngôn ngữ thông thường sang VSL và ngược lại. Điều này giúp đảm bảo rằng thông điệp không bị biến dạng và đảm bảo tính chính xác trong truyền đạt.

Mục tiêu tổng quát của bài toán là giúp cộng đồng người khiếm thính tại Việt Nam có cơ hội truyền đạt thông điệp, tham gia vào xã hội và học tập một cách dễ dàng và hiệu quả hơn, từ đó nâng cao chất lượng cuộc sống của họ và sự hiểu biết trong xã hội.

3. Phương pháp nghiên cứu

Phương pháp nghiên cứu cho bài toán dịch ngôn ngữ ký hiệu Việt Nam bao gồm các phần sau:

- *Thu thập dữ liệu*: Để xây dựng và đào tạo mô hình dịch, cần thu thập một bộ dữ liệu lớn về cặp câu hoặc văn bản tiếng Việt và tương ứng trong ngôn ngữ kí hiệu Việt Nam (VSL). Dữ liệu này phải phản ánh đa dạng về ngữ cảnh và chủ đề để đảm bảo tính đại diện.
- *Tiền xử lý dữ liệu*: Dữ liệu thu thập được thường cần được tiền xử lý để loại bỏ nhiễu và chuẩn hóa. Điều này có thể bao gồm việc làm sạch văn bản, phân đoạn câu, và chuyển đổi dữ liệu thành định dạng phù hợp cho quá trình đào tạo mô hình.
- *Xây dựng mô hình dịch máy*: Luận án sử dụng nhiều mô hình dịch máy khác nhau cho bài toán này, bao gồm mô hình dựa trên mạng nơ-ron (neural networks) như Transformer, Seq2Seq, mô hình dựa trên cấu trúc và luật ngữ pháp của ngôn ngữ ký hiệu. Mô hình sẽ được đào tạo trên dữ liệu đã thu thập và tiền xử lý để học cách dịch giữa hai ngôn ngữ.

- *Đánh giá và kiểm tra*: Để đảm bảo hiệu suất của mô hình, cần thực hiện các bước kiểm tra và đánh giá. Phương pháp đánh giá được sử dụng là đánh giá tự động bằng BLEU.
- *Tối ưu hóa và cải tiến*: Dựa trên kết quả đánh giá, mô hình và phương pháp dịch có thể được tối ưu hóa và cải tiến. Điều này bao gồm tinh chỉnh siêu tham số.

Tổng hợp lại, phương pháp nghiên cứu cho bài toán dịch ngôn ngữ ký hiệu Việt Nam là một quá trình bao gồm các bước từ thu thập dữ liệu, xây dựng mô hình, đánh giá, tối ưu hóa. Các phương pháp dựa trên mạng nơ-ron và dựa trên luật có thể giúp đảm bảo tính chính xác và hiệu quả của hệ thống dịch ngôn ngữ ký hiệu.

4. 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) 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], CT[3] [CT7].
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.

5. 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.

Trong toàn bộ luận án, khi đề cập tới dịch ngôn ngữ ký hiệu Việt Nam, các phương án triển khai và đánh giá được hiểu là ***quá trình dịch với đầu vào là văn bản trong tiếng Việt thông thường, đầu ra là dạng văn bản đúng cú pháp trong VSL.***

6. Cấu trúc của luận án

Nội dung chính của luận án như sau:

- 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 nơron trong dịch tự động ngôn ngữ ký hiệu Việt Nam, thực nghiệm và đánh giá các kết quả trên các phương pháp này.
- Kết luận: Đưa ra những kết quả đạt được của luận án; đánh giá ưu nhược điểm và định hướng phát triển.

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

Trong chương này, luận án trình bày những vấn đề tổng quan của bài toán dịch ngôn ngữ ký hiệu Việt Nam. Mục 1.1 là những vấn đề tổng quan về đặc điểm của VSL. Phần tiếp theo là những nghiên cứu liên quan về chủ đề dịch ngôn ngữ ký hiệu trên thế giới và những phân tích những ưu nhược điểm của các nghiên cứu này (mục 1.2). Đồng thời chỉ ra điểm mấu chốt của dịch ngôn ngữ ký hiệu Việt Nam trong bài toán dịch từ ngôn ngữ thông thường sang ngôn ngữ ký hiệu chính là vấn đề dịch từ văn bản tiếng Việt thông thường sang dạng văn bản đúng cú pháp trong VSL.

1.1. Tổng quan về ngôn ngữ ký hiệu

1.1.1. Lịch sử và phân loại ngôn ngữ ký hiệu trên thế giới

Ngôn ngữ ký hiệu được hình thành từ rất sớm gắn với sự phát triển ngôn ngữ thông thường. Năm 1620, Juan Bonet đã đưa ra luận thuyết đầu tiên được coi là tiền đề cho ngôn ngữ ký hiệu. 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 [1].

Các cộng đồng người khiếm thính trên thế giới đều có ngôn ngữ ký hiệu của riêng họ. Trong khi một số ngôn ngữ ký hiệu như ngôn ngữ ký hiệu Anh, Ngôn ngữ ký hiệu Mỹ, ngôn ngữ ký hiệu Ba Lan, ngôn ngữ ký hiệu Ấn Độ,v.v…được công nhận pháp lý thì cũng có một số mang tính chất địa phương. Có một quan niệm không chính xác là ngôn ngữ ký hiệu là một loại dùng chung trên toàn thế giới. Nhưng mỗi quốc gia lại có hơn một ngôn ngữ ký hiệu. Do những đặc điểm vùng miền về văn hoá xã hội nên chính trong ngôn ngữ ký hiệu Việt Nam cũng có những điểm khác biệt. Các chuyên gia về ngôn ngữ đã phân ra 3 vùng miền cho ngôn ngữ ký hiệu Việt Nam, đó là: Hà Nội, Hải Phòng và Thành phố Hồ Chí Minh. Tuy vậy thì kể cả là ngôn ngữ ký hiệu trên thế giới cũng có một số điểm tương đồng nhất định về đặc điểm cú pháp hay hình thái của biểu diễn một số từ ngữ. Ví dụ như từ "lái ô tô" thì đều thể hiện là hai tay giơ lên nắm không khí và làm động tác quay quay xoay vô lăng.

Mặc dù ngôn ngữ ký hiệu đã phát sinh một cách tự nhiên trong cộng đồng người khiếm thính bên cạnh ngôn ngữ nói, tuy nhiên chúng không liên quan đến ngôn ngữ nói và có cấu trúc ngữ pháp khác nhau ở cốt lõi. Ngôn ngữ ký hiệu có thể được phân loại theo cách nó phát sinh.

Ngôn ngữ ký hiệu địa phương là một ngôn ngữ bản địa mà thường phát sinh nhiều thế hệ trong một cộng đồng tương đối biệt lập với một tỷ lệ cao người khiếm thính, và được sử dụng bởi người khiếm thính và một phần đáng kể của cộng đồng nghe, gồm có gia đình và bạn bè của người khiếm thính. Lúc đầu, ngôn ngữ ký hiệu cộng đồng người khiếm thính thường không được biết đến bởi người nghe nói bình thường, trong nhiều trường hợp thậm chí các thành viên trong gia đình cũng không thể sử dụng thứ ngôn ngữ này. Tuy nhiên, chúng có thể phát triển, trong một số trường hợp trở thành một ngôn ngữ giảng dạy và nhận được sự công nhận chính thức, như trong trường hợp của ASL.

1.1.2. Đặc điểm về cú pháp trong câu ngôn ngữ ký hiệu Việt Nam

Trong VSL, cũng tương tự như các ngôn ngữ ký hiệu khác trên thế giới đều có 2 đặc điểm quan trọng nhất đó chính là sự giản lược và nhấn mạnh trọng tâm. Điều này là do trong đặc trưng tư duy của người khiếm thính và vì vậy có ảnh hưởng đến cách biểu đạt ngôn ngữ trong cú pháp của câu ngôn ngữ ký hiệu: sự rút gọn một số thành phần trong câu và sự sắp xếp trật tự các từ trong câu có sự khác biệt với ngôn ngữ thông thường. Vốn từ vựng của người khiếm thính cũng hạn chế so với thông thường nên những thành phần được coi là không quá quan trọng và không mang nhiều ý nghĩa trong câu sẽ được giản lược đi [2]. Việc giản lược này được phân tích kỹ ở phần tiếp theo. Ngoài ra, với đặc điểm nhấn mạnh trọng tâm, người khiếm thính có xu hướng đưa các từ quan trọng trong câu lên phía trước, khiến cho trật tự cú pháp của câu ngôn ngữ ký hiệu sẽ bị đảo lộn so với thông thường. Việc đảo trật tự cú pháp cũng dựa trên từng loại câu. Có thể liệt kê ra như sau:

Câu tường thuật: thường được lược bỏ một số giới từ, liên từ, đưa danh từ lên trước số đếm và động từ, ...

Câu nghi vấn: Đại từ nghi vấn thường được đặt ở cuối câu và đưa thông tin nghi vấn lên trên để tập trung chú ý. Thông thường câu nghi vấn có thể kèm thêm sự biểu đạt trên khuôn mặt. Một điểm đáng chú ý là các biểu đạt khuôn mặt được sử dụng khá nhiều trong ngôn ngữ ký hiệu

Câu phủ định: Từ phủ định trong câu thường được đặt cuối câu, giống như đặc điểm mà ta vừa phân tích ở câu nghi vấn.

Câu mệnh lệnh: từ tình thái là một thành phần chắc chắn được giản lược trong câu vì nếu bỏ ra khỏi câu, ý nghĩa hầu như vẫn được giữ nguyên để có thể hiệu được, trật tự từ trong câu mệnh lệnh cũng bị đảo vị trí so với ngôn ngữ tiếng Việt thông thường.

Do vậy, đặc trưng cơ bản trong việc chuyển đổi câu Tiếng việt thông thường sang dạng câu đúng cú pháp trong VSL sẽ bao gồm 2 yếu tố: Rút gọn và thay đổi trật tự cú pháp. Những đặc điểm ngôn ngữ học đặc trưng này sẽ được phân tích cụ thể.

A. Rút gọn giới từ, liên từ và từ tình thái

Giới từ được dùng để đánh dấu quan hệ chính phụ.Quan hệ chính phụ ở đây có thể là giữa một ngữ danh từ với định ngữ của nó ,giữa một ngữ vị từ với bổ ngữ của nó, giữa câu với trạng ngữ của nó.

Liên từ thì thông thường được hình dung là từ dùng để liên kết các ngữ đoạn (ngữ, cấu trúc đề thuyết) đẳng lập với nhau. Dù liên từ có thể vẫn biểu đạt quan hệ ngữ nghĩa giữa các ngữ đoạn, các câu lại với nhau nhưng nó không đảm nhiệm vai trò đánh dấu vai nghĩa. Và vì vậy, chức năng nổi bật của nó vẫn chỉ là nối kết các ngữ đoạn, các câu lại với nhau để diễn đạt mối quan hệ ý nghĩa giữa các thành phần này. Như vậy, giới từ là liên từ chính là thành phần phụ không mang nhiều ý nghĩa trong câu, nghĩa là khi ta bỏ nó khỏi câu thì câu rút gọn vẫn có thể hiểu được. Bởi sự hạn chế về mặt từ vựng và ngữ nghĩa nên trong câu VSL sẽ giản lược giới từ và liên từ trong câu. Điều này phù hợp với đặc điểm ngôn ngữ của người khiếm thính. Bảng 1.1 trình bày một số mẫu câu rút gọn giới từ và liên từ.

Tình thái từ là những từ được thêm vào câu để cấu tạo câu theo mục đích nói (nghi vấn, cầu khiến, cảm thán) và để biểu thị các sắc thái tình cảm của người đó. Đối với người khiếm thính, việc biểu thị sắc thái tình cảm hay cấu tạo câu theo mục đích nói thông thường sẽ dùng biểu cảm khuôn mặt và một số dấu hiệu nhất định. Vì vậy mà trong ngôn ngữ ký hiệu không có những ký hiệu để biểu đạt các từ tình thái này.

Trong phần này, luận án nghiên cứu các vấn đề liên quan đến từ tình thái và thu thập các luật rút gọn từ tình thái trong câu thông thường để biến đổi sang dạng

câu trong ngôn ngữ ký hiệu. Các tình thái từ là những từ biểu lộ thái độ tình cảm của người nói (người viết) đối với nội dung của câu hoặc đối với người cùng tham gia hoạt động giao tiếp (người nghe. người đọc). Các tình thái từ không thể đóng vai trò thành phần cấu tạo trong cụm từ hay trong câu, chúng chỉ được dùng trong câu để bày tỏ thái độ tình cảm.

Bảng 1.1. Một số mẫu câu rút gọn giới từ và liên từ

Câu thông thường	**Câu rút gọn liên từ và giới từ**
Viết ***bằng*** bút chì	Viết bút chì
Ăn ***và*** mặc là nhu cầu ***của*** mọi người	Ăn mặc là nhu cầu mọi người
Tôi *và* anh đi học	Tôi anh đi học
Anh ăn cháo ***hay*** ăn cơm?	Anh ăn cháo ăn cơm?
Mặc dầu trời mưa, tôi ***vẫn*** đi học	Trời mưa, tôi đi học
Lấy **hộ** chị quyển sách	Lấy chị quyển sách
Buổi sáng anh dắt xe ***giúp*** tôi ra cổng	Buổi sáng anh dắt xe tôi ra cổng
Áo ***của*** anh màu xanh	Áo anh màu xanh

Như vậy, trong ngôn ngữ tiếng Việt thông thường, khi rút gọn sang dạng ngôn ngữ ký hiệu, ta lược bỏ các từ tình thái trong câu. Các từ tình thái như đã liệt kê ở trên sẽ được loại bỏ câu theo một cấu trúc xác định dựa trên ngữ nghĩa.

B. Đặc điểm về trật tự cú pháp trong câu VSL

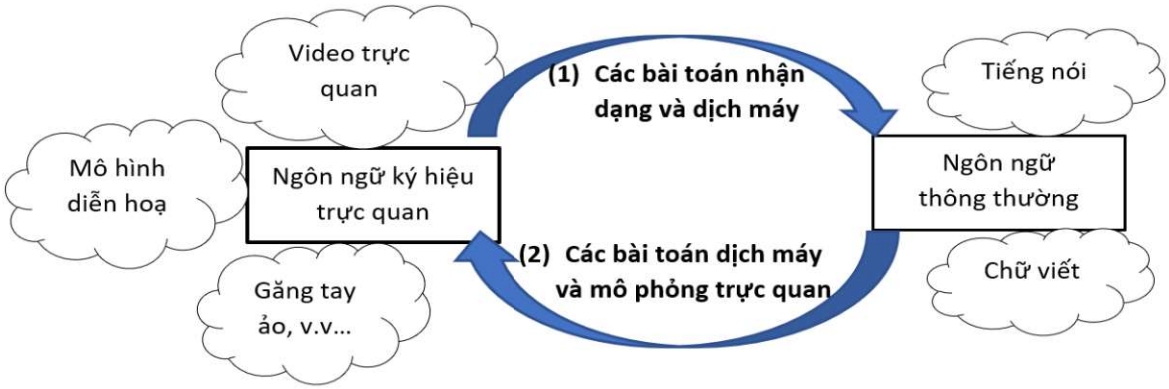
Tính chất rút gọn trong câu VSL khiến cho bài toán chuyển đổi câu tiếng việt sang dạng đúng trong VSL gần giống như bài toán tóm tắt văn bản. Tuy nhiên đặc trưng khác biệt so với bài toán tóm tắt văn bản là vấn đề trật tự cú pháp trong câu VSL. Do những đặc điểm đặc trưng của ngôn ngữ, thông tin chính được nhấn mạnh và thường đưa lên trước nên cú pháp câu VSL có trật tự cú pháp khác so với câu tiếng Việt thông thường [3].

Tóm lại, đặc trưng cơ bản nhất của các ngôn ngữ ký hiệu trên thế giới nói chung hay ngôn ngữ ký hiệu Việt Nam nói riêng thì chúng đều có đặc điểm về tính giản lược và nhấn mạnh trọng tâm. Điều đó khiến cho cú pháp trong câu ngôn ngữ ký hiệu có nhiều sự khác biệt với ngôn ngữ thông thường. Các đặc điểm này được phân tích, tổng hợp lại trong chương 3 liên quan đến việc dịch VSL theo cấu trúc.

1.2. Các nghiên cứu liên quan

Hầu hết các đặc điểm của bài toán dịch SL mang những tính chất tương đồng như vấn đề của các bài toán dịch ngôn ngữ khác. Chúng đều là quá trình sử dụng trí tuệ nhân tạo để tự động dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác mà không cần sự tham gia của con người.

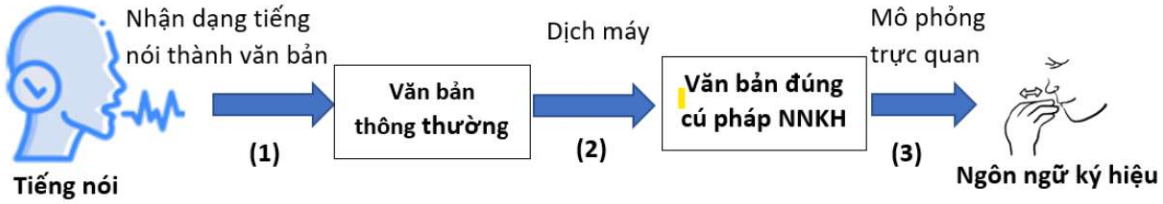
Vấn đề về dịch ngôn ngữ ký hiệu trên thế giới được chia thành 2 lớp bài toán. Một là dịch từ ngôn ngữ thông thường sang ngôn ngữ ký hiệu. Hai là dịch theo chiều ngược lại tức là từ ngôn ngữ ký hiệu sang dạng chữ viết hoặc giọng nói trong ngôn ngữ thông thường. Hình 1.1. miêu tả hai chiều của bài toán dịch ngôn ngữ ký hiệu.



Hình 1.1. Hai chiều của bài toán dịch ngôn ngữ ký hiệu.

Với những tiến bộ vượt bậc về khoa học công nghệ trong lĩnh vực công nghệ thông tin, trên thế giới đã có những hệ thống dịch ngôn ngữ kí hiệu ví dụ như: TESSA - Dịch từ tiếng nói sang ngôn ngữ kí hiệu Anh (BSL) [4]; trình dịch ViSiCAST là công cụ để dịch từ Tiếng Anh sang dạng ngôn ngữ kí hiệu Anh [5]; Dự án SignSynth sử dụng mô hình ASCII-Stokoe [6] ; Hệ thống ASL workbench là hệ thống dịch tự động văn bản sang dạng ngôn ngữ kí hiệu Mỹ [7]; Dự án TEAM là một hệ thống dịch từ văn bản sang dạng ngôn ngữ kí hiệu Mỹ sử dụng kĩ thuật cây đồng bộ ngữ pháp liền kề [8]. Các dự án như SignAloud [9], Kinect Sign Language Translator [10], SignAll và MotionSavvy [11], v.v… dịch các từ hoặc câu ngôn ngữ ký hiệu được mô tả bởi hình ảnh trực quan như video, ký hiệu trực tiếp từ gang tay ảo sang ngôn ngữ nói. 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 khiếm thính.

Quá trình dịch ngôn ngữ thông thường thành ngôn ngữ ký hiệu gồm các bước:



Hình 1.2. Quá trình dịch ngôn ngữ thông thường thành ngôn ngữ ký hiệu

Trong đó, (1) là quá trình dịch từ nhận dạng tiếng nói thành văn bản. Đã có nhiều nghiên cứu và các ứng dụng xử lý tốt phần công việc này, ví dụ như API của Google. (2) là quá trình xử lý từ văn bản thông thường sang dạng đúng cú pháp trong ngôn ngữ ký hiệu. (3) là quá trình mô phỏng từ dạng văn bản đúng cú pháp trong ngôn ngữ ký hiệu thành các biểu diễn như mô hình 3D hay các video, hình ảnh của ngôn ngữ ký hiệu.

Trong thủ tục này, bước thứ hai nhận được nhiều nhất sự chú ý do hoàn thành thông điệp được truyền tải. Thách thức cơ bản là ngôn ngữ ký hiệu nói chung có vốn từ vựng hạn chế so với sang ngôn ngữ nói/viết. Nếu bản dịch máy được thực hiện kém, thông báo hoàn chỉnh 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.

Những năm gần đây, dịch dựa trên cấu trúc vẫn được ứng dụng trong một số bài toán dịch ngôn ngữ ký hiệu. Tác giả Porta và các cộng sự nghiên cứu các cách tiếp cận dựa trên chuyển giao và áp dụng thuật toán tạo thứ tự từ để xử lý được định hướng theo chủ đề của ngôn ngữ ký hiệu Tây Ban Nha (LSE), tuân theo các thuật toán để che lấp các lỗ hổng ngữ nghĩa và từ vựng trong quá trình dịch thuật. Kết quả của nghiên cứu này là mô hình dịch thuật từ văn bản thông thường tạo ra văn bản dạng chú thích của LSE [12].

Trong một nghiên cứu về ngôn ngữ ký hiệu Bồ Đào Nha, tác giả mô tả một số kỹ thuật trích xuất thông tin được áp dụng trước khi được chuyển sang giai đoạn mô phỏng trực quan để phân tích. Ưu điểm của nghiên cứu này là sử dụng phần mềm mã nguồn mở tuy nhiên một điểm yếu được chỉ ra là do việc không chú trọng cú pháp ngôn ngữ ký hiệu dẫn đến việc dịch không có nhiều ý nghĩa thực tế [13].

Năm 2018, tác giả Kouremenos và các cộng sự đã ngiên cứu tạo ra các mô hình ngôn ngữ tốt với kho ngữ liệu lớn và chất lượng tốt ngôn ngữ ký hiệu Hy Lạp Đây cũng là nghiên cứu với kết quả đầu ra cuối là văn bản đúng cú pháp ngôn ngữ ký hiệu Hy Lạp [14].

Các nghiên cứu về dịch ngôn ngữ ký hiệu theo phương pháp tiếp cận dựa trên cấu trúc trong giai đoạn từ năm 1994 đến nay đã đạt được một số kết quả nhất định. Dù là một cách tiếp cận cổ điển trong dịch máy nhưng phương pháp này vẫn có những ưu điểm cho bài toán dịch ngôn ngữ ký hiệu. Dịch dựa trên luật là một phương án đơn giản và hiệu quả, phù hợp với ngôn ngữ ít tài nguyên như ngôn ngữ ký hiệu. Thêm nữa, những đặc trưng của việc dịch ngôn ngữ ký hiệu (mô hình ngôn ngữ đơn giản với mô hình xác suất hội tụ và mẫu câu đơn giản, tính chất rút gọn của ngôn ngữ ký hiệu) phù hợp với mô hình dịch theo luật.

Đối với phương pháp thống kê cổ điển thì mục tiêu dịch text-to-text trong dịch ngôn ngữ ký hiệu có những kết quả khả quan với nghiên cứu của dự án dịch Tiếng Anh sang ngôn ngữ ký hiệu tiếng Hà Lan (DSL) năm 2005 [15]. Dự án này là một trong những cách tiếp cận ban đầu với việc sử dụng giả thuyết đánh dấu cho phân đoạn văn bản tiếng Anh. Các phương pháp phân đoạn được sử dụng đã cung cấp các khối dữ liệu tương tự giúp căn chỉnh trong quá trình dịch. Nhưng nhược điểm của nghiên cứu này là kho ngữ liệu câu và từ vựng nhỏ với kết quả đầu ra là một dạng chú thích bằng văn bản sử dụng công cụ ELAN (EUDICO Linguistic Annotator).

Năm 2012, tác giả Lopez và các cộng sự công bố một nghiên cứu về dịch Tiếng Tây Ban Nha sang ngôn ngữ ký hiệu Tây Ban Nha (LSE). Với mô-đun tiền xử lý (sử dụng danh sách thẻ từ) của hệ thống được tích hợp vào cấu trúc dựa trên cụm từ. Mô hình tiền xử lý làm giảm sự thay đổi trong ngôn ngữ nguồn [16].

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. Trong nghiên cứu mới đây, bài toán dịch ngôn ngữ ký hiệu Thổ Nhĩ Kỳ sử dụng công cụ phân tích cú pháp cho thống kê trong miền dữ liệu đối với học sinh tiểu học [17]. Tuy nhiên nghiên cứu này không đề cập đến khả năng mở rộng của hệ thống.

Tóm lại với các phương pháp thống kê được áp dụng cho việc dịch text-to-text trong bài toán dịch ngôn ngữ ký hiệu thì hầu hết các nghiên cứu đều chỉ áp dụng trên một dữ liệu nhỏ. Với giới hạn của dữ liệu huấn luyện hệ thống có thể gây ảnh hưởng đến chất lượng dịch.

Ngoài ra, để hạn chế các nhược điểm và nâng cao hiệu quả của phương pháp dịch, một số nghiên cứu về dịch ngôn ngữ ký hiệu theo phương pháp kết hợp. Đây là các nghiên cứu với đầu vào là văn bản thông thường, đầu ra là các chú thích ngôn ngữ ký hiệu dưới dạng văn bản. Các phân tích và tổng hợp về ưu nhược điểm của những nghiên cứu này được trình bày trong bảng 1.2.

Bảng 1.2. Một số dự án sử dụng dịch máy kết hợp cho mục tiêu dịch text-to-text của bài toán dịch ngôn ngữ ký hiệu

Tài liệu	**Ngôn ngữ**	**Mô tả**	**Điểm mạnh**	**Hạn chế**	**Năm**
[18]	Tiếng Hy Lạp sang ngôn ngữ ký hiệu Hy Lạp	Kết hợp dịch dựa trên luật và dịch máy thống kê (RBMT+STM), tạo một kho văn bản song song lớn được sử dụng làm dữ liệu đào tạo	Quy trình không cần kiến thức ngữ pháp sâu về GSL. Sử dụng dữ liệu song ngữ hiệu quả để huấn luyện hệ thống dịch máy	Dữ liệu huấn luyện nhỏ và chỉ sử dụng cho tiếng Hy Lạp	2018
[19]	Tiếng Anh sang ngôn ngữ ký hiệu Mỹ	Xây dựng ngữ liệu nhân tạo bằng cách sử dụng các quy tắc phụ thuộc ngữ pháp	Các thuật toán của IBM được tăng cường bằng cách tích hợp khoảng cách Jaro-Winkler	Cần có sự chính xác và độ tin cậy cao trong quá trình thu thập và xử lý dữ liệu và chỉ sử dụng cho ASL	2019
[20]	Tiếng Ả Rập sang ArSL	Các quy tắc dịch thuật và cơ sở dữ liệu về các ký hiệu được sử dụng để dịch thuật, và các tên riêng được đánh vần chính xác hơn	Sự kết hợp giữa kiến thức ngôn ngữ và cơ sở dữ liệu về các dấu hiệu mang lại độ chính xác cao hơn cho bản dịch	Không có các đánh giá.	2019

Tài liệu	Ngôn ngữ	Mô tả	Điểm mạnh	Hạn chế	Năm
[21]	Thổ Nhĩ Kỳ sang ngôn ngữ ký hiệu Thổ Nhĩ Kỳ	Mô hình dịch sau khi áp dụng các quy tắ c, kết quả trung gian được đưa vào thành phần thống kê	Các quy tắc dành riêng cho ngôn ngữ làm tăng hiệu suất tổng thể của hệ thống	Dữ liệu huấn luyện nhỏ, chất lượng dịch hạn chế, hệ thống phức tạp	2019

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.

Một số nghiên cứu với cách tiếp cận liên ngôn ngữ cho dịch, sử dụng kết hợp học máy và học sâu cũng cho một số kết quả khả thi [22]. Tác giả Manzano và các cộng sự cho bài toán dịch từ tiếng Anh sang BSL đã đạt những kết quả tương đối tốt nhưng có hạn chế về số lượng từ vựng. Nghiên cứu này mô phỏng hình đại diện ảo bằng văn bản [23].

Năm 2021, tác giả Muhammad và các cộng sự nghiên cứu việc dịch các câu sang Ngôn ngữ ký hiệu Pakistan (PSL) cho người khiếm thính với biểu diễn trực quan bằng cách sử dụng ký tự ảo. Nghiên cứu này dựa trên cơ sở kiến thức lưu trữ cả kho văn bản của các từ PSL và dạng mã hóa của chúng trong hệ thống ký hiệu. Sign4PSL sử dụng văn bản tiếng Anh làm đầu vào, thực hiện dịch sang PSL thông qua ký hiệu ngôn ngữ ký hiệu và hiển thị cử chỉ cho người dùng bằng ký tự ảo. Dạng mã hoá trong hệ thống ký hiệu cũng có thể được coi như là một dạng văn bản đúng cấu trúc cú pháp trong ngôn ngữ ký hiệu. Nhưng đặc điểm về cấu trúc cú pháp của PSL chưa được xem xét cụ thể. [24].

Một số nghiên cứu mới đây tập trung vào mô hình hiện đại Transformer cho bài toán dịch ngôn ngữ ký hiệu đã cho những kết quả khả thi hơn [25]-[33].

Năm 2022, tác giả Galina Angelova và các cộng sự kiểm nghiệm các phương pháp và kỹ thuật, thí nghiệm trên cả hai tập dữ liệu song song của ngôn ngữ ký hiệu Đức (dữ liệu PHOENIX14T và kho văn bản DGS). Nghiên cứu này thử nghiệm hai kiến trúc NMT với việc tối ưu hóa các siêu tham số của chúng, một số phương thức

mã thông báo và hai kỹ thuật tăng cường dữ liệu (dịch ngược và diễn giải). Đạt được sự cải thiện đáng kể cho các mô hình được đào tạo trên hai tập dữ liệu tương ứng. Các mô hình RNN hoạt động tốt hơn các mô hình Transformer và phương pháp phân đoạn mà nghiên cứu đạt được kết quả tốt nhất là BPE, trong khi dịch ngược và diễn giải dẫn đến nhỏ nhưng không đáng kể [34].

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 khiếm thính ít được xem xét đến.

1.3. Bài toán dịch ngôn ngữ ký hiệu Việt Nam

Việc phát triển một máy dịch thuật từ ngôn ngữ tự nhiên sang dạng ngôn ngữ kí hiệu Việt Nam là một bài toán được quan tâm hiện nay. Hiện tại, dịch máy ngôn ngữ ký hiệu Việt Nam vẫn là một lĩnh vực nghiên cứu mới và chưa được khai thác triệt để. Cũng như các bài toán dịch ngôn ngữ ký hiệu khác trên thế giới, nhiều nghiên cứu về VSL tập trung vào bước thứ 2 của quá trình dịch – dịch từ văn bản thông thường sang dạng văn bản đúng cú pháp trong ngôn ngữ ký hiệu.

Bởi vậy, đã có một số nghiên cứu về VSL liên quan đến bài toán dịch tiếng Việt sang VSL với những kết quả khả quan nhưng bên cạnh đó cũng còn nhiều hạn chế. Tác giả Quách Luỹ Dạ và các cộng sự đưa ra một nghiên cứu dịch dựa trên từ loại dùng để xử lý văn bản dẫn đến độ chính xác cao hơn cho các bản dịch từ tiếng Việt sang chú thích ngôn ngữ ký hiệu dạng văn bản, tương đồng với văn bản có quan tâm đến cú pháp VSL. Tuy nhiên thuật toán còn nhiều hạn chế với thời gian xử lý lâu [35].

Nghiên cứu tiếp theo sử dụng cây quyết định để chuyển đổi các câu có cấu trúc thành các dạng câu đúng cú pháp VSL bằng cách rút gọn câu tiếng Việt thành câu ngắn gọn cho biểu diễn. Nghiên cứu này dịch từ văn bản tiếng Việt thành hình đại diện 3D sử dụng bản ghi HamNoSys. Tuy nhiên điểm hạn chế của nghiên cứu này là cơ sở dữ liệu nhỏ dẫn đến độ chính xác thấp [36].

Vào năm 2020 của các tác giả này đánh giá một thuật toán phân loại hiệu quả mà để tích hợp vào quy trình dịch thuật của VSL [37]. Tuy nhiên đây vẫn là một thử nghiệm nhỏ và chưa có các nghiên cứu sâu sắc về cú pháp của VSL.

Trong bài toán dịch VSL với đầu vào là câu tiếng Việt thông thường, đầu ra cuối cùng là hình ảnh, video, các mô hình diễn họa 3D. Nhưng một bước trung gian quan trọng của quá trình dịch là từ câu tiếng Việt thông thường sang câu dạng đúng cú pháp trong VSL. Bởi lẽ, VSL có một số đặc trưng cơ bản như tính giản lược, nhấn mạnh trọng tâm và thay đổi trật tự từ so với ngôn ngữ tiếng Việt thông thường. Ngoài ra, việc biểu diễn từ câu đúng cú pháp trong VSL sang các dạng hình ảnh, mô hình 3D đã có những phương pháp kỹ thuật được đề xuất và có kết quả tốt. Điều này có nghĩa là các thành phần trong câu được tách ra và lưu trữ nó trong từ điển dưới dạng 1 mã số sẽ có 2 thành phần là từ/cụm từ và cách biểu diễn nó bằng mô hình 3D. Việc chuyển động liên kết mềm mại giữa các thành phần trong câu được xử lý bằng một số kỹ thuật nội suy [38].

Trong các nghiên cứu về VSL hay các ngôn ngữ khác trên thế giới đã phân tích ở trên, ta thấy rằng ngoài các phương pháp dịch máy cổ điển và hiện đại được áp dụng hiệu quả với bài toán dịch ngôn ngữ thì có một vấn đề còn tồn tại nổi cộm. Đó chính là về dữ liệu. Theo Razieh Rastgoo, hầu hết các đề xuất các mô hình cho SLP được đánh giá trên bộ dữ liệu PHOENIX14T [39]. Bộ dữ liệu này chứa 8257 trình tự được được chú thích bằng cả văn bản và bản dịch ngôn ngữ nói. Đây là bộ dữ liệu cho Ngôn ngữ ký hiệu Đức. Có rất ít các bộ dữ liệu được công bố cho các nhà nghiên cứu sử dụng trong lĩnh vực này có thể kể tên như: Dicta-Sign (BSL), ASL-LEX (BSL), RWTH-Phoenix-2014T (DGS), KETI (KSL), How2Sign (BSL).

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.

Với các phân tích đã nêu trên, ta có:

- **Phát biểu bài toán:** Ngôn ngữ ký hiệu Việt Nam được sử dụng để ghi lại và truyền đạt thông tin cho người khiếm thính bằng các biểu tượng, ký hiệu hình ảnh, và các cử chỉ tay đặc thù. Tuy nhiên một vấn đề trọng tâm của việc dịch ngôn ngữ ký hiệu Việt Nam là chuyển đổi dạng đúng cú pháp vì VSL là một ngôn ngữ riêng với cú pháp đặc trưng của nó. Bài toán dịch tự động ngôn ngữ ký hiệu Việt Nam là quá trình biến đổi một câu dạng đúng cú pháp tiếng Việt thành một câu đúng cú pháp trong ngôn ngữ ký hiệu Việt Nam.
- **Đầu vào, đầu ra của bài toán:**
  - Đầu vào: Một câu dạng đúng cú pháp tiếng Việt.
  - Đầu ra: Một câu đúng cú pháp trong ngôn ngữ ký hiệu Việt Nam.
- **Kịch bản thử nghiệm**: Để thực hiện thử nghiệm cho bài toán dịch tự động ngôn ngữ ký hiệu Việt Nam, ta có thể thực hiện các bước sau:
  - Thu thập dữ liệu: Xây dựng một tập dữ liệu gồm các cặp câu song ngữ tiếng Việt - ngôn ngữ ký hiệu Việt Nam. Tập dữ liệu này cần phải đảm bảo tính chính xác cú pháp và sự tương đương giữa hai câu.
  - Tiền xử lý dữ liệu và làm giàu dữ liệu: Chuẩn hóa và làm sạch dữ liệu, loại bỏ các ký tự không cần thiết, đảm bảo sự phù hợp về định dạng và cú pháp cho cả hai ngôn ngữ. Đồng thời đề ra một phương pháp tăng cường dữ liệu để phù hợp với việc ứng dụng các mô hình dịch máy và đánh giá. Sử dụng điểm Perplexity cho đánh giá mô hình dữ liệu xây dựng.
  - Xây dựng mô hình dịch máy: Phát triển, cải tiến và sử dụng một số mô hình dịch máy phù hợp với bài toán như dịch dựa trên luật (rule-base), mô hình thống kê như IBM, mô hình sử dụng mạng noron như Seq2seq, Transfomer.
  - Chia tập dữ liệu: Phân chia tập dữ liệu thành tập huấn luyện, tập kiểm tra và tập đánh giá.

- Huấn luyện mô hình: Sử dụng tập huấn luyện để huấn luyện mô hình dịch máy.
- Đánh giá mô hình: Sử dụng tập kiểm tra và tập đánh giá để đánh giá hiệu suất của mô hình dịch máy. Dùng độ đo BLEU để đánh giá chất lượng dịch của mô hình

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. Một nội dung trọng tâm của chương là phân tích và đánh giá một số công trình nghiên cứu về dịch ngôn ngữ ký hiệu trên thế giới. Từ những phân tích 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.

CHƯƠNG 2
CÁC KIẾN THỨC CƠ SỞ

Trong chương này, luận án trình bày những kiến thức cơ sở được sử dụng trong các chương tiếp theo. Mở đầu, Mục 1.1 trình bày một số khái niệm cơ bản về dịch máy. Phần tiếp theo trình bày về mô hình dịch máy cụ thể là dịch dựa trên luật (rule-based). Phương pháp dịch này tuy cổ điển nhưng có những điểm được chỉ ra là phù hợp với bài toán. Luận án cũng giới thiệu về hai mô hình dịch máy thống kê được sử dụng nhiều trong các bài toán dịch ngôn ngữ ký hiệu trên thế giới là Transfomer và Seq2Seq. Bên cạnh đó, chương này cũng trình bày kiến thức cơ sở về điểm đánh giá các bản dịch máy sẽ được ứng dụng để đánh giá các mô hình được trình bày trong chương 3 và chương 4.

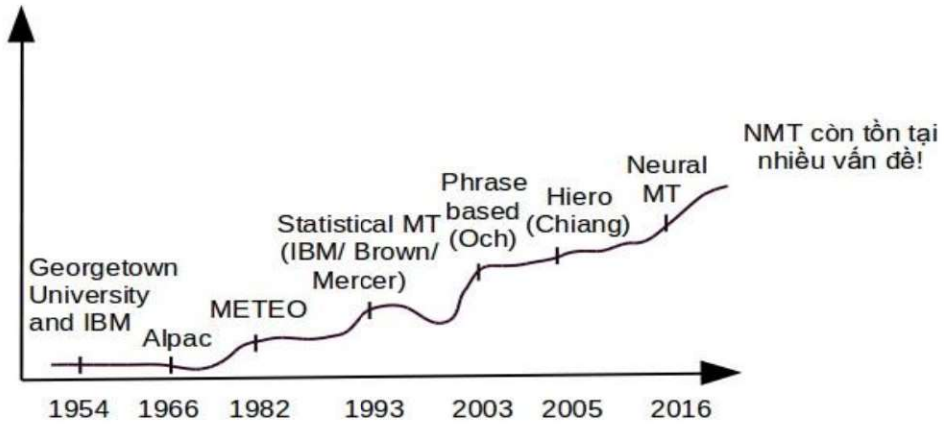
2.1. Kiến thức cơ sở về dịch máy

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.

Theo thống kê của Netcraft Ltd, số lượng người dùng có nhu cầu sử dụng các website cho việc dịch văn bản qua các ngôn ngữ khác nhau tăng lên rất nhanh. Cụ thể, cuối tháng 4 năm 2008 có 176 triệu website hỗ trợ chức năng dịch nhưng đến tháng 8 năm 2013 đã có 717 triệu website hỗ trợ chức năng này. Hiện nay, tồn tại nhiều hệ thống dịch lớn cho mục đích thương mại như Google dịch khoảng 100 tỉ từ mỗi ngày, mạng xã hội Facebook đưa ra thông điệp "When we turned MT off for some people, they went nuts!" nghĩa là "Khi tắt chức năng MT của một số người thì họ muốn điên lên!", eBay sử dụng chức năng dịch cho các giao dịch xuyên quốc gia.

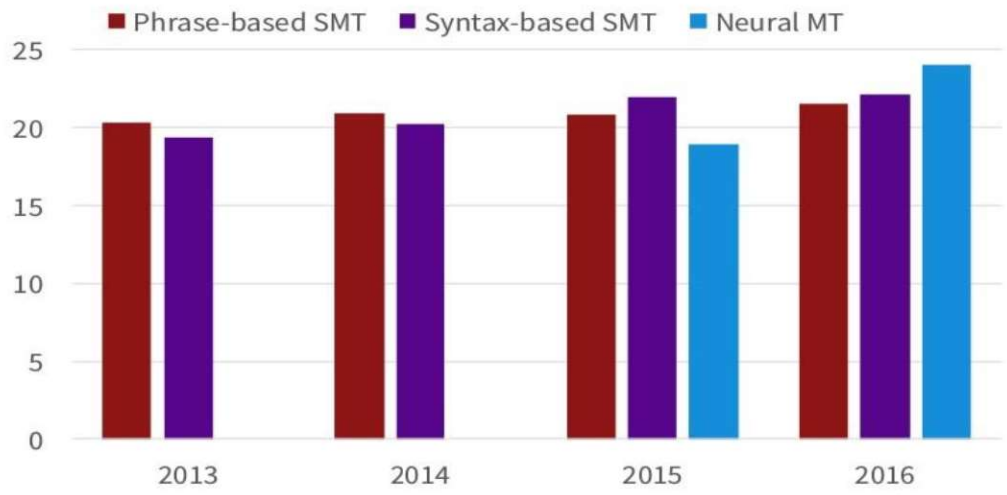
Trên thực tế vẫn chưa có hệ một hệ thống MT nào có đầy đủ chức năng dịch và có chất lượng cao. Các hệ thống dịch lớn Google Translate và Bing Translator phần nào đã đáp ứng được nhu cầu của người dùng ở một mức độ nhất định trên một số cặp ngôn ngữ nào đó nhưng vẫn chưa phải là các hệ thống dịch đầy đủ và đạt chất lượng cao. Việc xây dựng một hệ thống dịch với đầy đủ chức năng và có chất lượng cao vẫn còn là một mục tiêu xa vời mặc dù dịch máy đã ra đời hơn một nửa thế kỷ.

Quá trình phát triển của MT được minh họa qua các giai đoạn như hình 1.3. MT bắt đầu xuất hiện từ năm 1949 bởi IBM, nhưng cho đến những năm 1954 nhóm nghiên cứu của Đại học Georgetown đã công bố hệ thống MT đầu tiên dựa trên việc kế thừa các nghiên cứu của IBM. Năm 1962, chính phủ Mỹ đã thành lập một cộng đồng tư vấn về dịch tự động (Automatic Language Processing Advisory Committee) gọi tắt ALPAC để tiếp tục phát triển các nghiên cứu về MT. Năm 1966, ALPAC đã tuyên bố MT có tốc độ chậm, chất lượng dịch kém và chi phí cao gấp đôi so với con người, đồng thời họ đưa ra kết luận là không cần nghiên cứu thêm về MT nữa. Nhưng sau đó, các nghiên cứu về dịch máy vẫn tiếp tục diễn ra, điển hình năm 1982, trung tâm nghiên cứu về môi trường Canada đưa ra hệ thống dịch METEO phục vụ cho mục đích dịch các bản tin dự báo thời tiết. Các nghiên cứu về MT thời kỳ này chủ yếu dựa vào từ điển và các luật để sinh ra bản dịch đúng của các từ. Các nhà nghiên cứu luôn cố gắng khai thác tri thức về ngôn ngữ để cải thiện mô hình dịch của họ. Cho đến những năm 1990, phương pháp dịch máy thống kê bắt đầu xuất hiện và trở thành hướng nghiên cứu trọng tâm của thời kỳ này. Cách tiếp cận dịch thống kê sử dụng một kho dữ liệu chứa một tập các mẫu dịch trước để huấn luyện môt hình dịch. Bên cạnh đó, các nghiên cứu về MT đựa trên các luật và dựa trên tri thức vẫn tiếp tục được phát triển và được tích hợp trong các mô hình dịch thống kê ở nhiều nghiên cứu thời kỳ đó. Cách tiếp cận thống kê tiếp tục được phát triển mạnh mẽ trong nửa đầu thập niên tiếp theo cho đến khi đưa ra mô hình dịch thống kê dựa trên cụm từ (statistical machine translation), gọi tắt là SMT. Năm 2005, Chiang đưa ra hệ thống dịch Hiero là mô hình dịch dựa trên cụm từ phân cấp cho SMT với việc không đồng bộ ngữ cảnh ngữ pháp tự do (context-free grammar) gọi tắt là CFG [40]. Các mô hình SMT tiếp tục được nghiên cứu, phát triển và ứng dụng trong các hệ thống MT thương mại trong một thập niên gần đây.



Hình 2.1. Quá trình phát triển của MT

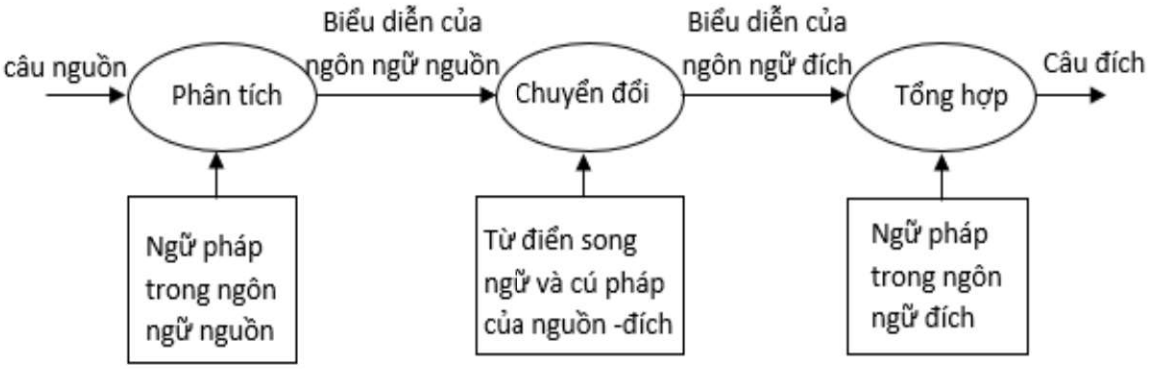
Mặc dù các mô hình SMT đem lại hiệu quả cho nhiều hệ thống dịch thương mại nhưng nó bị hạn chế bởi ngữ cảnh cục bộ và chất lượng dịch đạt tới độ bão hòa khi lượng dữ liệu huấn luyện đủ lớn, do đó các nghiên cứu về SMT trở lên bế tắc. Để khắc phục các nhược điểm của SMT, các nhà nghiên cứu về MT lại nỗ lực khai thác mô hình mạng nơron lần thứ hai cho MT và đưa ra giải pháp khắc phục hiện tượng thắt nút cổ chai (bottle-neck) như một cuộc cách mạng trong dịch máy với các nghiên cứi điển hình. Ý tưởng sử dụng mạng nơron cho MT được Ramon Neco và Mikel Forcada với một hệ thống dịch gồm một encoder và một decoder, tuy nhiên hệ thống dịch này gặp phải hiện tượng bùng nổ hoặc biến mất của giá trị gradient (gọi là hiện tượng bottle-neck) và bị ngưng lại những năm sau đó [41]. Năm 2016, mô hình dịch dựa trên mạng nơron (Neural Machine Translation) gọi tắt là NMT cho kết quả cao hơn các mô hình SMT trên cùng tập dữ liệu 1.2 và trở thành cách tiếp cận chính cho các nghiên cứu về MT hiện nay [42].



Hình 2.2. So sánh kết quả dịch dựa trên SMT và NMT

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 [43]. Tuy nhiên, do sự đa dạng về ngữ pháp của các cặp ngôn ngữ làm cho các hệ thống dịch gặp nhiều khó khăn bởi hiện nay chưa có được tập các luật đầy đủ cho mọi cặp ngôn ngữ. Mặc dù vậy, kỹ thuật dịch này vẫn phù hợp cho một số hệ thống MT nhỏ và yêu cầu tài nguyên hữu hạn. VSL chính là một ngôn ngữ ít tài nguyên như vậy. Hầu hết các hệ thống dịch máy dựa trên quy tắc tạo ra bản dịch bằng cách phân tích cú pháp văn bản nguồn, tạo ra một biểu diễn tượng trưng trung gian của nó, và sau đó tạo bản dịch cuối cùng trong ngôn ngữ đích. Cần áp dụng ánh xạ giữa các mục từ vựng được lưu trữ trong từ điển cũng như chuyển các quy tắc để giải thích sự khác biệt về cấu trúc giữa hai ngôn ngữ . 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 2.3. Sơ đồ dịch máy dựa trên luật.

Các hệ thống RBMT đầu tiên được phát triển vào đầu những năm 1970. Các bước quan trọng nhất của sự phát triển này là sự xuất hiện của các hệ thống RBMT : Systran [44]; Hệ thống MT Nhật Bản [45], EUROTRA (Eurotra). Ngày nay, các hệ thống RBMT phổ biến khác bao gồm: Apertium [46], GramTrans [47].

2.2.1. Các hướng tiếp cận chính

Các nghiên cứu dựa theo cách tiếp cận này được phân loại theo ba hướng tiếp cận chính là dịch trực tiếp, dịch chuyển đổi và dịch liên ngữ.

Dịch trực tiếp (Direct machine translation). Đây là một cách tiếp cận khá đơn giản, được thực hiện bằng cách thay các từ trong văn bản đầu vào bằng từng từ trong bản đầu ra rồi sinh ra kết quả theo đúng thứ tự ban đầu. Nhưng khi dùng cách tiếp cận này với những cặp song ngữ khác biệt lớn về cấu trúc cú pháp và sự đa nghĩa của từ vựng thì hiệu quả dịch rất thấp. Kiến trúc này bắt đầu cho quá trình phát triển của dịch máy và chỉ đạt hiệu quả trong một số ngữ cảnh hẹp hoặc bài toán dịch với chất lượng không yêu cầu cao.

Dịch chuyển đổi (Transfer machine translation). Với cách tiếp cận này, tiến trình dịch gồm hai mức là chuyển đổi cú pháp và chuyển đổi ngữ nghĩa thông qua việc thực hiện chuyển đổi các tri thức ngôn ngữ từ ngôn ngữ nguồn sang ngôn ngữ đích (các tri thức như từ, cú pháp, nghĩa, cách sử dụng...) thông qua một tập các luật. Các hệ thống dịch dựa trên cách tiếp cận này có độ chính xác cũng như khả năng giải quyết nhập nhằng tốt hơn nhiều so với kiến trúc dịch trực tiếp, tuy nhiên chúng thường đòi hỏi tốn nhiều công sức trong việc thực hiên chuyển đổi tri thức ngôn ngữ cho từng cặp ngôn ngữ.

Dịch qua ngôn ngữ trung gian (Interlingual machine translation). Kiến trúc này sử dụng một ngôn ngữ làm trung gian cho việc dịch giữa các cặp ngôn ngữ nguồn và ngôn ngữ đích. Ngôn ngữ trung gian không phụ thuộc vào tri thức của ngôn ngữ nguồn hay ngôn ngữ đích.

Cách tiếp cận chính của hệ thống RBMT dựa trên việc liên kết cấu trúc của câu đầu vào đã cho với cấu trúc của câu đầu ra được yêu cầu, nhất thiết phải bảo toàn ý nghĩa duy nhất của chúng.

2.2.2. Nguyên tắc cơ bản của RBMT

Ví dụ sau có thể minh họa nguyên tắc cơ bản của RBMT:

“A girl eats an apple.”

Ngôn ngữ nguồn = tiếng Anh;

Ngôn ngữ mục tiêu theo yêu cầu = Tiếng Đức

Để có được bản dịch tiếng Đức của câu tiếng Anh này, ta cần:

- Một từ điển sẽ ánh xạ mỗi từ tiếng Anh sang một từ tiếng Đức thích hợp.
- Các quy tắc biểu diễn cấu trúc câu tiếng Anh thông thường.

- Các quy tắc biểu diễn cấu trúc câu tiếng Đức thông thường.
- Các quy tắc mà theo đó người ta có thể liên hệ hai cấu trúc này với nhau.

Theo đó, các giai đoạn dịch như sau:

1. Nhận thông tin cơ bản của mỗi từ trong câu nguồn: xác định từ loại cho mỗi từ trong câu
2. Nhận thông tin cú pháp về động từ trong câu
3. Phân tích cú pháp câu nguồn: Thường chỉ phân tích cú pháp một phần là đủ để đi đến cấu trúc cú pháp của câu nguồn và ánh xạ nó vào cấu trúc của câu đích.
4. dịch các từ tiếng Anh sang tiếng Đức:

   a (từ loại = mạo từ) => ein (= mạo từ)

   girl (từ loại = danh từ) => Mädchen (từ loại = danh từ)

   eat (từ loại = động từ) => essen (từ loại = động từ)

   an (từ loại = mạo từ) => ein (= mạo từ)

   apple (từ loại = danh từ) => Apfel (từ loại = danh từ)
5. Ánh xạ các mục từ điển thành các hình thức được chọn lọc thích hợp:

A girl eats an apple. => Ein Mädchen isst einen Apfel.

2.2.3. Các thành phần của một hệ thống RBMT

Các thành phần của một hệ thống RBMT bao gồm:

- Một máy phân tích hình thái ngôn ngữ nguồn (SL) - phân tích một từ ngôn ngữ nguồn và cung cấp thông tin hình thái học;
- Trình phân tích cú pháp SL - là trình phân tích cú pháp phân tích các câu ngôn ngữ nguồn;
- Một trình dịch - được sử dụng để dịch một từ ngôn ngữ nguồn sang ngôn ngữ đích;
- Trình tạo hình thái TL - hoạt động như một trình tạo các từ ngôn ngữ đích thích hợp cho thông tin ngữ pháp nhất định;
- Trình phân tích cú pháp TL - hoạt động như một hệ thống soạn thảo các câu ngôn ngữ đích phù hợp;
- Một số từ điển - cụ thể hơn là tối thiểu ba từ điển:

- Một từ điển SL - cần thiết cho bộ phân tích hình thái ngôn ngữ nguồn để phân tích hình thái học,
- Từ điển song ngữ - được người dịch sử dụng để dịch các từ ngôn ngữ nguồn thành các từ ngôn ngữ đích,
- Một từ điển TL - cần thiết bởi trình tạo hình thái ngôn ngữ đích để tạo ra các từ ngôn ngữ đích. [48]

2.2.4. Ưu và nhược điểm của RBMT

Các ưu điểm của RBMT có thể kể đến là:

- Không cần văn bản song ngữ. Điều này giúp cho việc có thể tạo hệ thống dịch cho các ngôn ngữ không có văn bản chung, hoặc thậm chí không có dữ liệu số hóa.

- Miền dữ liệu độc lập. Các quy tắc thường được viết theo cách độc lập với miền dữ liệu, do đó, phần lớn các quy tắc sẽ "hoạt động" trong mọi miền và chỉ một số trường hợp cụ thể trên mỗi miền có thể cần các quy tắc được viết cho chúng.

- Mọi lỗi đều có thể được sửa chữa bằng quy tắc hướng đích. Điều này trái ngược với các hệ thống thống kê nơi các biểu mẫu không thường xuyên sẽ bị loại bỏ theo mặc định.

- Vì tất cả các quy tắc đều được xây dựng thủ công nên có thể dễ dàng gỡ lỗi hệ thống dựa trên quy tắc để xem chính xác vị trí một lỗi nhất định xâm nhập vào hệ thống và đánh giá được chất lượng bản dịch.

- Khả năng tái sử dụng. Bởi vì hệ thống RBMT thường được xây dựng từ phân tích ngôn ngữ nguồn được đưa đến bước chuyển giao và trình tạo ngôn ngữ đích, phần phân tích ngôn ngữ nguồn và phần tạo ngôn ngữ đích có thể được chia sẻ giữa nhiều hệ thống dịch, chỉ yêu cầu bước chuyển là chuyên biệt. Ngoài ra, phân tích ngôn ngữ nguồn cho một ngôn ngữ có thể được sử dụng lại để khởi động một phân tích ngôn ngữ có liên quan chặt chẽ.

Những nhược điểm của RBMT:

- Yêu cầu một số lượng từ điển lớn trong khi việc xây dựng từ điển mới rất khó khăn vì thường phải xây dựng một cách thủ công.

- Một số thông tin ngôn ngữ vẫn cần được thiết lập theo cách thủ công.

- Các tương tác quy tắc trong các hệ thống lớn rất khó xử lý, sự mơ hồ và các cách diễn đạt thành ngữ.

- Không thể thích ứng với các miền mới. Mặc dù các hệ thống RBMT thường cung cấp một cơ chế để tạo ra các quy tắc mới và mở rộng và điều chỉnh từ vựng, nhưng các thay đổi thường rất tốn kém và kết quả thường không tốt như mong muốn.

Với những phân tích trên, ta thấy rằng tuy RBMT là một phương pháp cổ điển và ít được sử dụng hiện nay, nhưng đây vẫn hoàn toàn có thể là một phương pháp phù hợp được sử dụng trong các bài toán dịch với ngôn ngữ ít tài nguyên. Vì vậy mà nó phù hợp với bài toán dịch tiếng Việt – VSL. Các phần thực nghiệm và đánh giá phương pháp này với bài toán sẽ được trình bày cụ thể trong chương 3.

2.3. 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.

$$e* = argmax_e p(e|f) \tag{2.1}$$

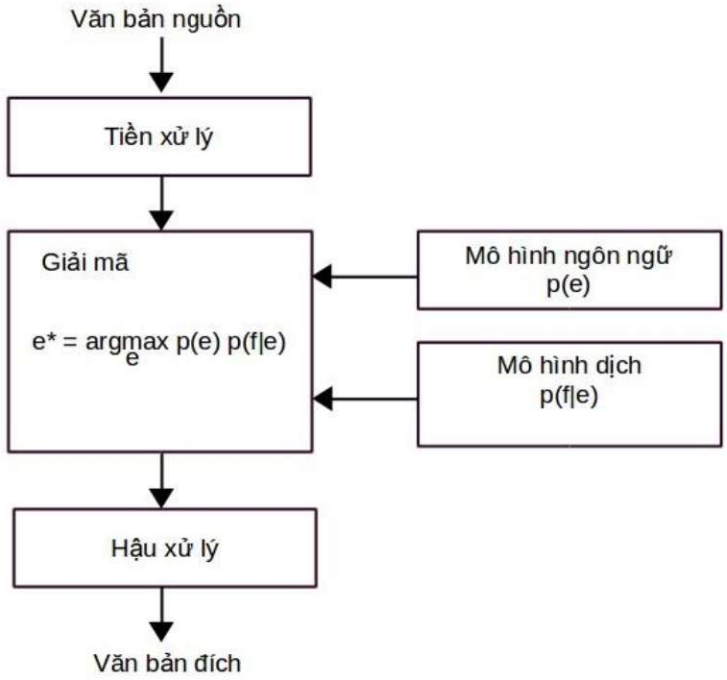
Sử dụng công thức Bayes:

$$p(e|f) = \frac{p(e)p(f|e)}{p(f)} \tag{2.2}$$

Do $p(f)$ không thay đổi khi so sánh các câu e_i khác nhau nên công thức 2.2 có thể được viết lại như sau:

$$e* = argmax_e p(e)p(f|e) \tag{2.3}$$

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 $p(f|e)$.



Hình 2.4. Dịch máy dựa trên mô hình SMT

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. Có nhiều phương pháp khác nhau để mô hình hóa quá trình dịch và được chia làm ba cách tiếp cận chính là dịch dựa trên từ (word-based), dịch dựa trên cụm từ (phrase-based) và dịch dựa trên cú pháp (syntax-based).

Mô hình SMT được minh họa như hình 2.5. Trong đó, pha giải mã là bài toán tìm kiếm thông thường, dùng để tìm kiếm câu đích e phù hợp nhất tương ứng với câu nguồn f. Các thuật toán tìm kiếm phổ biến thường được sử dụng nhất để giải quyết bài toán này như beam search, Viterbi Beam, A* stack.

Mô hình IBM là một trong những mô hình dịch máy thống kê đầu tiên được giới thiệu vào những năm 1990. Mô hình này được đặt tên theo tên của IBM, công ty đã phát triển nó. Mô hình IBM dịch từ ngôn ngữ nguồn sang ngôn ngữ đích bằng cách sử dụng một bộ sưu tập các cặp câu song ngữ (parallel corpus) để xác định xác suất của một từ hoặc cụm từ trong ngôn ngữ nguồn tương ứng với một từ hoặc cụm từ trong ngôn ngữ đích.

Mô hình IBM sử dụng một số tham số để tính toán xác suất của từng cặp từ song ngữ. Một số tham số này bao gồm xác suất chuyển đổi (translation probability) để xác định xác suất chuyển đổi từ một từ hoặc cụm từ trong ngôn ngữ nguồn sang một từ hoặc cụm từ trong ngôn ngữ đích và xác suất phù hợp (alignment probability) để xác định xác suất tương ứng giữa các từ trong hai câu.

Mô hình IBM đã mở đầu cho sự phát triển của các mô hình dịch máy thống kê và trở thành cơ sở cho nhiều mô hình dịch máy hiện đại khác. Tuy nhiên, những mô hình này đã phát triển thêm nhiều tính năng và cải tiến hơn so với mô hình IBM ban đầu.

2.4. Dịch máy dựa trên mạng rơron

Cách tiếp cận dịch dựa trên SMT đã đem lại những thành công lớn trong dịch máy nhưng vẫn gặp phải vấn đề ngữ cảnh cục bộ do quá trình dịch dựa trên cụm từ. Ngoài ra, cách tiếp cận này bị ảnh hưởng lớn bởi sự khác biệt về cấu trúc ngữ pháp giữa các cặp ngôn ngữ khác nhau nên nó cần thêm mô hình đảo trật từ (reorder word model). Như vậy, SMT gồm nhiều thành phần rời rạc được tích hợp với nhau, mỗi thành phần lại bao gồm một tập các tham số riêng làm quá trình xây dựng và phát triển một hệ thống dịch hoàn chỉnh và tối ưu gặp nhiều khó khăn và bế tắc khi đạt đến một ngưỡng nhất định. Khi này, cách tiếp cận NMT lại mở ra một hướng phát triển mới cho MT.

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 [49]. 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 có dạng:

$$h_t = f(h_{t-1}, x_t), \tag{2.4}$$

Trong đó, f là một hàm trừu tượng để tính toán trạng thái ẩn tại thời điểm t từ đầu vào x_t tại thời điểm đó và trạng thái ẩn trước đó h_{t-1}. Trạng thái ẩn ban đầu h_0 thường được khởi tạo là 0. Dạng phổ biến của hàm f thường được chọn là dạng *sigmoid* hoặc *tanh* như sau:

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1}) \quad (2.5)$$

Tại thời điểm t, một RNN có thể cho đầu ra là các giá trị rời rạc (thí dụ như các số thực). Trong trường hợp đầu ra Y là các giá trị rời rạc, phân bố xác suất p trên tập đầu ra Y là:

$$s_t = W_{hy}h_t \quad (2.6)$$

$$p_t = softmax(s_t) \quad (2.7)$$

$W_{hy} \in R^{(Y)xd}$với với d là số chiều của trạng thái ẩn trong RNN. Thông thường, với một tập Y lớn, ma trận vectơ trong công thức 2.6 sẽ bị hiện tượng thắt nút cổ chai (bottleneck) trong RNN và trở thành một thách thức lớn trong mô hình ngôn ngữ sử dụng mạng nơron cũng như trong MT. Hàm softmax sẽ chuyển vector s_t vào trong không gian xác suất p_t, mỗi xác suất p_t ứng với mỗi phần từ $y \in Y$ được tính như công thức 2.8.

$$p_t(y) = \frac{e^{-i\omega t}}{\sum_{y' \in Y} e^{St(y')}} \quad (2.8)$$

Các công thức 2.5, 2.6 chứa một tập các ma trận trọng số θ của RNN, bao gồm W_{xh} kết nối đầu vào, W_{hh} kết nối hồi quy và W_{hy} kết nối đầu ra. Quá trình huấn luyện là quá trình cập nhật giá trị trọng số của các ma trận này. Bằng cách kết nối các xác xuất đầu ra của các $y \in Y$, ta có thể tính được xác suất có điều kiện $p(y)$ như sau:

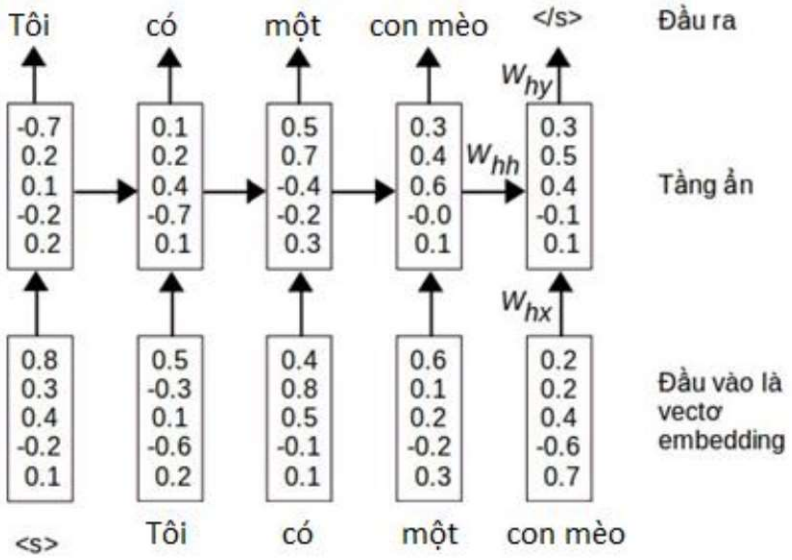
$$p(y) = \prod_{t=1}^{T} p(y_t | y_{t-1}, y_{t-2}, \dots, y_1) \quad (2.9)$$

Mô hình ngôn ngữ sử dụng mạng nơron hồi quy

Với ý tưởng tính toán xác suất có điều kiện trên các đầu ra của RNN như công thức 2.9, mạng RNN được ứng dụng trong các mô hình ngôn ngữ để mô hình hóa xác xuất của ngôn ngữ đích.

Đầu vào của mô hình là tập các câu $y \in Y$ thuộc một ngôn ngữ, với mỗi câu y chứa tập các từ ($y_1, y_2, \dots, y_T$) hệ thống tính toán xác suất có điều kiện của từ y_t so với trình tự các từ trước đó $y_{t-1}, y_{t-2}, \dots, y_1$ theo công thức 2.9. Các câu đầu vào được bắt đầu bởi một ký hiệu đặc biệt <s>, thí dụ x = <s>, “I”,“am”, “a”, “teacher”. Với mục tiêu của mô hình ngôn ngữ là dự đoán từ tiếp theo nên trình tự ký hiệu đầu ra được

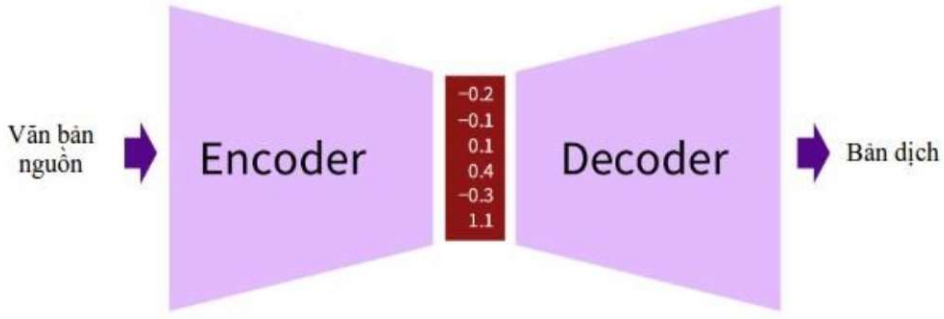
dịch đi một từ so với ký hiệu bắt đầu và kết thúc như hình 2.4. Ký hiệu đánh dấu kết thúc câu là </s>. Một từ y_i trong câu đầu vào có thể được biểu diễn bởi một vectơ one-hot $y_i \in R^{(V)}$. Với V là tập từ vựng, nếu V lớn thì các ma trận W_{hx} sẽ lớn và không tồn tại mối quan hệ giữa các từ. Do vậy, các y_i thường được huấn luyện từ một mô hình Embedding [50].



Hình 2.5. Mô hình ngôn ngữ sử dụng mạng RNN

2.4.1. 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 [51]. Kiến trúc encoder-decoder được minh họa như hình 2.6:



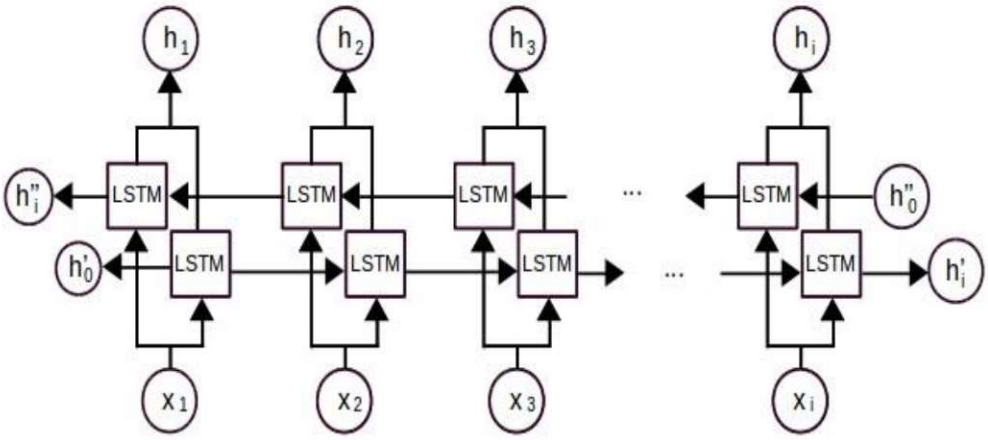
Hình 2.6. Kiến trúc encoder-decoder sử dụng mạng RNN

Trong mô hình seq2seq, các mạng RNN của Encoder và Decoder hoạt động đồng thời trong quá trình huấn luyện, trong đó:

Encoder thường sử dụng mạng RNN hai chiều làm nhiệm vụ mã hóa các một tập các câu đầu vào từ văn bản của ngôn ngữ nguồn $X = (x_1,x_2,...,x_T)$ vào trong các trạng thái ẩn $h = (h_1,h_2,...,h_T)$. RNN ở đây có thể là một chiều hoặc hai chiều. Tại mỗi thời điểm t, trạng thái ẩn h_t được tính toán từ sự kết hợp của hai vectơ $[h'_t, h''_t]$, với $h'_t = f(x_t, h'_{t-1})$là trạng thái ẩn trước đó và$h''_t = f(x_t, h''_{t+1})$ là trạng thái ẩn tiếp theo. Hàm f thường được tính toán từ các thành phần GRU (Gated Recurrent Units) hoặc LSTM (Long Short Term Memory).

Mạng RNN hai chiều là một hệ thống sử dụng hai RNN độc lập hoạt động cùng nhau được minh họa trong hình 2.7 và được gọi tắt là BRNN. Trong kiến trúc BRNN, một RNN ghi lại ngữ cảnh theo chiều thuận từ trái sang phải với trạng thái ẩn khởi tạo làh'_0 ban đầu, một RNN khác ghi lại ngữ cảnh của câu theo chiều ngược lại từ phải sang trái với trạng thái khởi tạo ban đầu làh''_0. Các trạng thái ban đầu thường được khởi tạo bằng 0. Vectơ h_t được sử dụng để tính toán ngữ cảnh c_t trong thành phần attention Như vậy, với một Encoder hai chiều, hệ thống dịch có thể ghi được toàn bộ ngữ cảnh của câu cho việc dự đoán một từ mục tiêu x_t tại thời điểm t trong quá trình huấn luyện. Các trạng thái ẩn đầu ra của Encoder được đưa vào làm trạng thái khởi tạo của Decoder.

Thông thường các hệ thống dịch sử dụng một Encoder nhiều lớp, khi đó, các trạng thái ẩn h_t là đầu ra của một lớp trước đó sẽ là đầu vào của lớp tiếp theo.



Hình 2.7. Encoder hai chiều sử dụng các mạng RNN

Decoder nhận một trình tự các câu đầu vào từ văn bản của ngôn ngữ đích $Y = (y_1,y_2,...,y_m)$ và sinh ra bản bản dịch $Y' = (y'_1, y'_2, ..., y'_{m'})$ từ các trạng thái ẩn h phía

encoder. Tại thời điểm i, xác suất có điều kiện cho mỗi từ y_i^0 thuộc tập từ vựng của ngôn ngữ đích V_y được tính toán như sau:

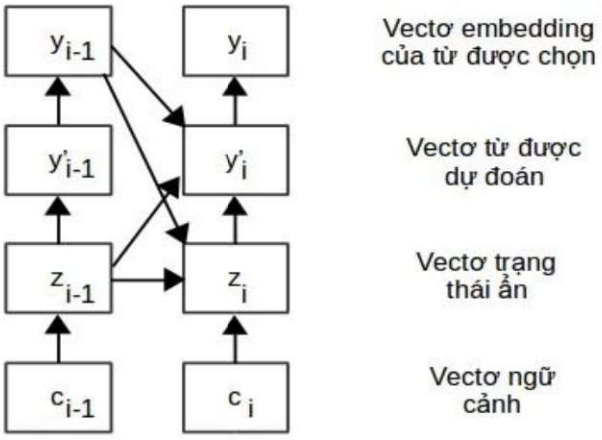
$$p(y_i'|y_{<i}', h) = softmax(y_{i-1}', z_i, c_i), \tag{2.10}$$

Với z_i là trạng thái ẩn thứ i của decoder và được tính toán từ sự kết hợp của trạng thái ẩn trước đó z_{i-1}, vectoy_{i-1}' và vectơ ngữ cảnh nguồn c_i theo công thức 2.8.

$$z_i = f(z_{i-1}, y_{i-1}, c_i), \tag{2.11}$$

Hình 2.8 minh họa quá trình tính toán các trạng thái ẩn và dự đoán trên decoder. Hàm f trong công thức 2.11 lại một lần nữa được tính toán từ các thành phần GRU hoặc LSTM. Thông thường, nếu encoder sử dụng thành phần LSTM thì decoder cũng sử dụng LSTM. Hàm softmax trong công thức 2.10 chuyển xác suất dự đoán của các đầu ra y_i^0 vào không gian xác suất có tổng các giá trị bằng 1. Trạng thái ban đầu của decoder được khởi tạo từ trạng thái đầu ra của encoder. Cũng giống như mô hình ngôn ngữ, các từ của văn bản nguồn và văn bản đích thường được huấn huyện từ mô hình word2vec để sinh ra các Word Embedding (Word Embedding là một vectơ số thực biểu diễn từ trong tập từ vựng). Số lớp của encoder luôn bằng số lớp của Decoder.

Các mạng nơron truyền thống tồn tại hai vấn đề chính khi huấn luyện với các câu dài, đó là giá trị gradient bị bùng nổ hoặc triệt tiêu hoàn toàn được chỉ ra trong nghiên cứu của Bengio và cộng sự (1994) [52]. Cụ thể là giá trị của ma trận W_{hh} trong 2.2 bị bùng nổ hoặc bị triệt tiêu trong quá trình huấn luyện. Giá trị gradient bị bùng nổ được hiểu là giá trị gradient tăng lên theo hàm mũ trong quá trình huấn luyện làm cho việc học là không thể. Ngược lại, giá trị gradient triệt tiêu khi nó nhanh chóng tiến tới 0, do đó việc điều chỉnh thuật toán lan truyền ngược không thể ghi lại ngữ cảnh dài trong câu. Để khắc phụ hiện tượng này, các nhà nghiên cứu đã đưa ra nhiều biến thể của mạng RNN.



Hình 2.8. Minh họa quá trình tính toán các trạng thái ẩn và dự đoán trên decoder

Ưu điểm của mô hình Seq2Seq:

- Khả năng xử lý chuỗi đầu vào và đầu ra có độ dài khác nhau: Mô hình Seq2Seq có thể xử lý các chuỗi có độ dài khác nhau, làm cho nó phù hợp cho nhiều ứng dụng, như dịch máy và tổng hợp văn bản.
- Dễ dàng thích nghi với các nhiệm vụ NLP: Seq2Seq có thể được sử dụng cho nhiều nhiệm vụ NLP khác nhau như dịch máy, tạo chú thích cho hình ảnh, tổng hợp văn bản, và nhiều ứng dụng khác.
- Mô hình mạnh mẽ với dữ liệu lớn: Khi được đào tạo với dữ liệu lớn, Seq2Seq có thể tạo ra các kết quả ấn tượng và chính xác cho nhiều nhiệm vụ NLP.

Nhược điểm của mô hình Seq2Seq:

- Khó khăn trong việc xử lý các chuỗi dài: Seq2Seq có thể gặp khó khăn khi xử lý các chuỗi có độ dài rất lớn, vì nó có thể bị mất thông tin quá lâu.
- Sự mất mát thông tin: Mô hình Seq2Seq thường gặp phải sự mất mát thông tin khi mã hóa thông tin từ chuỗi đầu vào và giải mã nó thành chuỗi đầu ra. Điều này có thể dẫn đến việc mô hình không hiểu rõ ngữ nghĩa của chuỗi.
- Khó khăn trong việc học các mối quan hệ xa: Seq2Seq có thể gặp khó khăn trong việc học các mối quan hệ xa giữa các từ hoặc phần tử trong chuỗi, đặc biệt là khi chuỗi đầu vào có độ dài lớn.

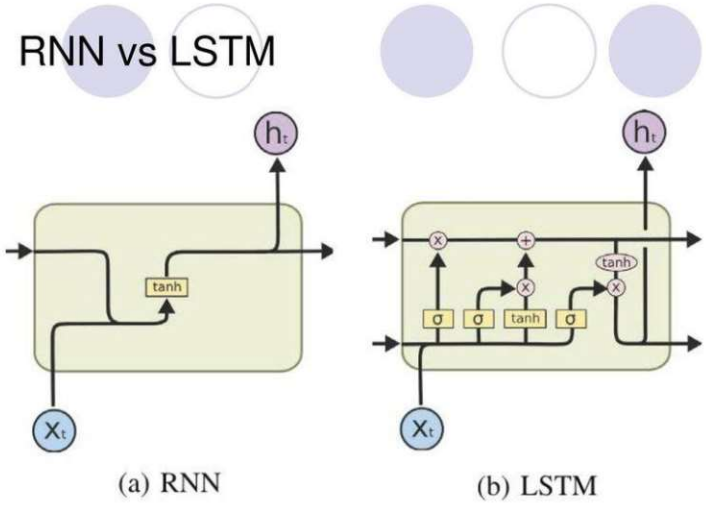
- Khả năng tạo ra kết quả không chính xác: Mô hình Seq2Seq có thể sản xuất kết quả không chính xác khi đối mặt với các trường hợp khó khăn hoặc hiếm gặp, và có thể cần sự điều chỉnh để cải thiện chất lượng đầu ra.
- Yêu cầu dữ liệu đào tạo lớn: Để đạt được hiệu suất tốt, mô hình Seq2Seq thường đòi hỏi một lượng lớn dữ liệu đào tạo, điều này có thể là một thách thức trong các trường hợp thiếu dữ liệu.

Tóm lại, mô hình Seq2Seq có nhiều ưu điểm và nhược điểm, và hiệu suất của nó phụ thuộc vào cách đào tạo và cấu hình cụ thể.

2.4.2. Mô hình Transformer

Transformer là một mô hình nổi tiếng gần đây trong cộng đồng xử lý ngôn ngữ tự nhiên và tạo ra những bước ngoặt lớn trong những bài toán dịch máy. Đối với bài toán dịch tự động VSL, ta có thể coi là một bài toán dịch với ngôn ngữ ít tài nguyên bởi những đặc điểm hạn chế về ngữ pháp và từ vựng của loại ngôn ngữ này. Trước đây các tác vụ dịch máy (Machine Translation) sử dụng kiến trúc Recurrent Neural Networks (RNNs) là chủ yếu. Nhưng các nhà nghiên cứu dịch máy đều có thể nhận thấy nhược điểm của phương pháp này là rất khó bắt được sự phụ thuộc xa giữa các từ trong câu và tốc độ huấn luyện chậm do phải xử lý input tuần tự. Transformers đã giải quyết được 2 vấn đề này. Và các biến thể của nó như BERT, GPT-2 tạo ra sự hiện đại mới cho các tác vụ liên quan đến xử lý ngôn ngữ tự nhiên.

Transformers được thiết kế để khắc phục tất cả các điểm yếu của mô hình Seq2Seq. Trong khi Seq2Seq nhận tuần tự các đầu vào và trả lại các đầu ra một cách tuần tự. Điều này khiến cho thời gian huấn luyện mô hình rất chậm. Người ta đã thử tìm một số cách khắc phục nhưng vấn đề căn bản là tốc độc huấn luyện phụ thuộc chủ yếu vào việc sử dụng CPU chứ không hề tận dụng được khả năng tính toán song song của GPU để tăng tốc độ train cho các mô hình ngôn ngữ. Với việc xử lý các câu dài, Transformer cũng thể hiện sự vượt trội khi xử lý chúng so với mô hình Seq2Seq



(a) RNN (b) LSTM

Hình 2.9. RNN và LSTM

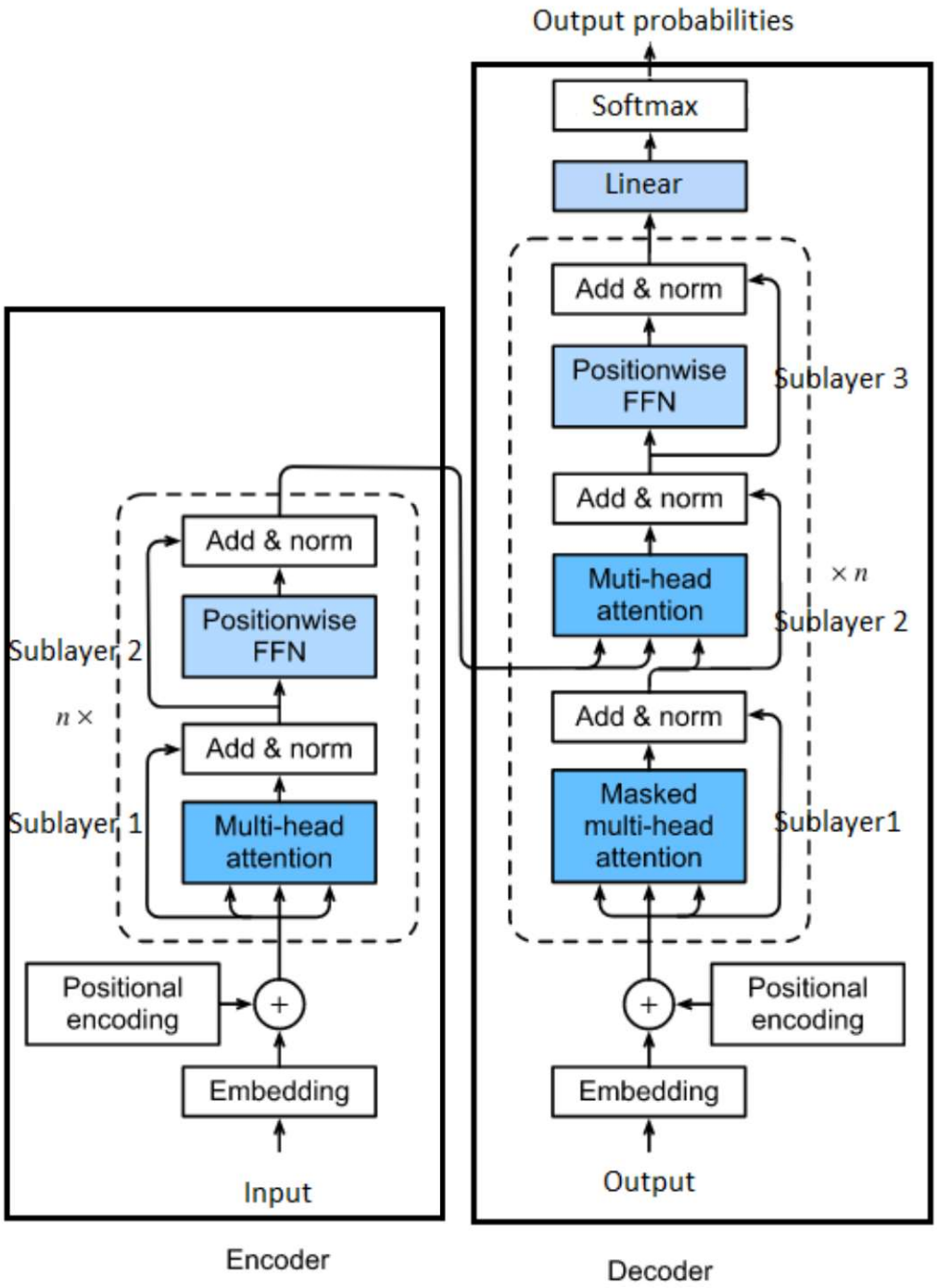
Đặc điểm giống của RNNs và Transformer đó chính là việc sử dụng 2 phần mã hoá và giải mã. Trong khi mô hình Transformer với đầu vào được đẩy vào để xử lý song song thì mô hình kia cần phải làm tuần tự. Chính điều này đã xoá đi khái niệm timestep trong Transformers và khiến nó vượt trội hơn hẳn về thời gian huấn luyện các mô hình xử lý. Kiến trúc của mô hình được minh hoạ trong hình 2.10.

Cụ thể, để tính toán vectơ truy vấn, vectơ giá trị và vectơ khóa, ta sử dụng các ma trận trọng số tương ứng: ma trận trọng số W_Q cho vectơ truy vấn, ma trận trọng số W_K cho vectơ khóa và ma trận trọng số W_V cho vectơ giá trị. Sau đó, chúng ta tính toán ma trận trọng số self-attention bằng cách lấy tích vô hướng giữa vectơ truy vấn và vectơ khóa, sau đó chuẩn hóa kết quả bằng hàm softmax để có được ma trận trọng số chuẩn hóa:

$$Attention(Q, K, V) = softmax_k\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2.12)$$

Ở đây, d_k là số chiều của vectơ khóa (key vector). Bằng cách nhân ma trận trọng số self-attention với vectơ giá trị (V), ta có thể tính toán biểu diễn tổng hợp cho từng từ trong câu. Với mỗi từ trong câu, ta sẽ có một ma trận trọng số self-attention và một biểu diễn tổng hợp tương ứng. Các biểu diễn tổng hợp này được kết hợp lại để tạo thành biểu diễn đầu ra cuối cùng của self-attention layer trong mô hình Transformer.

Một điểm quan trọng của self-attention là nó cho phép mô hình tập trung vào các từ quan trọng trong câu bằng cách thay đổi ma trận trọng số self-attention. Các từ quan trọng sẽ có giá trị trọng số cao hơn, trong khi các từ không quan trọng sẽ có giá trị trọng số thấp hơn. Điều này giúp mô hình hiểu được các mối quan hệ và phụ thuộc giữa các từ trong câu.



Hình 2.10. Kiến trúc của Transformers

Tóm lại, self-attention là một phần quan trọng trong kiến trúc Transformer, giúp mô hình xử lý thông tin từ ngữ tự nhiên bằng cách tổng hợp thông tin từ toàn bộ ngữ cảnh của câu. Qua việc tính toán các vectơ truy vấn, vectơ khóa và vectơ giá trị, kết hợp với ma trận trọng số self-attention, mô hình có khả năng tìm hiểu mối quan hệ phụ thuộc giữa các từ và tạo ra biểu diễn cuối cùng cho từng từ trong câu.

2.5. Đánh giá chất lượng bản dịch máy

2.5.1. Khái quát về đánh giá chất lượng bản dịch máy

Để đánh giá chất lượng bản dịch máy, thông thường sử dụng các giá trị như độ chính xác, độ hoàn thiện, độ tương đồng và độ tự nhiên.

- Độ chính xác (Accuracy): Là tỷ lệ phần trăm của các từ hoặc câu được dịch chính xác đúng so với bản gốc.
- Độ hoàn thiện (Fluency): Là khả năng của bản dịch để sử dụng các từ và cú pháp một cách chính xác và tự nhiên trong ngôn ngữ đích.
- Độ tương đồng (Similarity): Là mức độ giống nhau giữa bản dịch và bản gốc, bao gồm cả cấu trúc và nội dung.
- Độ tự nhiên (Naturalness): Là khả năng của bản dịch để nghe có vẻ tự nhiên như người nói bản ngữ.

Mỗi giá trị trên đều quan trọng trong việc đánh giá chất lượng của bản dịch máy và cần được xem xét cẩn thận để đảm bảo bản dịch đạt được tiêu chuẩn chất lượng cao. Các điểm đánh giá được tính toán bằng cách so sánh bản dịch máy với bản gốc và đưa ra đánh giá dựa trên các tiêu chí được xác định trước đó. Để tính toán các điểm đánh giá này, người ta thường sử dụng các phương pháp khác nhau như đo đạc tự động hoặc đánh giá thủ công.

- Đo đạc tự động: Đây là phương pháp đánh giá tự động dựa trên các thuật toán và công cụ phân tích ngôn ngữ tự nhiên (NLP). Ví dụ: BLEU, METEOR, TER, ROUGE.
- Đánh giá thủ công: Đây là phương pháp đánh giá bằng cách có người đánh giá chất lượng của bản dịch máy dựa trên các tiêu chí như độ chính xác, độ hoàn thiện, độ tương đồng và độ tự nhiên. Phương pháp này có thể mang lại kết quả chính xác hơn, nhưng đòi hỏi nhiều thời gian và chi phí hơn.

2.5.2. Điểm đánh giá BLEU

Phương pháp đánh giá tự động phổ biến nhất là BLEU (Bilingual Evaluation Understudy). Phương pháp này được sử dụng rộng rãi trong cộng đồng nghiên cứu về dịch máy và đã trở thành tiêu chuẩn đánh giá cho các hệ thống dịch máy. 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.

Để triển khai phương pháp đánh giá này ta cần có các bản dịch chính xác được thực hiện bởi chuyên gia (con người) rồi đem đi so sánh với bản dịch máy để thu được những chỉ số tương quan giữa hai bản dịch này. Sau đó chất lượng của hệ thống dịch được đánh giá dựa trên chỉ số được gọi là điểm BLEU.

Kết quả được tính toán là đo sự trùng khớp các n-grams (dãy ký tự gồm n từ hoặc ký tự) từ kho dữ liệu của kết quả dịch và kho các bản dịch tham khảo có chất lượng cao. Việc tính toán sự trùng khớp này có tính đến cả thứ tự của các từ trong câu. Giải thuật của IBM đánh giá chất lượng của hệ thống dịch qua việc trùng khớp của các n-grams đồng thời nó cũng dựa trên cả việc so sánh độ dài của các bản dịch.

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

$$score = exp\left\{\sum_{i=1}^{N} w_i \, log(P_i) - \max\left(\frac{L_{\text{ref}}}{L_{\text{tra}}} - 1{,}0\right)\right\} \qquad (2.13)$$

- $P_i = \frac{\sum_j NR_j}{\sum_j NT_j}$
- 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

Công thức này đánh giá sự trùng khớp của các n-grams giữa đoạn dịch của máy tính và đoạn dịch tham chiếu, đồng thời điều chỉnh dựa trên độ dài của đoạn dịch. Kết quả BLEU là một giá trị số từ 0 đến 1, với 1 là độ chính xác hoàn hảo và 0 là độ chính xác tệ nhất.

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 [54]. Do vậy luận án chọn phương pháp đánh giá bản dịch BLEU cho bài toán dịch.

2.5.3. Điểm đánh giá hiệu suất mô hình ngôn ngữ Perplexity

Một vấn đề quan trọng đặt ra cho bài toán trong luận án này là việc xây dựng các kho ngữ liệu cho việc đánh giá các mô hình đề xuất. Bởi vậy, việc xem xét đánh giá các kho ngữ liệu này là cần thiết. 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 [55]. 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. Perplexity trong mô hình ngôn ngữ n-gram được tính bằng công thức sau:

$$Perplexity(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots w_N,)}} \qquad (2.14)$$

Trong đó:

- N là số từ trong đoạn văn bản kiểm tra.
- $P(w_1, w_1, \dots w_1,)$ là xác suất của đoạn văn bản kiểm tra trong mô hình ngôn ngữ n-gram.
- $\sqrt[N]{\dots}$ là lấy căn bậc N, trong đó N là số từ trong đoạn văn bản kiểm tra. Công thức này giúp chuẩn hóa perplexity sao cho không phụ thuộc vào kích thước của đoạn văn bản.

Perplexity càng nhỏ thì mô hình càng tốt, tức là mô hình có khả năng dự đoán chuỗi từ mới tốt hơn. Trong các mô hình ngôn ngữ n-gram, perplexity cũng thường được sử dụng để so sánh giữa các mô hình khác nhau để đánh giá hiệu quả của chúng trong dự đoán ngôn ngữ [56]. Độ phức tạp thấp nhất đã được công bố năm

1992 trên kho dữ liệu Brown Corpus (1 triệu từ tiếng Anh Mỹ thuộc các chủ đề và thể loại khác nhau) với giá trị thực tế là khoảng 247, tương ứng với một cross-entropy $\log_2 247 = 7{,}95$ bit mỗi từ hoặc 1,75 bit mỗi chữ cái sử dụng mô hình 3-gram. Thường có thể đạt được mức độ phức tạp thấp hơn đối với các kho ngữ liệu chuyên biệt hơn, vì chúng dễ dự đoán hơn. Bảng 4.2. dưới đây là chỉ số perplexity của một số kho ngữ liệu phổ biến được tính bằng mô hình ngôn ngữ 3-gram:

Bảng 2.1. Chỉ số perplexity của một số kho ngữ liệu phổ biến

Kho ngữ liệu	**Chỉ số Perplexity trung bình**
WikiText-103	109-113
Penn Treebank	110-120
Common Craml	600-800

Điểm perplexity của một kho ngữ liệu phụ thuộc vào nhiều yếu tố như kích thước của kho ngữ liệu, độ phức tạp của cấu trúc ngôn ngữ, độ phong phú của từ vựng, v.v. Trong nhiều trường hợp, điểm perplexity sẽ tăng theo kích thước của kho ngữ liệu, đặc biệt là khi kích thước của kho ngữ liệu tăng lên đáng kể. Tuy nhiên, sự tăng này không phải lúc nào cũng xảy ra và có thể bị giới hạn bởi độ phức tạp của cấu trúc ngôn ngữ hoặc độ phong phú của từ vựng. Vì vậy, việc tính toán perplexity của một kho ngữ liệu không phải là một cách để đánh giá kích thước của nó, mà là một cách để đo lường độ chính xác của mô hình ngôn ngữ được huấn luyện trên kho ngữ liệu đó. Nếu mô hình ngôn ngữ đạt được perplexity thấp trên kho ngữ liệu lớn, điều đó cho thấy mô hình có khả năng dự đoán tốt hơn trên các dữ liệu mới, và do đó có thể có hiệu suất tốt hơn trong nhiều ứng dụng thực tế.

2.6. Kết luận chương

Chương 2 trình bày các kiến thức 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; điểm Perplexity- điểm đánh giá độ tương đồng cho các kho ngữ liệu mà luận án xây dựng phục vụ cho việc đánh giá các mô hình dịch đề xuất.

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

Trong chương này, luận án đề xuất một mô hình giải quyết bài toán dịch máy đã xác định rõ mục tiêu ban đầu. Đó là mô hình dịch dựa trên luật (rule-based). Tuy đây là một phương pháp cổ điển nhưng tại xuất phát điểm của những nghiên cứu liên quan đến dịch ngôn ngứ ký hiệu và thời điểm hiện tại, phương pháp này vẫn được đánh giá là có hiệu quả đối với lớp các bài toán cho xử lý ngôn ngữ ít tài nguyên. Đây cũng chính là một đặc điểm quan trọng của VSL. Phần cuối chương trình bày những kết quả đánh giá thực nghiệm mô hình này cho bài toán dịch thông qua điểm đánh giá chất lượng bản dịch BLEU trên các tập kiểm tra được xây dựng trên một số miền dữ liệu khác nhau.

3.1. Giới thiệu về bài toán

Với các phân tích và đánh giá ở chương 2, phương pháp tiếp cận dựa trên cấu trúc (rule-based) là phù hợp và hiệu quả cho bài toán dịch ngôn ngữ ký hiệu Việt Nam với mục tiêu đề ra là dịch câu đúng cú pháp tiếng Việt sang dạng câu đúng cú pháp trong VSL. Phương pháp này sử dụng các quy tắc ngữ pháp và từ điển để chuyển đổi các câu từ ngôn ngữ nguồn sang ngôn ngữ đích. Để thực hiện bài toán dịch máy dựa trên luật, cần có các yếu tố sau:

- Bộ từ điển: Đây là bộ từ điển chứa các cặp từ tương ứng giữa ngôn ngữ nguồn và ngôn ngữ đích. Các cặp từ này được sắp xếp theo quy tắc và có thể có nhiều dạng biểu thức để biểu diễn các trường hợp khác nhau.
- Quy tắc ngữ pháp: Đây là các quy tắc và luật ngữ pháp đặt ra để thực hiện việc chuyển đổi câu từ ngôn ngữ nguồn sang ngôn ngữ đích. Những quy tắc này có thể dựa trên cấu trúc ngữ pháp, từ loại, hoặc các thông tin văn bản khác. Đồng thời cần các kiến thức sâu về ngôn ngữ nguồn và ngôn ngữ đích để hiểu và áp dụng đúng các quy tắc và từ điển phù hợp.
- Mô hình dịch máy: Dựa vào quy tắc và từ điển, xây dựng một mô hình dịch máy dựa trên luật để thực hiện việc dịch từ ngôn ngữ nguồn sang ngôn ngữ đích. Mô hình này sẽ xử lý các quy tắc và từ điển để tạo ra các bản dịch.

- Đánh giá và cải thiện: Sau khi xây dựng mô hình dịch máy dựa trên luật, cần tiến hành đánh giá hiệu suất của mô hình và cải thiện các quy tắc và từ điển nếu cần thiết để tăng độ chính xác và hiệu quả của hệ thống dịch.

Bởi vậy, động lực trong bài toán dịch VSL đặt ra với việc ứng dụng mô hình dịch dựa trên cấu trúc xuất phát từ những yếu tố trên. Ban đầu, luận án tiến hành xây dựng các tập cơ sở dữ liệu ban đầu cho bài toán bao gồm: tập từ điển VSL-Lexicon; tập dữ liệu "song ngữ" Vie-VSL10K bao gồm 10.000 cặp câu tiếng Việt – Ngôn ngữ ký hiệu Việt Nam. Từ đó xây dựng các quy tắc cú pháp dựa trên vấn đề tổng hợp luật từ việc nghiên cứu các đặc điểm về cú pháp của VSL. Hệ thống dịch dựa trên luật bao gồm các thành phần chính là: các quy tắc được tổng hợp từ phân tích cú pháp; cùng với tập từ điển mà ở đây chủ yếu là ánh xạ 1-1, còn lại là ánh xạ của nhóm các từ đồng nghĩa; hệ thống so khớp luật; tập dữ liệu chờ bổ sung luật mới nếu không so khớp luật thành công. Từ đó với đầu vào là một câu tiếng Việt thông thường sẽ thu được đầu ra là một câu đúng cú pháp VSL hoặc trả kết quả là giữ nguyên câu gốc với thông báo không tìm thấy luật, đẩy dữ liệu vào tập dữ liệu chờ.

3.2. Xây dựng cơ sở dữ liệu ban đầu cho bài toán

3.2.1. Tập từ điển VSL-Lexicon.

Như đã phân tích ở chương 1, vấn đề cấp thiết và quan trọng trong bài toán dịch là cần có bộ dữ liệu để thực nghiệm và đánh giá với các mô hình, phương pháp dịch máy. Công việc ban đầu triển khai của luận án bao gồm việc xây dựng cơ sở dữ liệu phù hợp với bài toán này.

Với tính chất đặc trưng của VSL là hạn chế về từ vựng so với tiếng Việt, tổng số lượng đơn vị từ vựng (chữ cái, số, từ, cụm từ) được dùng trong VSL hiện nay khoảng trên 6000 đơn vị. Luận án thu thập thông tin trên một số nguồn đáng tin dậy được sử dụng trong cộng đồng người khiếm thính Việt Nam. Đầu tiên là dựa trên một số tài liệu phát hành nội bộ dùng trong cộng đồng người khiếm thính của câu lạc bộ người khiếm thính Thái Nguyên, trung tâm giáo dục trẻ em thiệt thòi dạng điếc câm của Thái Nguyên, Hà Nội, Hải Phòng. Nguồn thứ 2 là dựa trên sản phẩm nghiên cứu và công bố cho cộng đồng của tiến sĩ Cao Thị Xuân Mỹ và các cộng sự về từ điển ngôn ngữ ký hiệu.

Từ năm 2017, nghiên cứu sinh đã xây dựng kho từ điển VSL từ các nguồn trên (có tham vấn ý kiến chuyên gia và cộng đồng người khiếm thính). Tổng cộng thu thập được 3053 đơn vị ngôn ngữ. Hiện nay số lượng từ và cụm từ được bổ sung liên tục và đang có 6176 kí tự/từ/cụm từ được biểu diễn bằng ngôn ngữ ký hiệu. Từ đó xây dựng được từ điển VSL được đặt tên **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. Do vấn đề khó khăn khi phải sản xuất các mô hình diễn hoạ thủ công với khối lượng công việc rất lớn nên hiện tại trong VSL-Lexicon chỉ mới có 200 mô hình. Các mô hình được lưu dưới dạng file .FBX. Đối với định dạng tệp tin ".FBX", ta có thể xuất mô hình 3D với tất cả các animation, chuyển động, rigging và các thông số khác được lưu trữ trong tệp. Định dạng tệp ".FBX" được hỗ trợ bởi nhiều phần mềm 3D khác nhau và là định dạng tệp chuẩn được sử dụng trong Unity. 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	Từ 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, tử nạn,	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 trong CSDL

Thông thường trong một mô hình dịch dựa trên luật, ta cần từ điển song ngữ - đây là một tài nguyên quan trọng để xây dựng một mô hình dịch máy dựa trên luật. Từ điển này sẽ bao gồm các cặp từ/cụm từ tương ứng giữa ngôn ngữ nguồn và ngôn ngữ đích. Nhưng đối với việc dịch Vie-VSL, hầu hết các đơn vị từ vựng đều ánh xạ 1-1, tức là 1 từ bên câu tiếng Việt đồng nhất với 1 từ trong VSL. Trong một số

trường hợp từ vựng tiếng Việt không có trong VSL thì ta thay thế bằng các từ đồng nghĩa (một số trường hợp đặc biệt như tên riêng thì không thay thế mà lưu trữ lại để thể hiện dưới dạng đánh vẫn). Đó là sự khác biệt rõ nét nhất của từ điển VSL trong bài toán dịch Vie-VSL so với các loại từ điển thông thường như từ điển Việt – Anh, từ điển Nhật- Việt.

Hình 3.1. Hình ảnh về mô hình 3D mã số VSL0153 trong VSL-Lexicon

3.2.2. Bộ dữ liệu song ngữ Vie-VSL10k

Đối với mô hình dịch dựa trên luật, ngoài thành phần từ điển thì cần có một tài nguyên quan trọng là các quy tắc ngữ pháp. Một mô hình dịch máy dựa trên luật sẽ sử dụng các quy tắc ngữ pháp để phân tích và dịch câu. Các quy tắc này sẽ được xác định trước và được cấu trúc theo các ngữ pháp của các ngôn ngữ đang được dịch. Đối với bài toán trong luận án này, việc xây dựng các quy tắc ngữ pháp dựa trên việc phân tích các cú pháp của một bộ dữ liệu song ngữ Vie-VSL. 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. Một phần dữ liệu có nguồn từ các cặp câu Vie- VSL trong nghiên cứu của Tiến sĩ ngôn ngữ học Cao Thị Xuân Mỹ và các cộng sự. Một phần dữ liệu được lấy từ kho dữ liệu BTEC traveling bao gồm các câu trong miền giao tiếp của dữ liệu song ngữ Anh-Việt. Các dữ liệu này 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 và xem xét cuối cùng bởi Tiến sĩ Vũ Thị Hải Hà- Viện ngôn ngữ học -Viện hàn lâm khoa học và xã hội Việt Nam. 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 3.2

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	Từ hay tiếng nước ngoài (hay từ vay mượn)	B	0	0
16	Từ viết tắt	Y	0	0
17	Yếu tố cấu tạo từ	S	10	0
18	Các từ không phân loại được	X	322	0

3.3. Vấn đề tổng hợp luật.

Với các đặc điểm đặc trưng về cú pháp trong VSL đã trình bày trong chương 1. Có một số đặc điểm rút gọn và chuyển đổi cú pháp của câu trong VSL được tổng hợp lại như sau:

3.3.1. Tính chất rút gọn trong câu VSL

Các từ rút gọn (stop-words) bị loại bỏ trong câu VSL được thống kê trong bảng 3.3. Phương pháp loại bỏ các từ rút gọn này được thực hiện đơn giản dựa trên việc xây dựng từ điển Stop-words và tần xuất xuất hiện trong câu.

Bảng 3.3. Các từ được rút gọn trong câu VSL

Từ loại	Ví dụ
Định từ	mỗi, từng, mọi, cái; các, những, mấy
Phụ từ	đã, sẽ, đang, vừa, mới, từng, xong, rồi; rất, hơi, khí, quá, là
Tiểu từ tình thái	à, a, á, ạ, ấy, chắc, chăng, cho, chứ, có, nhỉ, nhé, chứ, vậy, đâu, hả, hử
Từ cảm thán	ơi, vâng, dạ, bẩm, thưa, ừ, ôi, trời ơi, ô, ủa, kìa, ái, ối, than ôi, hỡi ôi, eo ôi, ôi giời ôi, …
Trợ từ nhấn mạnh	cả, chính, đích, đúng, chỉ, những, đến, tận, ngay,...
Động từ tình thái	nên, cần, phải, cần phải, có thể, bị, được, mắc phải, trông, mong, chúc, ước, cầu, muốn, dám, định, nỡ, thôi, đành, …
Giới từ chính danh	tại, bởi, vì, từ, tuy, mặc dầu, nếu, dù…

3.3.2. Tập hợp đặc điểm cú pháp câu VSL

Tính chất rút gọn trong câu VSL khiến cho bài toán chuyển đổi câu tiếng việt sang dạng đúng trong VSL gần giống như bài toán tóm tắt văn bản. Tuy nhiên đặc trưng khác biệt so với bài toán tóm tắt văn bản là vấn đề trật tự cú pháp trong câu VSL. Do những đặc điểm đặc trưng của ngôn ngữ, thông tin chính được nhấn mạnh và thường đưa lên trước nên cú pháp câu VSL có trật tự cú pháp khác so với câu tiếng Việt thông thường.

Các đặc điểm về trật tự cú pháp trong VSL được tổng kết lại như sau:

Quy tắc 1 : Thay đổi trật tự danh từ và số từ:

Bảng 3.4. Cấu trúc chuyển đổi trật tự của danh từ- số từ trong câu VSL (a)

	Tiếng Việt	**Câu đúng cú pháp VSL**
Cấu trúc	Số đếm + danh từ	Danh từ + số đếm
Ví dụ	Hai quả táo	Quả táo hai
Cấu trúc	Danh từ + số thứ tự	Danh từ + số thứ tự
Ví dụ	Người thứ nhất	Người thứ nhất

Quy tắc 2 : Thay đổi trật tự động từ và từ phủ định trình bày trong bảng

Bảng 3.5. Cấu trúc chuyển đổi trật tự của động từ - từ phủ định trong câu VSL

	Tiếng Việt	**Câu đúng cú pháp VSL**
Cấu trúc	Từ phủ định + động từ	Động từ + từ phủ định
Ví dụ	Không ăn	Ăn không

Quy tắc 3 : Cấu trúc chuyển đổi trật tự từ của câu đơn trong VSL

Bảng 3.6. Cấu trúc chuyển đổi trật tự của động từ - từ phủ định trong câu VSL

	Tiếng Việt	**Câu đúng cú pháp VSL**
Cấu trúc	Chủ ngữ + động từ + bổ ngữ	Chủ ngữ + bổ ngữ + động từ
Ví dụ	Cô ấy ăn táo	Cô ấy táo ăn

Quy tắc 4 : Cấu trúc chuyển đổi trật tự từ của câu nghi vấn trong VSL. Riêng đối với VSL, những từ để hỏi luôn đứng ở vị trí cuối câu hỏi*:*

Bảng 3.7. Cấu trúc chuyển đổi trật tự từ của câu nghi vấn trong VSL (a)

	Tiếng Việt	**Câu đúng cú pháp VSL**
Cấu trúc	Chủ ngữ (từ để hỏi)+ vị ngữ + bổ ngữ ?	Bổ ngữ + vị ngữ + Chủ ngữ (từ để hỏi)
Ví dụ	Ai ăn táo?	Táo ăn ai?
Cấu trúc	Chủ ngữ + vị ngữ + từ để hỏi + bổ ngữ?	Chủ ngữ + bổ ngữ + vị ngữ + từ để hỏi?
Ví dụ	Cường ăn mấy quả táo?	Cường táo ăn mấy?

Quy tắc 5: Cấu trúc chuyển đổi trật tự từ của câu phủ định trong VSL. Tiếng Việt có nhiều dạng phủ định: Phủ định hoàn toàn, phủ định bộ phận. Trong cấu trúc phủ định bộ phận: phủ định động từ thì phủ định đứng trước động từ mà nó phủ định. Còn trong NGÔN NGỮ KÝ HIỆU thì từ phủ định này luôn đứng sau động từ và đứng ở cuối câu.

Bảng 3.8. Cấu trúc chuyển đổi trật tự từ của câu phủ định trong VSL

	Tiếng Việt	**Câu đúng cú pháp VSL**
Cấu trúc	Chủ ngữ + từ phủ định + vị ngữ	Chủ ngữ + vị ngữ + từ phủ định
Ví dụ	Cường không ăn táo.	Cường táo ăn không.

3.2.3. Vấn đề phân tích cú pháp và trích rút luật

Phân tích cú pháp văn bản nguồn, tạo ra một biểu diễn tượng trưng trung gian của nó, và sau đó tạo bản dịch cuối cùng trong ngôn ngữ đích là mục tiêu của phương pháp dich dựa trên luật. Trong luận án này sử dụng bộ công cụ phân tích cú pháp là sản phẩm nghiên cứu của Tiến sĩ Nguyễn Phương Thái và các cộng sự cho bài toán [57].

Quá trình tiền xử lý bao gồm chuẩn hóa dữ liệu vào cùng với bộ công cụ tách từ và gán nhãn từ loại VietWS. Bộ công cụ này được sử dụng rộng rãi trong cộng đồng xử lý tiếng Việt, được phát triển trong đề tài VLSP với độ chính xác đạt 97%. Bảng 3.9 mô tả một số ví dụ về kết quả tách từ.

Bảng 3.9. Kết quả tách từ

Câu ban đầu	**Tách từ sử dụng công cụ VietWS**
Thái Nguyên nổi tiếng là tỉnh có trà ngon nhất Việt Nam.	Thái_Nguyên nổi_tiếng là tỉnh có trà ngon nhất Việt_Nam .
Hôm nay tôi đi học.	Hôm_nay tôi đi học .
Con cái là niềm tự hào và hạnh phúc của cha mẹ.	Con_cái là niềm tự_hào và hạnh_phúc của cha_mẹ .
Tôi ăn hai quả táo xanh.	Tôi ăn hai quả táo xanh.
Nhà tôi ở Thái Nguyên.	Nhà tôi ở Thái_Nguyên.

Phân tích cú pháp câu văn bản tiếng Việt sẽ cho chúng ta cấu trúc cú pháp của câu dưới dạng cấu trúc cây. Mỗi từ loại trong cây đều được gán nhãn. Các nhãn trong phân tích cú pháp bao gồm: nhãn từ loại, nhãn thành phần cú pháp, nhãn cụm từ và nhãn mệnh đề.

Tập nhãn từ loại chỉ chứa thông tin về từ loại cơ sở mà không bao gồm các thông tin như hình thái, phân loại con, v.v. Tập nhãn từ loại liệt kê trong Bảng 3.10, tổng số nhãn là 18.

Bảng 3.10. Nhãn từ loại

STT	Tên	Chú thích	Ví dụ
1	N	Danh từ	Con người, sông ngòi, núi rừng, v.v…
2	Np	Danh từ riêng	Thái Nguyên, trường Đại học Công nghệ Thông tin,…
3	Nc	Danh từ chỉ loại	con, cái, đứa,
4	Nu	Danh từ đơn vị	Cân, mét, đồng, lít,…
5	V	Động từ	Ăn, chơi, đọc, thích, yêu,…
6	A	Tính từ	To, cao, ngon, đẹp,…
7	P	Đại từ	Tôi, cô ấy, chúng nó,…
8	L	Định từ	mỗi, từng, mọi, cái; các, những, mấy
9	M	Số từ	Hai, ba, nửa, rưỡi,
10	R	Phụ từ	đã, sẽ, đang, vừa, mới, từng,…
11	E	Giới từ	trên, dưới, trong, ngoài,…
12	C	Liên từ	và, với, cùng, vì vậy, nhưng,...
13	I	Thán từ	ôi, chao, a ha
14	T	Trợ từ, tiểu từ, từ tình thái	à, a, á, ạ, ấy, chắc, chăng, cho, chứ, có
15	B	Từ hay tiếng nước ngoài (hay từ vay mượn)	Internet, email, video, chat
16	Y	Từ viết tắt	OPEC, WTO, HIV
17	S	Yếu tố cấu tạo từ	bất, vô, gia, đa
18	X	Các từ không phân loại được	

Nhãn thành phần cú pháp: Cụm từ và mệnh đề là các thành phần cú pháp cơ bản và được mô tả bằng nhãn thành phần cú pháp. Đó chính là những thành phần cơ bản trên cây cú pháp. Nhưng do sự khác biệt của các môn ngữ thì thường tập nhãn cú pháp của các ngôn ngữ khác nhau sẽ khác nhau ở một tỉ lệ nhất định. Bảng 3.11 liệt kê tập nhãn cụm từ và Bảng 3.12 là nhãn mệnh đề .

Bảng 3.11. Tập nhãn cụm từ

STT	Tên	Chú thích
1	NP	Cụm danh từ
2	VP	Cụm động từ
3	AP	Cụm tính từ
4	RP	Cụm phụ từ
5	PP	Cụm giới từ

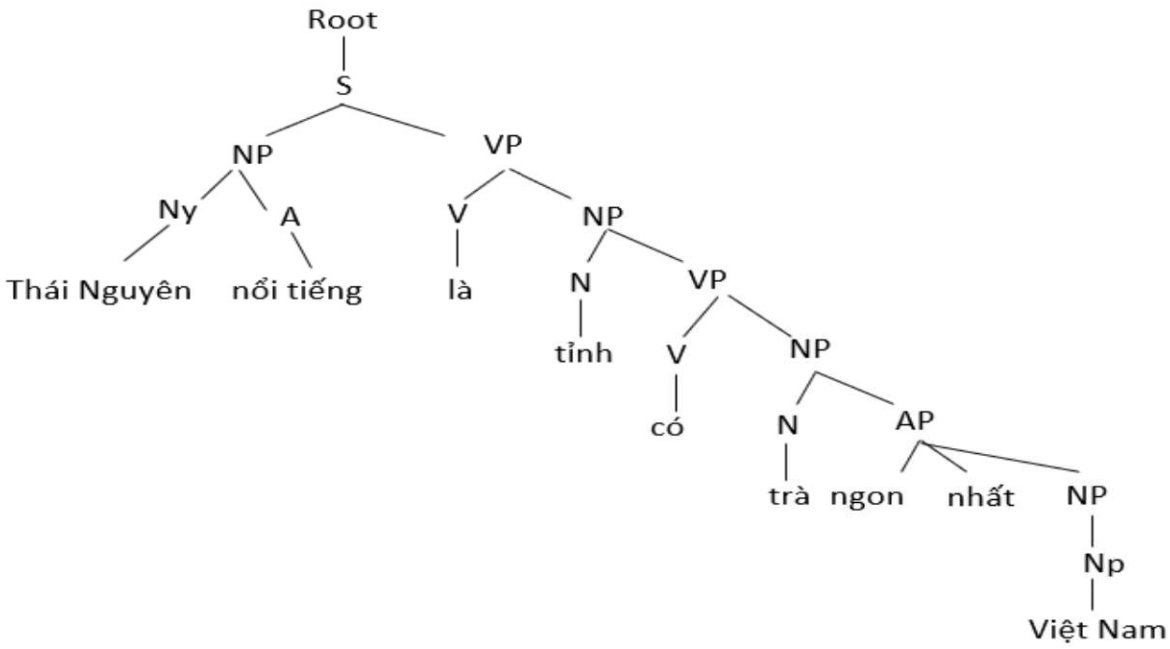
STT	Tên	Chú thích
6	QP	Cụm từ chỉ số lượng
7	MDP	Cụm từ tình thái
8	UCP	Cụm từ gồm hai hay nhiều thành phần không cùng loại được nối với nhau bằng liên từ đẳng lập
9	LST	Cụm từ đánh dấu đầu mục của danh sách
10	WHNP	Cụm danh từ nghi vấn (ai, cái gì, con gì, v.v…)
11	WHAP	Cụm tính từ nghi vấn (lạnh thế nào, đẹp ra sao,v.v..)
12	WHRP	Cụm tính từ nghi vấn khi hỏi về thời gian, nơi chốn, v.v…
13	WHPP	Cụm giới từ nghi vấn (với ai, bằng cách nào, v.v…)

Bảng 3.12. Nhãn mệnh đề

STT	Tên	Chú thích
1	S	Câu trần thuật (khẳng định hoặc phủ định)
2	SQ	Câu hỏi
3	S-EXC	Câu cảm thán
4	S-CMD	Câu mệnh lệnh
5	SBAR	Mệnh đề phụ kết (bổ nghĩa cho danh từ, động từ và tính từ)

Việc xây dựng cây cấu trúc cú pháp cho dữ liệu 10.000 câu tiếng Việt được tiến hành với công cụ phân tích cú pháp có được. Một ví dụ về câu được phân tích cấu trúc cú pháp thể hiện ở hình 3.2.

Ví dụ: Câu tiếng Việt: "*Thái Nguyên nổi tiếng là tỉnh có trà ngon nhất Việt Nam.*", bước đầu tiên sẽ được tách từ từ công cụ VietWS, ta thu được kết quả là câu: "*Thái_Nguyên nổi_tiếng là tỉnh có trà ngon nhất Việt_Nam .*". Cho kết quả này là đầu vào của công cụ Phân tích cú pháp PARSE, ta được cấu trúc cây với kết quả phân tích cú pháp gồm các thành phần là : *"((S (NP (Ny Thái_Nguyên) (A nổi_tiếng)) (VP (V là) (NP (N tỉnh) (VP (V có) (NP (N trà) (AP (A ngon) (R nhất) (NP (Np Việt_Nam))))))) (. .)))*"



Hình 3.2. Cây cú pháp khi phân tích câu bằng công cụ PARSE

Luận án trình bày việc áp dụng các công cụ này 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)) (..)

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
6	((S (NP (P tôi)) (VP (R không) (V đi)) (..)))	3	((S (NP (P tôi)) (VP (V đi) (R không)) (..)))	((S (NP (P)) (VP (R) (V)) (..))) → ((S (NP (P)) (VP (V) (R)) (..))) (..)
8	((S (NP (P tôi)) (VP (R không) (V chơi)) (..)))	3	((S (NP (P tôi)) (VP (V chơi) (R không)) (..)))	
9	S (NP (P Tôi)) (VP (V thích) (NP (N mèo))) (..)	4	S (NP (P Tôi)) (VP (N mèo) (V thích)) (..)	S (NP (P)) (VP (V) (NP (N)) (..) → S (NP (P)) (VP (N) (V))
10	SQ (NP (P Ai)) (VP (V biết) (VP (V bơi))) (? ?)	5	SQ (VP (V Biết) (VP (V bơi) (NP (P ai)))) (? ?)	SQ (NP (P)) (VP (V) (VP (V))) (? ?) → SQ (VP (V) (VP (V) (NP (P)))) (? ?)
11	S (NP (P Tôi)) (VP (R không) (V thích) (NP (N rắn))) (..)	3 và 4	S (NP (P tôi)) (A rắn) (V thích) (R không)) (..)	S (NP (P)) (VP (R) (V) (NP (N)) (..) → S (NP (P)) (A) (V) (R)) (..)
12	SQ (NP (M Một) (N năm)) (VP (V có) (NP (L mấy) (N mùa))) (? ?)	1 và 5	SQ (NP (M Một) (N năm) (N mùa) (L mấy)) (? ?)	SQ (NP (M) (N)) (VP (V) (NP (L) (N))(? ?)→ SQ (NP (M) (N) (N) (L)) (? ?)
13	SQ (NP (M Một) (N tuần)) (VP (V có) (NP (L mấy) (N ngày))) (? ?)	1 và 5	SQ (NP (M Một) (N tuần) (N ngày) (L mấy)) (? ?)	
14	SQ (NP (M Một) (N năm)) (VP (V có) (NP (L mấy) (N tháng))) (? ?)	1 và 5	SQ (NP (M Một) (N năm) (N tháng) (L mấy))) (? ?)	
	…	…	…	…

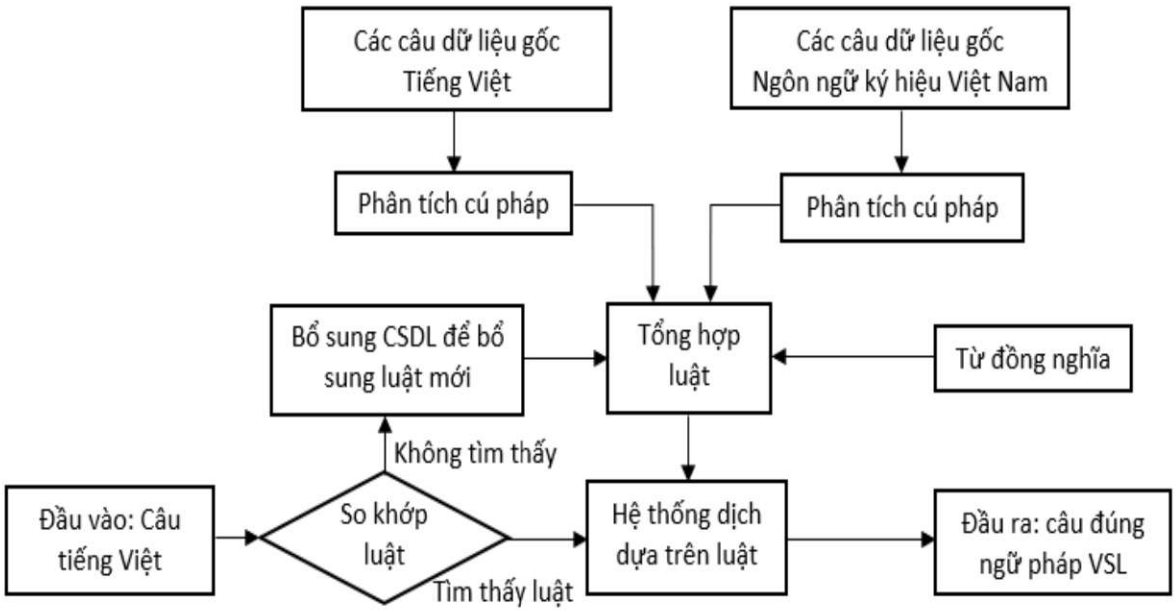
Từ 8025 luật được trích rút từ kho dữ liệu 10000 cặp câu song ngữ, ta tiến hành xây dựng hệ thống dịch máy dựa trên luật. Hiệu quả của phương pháp dịch này được phân tích và đánh giá ở phần sau. Tham khảo 8025 luật tại https://github.com/BichDiep/rules-VSL.git.

3.4. Xây dựng hệ thống dịch dựa trên luật

Với nội dung xây dựng cơ sở dữ liệu ban đầu cho bài toán ở 3.2, ta xây dựng hệ thống dịch theo luật dựa trên cơ sở dữ liệu này. Bước tiền xử lý tách từ được sử dụng trong hệ thống ở bước tiền xử lý cơ sở dữ liệu gốc và tiền xử lý cho câu đầu

vào của hệ thống dịch. Với dữ liệu ban đầu bao gồm 10000 cặp câu song ngữ tiếng Việt – ngôn ngữ ký hiệu Việt Nam (Vie-VSL) đã được xây dựng, ta tiến hành phân tích cú pháp với công cụ phân tích cú pháp Tiếng Việt. Kết hợp với dữ liệu từ đồng nghĩa trong từ điển Ngôn ngữ ký hiệu tiếng Việt, ta tiến hành tổng hợp luật để đưa ra hệ thống dịch dựa trên luật. Như vậy, khi cần dịch một câu tiếng Việt bất kỳ ta tiến hành phân tích cú pháp câu và so khớp luật. Nếu tìm thấy luật tương ứng, câu đầu vào sẽ được xử lý qua hệ thống dịch luật và đưa ra kết quả là câu dạng đúng cú pháp trong ngôn ngữ ký hiệu. Trong trường hợp không tìm thấy luật, câu được dịch giữ nguyên và thêm vào trong cơ sở dữ liệu chờ. Tại đây, bộ cơ sở dữ liệu sẽ được sinh thêm luật mới để bổ sung vào tập luật sau.

Quy trình xây dựng hệ thống và dịch dựa trên luật được mô tả theo sơ đồ hình 3.3.



Hình 3.3. Quy trình xây dựng hệ thống dịch máy theo luật

Ví dụ 1:

Câu dữ liệu đầu vào: "*Bến Tre nổi tiếng là tỉnh có dừa ngon nhất Việt Nam*" , ta có phân tích cú pháp của câu: *"((S (NP (Ny Bến _Tre) (A nổi_tiếng)) (VP (V là) (NP (N tỉnh) (VP (V có) (NP (N dừa) (AP (A ngon) (R nhất) (NP (Np Việt_Nam)))))))) (. .))).* Quá trình so khớp luật tìm thấy luật 128 phù hợp.

Luật 128: *"((S (NP (Ny) (A)) (VP (V (NP (N (VP (V) (NP (N) (AP (A) (R) (NP (Np)))))))) (. .))) → ((S (NP (Np)) (AP (A) (NP (N) (AP (A) (R) (NP (Np))))) (. .)))*

Từ đó hệ thống dịch tham chiếu để chuyển cú pháp và rút gọn thành phần trong câu theo luật 128 thu được kết quả câu đầu ra: "*Bến Tre nổi tiếng dừa ngon nhất Việt Nam*".

Trong sơ đồ hệ thống trên, 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 = Ø; (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′
```

Ví dụ 2:

Câu dữ liệu đầu vào: "*Mũi thuyền in một nét mơ hồ lòe nhòe vào bầu sương mù trắng như sữa có pha đôi chút màu hồng hồng do ánh mặt trời chiếu vào.*", ta có phân tích cú pháp của câu là*: ((S (NP (N Mũi) (N thuyền)) (VP (VP (V in) (NP (M một) (N nét) (A mơ_hồ) (VP (V lòe) (V nhòe) (PP (E vào) (NP (N bầu) (N sương_mù) (AP (A trắng) (C như) (NP (N sữa)))))))) (VP (V có) (VP (V pha) (NP (L đôi_chút) (N màu) (N hồng_hồng) (SBAR (E do) (S (NP (Nc ánh) (N mặt_trời)) (VP (V chiếu) (R vào)))))))) (. .))).* Quá trình so khớp không tìm thấy luật. Do vậy kết quả trả về là câu gốc được giữ nguyên. Đồng thời được bổ sung vào tập dữ liệu chờ để xem xét.

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

Cài đặt chương trình và cấu hình máy sử dụng cho việc cài đặt hệ thống dịch dựa trên luật bao gồm:

- Ngôn ngữ lập trình: Python 3.1

- Phần mềm hỗ trợ tích hợp bao gồm VietWS và Parsing để xử lý và phân tích ngôn ngữ tiếng Việt
- Trình biên dịch Visual Studio Code để phát triển và chạy chương trình.

Cấu hình máy tính cá nhân của nghiên cứu sinh bao gồm các thông số cơ bản như:

- Hệ điều hành: Windows 11 Home Single Language 64-bit.
- CPU: Core i7-1165G7; 2,8Ghz (8 CPUs).
- RAM :8129MB.
- Ổ cưng lưu trữ: Đảm bảo có đủ không gian lưu trữ để lưu trữ tập dữ liệu, từ điển và mã nguồn chương trình.

Để đánh giá hiệu quả của phương pháp dịch dựa trên luật cho bài toán dịch Vie-VSL, luận án thực hiện đánh giá trên 3 tập kiểm tra được chuẩn bị. Ngoài tập dữ liệu trong miền các câu giao tiếp thông thường, luận án này cũng lựa chọn một số dữ liệu trên miền khác như: văn học, kỹ thuật và y học để xây dựng tập kiểm tra đánh giá toàn diện cho phương pháp dịch mà luận án này xây dựng. Bảng dưới đây liệt kê thông số về tập dữ liệu thử nghiệm.

Bảng 3.14. Thông số của tập dữ liệu thử nghiệm hệ thống

TT	Tên miền	Số câu	Độ dài trung bình câu	Tổng lượng từ vựng
1	Miền câu giao tiếp thông thường	200	9,8	1245
2	Miền câu trong lĩnh vực y học	100	12,1	986
3	Miền câu trong lĩnh vực kỹ thuật	100	14,4	1027
4	Miền câu trong lĩnh vực văn học	100	17,7	1325

Trong đó độ dài trung bình câu được tính bằng số lượng đơn vị từ vựng trung bình trên một câu. Tổng số lượng từ vựng trên mỗi miền được tính không bao gồm các từ vựng trùng nhau trên mỗi miền dữ liệu.

Tập thứ nhất bao gồm câu trong miền những câu giao tiếp thông thường. Đây cũng chính là miền được chọn để xây dựng dữ liệu. Bởi đây là miền dữ liệu gần gũi và thông dụng nhất đối với người khiếm thính. Chúng cũng là miền dữ liệu hữu ích và có ý nghĩa cho bài toán dịch. Thêm nữa, các câu trong tập câu giao tiếp thường là câu đơn giản, tỉ lệ từ vựng thuộc tập từ điển nhiều hơn là các tập còn lại.

Tập kiểm tra này bao gồm các dữ liệu khác so với tập dữ liệu huấn luyện. Điểm BLEU của hệ thống dịch luật đối với miền này đạt điểm rất cao, vượt trội hơn so với điểm đánh giá các cặp song ngữ khác. Điểm BLEU tham chiếu với một số mô hình dịch của các cặp song ngữ Tiếng anh- Bồ Đào Nha với điểm BLEU cao nhất là 19 điểm; Tiếng Anh – Tiếng Pháp 22,52 điểm; Tiếng Anh – tiếng Tây Ban Nha 29 điểm [58]; Việt- Nhật là 23,7 điểm [59]; Tiếng Anh – Tiếng Việt là 45,47 điểm, Tiếng Việt – Tiếng Anh là 40,57 điểm [60].

Bảng 3.15. Điểm BLEU đánh giá trên tập kiểm tra dữ liệu miền các câu trong y học

Câu gốc	**Dịch máy**	**Dịch bởi chuyên gia VSL**	**Điểm BLEU**			
			1-gram	**2-gram**	**3-gram**	**4-gram**
Điều trị bệnh ***van động mạch chủ*** với kỹ thuật mổ tim mở ít ***xâm lấn***.	Điều trị bệnh ***van động mạch chủ*** với kỹ thuật mổ tim mở ít ***xâm lấn***.	Điều trị bệnh ***cửa mạch máu*** kỹ thuật mổ tim mở ít ***can thiệp bên trong***.	62.5	46.67	35.71	23.08
Trước khi thực hiện phẫu thuật, người bệnh sẽ được gây mê ***toàn thân*** giúp bạn ngủ và không có cảm giác đau.	Trước phẫu thuật, người bệnh sẽ được gây mê ***toàn thân*** giúp bạn ngủ và không có cảm giác đau .	Trước phẫu thuật, người bệnh gây mê ***cơ thể*** ngủ ***tự nhận biết*** đau không.	66.67	50	38.46	33.33
Bệnh tim có nguy cơ cao, ***suy tim*** rất nặng.	Bệnh tim nguy cơ cao, ***suy tim*** nặng.	Bệnh tim nguy cơ cao, ***giảm hoạt động tim*** nặng.	70	55.56	37.5	28.57
Bệnh nhân bị ***béo phì.***	Bệnh nhân ***béo phì.***	Bệnh nhân ***béo nhiều.***	75	66.67	50	0

Các câu thuộc miền y học được chọn lựa ngẫu nhiên trên một trang web về y tế của một bệnh viện nổi tiếng tại Việt Nam. Rất nhiều thuật ngữ y khoa chưa có trong tập từ điển VSL hiện tại. Bởi vậy, theo hệ thống dịch máy dựa trên luật xây dựng thì những từ này giữ nguyên và được biểu diễn dưới dạng đánh vần từng chữ

để người khiếm thính có thể hiểu. Tuy nhiên, khi các chuyên gia về ngôn ngữ ký hiệu dịch thủ công thì những từ này sẽ được chuyển đổi thành các từ đồng nghĩa hoặc gần nghĩa trong tập từ điển VSL hiện có để người khiếm thính có thể hiểu được dễ dàng hơn.

Các câu thuộc miền kỹ thuật được lựa chọn từ cuốn sách "Cẩm nang kỹ thuật cơ khí" của tác giả Nguyễn Văn Huyền (Nhà xuất bản Xây dựng -2010). Dữ liệu trong tập kiểm tra này bao gồm 100 câu lựa chọn ngẫu nhiên trong tài liệu. Trong đó bao gồm nhiều thuật ngữ kỹ thuật trong lĩnh vực cơ khí.

Tập thứ 4 với dữ liệu đầu vào từ văn bản trong tác phẩm văn học Việt Nam "Chiếc thuyền ngoài xa" của tác giả Nguyễn Minh Châu. Câu đích được dịch thủ công bởi chuyên gia ngôn ngữ ký hiệu. Đối với tập dữ liệu này, rất nhiều từ trong câu gốc tồn tại hoặc có đồng nghĩa trong VSL. Thêm nữa, sự phức tạp của câu gốc nếu được dịch ra VSL được giản lược rất nhiều.

Bảng 3.16. Điểm BLEU đánh giá trên tập kiểm tra dữ liệu miền các câu trong văn học

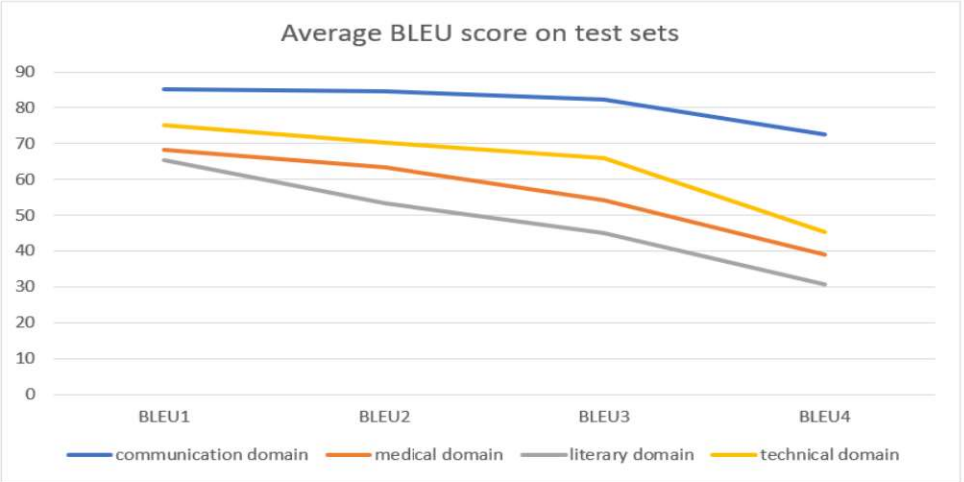
Câu gốc	**Dịch máy**	**Dịch bởi chuyên gia VSL**	**Điểm BLEU**			
			1-gram	**2-gram**	**3-gram**	**4-gram**
Lúc bấy giờ trời đầy ***mù*** từ ngoài biển bay vào.	Lúc bấy giờ trời đầy ***mù*** từ ngoài biển bay vào.	Bây giờ trời đầy ***mù sương*** biển bay vào.	88.88	75.00	57.14	33.33
Lại ***lác đác*** mấy hạt mưa.	Lại ***lác đác*** mấy hạt mưa.	Hạt mưa ít.	17.11	0	0	0
Mũi thuyền in một nét ***mơ hồ lòe nhòe*** vào bầu sương mù trắng như sữa có ***pha*** đôi chút màu hồng hồng do ánh mặt trời chiếu vào.	Mũi thuyền in một nét ***mơ hồ lòe nhòe*** vào bầu sương mù trắng như sữa có ***pha*** đôi chút màu hồng hồng do ánh mặt trời chiếu vào.	Mũi thuyền in hình sương mù trắng ***rõ không lẫn lộn*** màu hồng mặt trời chiếu vào.	49.60	39.52	28.11	15.06

Điểm BLEU đánh giá bản dịch trong dịch tự động Vie-VSL đối với các tập dữ liệu như trong bảng 3.17. So sánh điểm BLEU giữa các miền khác nhau trong hình 3.7. Với các giá trị BLEU1 là điểm BLEU trung bình của các câu có độ dài dưới 3 từ, BLEU2 là điểm BLEU trung bình của các câu có độ dài 4 từ, BLEU3 là điểm BLEU trung bình của các câu có độ dài 5 từ, BLEU4 là điểm BLEU trung bình của các câu có độ dài trên 5 từ.

Bảng 3.17. 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 y học	55.72
Data set 3: Miền các câu trong kỹ thuật	64.13
Data set 4: Miền các câu trong văn học	48.68
Trung bình	62.55

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 hai 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.



Hình 3.4. Thống kê điểm BLEU trung bình trên các tập kiểm tra.

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.

Các thí nghiệm tương tự trong một số nghiên cứu dịch máy ngôn ngữ ký hiệu khác trên thế giới cho thấy dịch dựa trên quy tắc dù cổ điển nhưng vẫn là một lựa chọn thích hợp cho lớp bài toán này. Kanis [61] trong nghiên cứu về dịch máy ngôn ngữ ký hiệu tiếng Séc ứng dụng thử nghiệm với tập huấn luyện 12.616 câu với dịch dựa trên luật đạt 81 điểm BLEU. Tương tự, trong nghiên cứu của Dimitrios Kouremenos và các cộng sự [62] trong trường hợp dịch ngôn ngữ ký hiệu Hy Lạp. Điểm BLEU 3-gram thu được là 85 với bộ dữ liệu huấn luyện và kiểm tra nhỏ bao gồm 900 câu và 109 luật.

Tuy nhiên, đường cơ sở được báo cáo với bộ công cụ mã nguồn mở để dịch máy thống kê Moses [63] trong dịch ngôn ngữ ký hiệu Đức là 18 điểm BLEU với tập huấn luyện gồm 2.565 câu và tập kiểm tra gồm 512 câu. Bằng cách kết hợp một số hệ thống, cuối cùng họ đã đạt được BLEU là 23,4. Ở đây cần lưu ý rằng sự khác biệt giữa các kết quả này là do sự khác biệt giữa những đặc điểm cú pháp tiếng Đức và các ngôn ngữ khác. Rõ ràng, sự ánh xạ của đơn vị từ vựng tiếng Séc và tiếng Việt hay tiếng Hy Lạp đến ngôn ngữ ký hiệu có nhiều điểm tương đồng và giống nhau hơn. Trong khi từ tiếng Đức tới ngôn ngữ ký hiệu Đức thì không như vậy.

Hơn nữa, kết quả xác nhận rằng sự khan hiếm dữ liệu và miền thưa thớt khiến các phương pháp tiếp cận dựa trên dữ liệu hoạt động kém hơn so với các hệ thống dựa trên quy tắc. Bởi vậy, không có một mẫu số chung nào cho bài toán dịch ngôn ngữ ký hiệu trên tiếp cận cấu trúc cụ thể là dịch dựa trên luật cho tất cả các ngôn ngữ ký hiệu trên thế giới, tuy nhiên phương pháp được chọn cho bài toán dịch ngôn ngữ ký hiệu Việt Nam đã đạt được kết quả tốt.

3.6. 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

Trong chương này, luận án đề xuất một phương pháp làm giàu dữ liệu đơn giản và hiệu quả cho bài toán. Ý tưởng đề xuất 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. Dữ liệu ban đầu được xây dựng bởi một phương pháp tự động một phần và được đánh giá thủ công đã được trình bày cụ thể ở chương 3. Cuối chương là những phân tích đánh giá quá trình thực nghiệm thuật toán làm giàu dữ liệu. Kết quả đạt được là 2 bộ dữ liệu Vie-VSL10k và Vie-VSL60k gồm các cặp câu song ngữ tiếng Việt – câu đúng cú pháp trong VSL để áp dụng cho các mô hình dịch thống kê và dựa trên mạng noron trong chương 5.

4.1. Giới thiệu chung về kỹ thuật làm giàu dữ liệu trong dịch máy

Kỹ thuật làm giàu dữ liệu trong dịch máy là quá trình tạo thêm các cặp câu dịch trong tập dữ liệu huấn luyện nhằm nâng cao hiệu suất và chất lượng của hệ thống dịch máy. Điều này có ý nghĩa quan trọng vì việc có một lượng dữ liệu đủ lớn và đa dạng là yếu tố quan trọng để đạt được kết quả dịch chính xác và tự nhiên.

Có nhiều phương pháp và kỹ thuật để làm giàu dữ liệu trong dịch máy. Dưới đây là một số phương pháp phổ biến:

- Dịch ngược (Back-translation): Sử dụng mô hình dịch máy để dịch các câu từ ngôn ngữ nguồn sang ngôn ngữ đích. Sau đó, dùng mô hình dịch ngược để dịch lại các câu đã dịch sang ngôn ngữ đích về ngôn ngữ nguồn. Quá trình này tạo ra thêm các cặp câu dịch mới cho huấn luyện [64] [65].
- Đồng ngữ (Monolingual data): Sử dụng dữ liệu ngôn ngữ đích không có cặp câu dịch nguồn tương ứng. Một phương pháp phổ biến là sử dụng mô hình dịch máy để tạo ra các dự đoán dịch cho các câu trong ngôn ngữ đích. Các cặp câu gốc và dự đoán dịch sau đó được sử dụng để làm giàu dữ liệu [66] .
- Thay đổi từ vựng (Lexical substitution): Thay đổi một số từ trong câu bằng các từ tương đương hoặc từ khác có cùng ý nghĩa. Việc này giúp tạo ra các biến thể câu với ngữ cảnh và từ vựng khác nhau, mở rộng phạm vi của dữ liệu huấn luyện [67].

- Tổng hợp từ (Word synthesis): Tạo ra các từ mới bằng cách kết hợp các từ có sẵn trong từ điển hoặc sử dụng phương pháp tổng hợp từ dựa trên mô hình ngôn ngữ. Các từ mới này được sử dụng để tạo ra các câu mới trong quá trình làm giàu dữ liệu [68].
- Ngoài ra, còn có một số phương pháp tổng hợp áp dụng kỹ thuật mới dùng để cải tiến dữ liệu cho ngôn ngữ ít tài nguyên trong một số nghiên cứu mới đây [69].

Quá trình làm giàu dữ liệu giúp mở rộng tập dữ liệu huấn luyện và đa dạng hóa ngữ cảnh, từ vựng, cấu trúc câu. Điều này có thể cải thiện khả năng dịch của hệ thống dịch máy. Đối với bài toán dịch ngôn ngữ ký hiệu – được coi là một trong số các ngôn ngữ ít tài nguyên, việc xây dựng các bộ dữ liệu thực nghiệm và các kỹ thuật làm giàu dữ liệu là một trong những công việc được quan tâm nhất.

Các bộ dữ liệu công bố cho cộng đồng nghiên cứu trong lĩnh vực dịch máy chủ đề dịch ngôn ngữ ký hiệu sử dụng hiện chỉ có ở một vài ngôn ngữ như tiếng Anh, tiếng Đức. Các bộ dữ liệu sử dụng cho các nghiên cứu ngôn ngữ ký hiệu khác không có sẵn hoặc là dữ liệu nhỏ [70]. Một số thông số các bộ dữ liệu cho dịch ngôn ngữ ký hiệu được liệt kê trong bảng 4.1.

Bảng 4.1. Liệt kê một số bộ dữ liệu trong các nghiên cứu của lĩnh vực dịch máy chủ đề dịch ngôn ngữ ký hiệu

Dữ liệu	Ngôn ngữ	Cấp độ	Năm
ASLLVD [71]	ASL	Từ vựng	2008
ATIS Corpus [72]	Đa ngôn ngữ	Câu hoàn chỉnh	2008
Dicta-Sign [73]	ASL	Từ vựng	2012
ASL-LEX [74]	ASL	Từ vựng	2016
RWTH-Phoenix-2014T [75]	DGS	Câu hoàn chỉnh	2018
KETi [76]	KSL	Câu hoàn chỉnh	2019
How2Sign [77]	ASL	Câu hoàn chỉnh	2021
OpenSubtitles [78]	Đa ngôn ngữ	Câu hoàn chỉnh	2016
Multi30K [79]	BSL, DGS	Câu hoàn chỉnh	2016
ASPEC [80]	JSL	Câu hoàn chỉnh	2016
MUSE [81], [82]	Đa ngôn ngữ	Từ vựng	2017
MTNT [83]	JSL, LSP	Câu hoàn chỉnh	2018

Ta thấy rằng đối với ngôn ngữ ký hiệu Việt Nam, hiện chưa có một cơ sở dữ liệu nào có thể truy cập công khai sử dụng cho mục đích là dữ liệu cơ sở cho các nghiên cứu dịch tự động VSL. Do vậy, nghiên cứu sinh đề xuất một phương pháp làm giàu dữ liệu dựa trên bộ dữ liệu cơ sở được xây dựng trong phần chương 3 của luận án này. Cơ sở đề xuất của phương pháp và các thực nghiệm được trình bày ở phần tiếp theo.

4.2. Cơ sở của phương pháp đề xuất

Ở chương 3, luận án đã trình bày một số kết quả đạt được nhất định với các phương pháp và mô hình dịch trên một tập dữ liệu đã xây dựng. Vấn đề đặt ra là với xu hướng dịch máy phát triển cùng với các mô hình dịch hiện đại tiên tiến cần được đánh giá và thử nghiệm trên một tập dữ liệu đủ lớn. Song song với việc xây dựng và đánh giá thủ công bởi một số chuyên gia về ngôn ngữ thì việc sinh ra tập dữ liệu tự động đủ lớn là một việc quan trọng và cần thiết. Ý tưởng phần này chính là làm giàu dữ liệu dựa trên hệ thống WordNet.

Mạng từ tiếng Anh (WordNet) là một bộ dữ liệu ngữ nghĩa ở mức từ vựng, thể hiện quan hệ về nghĩa giữa các từ với nhau. WordNet bao gồm ba bộ dữ liệu riêng biệt, một bộ của danh từ, một bộ của động từ, một bộ của tính từ và trạng từ. Mạng WordNet được tổ chức theo mô hình cây như mô tả ở hình 4.1, mỗi node chứa một từ nguyên mẫu (lemma) cùng với tập các từ đồng nghĩa với nó (synset). Mạng WordNet chỉ thể hiện quan hệ về ngữ nghĩa chứ không thể hiện quan hệ về ngữ âm hay hình thái [84]. Tính đến phiên bản 3.0, bộ dữ liệu WordNet dành cho tiếng Anh đã có khoảng 117000 danh từ, 11400 động từ, 22000 tính từ và 4600 trạng từ.

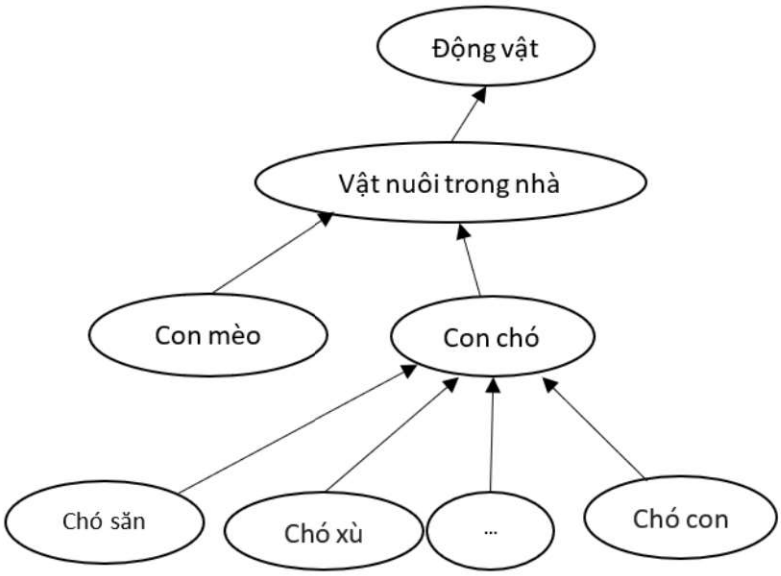
Mạng từ tiếng Anh là một bộ dữ liệu ngữ nghĩa quan trọng trong xử lý ngôn ngữ tự nhiên. Nó cung cấp một mô hình ngữ nghĩa của tiếng Anh, bao gồm các mối quan hệ giữa các từ. Một trong những mối quan hệ quan trọng trong WordNet là quan hệ thượng danh và hạ danh. Quan hệ thượng danh và hạ danh là một mối quan hệ giữa hai từ, trong đó một từ là từ thượng danh và từ còn lại là từ hạ danh. Từ thượng danh là từ chỉ một khái niệm rộng hơn, bao hàm khái niệm của từ hạ danh. Ví dụ, từ "xe cơ giới" là từ thượng danh của từ "xe ô tô", vì "xe cơ giới" là khái niệm rộng hơn bao hàm cả khái niệm "xe ô tô".

Hình 4.1. Cấu trúc phân cấp trong WordNet

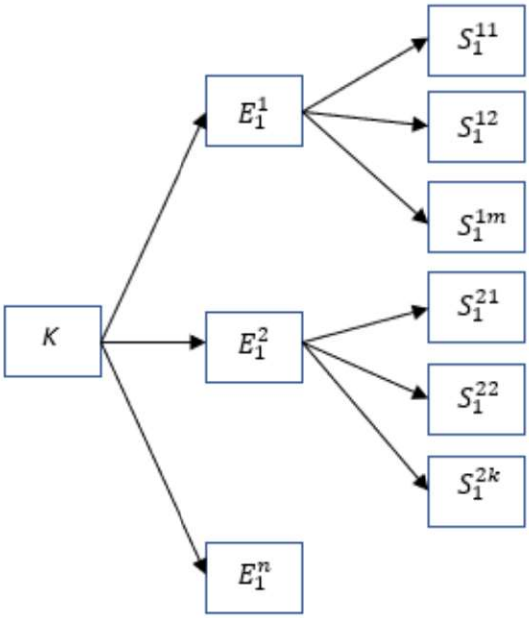
Quan hệ thượng danh và hạ danh có thể được sử dụng để giải quyết nhiều bài toán liên quan đến tiếng Anh, chẳng hạn như: tìm từ đồng nghĩa, trái nghĩa, phân loại từ. Xuất phát từ đó, ý tưởng sử dụng cấu trúc thượng danh và hạ danh của tiếng Anh có thể được áp dụng để giải quyết các bài toán liên quan đến tiếng Việt. Cụ thể, trong bài toán làm giàu dữ liệu cho dịch Vie-VSL có thể sử dụng cấu trúc này để tìm các từ đồng nghĩa và gần nghĩa để thay thế cho 1 từ trong câu tạo ra câu 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.

Việc phân tích cú pháp văn phạm có thể xét được tính đúng đắn về văn phạm nhưng lại không thể kiểm tra chính xác sự đúng đắn về ngữ nghĩa. Ví dụ ta xét một câu "*cái tủ ăn thịt gà*", nếu xét các thành phần chủ ngữ, vị ngữ, động từ trong câu thì không sai về mặt cú pháp. Nhưng xét về mặt ngữ nghĩa thì câu này thiếu tính hợp lý về logic. Nếu câu trên thay bằng câu: "*con chó ăn thịt gà*" thì về mặt ngữ nghĩa sẽ là phù hợp hơn. Vậy làm sao để xét được một câu có thể hợp lý và biết được "*cái tủ*" hay "*con chó*" có thể ăn được "*thịt gà*". Vấn đề này có thể giải quyết được bằng cách sử dụng quan hệ hạ danh-thượng danh trong WordNet. Giả sử ra có một heuristic là chỉ có "*động vật*" mới có thể thực hiện động từ "*ăn*". Như vậy, để kiểm tra một vật có biết ăn hay không ta sẽ kiểm tra xem nó có phải "*động vật*" hay không bằng cách duyệt các thượng danh của nó. Bằng cách duyệt ngược về các thượng danh, ta dễ dàng kiểm tra được là "*con chó*" có thể thực hiện

hành động "*ăn*" còn "*cái tủ*" thì không thể. Tương tự, ta có thể thêm các ràng buộc về ngữ nghĩa để kiểm tra tính đúng đắn về ngữ nghĩa trong câu. Từ đó có thể sinh ra câu mới bằng cách thay thế các từ có cùng thượng danh.



Hình 4.2. Cấu trúc thượng danh và hạ danh đối với từ khoá "con chó".



Hình 4.3. Minh hoạ các tiêu chuẩn với tập Synset E_i^j.

Ở đây sử dụng 3 tiêu chuẩn:

Tiêu chuẩn anh em: áp dụng khi các tập Synset S_i^j đều có các synset là anh em với nhau (có cùng synset cha (hypernym)). Khi đó synset $\{E_1^1, E_1^2, \dots\}$ được chọn là các synset anh em này. Với *SV* là tập hợp các synset được chọn là các synset anh em dựa trên tiêu chuẩn này, ta có:

$$SV = \{S_i^{jk} / S_i^{jk} \in S_i^j (\forall j: 0 \le j \le n_i^j) : \exists S_p : (S_P \; is\_hyper \; S_i^{jk})\} \qquad (4.1)$$

Tiêu chuẩn cha con: áp dụng khi trong các tập synset S_i^j có một synset là cấp trên của các synset còn lại (chỉ cần mỗi tập synset còn lại có một synset là cấp dưới của synset cấp trên nói trên). Khi đó synset $\{E_1^1, E_1^2, \dots\}$ được chọn là các synset anh em này. Với SV là tập hợp các synset được chọn là các synset cha con dựa trên tiêu chuẩn này, ta có:

$$SV = \{S_i^{jk} / \exists S_p \in S_i^h \ (h \in [1 \dots n_i^j])\ S_i^{jk} \in S_i^h (\forall j: 0 \le j \le n_i^j, \quad j \ne h): (S_P\ is\_hyper\ S_i^{jk})\} \tag{4.2}$$

Tiêu chuẩn ông cháu: áp dụng khi trong các tập synset S_i^j có một synset là cấp trên của các synset còn lại (chỉ cần mỗi tập synset còn lại có một synset là cấp dưới của synset cấp trên nói trên). Khi đó synset $\{E_1^1, E_1^2, \dots\}$ được chọn là các synset cấp dưới này. Với SV là tập hợp các synset được chọn là các synset ông cháu dựa trên tiêu chuẩn này, ta có:

$$SV = \{S_i^{jk} / \exists S_g \in S_i^h \ (h \in [1 \dots n_i^j]), S_i^{jk} \in S_i^j (\forall j: 0 \le j \le n_i^j, j \ne h)(S_g\ is\_dist\_hyper\ S_i^{jk})\} \tag{4.3}$$

Trong đó:

- A *is_hyper B* có ý nghĩa A là thượng danh B còn B là hạ danh của A.

- *A is_dist_hyper B* có ý nghĩa A là thượng danh của một từ mà từ đó là thượng danh của B.

Quan hệ thượng danh trong mạng từ là một mối quan hệ giữa hai từ, trong đó một từ là từ thượng danh và từ còn lại là từ hạ danh. Từ thượng danh là từ chỉ một khái niệm rộng hơn, bao hàm khái niệm của từ hạ danh.

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.

Bởi vậy, dựa vào cấu trúc thượng danh và hạ danh cùng một số đặc điểm khác của WordNet, ta có thể sinh dữ liệu mời bằng cách thay thế từ trong câu cũ.

4.3. Quy trình làm giàu dữ liệu

Dựa trên những tính chất và đặc điểm của WordNet đối các ràng buộc về ngữ nghĩa để kiểm tra tính đúng đắn về ngữ nghĩa trong câu, ta có quy trình xây dựng dữ liệu mới từ một câu S ban đầu.

Đầu tiên là quá trình tiến hành tách từ. Sau đó tìm kiếm các từ có trong cơ sở dữ liệu mạng từ Tiếng Việt, chọn danh từ X trong câu vừa tách, tìm thượng danh W của nó trong WordNet Tiếng Việt. Bước tiếp theo là tìm tập hợp X_i là các từ hạ

danh của W. Có thể tiếp tục tìm các từ Y_i là hạ danh X_i và lặp lại bước này nếu các từ tìm được tiếp tục tồn tại hạ danh. Cuối cùng thay thế X bằng các từ hạ danh tìm được ở các bước trên để sinh ra câu mới S'.

Tài nguyên dùng trong quy trình này là bộ WordNet tiếng Việt bộ dữ liệu WordNet tiếng việt của cộng đồng Xử lí Ngôn ngữ và Tiếng nói tiếng Việt VLSP (Association for Vietnamese Language and Speech Processing). Bộ dữ liệu này bao gồm 10.000 đơn vị từ vựng chính. Mỗi đơn vị này bao gồm thông tin về các thuộc tính của nó như từ dịch nghĩa tiếng Anh, từ đồng nghĩa, từ trái nghĩa trong tiếng Việt, cấu trúc thượng danh và hạ danh. Cấu trúc lưu trữ của một đơn vị từ vựng được mô tả trong WordNet Tiếng Việt như sau:

Ví dụ 1: Cấu trúc danh sách từ "*chó nhà*"

{"antonyms": "", "attributes": "", "causes": "", "danh sach tu": "'chó nhà', 'chó nuôi'", "entailments": "", "hypernym": "hunting_dog.n.01", "hyponyms": "", "index": "3957", "instance_hypernyms": "", "instance_hyponym": "", "member_holonyms": "pack.n.06", "member_meronyms": "", "nghia": "'giống chó được dùng nuôi trong nhà'", "part_holonyms": "", "part_meronyms": "", "region_domain": "", "similar_tos": "", "substance_holonyms": "", "substance_meronyms": "", "ten synset": "hound.n.01", "the loai": "noun.animal", "topic_domain": "", "tu loai": "n", "usage_domain": "", "verb_groups": "", "vi du": """}

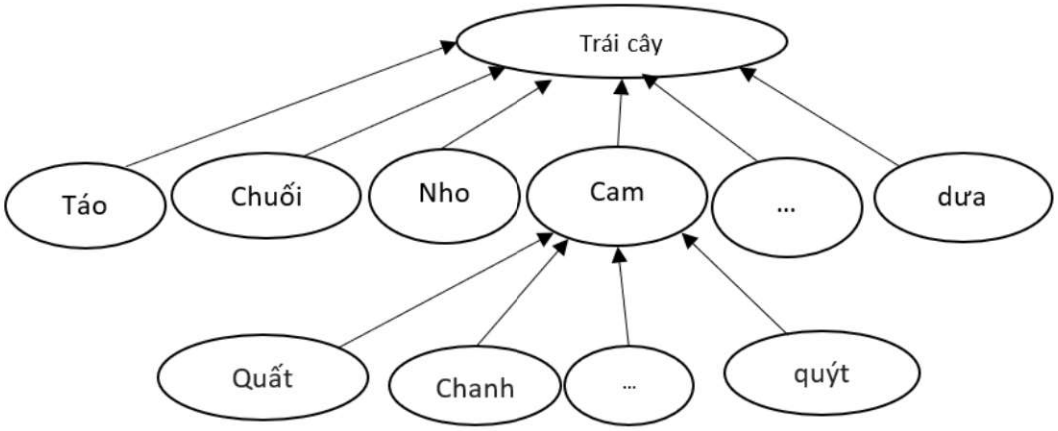
Trong một cấu trúc cụ thể như trên, ta thấy từ "*chó nhà*" đồng nghĩa với từ "*chó nuôi*" với một số đặc điểm quan trọng được sử dụng như: từ loại là danh từ, thượng danh của nó là "*hunting_dog.n.01",* danh sách các từ hạ danh là trống.

Ngoài ra, còn một số nhóm từ vựng được xây dựng thủ công bổ sung thêm trong bộ dữ liệu WordNet Tiếng Việt với cũng cấu trúc để phục vụ cho bài toán.

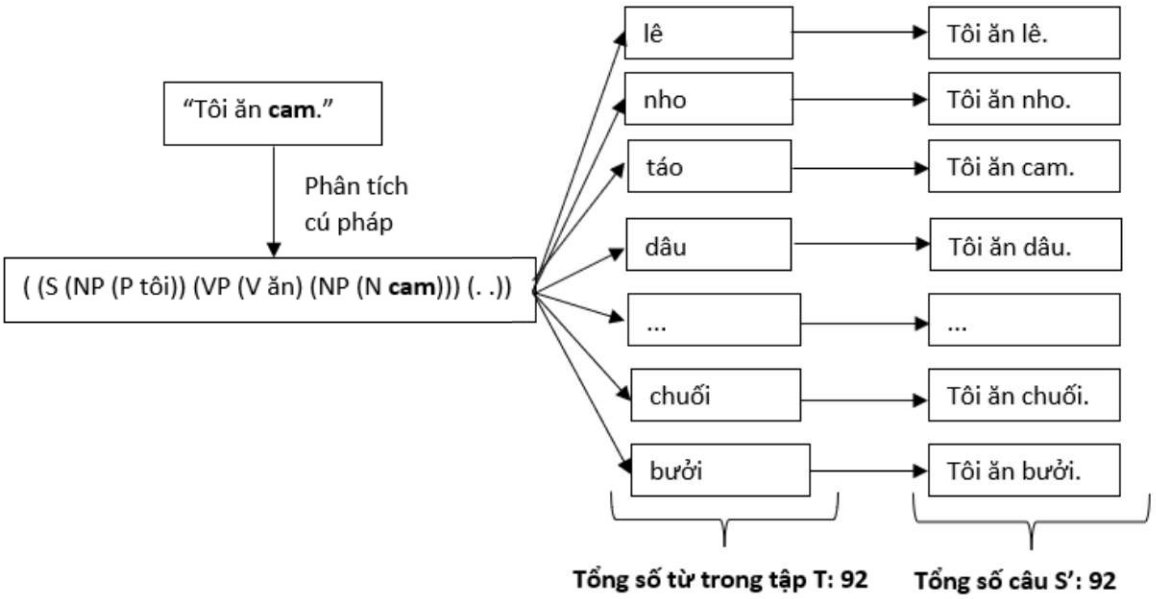
Ví dụ 2: cấu trúc của từ "cam" được lưu trong WordNet Tiếng Việt, với từ loại danh từ, thượng danh là trái cây, hạ danh bao gồm: *citrange.n.02 - cam lai, citron.n.01- thanh yên, grapefruit.n.02- bưởi chùm, kumquat.n.02 - quất, lemon.n.01- chanh vàng, lime.n.06 - chanh xanh, pomelo.n.02 - bưởi*

{"antonyms": "", "attributes": "", "causes": "", "danh sach tu": " ‘cam’, 'quả cam’, 'trái cam’ ", "entailments": "", "hypernym": "fruit_tree.n.01", "hyponyms": " citrange.n.02, citron.n.01, grapefruit.n.01, kumquat.n.01, lemon.n.01, pomelo.n.01, grapefruit.n.02, lime.n.06", "index": "7304", "instance_hypernyms": "", "instance_hyponym": "", "member_holonyms": "", "member_meronyms": "", "nghia": "'quả trong số nhiều loại quả thuộc chi Cam chanh có vỏ dày và cùi mọng'; 'được trồng ở những vùng ấm áp'", "part_holonyms": "", "part_meronyms": "citrus.n.01", "region_domain": "", "similar_tos": "", "substance_holonyms": "", "substance_meronyms": "", "ten synset": "citrus.n.02", "the loai": "noun.plant", "topic_domain": "", "tu loai": "n", "usage_domain": "", "verb_groups": "", "vi du": """}

Với ví dụ trong câu “*Tôi ăn cam”*, ta tìm được thượng danh của “*cam*” là “*trái cây* ”. Sau đó ta tìm ra các loại trái cây có cùng thượng danh với “*cam”* (tiêu chuẩn anh em) hoặc các từ là hạ danh của từ cùng cấp với “*cam”* (tiêu chuẩn cha con). Từ đó có thể sinh ra các câu đảm bảo được về ngữ nghĩa như: “ *tôi ăn chuối*” , “*tôi ăn táo*”, v.v… Tập T những từ có thể thay thế được táo có nút gốc là “*trái cây*” bao gồm 92 từ có thể thay thế. Như vậy từ một câu sinh ra 92 câu mới.



Hình 4.4. Cấu trúc thượng danh đối với từ khoá “cam”.



Hình 4.5. Ví dụ về xây dựng tập T và sinh dữ liệu mới.

Tiếp theo là tiến hành thực nghiệm phương pháp trên một số dữ liệu được xây dựng ban đầu và làm giàu dữ liệu bằng phương pháp đã đề xuất.

Mô tả các bước của thuật toán

Đầu vào: câu S.

Đầu ra: tập hợp các câu S' là các câu sinh ra từ câu gốc S. Câu S' đảm bảo điều kiện là tương đồng về mặt cú pháp với câu S.

- Bước 1: Phân tách các từ W trong câu S. Việc thực hiện phân tách các từ này sử dụng công cụ VietWS là công cụ được công bố và sử dụng trong cộng đồng xử lí văn bản và tiếng nói tiếng Việt (VLSP).
- Bước 2: Tìm kiếm tập X bao gồm n thượng danh của W
- Bước 3: Xét tất cả tập X, tìm Xi là hạ danh của X và thêm vào tập T
- Bước 4: Lặp lại việc tìm kiếm hạ danh của từng phần tử trong T cho đến khi còn tồn tại hạ danh, sau đó thêm các phần tử này vào tập T
- Bước 5: Thay thế W trong S bởi từng phần tử trong tập T thu được câu S'
- Bước 6: Trả về kết quả là tập các câu S' thu được.

Thuật toán làm giàu dữ liệu được mô tả bởi mã giả như sau:

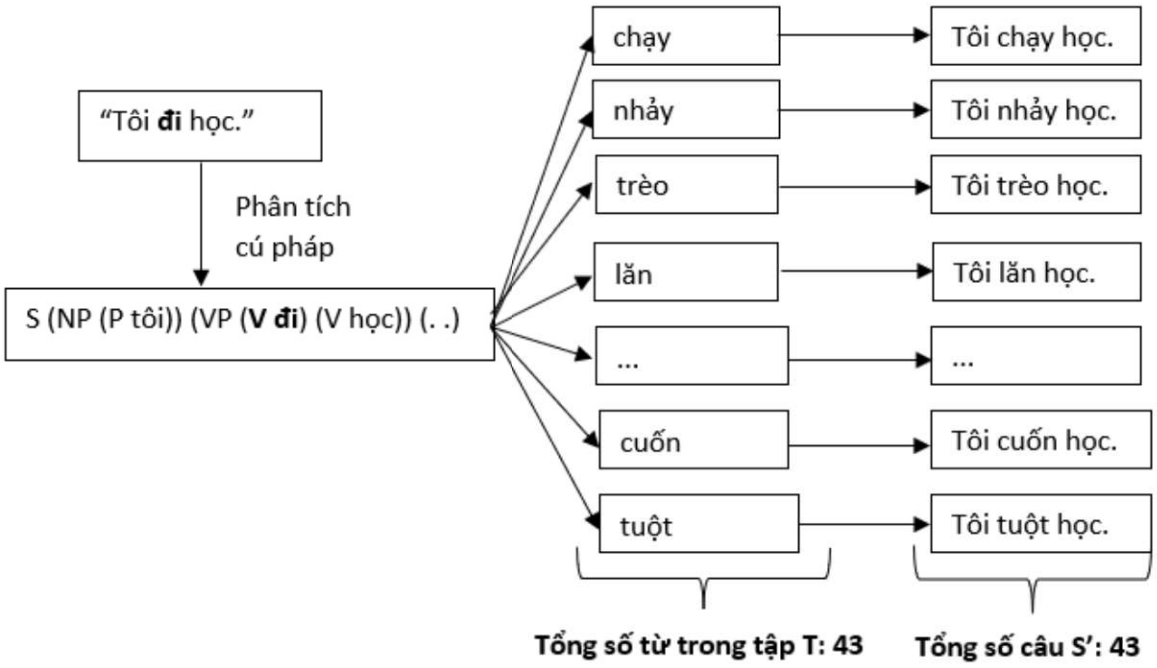
```
Algorithm: Data-Augment-VSL
Input: Sentences S
Output: Set of sentences S' are generated based on S.
1: Split W word ∈ S
2: X ⇐W. hypernyms()
   n = len(X);
3: For i=1,n do
       Xi⇐X.hyponyms()
       Add Xi to set T
4:  While !∃ Xi.hyponyms:
      Yi ⇐ Xi.hyponyms()
      Add Yi to set T
5: S'= Replace(W,Ti);
6: Return Set of sentences S'
```

Thuật toán thực hiện việc phân tách các từ trong câu S và thay thế bằng các từ có cùng tính chất và các tiêu chuẩn đã xét ở trên để đảm bảo về mặt ngữ nghĩa cho câu mới sinh.

4.4. 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.



Hình 4.6. Ví dụ về xây dựng tập T và sinh dữ liệu không phù hợp với động từ.

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 4.2. 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 4.2. 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

Từ loại	Nhóm	Ví dụ	T	W_S	W'_S
Danh từ	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, hình vuông,…	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. Kho ngữ liệu làm giàu công bố tại: https://github.com/BichDiep/VSL-DATA-AUGMENTATION.

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 ta thấy rằng, với 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.

Dữ liệu đã được làm giàu để huấn luyện các mô hình và thử nghiệm trên các tập test. Điểm BLEU cũng là một tiêu chí để so sánh hiệu quả các mô hình dịch. Ngoài ra chúng cũng dùng để so sánh giữa mô hình với dữ liệu gốc và mô hình với dữ liệu làm giàu. Phần đánh giá kết quả của việc làm giàu dữ liệu thông qua các mô hình dịch sẽ được trình bày trong chương 5.

4.5. 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

Trong chương này, luận án đề xuất một số mô hình thống kê và những cải tiến áp dụng cho bài toán dịch. 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 chương là các phân tích và đánh giá các bộ dữ liệu với các mô hình dịch đề xuất.

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

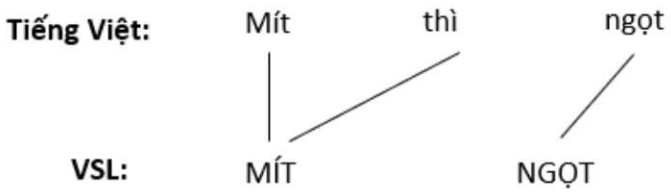
Trong phần này, luận án trình bày một mô hình đơn giản cho dịch máy ngôn ngữ ký hiệu dựa trên dịch từ vựng, dịch từ. Phương pháp này yêu cầu một từ điển ánh xạ các từ từ ngôn ngữ nguồn sang ngôn ngữ đích. Trong bài toán dịch Vie-VSL, từ điển ánh xạ này đơn giản hơn rất nhiều các bài toán dịch giữa các 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.

$$P_f = e \rightarrow P_f(e) \tag{5.1}$$

Trong đó: P_f là xác suất dịch từ vựng cho một lựa chọn bản dịch cụ thể. $P_f(e)$ là một giá trị liên quan đến độ tin cậy của bản dịch. Hàm số này sử dụng hàm mũ cơ số e để ước tính xác suất dịch từ vựng dựa trên giá trị $P_f(e)$. Khi $P_f(e)$ tăng lên, giá trị của P_f cũng tăng lên, và khi $P_f(e)$giảm, P_f cũng giảm theo cách phi tuyến tính.

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:

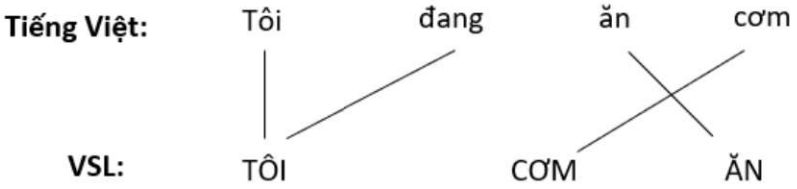


$$W: (1 \rightarrow 1; \quad 1 \rightarrow 2; \quad 2 \rightarrow 3)$$

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:



$$W: (1 \rightarrow 1; \quad 1 \rightarrow 2; \quad 2 \rightarrow 4; \quad 3 \rightarrow 3)$$

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_evới sự liên kết của từng từ VSL e_j sang một từ tiếng Việt từ f_i theo hàm căn chỉnh w: $(j \rightarrow i)$ như sau:

$$p(e, w|f) = \frac{\varepsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{w(j)}) \qquad (5.2)$$

Trong đó:

- $p(e, w|f)$: Đây là xác suất dịch một câu tiếng Việt e sang một câu VSL f với một bộ tham số w. Công thức này mô tả mối quan hệ giữa câu tiếng Việt và VSL thông qua bộ tham số w.
- ε: Đây là hằng số dương, được sử dụng để điều chỉnh tổng xác suất thành giá trị dự đoán. ε có giá trị nhỏ.
- l_f: Độ dài của câu tiếng Việt f, đo bằng số từ trong câu.
- l_f: Độ dài của câu VSL e, đo bằng số từ trong câu.
- $t(e_j|f_{w(j)})$: là một thành phần quan trọng trong công thức biểu thị xác suất dịch từ VSL e_j sang từ tiếng Việt $f_{w(j)}$, với $w(j)$ là một chỉ số của từ tiếng Việt trong câu f. Trong mô hình IBM Model 1, xác suất này được biểu diễn dựa trên mối quan hệ giữa từ VSL và từ tiếng Việt trong dữ liệu huấn luyện.
- $\prod_{j=1}^{l_e} t(e_j|f_{w(j)})$: Là tích của xác suất dịch từng từ VSL e_j sang từ tiếng Việt $f_{w(j)}$, với j chạy từ 1 đến l_e Công thức này tính toán xác suất tổng hợp cho toàn bộ câu VSL *e* dựa trên ánh xạ *w* và xác suất dịch từng từ riêng lẻ.

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*". Bảng 5.1. trình bày một số lần lặp với các xác suất dịch các từ tiếng Việt sang dạng văn bản VSL là 1/3=0.33.

Bảng 5.1. Một số lần lặp với các xác suất dịch
các từ tiếng Việt sang dạng văn bản VSL với mô hình IBM 1

e	f	Ban đầu	Lần 1	Lần 2	Lần 3	…	Lần 10
TÔI	tôi	0.33	0.52	0.66	0.75	…	1.00
TÔI	ăn	0.33	0.40	0.35	0.21	…	0.00
TÔI	cơm	0.33	0.40	0.35	0.21	…	0.00
ĂN	tôi	0.33	0.25	0.12	0.04	…	0.00
ĂN	ăn	0.33	0.38	0.40	0.42	…	0.50
ĂN	cơm	0.33	0.38	0.40	0.42	…	0.50
CƠM	tôi	0.33	0.25	0.12	0.04	…	0.00
CƠM	ăn	0.33	0.38	0.4	0.42	…	0.50
CƠM	cơm	0.33	0.60	0.72	0.77	…	1.00

Trong mô hình IBM 1 không có mô hình xác suất cho khía cạnh dịch thuật này. Kết quả là, theo mô hình IBM 1, xác suất dịch cho hai ví dụ được trích dẫn trước đó là như nhau. Mô hình IBM 2 giải quyết vấn đề căn chỉnh bằng một mô hình rõ ràng để căn chỉnh dựa trên vị trí của các từ đầu vào và đầu ra. Bản dịch của một từ đầu vào tiếng Việt ở vị trí i sang một từ VSL ở vị trí j được mô hình hóa bằng phân phối xác suất:

$$w(i|j, l_e, l_f) \tag{5.3}$$

Có thể xem dịch theo mô hình IBM 2 như một quy trình gồm hai bước với bước dịch từ vựng và bước căn chỉnh. Bước đầu tiên là dịch từ vựng như trong mô hình IBM 1, một lần nữa được mô hình hóa bằng xác suất dịch $t(e|f)$). Bước thứ hai là bước căn chỉnh. Chẳng hạn, dịch 'ăn' thành 'ĂN' có xác suất dịch từ vựng là $t(\text{ĂN}|\text{ăn})$) và xác suất căn chỉnh của $w(2, |4,4,3))$- từ VSL thứ 2 được căn chỉnh với từ tiếng Việt thứ 4. Lưu ý rằng chức năng căn chỉnh w ánh xạ từng từ đầu ra VSL j sang vị trí đầu vào tiếng Việt $w(i)$ và phân bố xác suất căn chỉnh cũng được thiết lập theo hướng ngược lại này. Hai bước này được kết hợp để tạo thành mô hình IBM 2.

$$p(e, w|f) = \varepsilon \prod_{j=1}^{l_e} t(e_j|f_{w(j)}) w(w(j)|j, l_e, l_f) \tag{5.4}$$

Trong mô hình IBM 3 tính đến mã thông báo NULL. Nói cách khác là có thể nhận được một từ bằng tiếng Việt không được dịch sang VSL. Xác suất tạo mã thông báo NULL (∅) là:

$$p(\emptyset_e) = \binom{l_e - \emptyset_0}{\emptyset_0} p_1^{\emptyset_0} p_2^{l_e - 2\emptyset_0} \tag{5.5}$$

Bảng 5.2. Một số lần lặp với các xác suất dịch
các từ tiếng Việt sang dạng văn bản VSL với mô hình IBM 3

e	f	Ban đầu	Lần 1	Lần 2	Lần 3
TÔI	tôi	0.75	0.82	0.96	1.00
TÔI	ăn	0.21	0.18	0.05	0.00
TÔI	cơm	0.21	0.12	0.02	0.00
ĂN	tôi	0.04	0.01	0.00	0.00
ĂN	ăn	0.42	0.45	0.48	0.50
ĂN	cơm	0.42	0.45	0.48	0.50

e	f	Ban đầu	Lần 1	Lần 2	Lần 3
CƠM	tôi	0.04	0.02	0.00	0.00
CƠM	ăn	0.42	0.45	0.48	0.50
CƠM	cơm	0.77	0.83	0.90	1.00

Một trong những vấn đề điển hình của học máy là khi ước tính mô hình từ dữ liệu không đầy đủ. Vì vậy, luận án này dụng thuật toán tối ưu hoá EM (Expectation-Maximization) để giải quyết tình trạng này. Đây là một phương pháp học lặp đi lặp lại nhằm lấp đầy các khoảng trống trong dữ liệu và đào tạo một mô hình theo các bước xen kẽ. Vì vậy ở đây áp dụng EM cho mô hình IBM 1, 2 và 3.

Các từ trong VSL hầu hết là đồng nhất với văn bản viết bằng tiếng Việt. Vì vậy, sử dụng kỹ thuật so khớp chuỗi (String Matching) để học dữ liệu sẽ nhanh chóng, hiệu quả và phù hợp với bài toán này. So khớp chuỗi bao gồm việc tìm một hoặc tổng quát hơn là tất cả các lần xuất hiện của một chuỗi trong một văn bản với một chuỗi khác. Một mẫu được ký hiệu là $x = x[0 \dots m-1)$ chiều dài của x bằng m. Một câu văn bản được ký hiệu là $y = y[0 \dots n-1)$ chiều dài của y bằng n. Cả hai chuỗi được xây dựng trên một bộ ký tự hữu hạn được biểu thị bằng bảng chữ cái với kích thước bằng nhau. Một số thuật toán và phương pháp tồn tại như khoảng cách Jaro-Winkler sẽ được sử dụng trong quá trình căn chỉnh từ trong dịch máy ngôn ngữ ký hiệu thống kê.

Khoảng cách Jaro–Winkler [50] là thước đo mức độ giống nhau giữa hai chuỗi. Nó là một biến thể của thước đo khoảng cách Jaro và chủ yếu được sử dụng trong lĩnh vực liên kết bản ghi. Khoảng cách Jaro–Winkler cho hai chuỗi càng cao thì các chuỗi càng giống nhau. Nó được thiết kế và phù hợp nhất cho các chuỗi ngắn chẳng hạn như tên người. Điểm số được chuẩn hóa sao cho 0 tương đương với không tương đồng và 1 là khớp chính xác. Khoảng cách Jaro d_j của hai chuỗi đã cho S_1 và S_2 là:

$$d_j = \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) \tag{5.6}$$

Trong đó:

- d_j là khoảng cách Jaro–Winkler giữa hai chuỗi $S1$ và $S2$.

- m là số ký tự trong mẫu x mà có thể tìm thấy một ký tự tương ứng trong câu văn bản y, bất kể chúng được sắp xếp theo thứ tự nào.

- t là số lần chuyển đổi vị trí

- $|S_1|$ là độ dài (số ký tự) của chuỗi $S1$.

- $|S_2|$ là độ dài (số ký tự) của chuỗi $S2$.

Khoảng cách Jaro–Winkler là một giá trị nằm trong khoảng từ 0 đến 1. Điểm số 0 tương đương với hai chuỗi không tương đồng (không giống nhau), và điểm số 1 tương đương với việc hai chuỗi khớp hoàn toàn (giống nhau). Công thức này nhấn mạnh sự tương đồng giữa các ký tự phù hợp, độ dài của các chuỗi và sự sắp xếp sai lệch thông qua tham số t.

Khoảng cách Jaro–Winkler sử dụng thang đo p mang lại thứ tự thuận lợi hơn cho các chuỗi khớp ngay từ đầu với độ dài tiền tố đã thiết lập là l. Khoảng cách Jaro–Winkler của hai chuỗi S_1 và S_2 là:

$$d_w = d_j + l.p.(1 - d_j) \qquad (5.7)$$

Trong đó:

- d_j là khoảng cách Jaro giữa S_1 và S_2.

- l là độ dài của tiền tố chung ở đầu chuỗi tối đa 4 ký tự.

- p là hệ số tỷ lệ không đổi cho số điểm được điều chỉnh tăng lên khi có các tiền tố chung. p không được vượt quá 0.25, nếu không khoảng cách có thể lớn hơn 1. Tiêu chuẩn giá trị của hằng số này trong nghiên cứu của Winkler là $p = 0.1$.

Bảng 5.3. Khoảng cách Jaro và khoảng cách Jaro-Winkeler

S1	**S2**	**Khoảng cách Jaro**	**Khoảng cách Jaro-Winkeler**
TÔI	tôi	1.00	1.00
TÔI	ăn	0.00	0.00
ĂN	cơm	0.42	0.48
ĂN	ăn	1.00	1.00
CƠM	cơm	1.00	1.00

Tối ưu hoá EM trong mô hình IBM-1

Từ công thức:

$$p(w|e,f) = \frac{p(e,w|f)}{p(e|f)} \tag{5.8}$$

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

$$p(w|e,f) = \frac{\alpha.p(e,w|f) + (1-\alpha).d_w(e,f)}{p(e|f)} \tag{5.9}$$

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 với thử nghiệm bộ kiểm tra một kho dữ liệu nhỏ. Lưu ý rằng kho văn bản chứa hai từ có giá trị tương tự nhưng không có cùng ngữ nghĩa và vai trò *"ăn"* và *"cơm"*. Bảng tiếp theo trình bày kết quả thí nghiệm.

Bảng 5.5. Kết quả xác suất dịch với mô hình IBM 2 có tối ưu hoá

e	f	Ban đầu	Lần lặp 2	Lần lặp 2 và so khớp chuỗi
TÔI	tôi	0.75	0.83	0.99
TÔI	ăn	0.21	0.15	0.00
TÔI	cơm	0.21	0.15	0.01
ĂN	tôi	0.04	0.02	0.01
ĂN	ăn	0.42	0.65	0.92
ĂN	cơm	0.42	0.32	0.24
CƠM	tôi	0.04	0.02	0.01
CƠM	ăn	0.42	0.32	0.24
CƠM	cơm	0.77	0.87	0.99

Với dịch dựa trên cụm từ (Phrase-based Translation), mục đích là để giảm bớt các hạn chế của dịch dựa trên từ bằng cách dịch toàn bộ chuỗi từ, trong đó độ dài có thể khác nhau. Đầu tiên, ta sử dụng công cụ MOSES để liên kết cụm từ, sau đó khai thác công cụ giải mã. Đầu vào là một câu tiếng Việt. Vai trò của bộ giải mã là tìm ra bản dịch tốt nhất. Mô hình xác suất cho dịch dựa trên cụm từ là:

$$e_{max} = argmax_e \prod_{i=1}^{l} \emptyset(\overline{f_i}|\,\overline{e_i}) d(start_i - end_{i-1} - 1) P_{LM}(e) \qquad (5.10)$$

Trong đó:

- Dịch cụm từ: chọn cụm từ $\overline{f_i}$ để dịch thành cụm từ $\overline{e_i}$. Điểm $\emptyset(\overline{f_i}|\,\overline{e_i})$ được tra cứu từ bảng dịch cụm từ.
- Sắp xếp lại: Cụm từ trước kết thúc bằng end_{i-1}, cụm từ hiện tại bắt đầu bằng $start_i$. Cần tính toán $d(start_i - end_{i-1} - 1)$
- Mô hình ngôn ngữ: Đối với mô hình n-gram, cần theo dõi (n-1) từ cuối. Tính toán điểm $P_{LM}(X_i|X_{i-(n-1)} \dots X_{i-1})$ để thêm từ X_i
- P_{LM} là xác suất ngôn ngữ (Language Model Probability) của câu dịch. Nó biểu thị xác suất của câu dịch đối với ngôn ngữ đích, dựa trên mô hình ngôn ngữ (language model) được sử dụng.

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.

Thông số cấu hình phần cứng, phần mềm thực nghiệm cho mô hình là:

- Cấu hình phần cứng sử dụng máy tính cá nhân của nghiên cứu sinh:
  - Hệ điều hành: Windows 11 Home Single Language 64-bit.
  - CPU: Core i7-1165G7; 2,8Ghz (8 CPUs).
  - RAM :8129MB.
- Phần mềm: Sử dụng Python để triển khai mô hình IBM. Thư viện hỗ trợ: NumPy để thao tác với dữ liệu.
- Dữ liệu huấn luyện:
  - Số lượng cặp câu song ngữ: Bộ dữ liệu Vie-VSL10k và Vie-VSL60k bao gồm các cặp câu tiếng Việt và ngôn ngữ ký hiệu Việt Nam, cùng với các tương ứng từ điển.
- Tham số đào tạo:
  - Số lần lặp (epoch): 30 lần - đảm bảo độ hội tụ của mô hình.

Điểm đánh giá chất lượng bản dịch của mô hình được trình bày cụ thể trong phần 5.4. so sánh phân tích và đánh giá thực nghiệm các mô hình.

5.2. Mô hình Sequence to Sequence cho bài toán

Mô hình Sequence to Sequence (Seq2Seq) là một trong những mô hình thành công nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên. Mô hình này có nhiều ưu điểm như: Có thể áp dụng cho nhiều tác vụ khác nhau đặc biệt có thể được sử dụng để giải quyết các vấn đề liên quan đến xử lý ngôn ngữ tự nhiên như dịch máy, tổng hợp văn bản, tóm tắt văn bản, hỏi và trả lời tự động, và nhiều ứng dụng khác; Có khả năng học cách biến đổi từ dữ liệu huấn luyện: Seq2Seq cho phép học cách chuyển đổi từ một loại dữ liệu sang loại dữ liệu khác; Dễ dàng mở rộng: Seq2Seq có thể dễ dàng mở rộng để xử lý dữ liệu đầu vào và đầu ra có kích thước khác nhau; Độ chính xác cao: Seq2Seq có khả năng sinh ra các đầu ra chính xác và tự nhiên, đặc biệt là trong các tác vụ dịch máy và tổng hợp văn bản; Có thể kết hợp với các mô hình khác: Seq2Seq có thể được kết hợp với các mô hình khác để cải thiện hiệu suất và độ chính xác của mô hình. Ví dụ: kết hợp với mô hình Attention. Như vậy đối với bài toán cho bài toán dịch câu Tiếng Việt sang câu dạng đúng cú pháp trong VSL thì sử dụng mô hình Seq2Seq là một phương án khả thi.

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 embedding 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).*

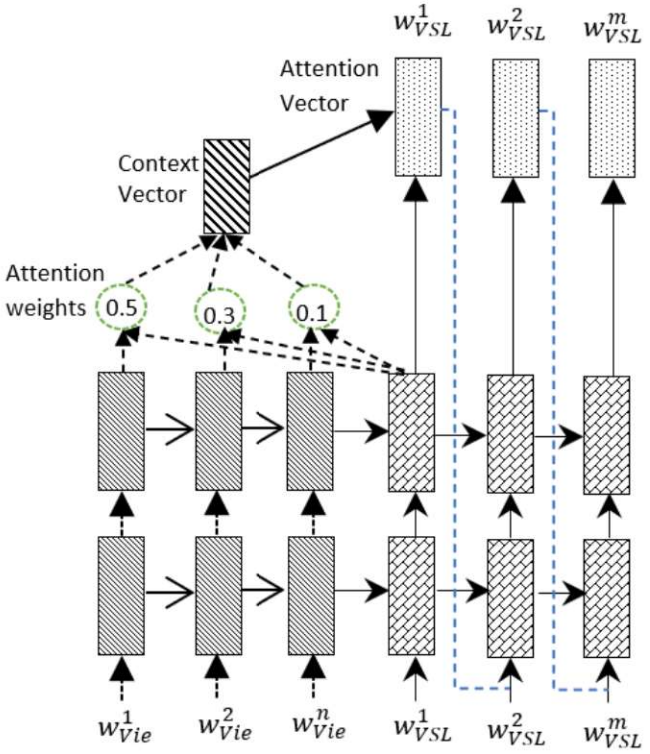
$$\alpha_t = \frac{exp\left(score\left(h_t, \overline{h_s}\right)\right)}{\sum_{s\prime=1}^{S} exp\left(score\left(h_t, \overline{h_s}\right)\right)} \tag{5.11}$$

Vectơ ngữ cảnh:

$$c_t = \sum_s \alpha_{ts} \overline{h_s}$$

Với

- s: là chỉ số bộ mã hoá
- T là chỉ số bộ giải mã
- α_{ts} là trọng số chú ý.
- h_s là trình tự của các kết quả đầu ra bộ mã hóa được tham chiếu.
- $h_t$là trạng thái các bộ giải mã tham dự vào chuỗi .
- c_t là kết quả vector ngữ cảnh.



Hình 5.3. Mô hình bộ mã hoá và giải mã trong bài toán dịch Vie-VSL

Trong quá trình giải mã, mô hình Seq2Seq sẽ thực hiện dự đoán cho mỗi từ hoặc ký tự tiếp theo trong chuỗi đầu ra. Logit là một giá trị số thực được sử dụng để biểu diễn xác suất của từng từ hoặc ký tự trong bước thời gian (time step) tương ứng. Thể hiện của dự đoán logit sẽ xuất hiện trong phần giải mã của mô hình Seq2Seq. Mỗi logit cho biết xác suất của một từ hoặc ký tự tiếp theo trong chuỗi đầu ra. Dựa trên các logit này, mô hình sử dụng một hàm softmax để tính toán xác suất có điều kiện và dự đoán từ hoặc ký tự tiếp theo. Hàm softmax sẽ áp dụng lên các logit để chuyển chúng thành các xác suất có tổng bằng 1, và từ đó mô hình chọn từ hoặc ký tự có xác suất cao nhất làm dự đoán cho bước thời gian đó.

Các phương trình:

1. Tính trọng số chú ý, αts, như một softmax qua chuỗi đầu ra của encoder.
2. Tính toán vectơ ngữ cảnh dưới dạng tổng trọng số của các đầu ra bộ mã hóa.

Cuối cùng là chức năng tính điểm (score function). Công việc của nó là tính toán điểm logit vô hướng cho mỗi cặp khóa-truy vấn. Việc triển khai vectơ hóa của lớp chú ý cho phép chuyển một loạt chuỗi các vectơ truy vấn và một loạt chuỗi các vectơ giá trị. Kết quả có được một loạt chuỗi các vectơ kết quả có kích thước bằng kích thước của các truy vấn.

Bộ giải mã

Công việc của bộ giải mã là tạo ra các dự đoán cho mã thông báo đầu ra tiếp theo.

1. Bộ giải mã nhận được đầu ra bộ mã hóa hoàn chỉnh.
2. Sử dụng RNN để theo dõi những gì nó đã tạo ra cho đến thời điểm hiện tại.
3. Bộ giải mã sử dụng đầu ra RNN của nó làm truy vấn để thu hút sự chú ý qua đầu ra của bộ mã hóa, tạo ra vectơ ngữ cảnh.
4. Nó kết hợp đầu ra RNN và vectơ ngữ cảnh bằng cách sử dụng Công thức 5.11 để tạo ra "vectơ attention ".
5. Nó tạo ra các dự đoán logit cho mã thông báo tiếp theo dựa trên "vectơ attention".

$$a_t = f(c_t, h_t) = \tanh\left(W_c[c_t; h_t]\right) \tag{5.12}$$

Các **bộ mã hóa** xử lý chuỗi đầu vào đầy đủ với một lần gọi duy nhất để hồi quy lặp lại.

Bộ giải mã có 4 đầu vào.

- new_tokens - Các dấu hiệu cuối cùng được tạo. Khởi tạo các bộ giải mã với "[START]" token.
- enc_output - được tạo ra bởi Encoder .
- mask - Một tensor boolean chỉ ra nơi tokens != 0
- state - Các trước state đầu ra từ các bộ giải mã (trạng thái nội bộ của RNN của bộ giải mã).

5.2.2. Huấn luyện mạng

- Cần một hàm xác định mất mát và trình tối ưu hóa.
- Chức năng đào tạo xác định cách cập nhật mô hình cho từng lô đầu vào /mục tiêu.
- Một vòng lặp đào tạo để thúc đẩy quá trình đào tạo và lưu các điểm kiểm tra.

Xác định hàm mất mát

Thực hiện bước đào tạo

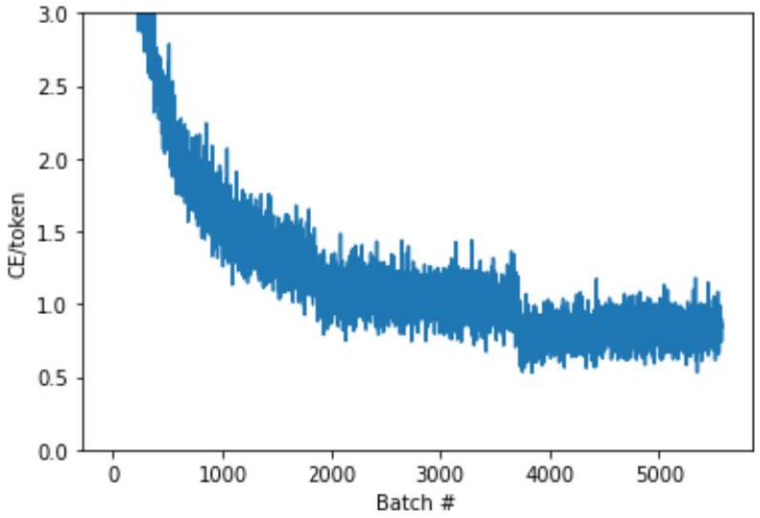
Nhìn chung việc thực hiện cho mô hình đào tạo bao gồm các bước như sau:

1. Nhận một loạt input_text, target_text từ kho ngữ liệu.
2. Chuyển đổi các đầu vào văn bản thô đó thành mã thông báo và mặt nạ.
3. Chạy bộ mã hóa trên input_tokens để có được những encoder_output và encoder_state.

4. Khởi tạo trạng thái bộ giải mã và mất mát.
5. Vòng qua target_tokens :
    a. Chạy bộ giải mã từng bước một.
    b. Tính toán sự mất mát cho mỗi bước.
    c. Tích lũy lỗ trung bình.
6. Tính gradient của sự mất mát và sử dụng tối ưu hóa để áp dụng bản cập nhật trên của mô hình trainable_variables .

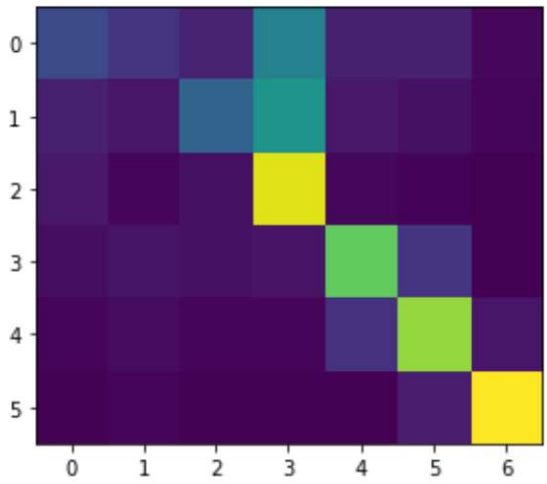
Kiểm tra bước đào tạo

Xây dựng một Mô hình đào tạo dịch , và cấu hình nó cho đào tạo bằng cách sử dụng một mô hình biên dịch.



5.2.3. Tiến trình dịch

Các mô hình được đào tạo, thực hiện một chức năng để thực hiện đầy đủ quá trình dịch.



Các câu ngắn thường hoạt động tốt, nhưng nếu đầu vào quá dài, mô hình sẽ mất tập trung theo đúng nghĩa đen và ngừng cung cấp các dự đoán hợp lý. Có hai lý do chính cho việc này.

1. Mô hình đã được đào tạo với việc buộc phải cung cấp đúng mã thông báo ở mỗi bước, bất kể dự đoán của mô hình. Mô hình có thể được thực hiện mạnh mẽ hơn nếu đôi khi nó được cung cấp các dự đoán của chính nó.

2. Mô hình chỉ có quyền truy cập vào đầu ra trước đó của nó thông qua trạng thái RNN. Nếu trạng thái RNN bị hỏng, không có cách nào để mô hình phục hồi. Transformers giải quyết điều này bằng cách sử dụng tự sự chú ý trong bộ mã hóa và giải mã.

Như vậy, 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.

Các thông số cơ bản hiệu quả cho mô hình Seq2seq với bài toán là:

- Batch size: 128
- Số epoch: 10
- Learning rate: 0.001-0.01
- Kiến trúc mô hình: LSTM với 3 lớp ẩn và số chiều ẩn là 256.
- Thời gian huấn luyện 4.5 giờ trên Google Colab, tốc độ huấn luyện trên CPU xấp xỉ 30-40 mẫu/giây.
- GPU: NVIDIA Tesla T4

Các thông số trên được thực nghiệm với mô hình Seq2seq có thể được xem là phù hợp với bài toán dịch máy ngôn ngữ ký hiệu Việt Nam đặt ra với dữ liệu huấn luyện được chọn. Dưới đây là giải thích về sự phù hợp của từng thông số đã lựa chọn:

1. Batch size: 128 là số lượng mẫu dữ liệu được sử dụng để cập nhật gradient trong mỗi lần huấn luyện. Batch size 128 là một giá trị phổ biến và phù hợp, vừa đủ để có đủ đại diện cho dữ liệu huấn luyện và vừa đủ để tận dụng hiệu năng tính toán song song trên GPU.

2. Số epoch: 10 chính là số lần mô hình được huấn luyện trên toàn bộ tập dữ liệu. Trong các thực nghiệm của luận án này, 10 là một giá trị hợp lý để đạt được một mức độ học tương đối trong bài toán dịch máy, do độ phức tạp của nhiệm vụ của bài toán dịch ngôn ngữ ký hiệu và kích thước dữ liệu với 10k-60k dữ liệu song ngữ được cung cấp..

3. Learning rate: quyết định tốc độ cập nhật các trọng số trong quá trình huấn luyện. Giá trị trong khoảng 0.001-0.01 là một phạm vi thông thường cho learning rate, cho phép mô hình học một cách ổn định và đồng thời tránh việc vượt quá bước cập nhật lớn.

4. Kiến trúc mô hình: Kiến trúc LSTM (Long Short-Term Memory) là một loại mạng nơ-ron truyền thẳng (feedforward neural network) với khả năng xử lý các chuỗi dữ liệu dài và xử lý hiện tượng "vanishing gradient". Hiện tượng đó xảy ra khi đạo hàm của hàm mất mát theo các tham số truyền lại từ các layer sau càng tiến về các layer trước, giảm dần đáng kể đến mức gần như không còn tác động đến quá trình cập nhật tham số. Kết quả là các layer đầu tiên trong mạng không được cập nhật hiệu quả và không học được các đặc trưng phức tạp của dữ liệu. Ba lớp ẩn và số chiều ẩn là 256 được lựa chọn dựa trên độ phức tạp của nhiệm vụ và khả năng tính toán có sẵn. Số lượng lớp ẩn và kích thước đại diện cho khả năng học và biểu diễn của mô hình.

5. Thời gian huấn luyện và tốc độ huấn luyện: Với GPU NVIDIA Tesla T4 là một GPU mạnh mẽ có thể tận dụng hiệu năng tính toán song song để cải thiện tốc độ huấn luyện. Với mô hình đã chọn, cùng với số lượng dữ liệu thực tế cung cấp cho mô hình, thời gian huấn luyện 4.5 giờ và tốc độ huấn luyện 30-40 mẫu/giây là các giá trị thực tế phù hợp với mô hình và môi trường huấn luyện cho bài toán này.

5.3. Mô hình Transformer cho bài toán dịch

Như đã phân tích ở chương 2, ưu điểm vượt trội của Transformer là khả năng tính toán song song cùng với GPU và khả năng xử lý tốt các câu dài. Như vậy thời gian huấn luyện mô hình sẽ có những cải thiện đáng kể. Đây chính là những nguyên nhân để mô hình này được đánh giá phù hợp với bài toán dịch ngôn ngữ ký hiệu Việt Nam. Luận án tiến hành ứng dụng mô hình này với các bước thực hiện trong quá trình bao gồm: mã hoá dữ liệu, huấn luyện dữ liệu, giải mã và sử dụng mô hình dịch để dịch câu tiếng Việt sang câu đúng cú pháp trong ngôn ngữ ký hiệu Việt

Nam. Cuối cùng là phần đánh giá hiệu quả bản dịch sẽ được trình bày chi tiết ở cuối chương cùng với sự so sánh các chỉ số đánh giá với mô hình Seq2Seq và cả việc phân tích, so sánh kết quả với mô hình dịch tương tự được áp dụng cho ngôn ngữ ký hiệu khác.

5.3.1. Quá trình mã hóa và giải mã

Đầu tiên là quá trình vector hoá các cặp câu song ngữ tiếng Việt – ngôn ngữ ký hiệu Việt Nam. Bộ dữ liệu của chúng tôi được lưu lại dưới dạng file văn bản. Việc vector hoá là quá trình chuyển đổi các cặp câu trong file văn bản thành các chuỗi mã hoá.

Để chuẩn bị dữ liệu cho mô hình huấn luyện, chúng tôi sử dụng công cụ tách từ VietWS để thu được kết quả là hai dạng văn bản đã tách từ (tokenizer), một cho tiếng Việt thông thường và một cho ngôn ngữ ký hiệu Việt Nam. Mã hoá là quá trình chuyển đổi câu văn bản thành các mã thông báo. Còn việc giải mã được thực hiện theo chiều ngược lại, tức là chuyển đổi các mã thông báo này trở lại thành văn bản mà con người có thể đọc được.

• Thiết lập đầu vào: Việc mã hoá các lô văn bản thô sử dụng một hàm tokenize_pairs với 2 tham số truyền vào là câu ngôn ngữ ký hiệu Việt Nam “vsl” và câu tiếng Việt “vi”. Thiết lập đầu vào này cần thiết cho việc huấn luyện áp dụng cho các biến đổi tập dữ liệu.

• Mã hóa vị trí - Positional Encoding : Vì đầu vào là tập các vector không có thứ tự nên một " Positional Encoding" được thêm vào để mô hình có thể biết được thông tin về vị trí của từ được vector hoá trong câu. Vector Embedding có thêm Vector Positional Encoding. Vecto Embedding đại diện cho một mã thông báo trong không gian d chiều nơi các mã thông báo có ý nghĩa tương tự sẽ gần nhau hơn. Nhưng Vecto Embedding không mã hóa vị trí tương đối của các vector từ trong câu. Chính về thế, các mã thông báo sẽ giống nhau về ý nghĩa và vị trí trong không gian nhiều chiều khi được thêm Positional Encoding. Công thức tính toán mã hóa vị trí như sau: (5.13)

$$PE_{(pos,2i)} = \sin(pos/1000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/1000^{2i/d_{model}})$$

• Mặt nạ nhìn trước (look-ahead mask) có nhiệm vụ giấu đi các mã thông báo sắp tới theo một thứ tự nhất định. Có nghĩa là look-ahead mask cho biết thông tin

về mục nhập không sử dụng để dự đoán mã thông báo thứ ba, ta cần sử dũng mã thông báo thứ nhất và mã thông báo thứ hai. Cách làm tương tự với việc dự đoán mã thông báo khác.

- Hàm chú ý (attention function) được sử dụng bởi transformer có ba đầu vào: Q (truy vấn), K (khoá), V (giá trị). Việc tính toán các Vector K, Q, V này sử dụng phương trình (2.12)

Vector K sử dụng cho hàm tính xác xuất softmax chuẩn hoá, và sẽ quyết định giá trị của vector Q. Đầu ra đại diện cho phép nhân của trọng số chú ý và vectơ V (giá trị). Điều này đảm bảo rằng các mã thông báo muốn tập trung vào được giữ nguyên trạng và các mã thông báo không liên quan sẽ bị loại bỏ.

5.3.2. Khởi tạo mô hình Transformer

Bộ mã hóa, bộ giải mã và một lớp tuyến tính cuối cùng là các thành phần chủ yếu trong mô hình. Đầu ra của bộ giải mã decoder là đầu vào của lớp tuyến tính và ta sẽ thu được giá trị đầu ra.

- 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) [85].
- Huấn luyện và kiểm tra:

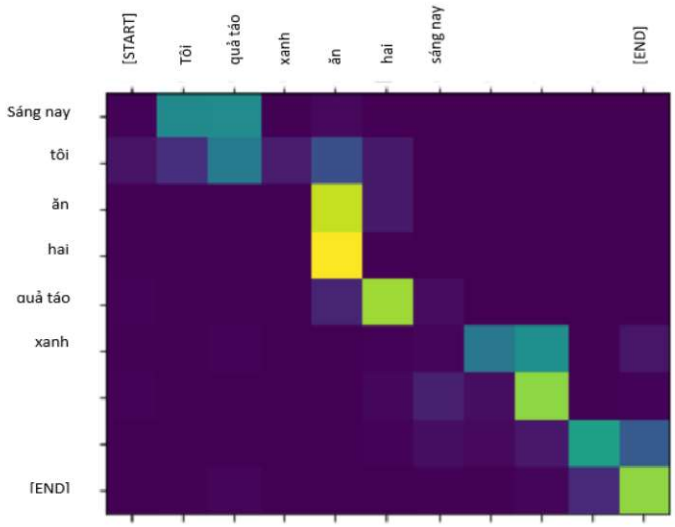
Sau mỗi bước huấn luyện việc lưu các checkpoint được thực hiện bằng cách tạo đường dẫn checkpoint và trình quản lý checkpoint sử dụng.

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: Encoder 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 decoder 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.



Hình 5.4. Bản đồ Attention

Mô hình Transformer cho bài toán dịch VSL trên Github tại địa chỉ https://github.com/BichDiep/transformer-vsl .

Thời gian và môi trường huấn luyện:

• Thời gian huấn luyện: khoảng 8 giờ với số lần lặp (epoch) là 30.

• Môi trường huấn luyện: cấu hình GPU Tesla T4 và RAM 16GB.

• Kích thước batch size: 64.

• Số lượng layer trong mô hình: 6.

• Số lượng head trong multi-head attention: 8.

• Số chiều của bộ mã hóa và giải mã: 512.

❖ **Các phân tích đánh giá về thời gian và môi trường huấn luyện trong bài toán:**

Trong quá trình thực nghiệm của bài toán này, bộ dữ liệu Vie-VSL-60k bao gồm 60.000 cặp câu "song ngữ" (câu tiếng Việt – câu đúng cú pháp VSL). Dữ liệu huấn luyện này cho bài toán dịch ngôn ngữ ký hiệu trên mô hình Transformer có thể được coi là một số lượng dữ liệu khá nhỏ so với một số bài toán dịch máy hoặc xử lý ngôn ngữ tự nhiên khác. Thường thấy dữ liệu huấn luyện cho bài toán đó có kích thước lớn hơn nhiều, có thể hàng triệu hoặc hàng tỷ câu dùng cặp ngôn ngữ. Điều này giúp mô hình dịch máy thông thường học được nhiều tri thức và đặc trưng ngôn ngữ hơn, cũng như giảm thiểu nguy cơ overfitting. Dữ liệu huấn luyện lớn có nhiều lợi ích, như giúp mô hình học tốt hơn và đạt được hiệu suất cao hơn.

Tuy nhiên, với dữ liệu dạng song ngữ dự đoán ký hiệu Việt Nam trong bài toán này, 60.000 cặp câu vẫn có thể giúp mô hình học các quy luật và mối quan hệ cơ bản giữa ngôn ngữ ký hiệu và văn bản tiếng Việt. Điều quan trọng là các thử nghiệm và điều chỉnh các thông số mô hình để đạt được hiệu suất tốt trên dữ liệu huấn luyện có sẵn cho thấy 6 layer trong mô hình là hợp lý nhất. Nếu tăng số lượng layer thì thời gian huấn luyện tăng đáng kể (từ 8h lên đến 12h) trong khi mô hình không đạt hiệu quả cao hơn (với điểm chất lượng bản dịch BLEU không thay đổi).

5.4. Đánh giá các kết quả thực nghiệm

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. BLEU là một phương pháp để đánh giá chất lượng của các tài liệu tự động dịch máy, do IBM đề xuất và được sử dụng làm thước đo đánh giá chính cho nghiên cứu về dịch máy. Trong các thực nghiệm này, chúng tôi đánh giá hiệu suất dịch bằng điểm BLEU bằng các tập lệnh Multi-BLEU.

Tập dữ liệu kiểm tra là tập dữ liệu được xây dựng ở chương 3 với các số liệu thống kê trong bảng 3.14 với tổng số câu kiểm tra là 500 câu (lĩnh vực giao tiếp, y học, kỹ thuật và văn học) với độ dài câu trung bình khoảng 12,8 đơn vị từ vựng.

Bảng 5.6. 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
5	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 khác. Còn với dữ liệu lớn hơn thì các mô hình còn lại cho kết quả vượt trội và tăng dần. Trong các mô hình được sử dụng thì hiện tại trong nghiên cứu của chúng tôi, mô hình Transformer là cho kết quả tốt hơn cả. Việc có độ đo BLEU vượt trội khi huấn luyện trên Viet-VSL-60K dùng Transformer có thể do một số nguyên nhân sau:

- Dữ liệu lớn hơn: Có một nguyên nhân đơn giản là dữ liệu huấn luyện Viet-VSL-60K lớn hơn so với Viet-VSL-10K, điều này cho phép mô hình học được và cải thiện khả năng dự đoán chính xác. Mặc dù
- Kiến trúc mạng: Transformer là một mô hình mạng nơ-ron phức tạp và có khả năng mô hình hóa tương quan phi ngữ cảnh, giúp nó xử lý hiệu quả.

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. Ví dụ như trong dịch tiếng Đức sang ngôn ngữ ký hiệu Đức [86] cũng áp dụng các mô hình dịch Seq2Seq và Transformer đối với tập dữ liệu kiểm tra và huấn luyện của họ. Kết quả tham chiếu trong bảng 5.6.

Như vậy, 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 hai ngôn ngữ.

Bảng 5.7. Tham chiếu điểm BLEU trên bài toán dịch ngôn ngữ ký hiệu khác

Mô hình dịch	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Vietnamese -VSL				
Rule-based	85.45	82.54	78.33	68.02
Seq2Seq	92.5	89.25	85.4	82.44
Transformer	**94.87**	**92.16**	**90.15**	**89.23**
German – German Sign Language				
Rule-based	54.19	39.26	28.44	20.63
Seq2Seq	86.7	79.5	73.2	65.9
Transformer	92.88	89.22	85.95	82.87

Vì bài toán dịch tự động ngôn ngữ ký hiệu Việt Nam đòi hỏi sự kết hợp giữa kiến thức về ngôn ngữ ký hiệu và các mô hình học máy. Mô hình thống kê truyền thống có thể sử dụng luật ngữ cố định dựa trên kiến thức, trong khi mạng Neural Networks có khả năng học tự động từ dữ liệu.

Mạng nơron và ưu điểm vượt trội: Từ các điểm mạnh của mạng Neural Networks, như Seq2Seq và Transformer đã phân tích cụ thể ở trên, trong việc dịch ngôn ngữ ký hiệu, thấy khả năng học các biểu đồ, ngữ cảnh và quan hệ ngôn ngữ ký hiệu thông qua việc sử dụng mạng nơron. Điều này giúp mô hình tự động hơn và có khả năng tự động cập nhật khi có dữ liệu mới. Transformer đã được chứng minh là hiệu quả trong nhiều nhiệm vụ dịch, và nó có khả năng học các quan hệ không tuyến tính và cấu trúc phức tạp trong ngôn ngữ ký hiệu.

Sự cải tiến và tối ưu hóa mô hình trong bài toán này là việc tạo ra tập dữ liệu đặc biệt cho ngôn ngữ ký hiệu, và tối ưu hóa các tham số cho mô hình cụ thể đã áp dụng cho mô hình Seq2Seq và Transformer trong ngôn ngữ ký hiệu Việt Nam.

5.5. 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 là một thách thức lớn 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 hệ thống ngôn ngữ đặc biệt dành cho người khiếm thính với cấu trúc cú pháp của một ngôn ngữ riêng biệt. Với sự phát triển của công nghệ và mô hình học máy, đã có một số nỗ lực trong việc phát triển các hệ thống dịch tự động ngôn ngữ ký hiệu Việt Nam. Tuy nhiên, dịch tự động ngôn ngữ ký hiệu Việt Nam vẫn còn nhiều thách thức do đặc điểm của ngôn ngữ này. Trong đó bài toán dịch theo chiều từ tiếng Việt sang VSL có ý nghĩa quan trọng với mục đích truyền tải kiến thức cho người khiếm thính. Trong các quá trình của bài toán đó, quá trình dịch văn bản tiếng Việt sang câu đúng cú pháp trong VSL được chú ý hơn cả

Với những vấn đề đã trình bày trong luận án về việc triển khai được một số mô hình dịch ứng dụng cho bài toán dịch tự động văn bản tiếng Việt sang dạng văn bản đúng cú pháp trong ngôn ngữ ký hiệu Việt Nam. Kết quả cho thấy các mô hình dịch đáp ứng được yêu cầu đặt ra. Với việc xây dựng một bộ dữ liệu cho bài toán dịch tuy chưa đầy đủ về các mô hình 3D diễn tả trực quan ngôn ngữ ký hiệu mà tập trung vào dịch câu tiếng Việt sang câu đúng cú pháp trong VSL nhưng cũng đã có nhiều ý nghĩa cho việc đánh giá mô hình dịch.

Các kết quả đạt được của luận án bao gồm:

- 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].

- 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]

- 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 **VSL-Lexicon**; dữ liệu "song ngữ" **Vie-VSL10k**, **Vie-VSL60k** công bố cho cộng đồng nghiên cứu sử dụng.

Với các đóng góp trên, luận án không chỉ đáp ứng mục tiêu cụ thể về lý luận trong lĩnh vực dịch tự động ngôn ngữ ký hiệu mà còn đóng góp cho nền tảng xử lý ngôn ngữ tự nhiên. Đặc biệt, trong bối cảnh cụ thể của việc dịch ngôn ngữ ký hiệu Việt Nam, những kết quả này có ý nghĩa lớn trong việc nâng cao tri thức xã hội, tạo cơ hội việc làm, và giúp người khiếm thính hòa nhập vào cộng đồng một cách dễ dàng hơn, vượt qua rào cản giao tiếp.

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 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.

TÀI LIỆU THAM KHẢO

[1] Cao Thị Xuân Mỹ, *Quá trình hình thành và phát triển ngôn ngữ kí hiệu*, Tạp chí KHOA HỌC ĐHSP TPHCM, Số 46, Trang 181-185, 2013.

[2] Đỗ Thị Hiên, *Ngôn ngữ ký hiệu của cộng đồng người khiếm thính Việt Nam: thực trạng và giải pháp*, Báo cáo tổng hợp đề tài nghiên cứu khoa học cấp bộ, Viện Khoa học xã hội Việt Nam, 2012.

[3] Phạm Thị Cơi, *Quá trình hình thành ngôn ngữ nói ở người điếc Việt Nam, Luận án Phó tiến sĩ khoa học Ngữ văn*, Viện Ngôn ngữ học, Tr. 31, 1988.

[4] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, "*Tessa, a system to aid communication with deaf people*", Proceedings of the fifth international ACM conference on Assistive technologies, 2002.

[5] J. A. Bangham, S. J. Cox, R. Elliot, J. R. W. Glauert, I. Marshall, S. Rankov, and M. Wells, "*Virtual signing: Capture, animation, storage and transmission – An overview of the ViSiCAST project*" , IEEE Seminar on Speech and language processing for disabled and elderly people, 2000.

[6] Angus Grieve-Smith, *SignSynth: A Sign Language Synthesis Application Using Web3D and Perl*, Conference: Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction, 2002.

[7] Bernd Krieg-Brückner, Jan Peleska, Ernst-Rüdiger Olderog, Alexander Baer, *The Uniform Workbench, A Universal Development Environment for Formal Methods*, Lecture Notes in Computer Science 1709, Springer 1999.

[8] L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, M. Palmer, "A Machine Translation System from English to American Sign Language", *Envisioning Machine Translation in the Information Future*, Vol. 1934, pp. 191-193, 2000.

[9] Hussein A., Abdul-Wahab. M. A, *SignAloud: A Glove-based System for Unobtrusive ASL Recognition*, ACM Conference on Human-Computer Interaction and Information Retrieval, 3(1), 1-6, doi: 10.1145/2984753.2984756, 2016.

[10] Zhang, J., Thangali, A., Li, Y., & Nevatia. R, *Kinect-based Sign Language Recognition and Translation*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 347-354, doi: 10.1109/CVPRW.2012.6239217, 2012.

[11] Máté. A., Varga. D., Szabó. L. et al, *A Sign Language Recognition and Translation Corpus From Hungarian Sign Language*, Language Resources and Evaluation Conference, 3563–3570, doi: 10.18653/v1/L17-1337, 2017.

[12] Porta J., et al, *A rule-based translation from written Spanish to Spanish sign language glosses*, Comput Speech Lang, 28(3), 788–811, DOI: 10.1016/j.csl.2013.10.003, 2014.

[13] Almeida. I, *Exploring challenges in avatar-based translation from European Portuguese to Portuguese sign language*, Master's Thesis Instituto Superior Técnico, 2014.

[14] Kouremenos D., et al., *A novel rule based machine translation scheme from Greek to Greek sign language: production of different types of large corpora and language models evaluation*, Comput. Speech Lang. 51, 110–135, doi 10.1016/j.csl.2018.04.001, 2018.

[15] Morrissey. S, Way.A., *An example-based approach to translating sign language*, In: Workshop example-based machine translation (MT X-05), pp. 109–116, 2005.

[16] Lopez-Ludena V., et al, *Automatic categorization for improving Spanish into Spanish Sign Language machine translation*, Comput. Speech Lang. 26(3), 149–167, DOI:10.1016/j.csl.2011.09.003, 2012

[17] Buz B, Gungor T, *Developing a statistical Turkish sign language translation system for primary school students.* In: IEEE International Symposium on Innovations in Intelligent SysTems and Applications, DOI:10.1109/INISTA.2019.8778246, 2016.

[18] Kouremenos, et al.., *Statistical machine translation for Greek to Greek sign language using parallel corpora produced via rule-based machine translation*, In: IEEE 31st International Conference on Tools with Artifcial Intelligence (ICTAI), pp. 1–15, 2018.

[19] Achraf, O., Jemni, M.: *Designing high accuracy statistical machine translation for sign language using parallel corpus: case study English and American sign language*. J. Inf. Technol. Res. 12(2), 134–158, doi:10.4018/JITR.20190 40108, 2019.

[20] Brour M., Benabbou A, *ATLASLang MTS 1: Arabic text language into Arabic sign language machine translation system*, In: 2nd International Conference on Intelligent Computing in Data Sciences, pp. 236–245, Doi:10.1016/j.procs. 2019.01.066, 2019.

[21]. Kayahan D., Gungor T, *A hybrid translation system from Turkish spoken language to Turkish sign language,* In: IEEE international symposium on innovations in intelligent systems and applications, pp. 1–6, Doi:10.1109/INISTA.8778347, 2019.

[22] Jenkins, J., & Rashad, S. *LeapASL: A platform for design and implementation of real time algorithms for translation of American Sign Language using personal supervised machine learning models*. Software Impacts, 12, Article 100302, 2022.

[23] Morrissey. S, *Assistive technology for deaf people: Translating into and animating Irish sign language*, In: Proceedings of the 12th International Conference on Computers Helping People with Special Needs, pp. 8–14, 2008.

[24] Muhammad Sanaullah1, Babar Ahmad, Muhammad Kashif, Tauqeer Safdar, Mehdi Hassan, Mohd Hilmi Hasan and Norshakirah Aziz, *A Real-Time Automatic Translation of Text to Sign Language* ,Computers - Materials & Continua, Tech Science Press, DOI:10.32604/cmc.2022.019420, 2021.

[25] Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. *Frozen pretrained transformers for neural sign language translation*. In 18th Biennial Machine Translation Summit (MT Summit 2021), pp. 88–97. Association for Machine Translation in the Americas, 2021.

[26] San Kim, Chang Jo Kim, Han-Mu Park, Yoonyoung Jeong, Jin Yea Jang, and Hyedong Jung. *Robust keypoint normalization method for korean sign language translation using transformer*. In 2020 International Conference on Information

and Communication Technology Convergence (ICTC), pp. 1303–1305. IEEE, 2020.

[27] Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. *Stochastic transformer networks with linear competing units: Application to end-to-end SL translatio*n. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11946–11955, 2021.

[28] Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, *R. Multi-channel transformers for multi-articulatory sign language translation*. In European conference on computer vision, pp. 301–319, Springer, 2020.

[29] Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. *Sign language transformers: Joint end-to-end sign language recognition and translation*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10023–10033, 2020.

[30] De Coster, M., D'Oosterlinck, K., Pizurica, M., Rabaey, P., Van Herreweghe, M., Dambre, J., et al. *Frozen pretrained transformers for neural sign language translation*. In 18th Biennial machine translation summit, pp. 88–97, 2021.

[31] Egea, S., McGill, E., & Saggion, H. *Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation*. In Proceedings of the 14th workshop on building and using comparable corpora, 2021.

[32] Kim, S., Kim, C. J., Park, H.-M., Jeong, Y., Jang, J. Y., & Jung, H. *Robust keypoint normalization method for Korean sign language translation using transformer*. In 2020 international conference on information and communication technology convergence ICTC, pp. 1303–1305, IEEE, 2020.

[33] Saunders, B., Camgoz, N. C., & Bowden, R., *Progressive transformers for endto-end sign language production*. In European conference on computer vision pp. 687–705, 2020.

[34] Galina Angelova, Eleftherios Avramidis and Sebastian Möller, *Using Neural Machine Translation Methods for Sign Language Translation*, 60th Annual Meeting of the Association for Computational Linguistics Student Research Workshop, pages 273 – 284, 2022

[35] Quach, L., Nguyen, C.-N.: *Conversion of the Vietnamese grammar into sign language structure using the example-based machine translation algorithm*. In: International Conference on Advanced Technologies for Communications, pp. 27–31, 2018.

[36] Da, Q.L., et al.: *Converting the vietnamese television news into 3D sign language animations for the deaf*. In: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 257. Springer, 2019.

[37] Quach, LD., Duong-Trung, N., Vu, AV., Nguyen, CN, *Recommending the Workflow of Vietnamese Sign Language Translation via a Comparison of Several Classification Algorithms*. In:Computational Linguistics, Communications in Computer and Information Science, vol 1215. Springer, 2020.

[38] Lê Sơn Thái, Đỗ Năng Toàn, Mã Văn Thu, Nguyễn Thị Bích Điệp, *Một kỹ thuật điều khiển động tác của con người trong thực tại ảo ứng dụng diễn hoạ ngôn ngữ ký hiệu Việt Nam*, Kỷ yếu hội thảo quốc gia Nghiên cứu cơ bản và ứng dụng CNTT FAIR 10, NXB Khoa học Tự nhiên và Công nghệ, 2017.

[39] N.C. Camgoz and S. Hadfield and O. Koller and H. Ney and R. Bowden, *RWTH-PHOENIX-Weather 2014 T: Parallel Corpus of Sign Language Video, Gloss and Translation*, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, 2018.

[40] David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, Michael Subotin, *The Hiero Machine Translation System: Extensions, Evaluation, and Analysis,* Human Language Technology Conference and Conference on Empirical Methods in Natural Language, pages 779–786, 2005.

[41] Neco R., Forcada Mikel, *Neural machine translation with an encoder-decoder architecture for rare-word processing*. Computational Linguistics and Intelligent Text Processing, 10461, 329-343, 2018.

[42] Wu Y., & Schuster M, *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv:1609.08144, 2016.

[43] Manning D., & Schütze H., *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[44] Senellart. J, *Systran: A history of machine translation*, John Benjamins Publishing Company, 2017.

[45] Wilks. Y, *Machine Translation: Its Scope and Limits,* Springer Science & Business Media, 2008.

[46] Forcada. M, Ginestí-Rosell. M., & Sánchez-Martínez. F, *The Apertium machine translation platform*, Computational Linguistics, 37(2), 309-318, 2011.

[47] Derczynski. L, Nielsen. H. S., & Søgaard. A, *Gramtrans: A rule-and example-based machine translation platform,* In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 53-56), 2013.

[48] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. *Distributed representations of words and phrases and their compositionality*, CoRR, abs/1310.4546, 2013.

[49] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, Sanjeev Khudanpu, *Recurrent neural network based language model*, Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), 2010.

[50] Winkler, W. E. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 354-359, 1990.

[51] Sutskever. I., Vinyals. O., & Le Q. V, *Sequence to sequence learning with neural networks*, In Advances in neural information processing systems, pp. 3104-3112, 2014.

[52] Bengio. Y., Simard. P., & Frasconi. P., *Learning Long-Term Dependencies with Gradient Descent is Difficult*, Advances in Neural Information Processing Systems pp. 1-10, 1994.

[53] Llion Jones, Aidan N. Gomez, Łukasz Kaiser, *Attention Is All You Need*, 31st Conference on Neural Information Processing Systems USA, 2017.

[54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, *BLEU: a Method for Automatic Evaluation of Machine Translation*, . Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[55] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin, *A Bit of Progress in Language Modeling*, Neural Networks, Volume 16, Issue 9, Pages 1329-1338, 2003.

[56] Jelinek. F., & Mercer. R. L, *Interpolated estimation of Markov source parameters from sparse data*, In Proceedings of the workshop on speech and natural language, Association for Computational Linguistics, pp. 357-366, 1980.

[57] Nguyễn Phương Thái và các cộng sự, " Đề tài SP8.5. Công cụ phân tích cú pháp Tiếng Việt", Mã số KC01.01.03/06-10, 2008.

[58] Bojar, Ondrej ; Chatterjee, Rajen ; Federmann, Christian et al. *Findings of the 2016 Conference on Machine Translation (WMT16)*. Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers. Berlin, Germany : Association for Computational Linguistics, pp. 131-198, 2016.

[59] Thi-Vinh Ngo, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha & Le-Minh Nguyen, *An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation - An empirical study of Chinese, Japanese to Vietnamese Neural MachineTranslation,* Artificial Intelligence, Issue 1, Taylor & Francis Volume 36, 2022.

[60] Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, Minh-Thang Luong, *MTet: Multi-domain Translation for English and Vietnamese*. CoRR abs/2210.05610, 2022.

[61] J. Kanis and J. Zahradil and F. Jurcıcek and L. Muller, *Czech-Sign Speech Corpus for Semantic based Machine Translation*, International Conference on Text, Speech and Dialogue, pp. 613–620, 2006.

[62] Dimitris Kouremenos, Stavroula-Evita Fotinea, Eleni , Klimis S. Ntalianis, *A prototype Greek text to Greek Sign Language conversion system*, Behaviour and Information Technology, Taylor & Francis, 2010.

[63] Hanke, T., Schulder, M., Konrad, R., & Jahn, E, *Extending the public DGS corpus in size and depth*. In Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the

service of the language community, Technological challenges and application perspectives (pp. 75–82), 2020

[64] Kamath, Chandrashekhar, et al, *Unsupervised Bilingual Lexicon Induction for Sign Language*, Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, pp 56-62, 2020.

[65] Sennrich, Rico, Barry Haddow, and Alexandra Birch, *Improving Neural Machine Translation Models with Monolingual Data*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 86-96, 2016.

[66] Zhang, Xingxing, et al, *Joint training for pivot-based neural machine translation*, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1723-1732, 2017.

[67] Xie, Ruochen, et al, *Unsupervised Lexical Paraphrasing via Adversarial Training*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 3682-3692, 2018.

[68] Lample, Guillaume, et al, *Phrase-Based & Neural Unsupervised Machine Translation,* Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 5039-5049, 2018.

[69] Thi-Vinh Ngo, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha & Le Minh Nguyen, *An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation*, Applied Artificial Intelligence, 36:1, 2101755, 2022.

[70] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, Mohammad Sabokrou, *All You Need In Sign Language Production*, Computer Vision and Pattern Recognition, 2022.

[71] V. Athitsos and C. Neidle and S. Sclaroff and J. Nash and A. Stefan and Q. Yuan and A. Thangali, *The American sign language lexicon video dataset*, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 1–8, 2018.

[72] J. Bungeroth and D. Stein and Ph. Dreuw and H. Ney and S. Morrissey and A. Way and L.V. Zijl, *The ATIS sign language corpus*, 6th International Conference on Language Resources and Evaluation, 2008.

[73] S. Matthes and Th. Hanke and A. Regen and J. Storz and S. Worseck and E. Efthimiou and A.L. Dimou and A. Braffort and J. Glauert and E. Safar, *Dicta-Sign–building a multilingual sign language corpus*, In 5th LREC. Istanbul, 2012.

[74] N.K. Caselli and Z.S. Sehyr and A.M. Cohen-Goldberg and K. Emmorey, *ASL-Lex: A lexical database for ASL*, Behavior Research Methods 49, pp. 784–801, 2017.

[75] RWTH-PHOENIX-Weather 2014 T: Parallel Corpus of Sign Language Video, Gloss and Translation Human Language Technology & Pattern Recognition Group RWTH Aachen University, Germany, 2014.

[76] S.K. Ko and Ch.J. Kim and H. Jung and Ch. Cho, *Neural Sign Language Translation Based on Human Keypoint Estimation*, Applied Sciences 9, 2019.

[77] A. Duarte and Sh. Palaskar and D. Ghadiyaram and K. DeHaan and F. Metze and J. Torres and X. GiroiNieto, *How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language*, Sign Language Recognition, Translation, and Production workshop, 2020.

[78] J. Tiedemann, *Finding Alternative Translations in a Large Corpus of Movie Subtitles*, Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 2016.

[79] D. Elliott and S. Frank and K. Simaean and L. Specia, ´ *Multi30K: Multilingual English-German Image Descriptions*, Proceedings of the 5th Workshop on Vision and Language, pp. 70–74, 2016.

[80] T. Nakazawa and M. Yaguchi and K. Uchimoto and M. Utiyama and E. Sumita and S. Kurohashi and H. Isahara, *ASPEC: Asian Scientific Paper Excerpt Corpus*, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016), pp. 2204-2208, 2016.

[81] Conneau, A. and Lample, G. and Ranzato, M. and Denoyer, L. and Jegou, H. ´ *Word Translation Without Parallel Data*, International Conference on Learning Representations, 2018.

[82] Lample and A. Conneau and L. Denoyer and M. Ranzato, *Unsupervised Machine Translation Using Monolingual Corpora Only*, International Conference on Learning Representations (ICLR), 2017.

[83] P. Michel and G. Neubig, *MTNT: A Testbed for Machine Translation of Noisy Text*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

[84] Fellbaum.C, *WordNet: An Electronic Lexical Database,* MIT press, VOL 13, 1998.

[85] Diederik P. Kingma, Jimmy Lei Ba, "Adam: a method for stochastic optimization", *International Conference on Learning Representations*, 2015.

[86] Kayo Yin, Jesse Read, *Better Sign Language Translation with STMC-Transformer*, Proceedings of the 28th International Conference on Computational Linguistics, 2020.