ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

– – – – – – – ★ – – – – – – –

Ngô Thị Vinh

CẢI TIẾN CHẤT LƯỢNG DỊCH MÁY CHO CẶP NGÔN NGỮ HẠN CHẾ TÀI NGUYÊN

Chuyên ngành: Khoa học máy tính

Mã số: 9480101.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội - 2023

Công trình này được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội

Người hướng dẫn khoa học: PGS. TS Nguyễn Phương Thái
GS. TS Nguyễn Lê Minh

Phản biện 1:.....
Phản biện 2:.....
Phản biện 3:.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc Gia chấm luận án tiến sĩ họp tại .....
vào hồi .....giờ .....ngày .....tháng.....năm.....

Có thể tìm hiểu luận án tại:
- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội.

Mục lục

MỞ ĐẦU

Dịch máy là một trong những bài toán quan trọng hàng đầu trong trong xử lý ngôn ngữ tự nhiên. Ngay từ những thập niên 1950 và 1960 của thế kỷ 20 các nhà khoa học đã sớm đề xuất các ý tưởng về việc xây dựng các hệ thống dịch máy tự động để đáp ứng nhu cầu dịch thuật trong thực tế.

Ban đầu các hệ thống dịch chủ yếu dựa vào luật, từ điển hoặc thông qua các mẫu ví dụ. Tiếp đó là cách tiếp cận dịch thống kê (Statistical Machine Translation - SMT) các luật dịch được học tự động từ kho dữ liệu song ngữ. Tuy nhiên, các văn bản đầu ra thường kém trôi chảy. Cho và cộng sự (2014) đã áp dụng thành công mô hình mạng nơron (Neural Machine Translation - NMT) với việc khắc phục hiện tượng bản dịch không trôi chảy của các hệ thống SMT. Tuy nhiên, trong NMT, dữ liệu song ngữ vẫn được coi là nguồn tài nguyên chính để phát triển các hệ thống dịch NMT hiện nay, nhưng trên thực tế chúng lại khan hiếm. Trong điều kiện khan hiếm tài nguyên song ngữ, các nhà nghiên cứu đã nỗ lực đề xuất nhiều giải pháp khác nhau để nâng cao chất lượng dịch máy nhưng chúng vẫn tồn tại các hạn chế như:

- Các phương pháp sinh dữ liệu song ngữ tổng hợp từ dữ liệu đơn ngữ có những hạn chế là chúng đòi hỏi phải có một hệ thống dịch ban đầu đủ tốt để sinh ra các bản dịch có chất lượng. Điều này khá khó khăn đối với các cặp ngôn ngữ ít tài nguyên hoặc không có tài nguyên song ngữ.

- Phương pháp sử dụng các mô hình ngôn ngữ lớn để khởi tạo hệ thống dịch đòi hỏi nhiều nỗ lực thu thập dữ liệu đơn ngữ, mặc dù chúng sẵn có. Ngoài ra, phương này bị hạn chế khi các tham số của mô hình ngôn ngữ bị điều chỉnh bởi mô hình dịch với lượng dữ liệu nhỏ trong quá trình huấn luyện dẫn đến hiệu năng dịch bị giới hạn.

- Phương pháp sử dụng dịch đa ngữ đòi hỏi cặp ngôn ngữ cha giàu tài nguyên phải có nhiều điểm tương đồng về mặt ngôn ngữ học (như cấu trúc ngữ pháp, từ vựng) với cặp ngôn ngữ ít tài nguyên hoặc các cặp ngôn ngữ trong cùng hệ thống dịch cũng phải có độ tương đồng nhất định về mặt ngôn ngữ học.

- Các phương pháp xử lý từ hiếm vẫn phụ thuộc vào các từ điển, hoặc chỉ cải thiện đáng kể việc dịch các từ hiếm nhưng không giải quyết triệt để, hoặc chưa hiệu quả trong tình huống ít tài nguyên song ngữ.

- Các phương pháp cải tiến mô hình dịch đòi hỏi thiết kế lại kiến trúc hệ thống, hàm mục tiêu. Điều này khá phức tạp và mất nhiều thời gian. Bên cạnh đó, các hệ thống này thường đòi hỏi thêm các tài nguyên bên ngoài như từ điển, mô hình ngôn ngữ để khởi tạo hệ thống.

- Ngoài ra, việc chọn lọc các dữ liệu từ nhiều miền khác nhau để cải thiện chất lượng dịch trên một miền cụ thể là rất cần thiết. Các phương pháp chọn lọc dựa vào từ điển, tập từ vựng thường làm mất ngữ cảnh câu trong khi các phương pháp chọn lọc dữ liệu sử dụng biểu diễn từ hoặc câu trong các mạng nơron phụ thuộc vào chất lượng dữ liệu, kiến trúc mạng và tốn nhiều thời gian.

Từ các nhược điểm vừa nêu, có thể thấy các phương pháp đã có chưa giải quyết triệt để các khía cạnh khác nhau của bài toán dịch máy trong điều kiện khan hiếm tài nguyên song ngữ. Dưới sự định hướng, hỗ trợ của các thầy hướng dẫn và các cộng sự, tác giả đã thực hiện luận án với đề tài *"Cải tiến chất lượng dịch máy cho cặp ngôn ngữ hạn chế tài nguyên"* nhằm đóng góp thêm các giải pháp để giải quyết các vấn đề còn tồn tại của bài toán này.

Đóng góp chính của luận án bao gồm:

- Đề xuất phương pháp sinh dữ liệu song ngữ tổng hợp trong điều kiện hạn chế tài nguyên song ngữ (công trình số 1).
- Đề xuất các hệ thống dịch đa ngữ giữa các ngôn ngữ có điểm tương đồng về mặt ngôn ngữ học và các phương pháp học sự tương tự giữa các đơn vị dịch trong không gian dịch đa ngữ (công trình 1, 4).
- Đề xuất các cách tiếp cận khác nhau để nâng cao chất lượng dịch các từ hiếm bao gồm: cải tiến quá trình giải mã, kết hợp các vectơ embedding tới xác suất dự đoán đầu ra, sử dụng các phụ tố tách từ có giám trong văn bản tiếng Anh, sử dụng quan hệ đồng nghĩa trong mạng từ Wordnet (Công trình 2, 3, 7).
- Đề xuất các hệ thống dịch sử dụng các phương pháp phân đoạn từ khác nhau và phương pháp phân đoạn từ theo cách tiếp cận học không giám sát cho văn bản tiếng Việt, đề xuất kết hợp các phương pháp nâng cao hiệu quả dịch máy trên miền cụ thể (công trình 5, 6, 8, 9).
- Bên cạnh các thử nghiệm cho cặp ngôn ngữ Anh-Việt, tác giả tập trung vào các cặp ngôn ngữ hạn chế tài nguyên và còn khá ít nghiên cứu như Trung-Việt, Nhật-Việt, Lào-Việt, Khmer-Việt, Pháp-Việt.
- Ngoài việc sử dụng các tập dữ liệu Anh-việt sẵn có, đã được công bố trước đó, tác giả công bố một số tập dữ liệu thu thập được trong quá trình thực hiện luận án cho mục đích nghiên cứu, bao gồm các tập dữ liệu song ngữ Anh-Việt, Pháp-Việt, Nhật-Việt, Trung-Việt.

Nội dung của luận án bao gồm 05 chương, trong đó, **Chương 1** trình bày tổng quan về dịch máy cho cặp ngôn ngữ hạn chế tài nguyên, **Chương 2** trình bày về phương pháp tăng cường dữ liệu tổng hợp cho cặp ngôn ngữ hạn chế tài nguyên, **Chương 3** trình bày về các phương pháp sử dụng hệ dịch đa ngữ cho các cặp ngôn ngữ tương đồng, **Chương 4** trình bày về các phương pháp đề xuất nhằm nâng cao chất lượng dịch các từ hiếm, **Chương 5** trình bày các đề xuất phân đoạn văn bản trong dịch máy và cải thiện chất lượng dịch máy theo miền.

Chương 1

TỔNG QUAN VỀ DỊCH MÁY CHO CẶP NGÔN NGỮ HẠN CHẾ TÀI NGUYÊN

1.1 Giới thiệu về bài toán dịch máy

Khái niệm về dịch máy (Machine Translation) hay còn gọi là dịch tự động được đề cập lần đầu tiên bởi Weaver năm 1955 dùng để chỉ *việc nghiên cứu sử dụng máy tính để dịch văn bản từ một ngôn ngữ sang này một ngôn ngữ khác.*

Tiến trình phát triển của dịch máy trải qua mô hình dịch dựa vào từ điển, luật (1970-2007), dựa vào mô hình dịch thống kê (2003-2015), dịch dựa vào mạng nơron (2014 đến nay).

1.2 Một số khái niệm được sử dụng trong luận án

Phần này trình bày các khái niệm sử dụng trong luận án bao gồm khái niệm về ngôn ngữ hạn chế tài nguyên, cặp ngôn ngữ hạn chế tài nguyên, từ hiếm, dịch đa ngữ, dữ liệu tổng hợp.

1.3 Phạm vi của luận án

Tác giả nghiên cứu các hệ thống dịch trên các cặp ngôn ngữ hạn chế tài nguyên có liên quan đến tiếng Việt, các hệ thống dịch đa ngữ từ nhiều ngôn ngữ sang một ngôn ngữ, các tập dữ liệu từ miền TED Talks, ALT, CCAligned, DongDu và được xây dựng từ đề tài KC-4.0.12/19-25.

1.4 Những thách thức trong mô hình dịch máy hiện nay

Thứ nhất là vấn đề khan hiếm tài nguyên song ngữ trong dịch máy.

Thứ hai là vấn đề dịch các từ hiếm.

Thứ ba là vấn đề lệch miền (out of domain) trong các hệ thống dịch.

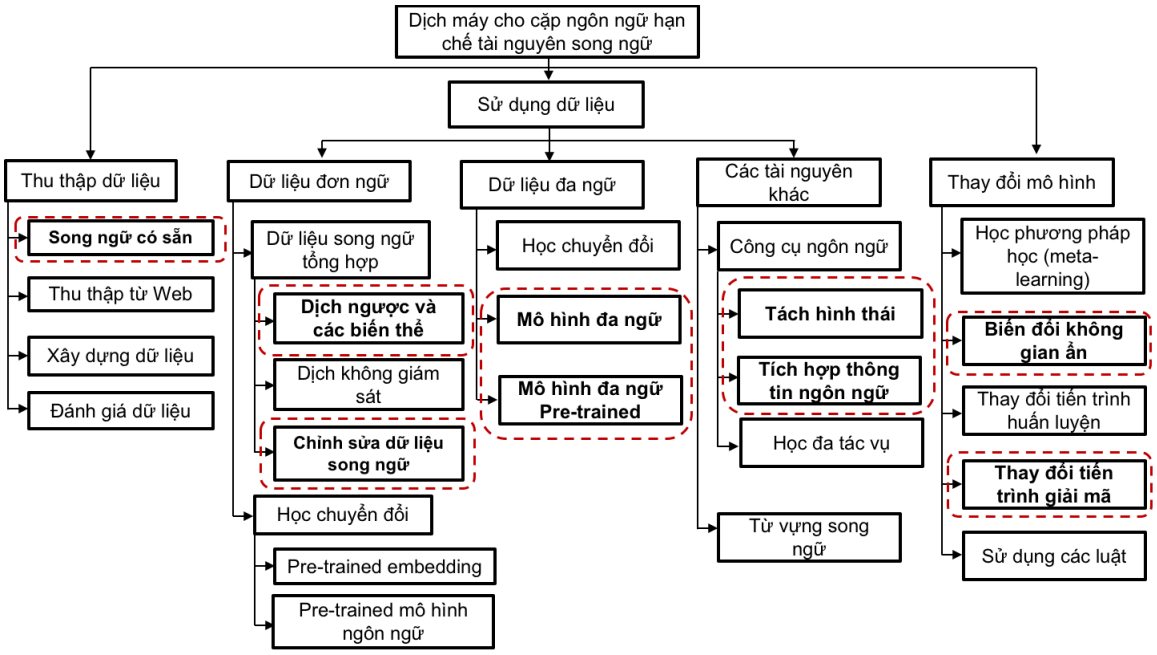
Thứ tư là vấn đề dịch câu dài, dịch ở mức tài liệu.

Thứ năm là vấn đề về tốc độ dịch chậm.

Ngoài ra, còn một số thách thức khác như vấn đề dịch các từ đa nghĩa, dịch các thành ngữ, dịch văn bản dạng bài nói, đánh giá tự động các hệ thống dịch máy cần gần với đánh giá của con người hơn, ...

1.5 Các hướng nghiên cứu chính về bài toán dịch máy hạn chế tài nguyên

Theo Haddow và cộng sự (2022), các hướng nghiên cứu chính cho bài toán hạn chế tài nguyên song ngữ được mô tả như trong Hình 1.2. Trong đó, các phương pháp được in đậm và bao quanh bởi các đường nét đứt có liên quan đến các phương pháp đề xuất trong luận án.



Hình 1.2. Tổng quát về các hướng nghiên cứu chính cho bài toán dịch máy hạn chế tài nguyên song ngữ.

1.5.1 Các phương pháp thu thập dữ liệu

Các phương pháp này bao gồm việc thu thập, xây dựng và đánh giá dữ liệu.

1.5.2 Các phương pháp dựa vào dữ liệu đơn ngữ

1.5.2.1 Sử dụng dữ liệu song ngữ tổng hợp

Điển hiền là phương pháp dịch ngược với việc sử dụng các mô hình dịch máy được huấn luyện trên tập dữ liệu song ngữ khởi đầu để dịch dữ liệu song ngữ.

1.5.2.2 Sử dụng dịch không giám sát

Dịch không giám sát là cách tiếp cận dịch dựa hoàn toàn vào dữ liệu đơn ngữ được thực nghiệm lần đầu tiên bởi Lample và cộng sự (2018).

1.5.2.3 Chỉnh sửa dữ liệu song ngữ có sẵn

Một số nghiên cứu đề xuất chỉnh sửa dữ liệu song ngữ sẵn có để sinh ra dữ liệu song ngữ tổng hợp như loại bỏ, di chuyển, thay thế hoặc sao chép các từ trong câu gốc phía nguồn hoặc đích để sinh ra câu tổng hợp trong khi giữ nguyên câu phía còn lại.

1.5.2.4 Sử dụng các mô hình được huấn luyện trước (pre-trained)

Cách tiếp cận này tận dụng các mô hình đã được huấn luyện trước (pre-trained model) trên một tập dữ liệu đơn ngữ lớn.

1.5.3 Các phương pháp dựa vào dịch đa ngữ

1.5.3.1 Học chuyển đổi

Phương pháp này huấn luyện hệ thống dịch trên các cặp ngôn ngữ cha giàu tài nguyên sau đó làm mịn trên cặp ngôn ngữ hạn chế tài nguyên.

1.5.3.2 Mô hình dịch đa ngữ

Cho phép nhiều cặp ngôn ngữ cùng chia sẻ thông tin trong một không gian chung.

1.5.3.3 Sử dụng mô hình đa ngữ pre-trained

Sử dụng các mô hình đa ngữ được huấn luyện trước như mBERT, mBART để nâng cao chất lượng dịch máy cho cặp ngôn ngữ hạn chế tài nguyên.

1.5.4 Các phương pháp dựa vào tài nguyên khác

Một số nghiên cứu cải thiện chất lượng dịch cho cặp ngôn ngữ hạn chế tài nguyên thông qua việc phân đoạn từ thành các sub-word, tích hợp thông tin đặc trưng ngôn ngữ, sử dụng học đa tác vụ bằng cách kết hợp tác vụ khác (như phân tích cú pháp) vào tác vụ dịch máy, hoặc sử dụng các từ vựng song ngữ.

1.5.5 Các phương pháp thay đổi mô hình

Sử dụng các phương pháp như meta-learning, biến đổi vectơ embedding, thay đổi kiến trúc hệ thống huấn, cải tiến quá trình giải mã, ...

1.6 Kiến trúc hệ thống dịch máy dựa trên mạng nơron

1.6.1 Kiến trúc hồi quy RNN

Phần này trình bày cơ bản về kiến trúc RNN cho bài toán dịch máy.

1.6.2 Kiến trúc Transformer

Phần này trình bày cơ bản về kiến trúc Transformer cho bài toán dịch máy.

1.6.3 Thiết lập hệ thống trong các thực nghiệm

Phần này mô tả các thiết lập cơ bản cho các kiến trúc RNN và Transformer trong các thực nghiệm của luận án.

1.7 Đánh giá hệ thống dịch máy

Phần này tác giả trình bày về độ đo BLEU và TER là các độ đo được sử dụng phổ biến trong dịch máy.

1.8 Hệ thống máy tính toán

Các thực nghiệm trong luận án được tiến hành trên máy chủ với cạc đồ hoạ NVIDIA GeForce GTX 1080 12GB VRAM, và GeForce RTX 3090 24GB VRAM.

1.9 Phương pháp lựa chọn dữ liệu TF-IDF

Phần này trình bày chi tiết về phương pháp lựa chọn dữ liệu dựa vào độ đo TF-IDF.

1.10 Tóm tắt chương

Trong chương này, tác giả trình bày khái quát về các khái miện, thách thức, mô hình, độ đo được sử dụng trong luận án.

Chương 2

CẢI TIẾN CHẤT LƯỢNG DỊCH BẰNG TĂNG CƯỜNG DỮ LIỆU TỔNG HỢP

Chương này trình bày phương pháp sinh dữ liệu tổng hợp được đề xuất bởi tác giả.

2.1 Đặt vấn đề

Tăng cường dữ liệu song ngữ tổng hợp là chiến lược được nhiều nghiên cứu quan tâm, điển hình là kỹ thuật dịch ngược, sao chép dữ liệu nguồn hoặc đích, loại bỏ, thay, thế, chèn các từ trong câu gốc để sinh ra câu giả. Các cách tiếp cận này phần lớn yêu cầu thêm các tài nguyên bổ sung cho việc sinh dữ liệu tổng hợp như mô hình dịch được huấn luyện trước, mô hình ngôn ngữ, bộ phân tích cú pháp, từ điển song ngữ.

Cách tiếp cận đề xuất trong luận án có ưu điểm so với các cách tiếp cận trước đó là:

— Không yêu cầu các mô hình dịch máy được huấn luyện trước nên tránh được việc sinh ra các dữ liệu tổng hợp kém chất lượng trong điều kiện hạn chế tài nguyên.

— Không yêu cầu sử dụng thêm các nguồn tài nguyên khác bên ngoài tập dữ liệu song ngữ sẵn có nên giảm bớt độ phức tạp cho mô hình và tránh được việc lan truyền lỗi từ các tác vụ phụ sang tác vụ dịch máy.

— Các kỹ thuật đề xuất trước đây thường thay đổi cấu trúc câu nguồn hoặc làm mất ngữ cảnh hay xáo trộn trật tự trong câu đích trong khi cách tiếp cận đề xuất trong luận án không làm thay đổi trật tự của câu.

2.2 Phương pháp đề xuất

Ý tưởng chính của phương pháp là sinh ra các đơn vị dịch nhân tạo từ các đơn vị dịch chuẩn trong câu đích, trong khi câu nguồn được giữa nguyên. Các bước chính của giải thuật như sau:

Bước 1. Từ tập các câu đơn ngữ trong văn bản đích, sinh ra từ vựng V_T chứa các đơn vị dịch chuẩn (không phân biệt ký tự, từ hay các sub-word) với tần số của chúng.

Bước 2. Từ các đơn vị dịch chuẩn, sinh ra các đơn vị dịch nhân tạo ký hiệu là ATU (Artificial Translation Units) bằng cách gán cho mỗi đơn vị dịch chuẩn một nhãn khác nhau và duy nhất sao cho các nhãn không trùng lặp với các đơn vị dịch chuẩn và thu được từ điển gán nhãn V_{ATU}. Chẳng hạn, tác giả sử dụng các ký hiệu $id0, id1, ..., idn$.

Bước 3. Với mỗi cặp câu song ngữ chuẩn, sinh ra các câu tổng hợp từ câu đích bằng cách thay thế các đơn vị dịch chuẩn có tần số lớn hơn ngưỡng *ths* bởi các đơn vị dịch nhận tạo ATU. Sau đó, các câu tổng hợp lại được ghép cặp với câu nguồn ban đầu để tạo thành cặp song ngữ tổng hợp.

Bảng 2.6. Kết quả thực nghiệm phương pháp tăng cường dữ liệu tổng hợp đề xuất trên tập dữ liệu **TED Talks** với ngưỡng tần số thay thế là 7 và so sánh với hệ thống dịch dịch ngược.

Chiều dịch	**Hệ thống dịch**	**Tập phát triển**	**Tập đánh giá**
Cn→ Vi	**Hệ thống cơ sở**	17.1	17.4
	Phương pháp đề xuất ($ths = 7$)	**17.2** (+0.1)	**17.9** (+ 0.5)
	Phương pháp dịch ngược	**18.0**(+0.9)	**18.5** (+ 1.1)
	Kết hợp phương pháp đề xuất và dịch ngược	**18.3**(+1.2)	**18.6**(+1.2)
Ja → Vi	**Hệ thống cơ sở** (**ja-kytea**)	14.1	15.1
	Phương pháp đề xuất (**ja-kytea**, $ths = 7$)	**16.8**(+2.7)	**18.0** (+2.9)
	Hệ thống cơ sở (**ja-spacy**)	14.9	15.9
	Phương pháp đề xuất (**ja-spacy**, $ths = 7$)	**17.6** (+2.7)	**18.4** (+2.5)
	Phương pháp dịch ngược (**ja-spacy**)	13.4 (-1.5)	14.3 (-1.6)
	Hệ thống cơ sở (**ja-mecab**)	14.0	15.4
	Phương pháp đề xuất (**ja-mecab**, $ths = 7$)	**18.1** (+4.1)	**19.4** (+4.0)
	Phương pháp dịch ngược (**ja-mecab**)	13.7(-0.3)	14.6(-0.8)
	Kết hợp phương pháp đề xuất ($ths = 7$)và dịch ngược (**ja-mecab**)	**15.4**(+1.4)	**16.6**(+1.2)

Bước 4. Nối tập dữ liệu tổng hợp với tập dữ liệu chuẩn để huấn luyện hệ thống dịch NMT.

2.3 Thực nghiệm và kết quả

2.3.1 Tập dữ liệu và tiền xử lý

Bảng 2.5 trong luận án thống kê chi tiết về tập dữ liệu được sử dụng trong thực nghiệm ở chương này.

2.3.2 Thiết lập hệ thống và huấn luyện

Tác giả sử dụng kiến trúc Transformer như mô tả trong mục 1.6.3.

2.3.3 Kết quả và phân tích

Các kết quả thực nghiệm được trình bày trong Bảng 2.6 và Bảng 2.7 sử dụng độ đo `SacreBLEU`. Phương pháp đề xuất đạt được sự cải thiện cao nhất +4.0 điểm BLEU và cho hiệu quả tương đương so với phương pháp dịch ngược.

Bảng 2.7. Kết quả thực nghiệm phương pháp tăng cường dữ liệu tổng hợp đề xuất trên tập dữ liệu **ALT**, với ngưỡng tần số thay thế là 7.

Chiều dịch	**Hệ thống dịch**	**Tập phát triển**	**Tập đánh giá**
Cn→ Vi	**Hệ thống cơ sở**	9.9	9.5
	Phương pháp đề xuất ($ths = 7$)	**11.8** (+1.9)	**11.6**(+2.1)
Ja → Vi	**Hệ thống cơ sở** (**ja-kytea**)	8.5	8.5
	Phương pháp đề xuất (**ja-kytea**, $ths = 7$)	**9.7**(+1.2)	**9.7**(+1.2)
	Hệ thống cơ sở (**ja-spacy**)	8.2	8.0
	Phương pháp đề xuất (**ja-spacy**, $ths = 7$)	**9.8** (+1.6)	**9.4** (+1.4)
	Hệ thống cơ sở (**ja-mecab**)	8.0	7.8
	Phương pháp đề xuất (**ja-mecab**, $ths = 7$)	**9.5** (+1.5)	**9.4**(+1.6)

2.4 Tóm tắt chương

Trong chương này, tác giả đề xuất cách tiếp cận tăng cường dữ liệu tổng hợp trong điều kiện hạn chế tài nguyên. Phương pháp đề xuất có một số ưu điểm so với các phương pháp trước đó không là yêu cầu thêm các tài nguyên khác như dữ liệu đơn ngữ, mô hình phân tích cú pháp, mô hình ngôn ngữ, từ điển song ngữ, ...

Chương 3

CẢI TIẾN CHẤT LƯỢNG DỊCH CHO CẶP NGÔN NGỮ HẠN CHẾ TÀI NGUYÊN DỰA VÀO DỊCH ĐA NGỮ

3.1 Đặt vấn đề

Dựa trên các đặc điểm về nguồn gốc từ vựng và bảng chữ cái, tác giả đề xuất hệ địch đa ngữ từ tiếng Trung, Nhật sang tiếng Việt và từ tiếng Anh, Pháp sang tiếng Việt với việc sử dụng các phương pháp phân đoạn khác nhau cho văn bản tiếng Nhật và các phương học tự động học các từ tương tự nhau trong không gian đa ngữ.

3.2 Hệ dịch từ tiếng Trung, Nhật sang tiếng Việt

3.2.1 Phương pháp đề xuất

Tác giả đề xuất sử dụng các phương pháp phân đoạn khác nhau cho văn bản tiếng Nhật với việc tận dụng mô hình pre-trained để đánh giá hiệu quả trong tình huống giàu tài nguyên.

3.2.2 Thử nghiệm và kết quả

3.2.2.1 Tập dữ liệu và tiền xử lý

Các tập dữ liệu được thống kê như trong Bảng 3.4 và Bảng 3.5 của luận án.

3.2.2.2 Thiết lập hệ thống và huấn luyện

Tác giả sử dụng kiến trúc Transformer với các thiết lập cơ bản như trong mục 1.6.3. Kích thước tập từ vựng là 60000 từ cho cả phía ngôn ngữ nguồn và ngôn ngữ đích. Để huấn luyện mô hình BERT, tác giả sử dụng mã nguồn được giới thiệu theo bài báo gốc của Devlin [1]. Các hệ thống dịch được huấn luyện sau 40 lần đi qua mỗi tập dữ liệu.

3.2.2.3 Kết quả và phân tích

Các kết quả thực nghiệm của phương pháp đề xuất được đánh giá thông qua độ đo SacreBLEU như Bảng 3.6, Bảng 3.7, Bảng 3.8 và Bảng 3.9.

3.3 Hệ dịch từ tiếng Anh, Pháp sang tiếng Việt

3.3.1 Phương pháp đề xuất

3.3.1.1 Phương pháp học các từ tương tự trong không gian đa ngữ

Phương pháp đề xuất được mô tả như sau:

Bước 1: Trích xuất danh sách các đơn vị dịch thuộc tiếng Anh, gọi là tập $\{A\}$ và k từ có tần số lớn nhất từ tập từ vựng của tiếng Pháp, gọi là tập $\{B\}$. Trong thực nghiệm, tác giả

1. https://github.com/google-research/bert

Bảng 3.6. Kết quả thực nghiệm trên hệ thống dịch đa ngữ từ tiếng Trung, Nhật sang tiếng Việt trên tập dữ liệu **TED Talks**.

STT	**Hệ thống**	**Cn → Vi**		**Ja → Vi**	
		dev	test	dev	test
1	Hệ thống cơ sở chỉ sử dụng song ngữ	17.1	17.4	14.9	15.9
2	Đa ngữ (ja-kytea)	18.2(+1.1)	**18.7**(+1.3)	15.2(+0.3)	16.4(+0.5)
3	Đa ngữ (ja-spacy)	**18.4** (+1.3)	18.2(+0.8)	14.7(-0.2)	16.3(+0.4)
4	Đa ngữ (ja-mecab)	18.0(+0.9)	18.0(+0.6)	**15.8**(+0.9)	**16.8**(+0.9)
5	Đa ngữ (ja-kytea) + Tăng cường dữ liệu tổng hợp ($ths = 7$)	17.9(+0.8)	18.4(+1.0)	16.8(+1.9)	17.9(+2.0)
6	Đa ngữ (ja-spacy) + Tăng cường dữ liệu tổng hợp ($ths = 7$)	**18.3**(+1.2)	**18.6**(+1.2)	**17.1**(+2.2)	**18.6** (+2.7)
7	Đa ngữ (ja-mecab) + Tăng cường dữ liệu tổng hợp ($ths = 7$)	17.7(+0.6)	18.4(+1.0)	16.8(+1.9)	18.5(+2.6)

Bảng 3.7. Kết quả thực nghiệm trên hệ thống dịch đa ngữ từ tiếng Trung, Nhật sang tiếng Việt trên tập dữ liệu **ALT**.

No.	**Hệ thống**	**Cn → Vi**		**Ja → Vi**	
		Tập phát triển	Tập đánh giá	Tập phát triển	Tập đánh giá
1	Hệ thống cơ sở (chỉ sử dụng song ngữ))	9.9	9.5	8.2	8.0
2	Đa ngữ (ja-spacy)	**10.3**(+0.4)	**10.3**(+0.8)	**9.2**(+1.0)	**9.4**(+1.4)
3	Đa ngữ (ja-spacy) + Tăng cường dữ liệu tổng hợp ($ths = 7$)	**11.4**(+1.5)	**11.3**(+1.8)	**10.4**(+2.2)	**9.9**(+1.9)

sử dụng $k = 15000$.

Bước 2: Tính toán độ tương tự giữa đơn vị dịch $t1_i$, $\forall\, t1_i \notin \{A \cup B\}$ và đơn vị dịch $t2_j$, $\forall\, t2_j \in \{A \cup B\}$ theo công thức 3.1.

$$score_i = min(d_j(e1_i, e2_j) \cdot e^{\cos(e1_i, e2_j)}) \tag{3.1}$$

với $j = 1..M$ và M là tổng số đơn vị dịch trong $A \cup B$; d là khoảng cách Euclid giữa vectơ embedding $e1_i$ của đơn vị dịch $t1_i$ và vectơ embedding $e2_j$ của đơn vị dịch $t2_j$. Khoảng cách Euclid được biểu diễn như công thức 3.2.

$$d_j(e1_i, e2_j) = \sqrt{\Sigma_{l=1}^{n}(x_l - y_l)^2} \tag{3.2}$$

Bảng 3.8. Kết quả thực nghiệm trên hệ thống dịch đa ngữ tích hợp mô hình BERT trên tập dữ liệu **TED Talks**.

STT	**Hệ thống**	**Cn → Vi**		**Ja → Vi**	
		Tập phát triển	Tập đánh giá	Tập phát triển	Tập đánh giá
1	Hệ thống cơ sở (chỉ sử dụng song ngữ)	17.1	17.4	14.9	15.9
2	Đa ngữ (ja-spacy) + BERT	16.9(-0.2)	17.2(-0.2)	**17.2**(+2.3)	**17.6**(+1.7)
3	Đa ngữ (ja-kytea) + BERT	**17.2**(+0.1)	**17.6**(+0.2)	**17.4**(+2.5)	**17.1**(+1.2)
4	Đa ngữ (ja-spacy) + BERT + Tăng cường dữ liệu giả ($ths = 7$)	16.5(-0.6)	17(-0.4)	**20.9**(+6.0)	**21.3**(+5.4)
5	Đa ngữ (ja-spacy) + BERT + Tăng cường dữ liệu giả ($ths = 7$) + Tiếp tục huấn luyện sau 70 epochs	15.8(-1.3)	16.1(-1.3)	**23.3**(+8.4)	**23.7**(+7.8)

Bảng 3.9. Kết quả thực nghiệm trên hệ thống dịch đa ngữ tích hợp mô hình BERT trên tập dữ liệu **ALT**.

No.	**Hệ thống**	**Cn → Vi**		**Ja → Vi**	
		Tập phát triển	Tập đánh giá	Tập phát triển	Tập đánh giá
1	Hệ thống cơ sở (chỉ sử dụng song ngữ)	9.9	9.5	8.2	8.0
2	Đa ngữ (ja-spacy) + BERT	**13.9**(+4.0)	**13.8**(+4.3)	**13.0**(+4.8)	**12.8**(+4.8)
3	Đa ngữ (ja-spacy) + BERT + Tăng cường dữ liệu tổng hợp ($ths = 7$)	**14.9**(+5.0)	**14.8**(+5.3)	**14.5**(+6.3)	**13.8**(+5.8)
4	Đa ngữ (ja-spacy) + BERT + Tăng cường dữ liệu tổng hợp ($ths = 7$) + Tiếp tục huấn luyện sau 70 epochs	**15.6**(+5.7)	**15.7** (+6.2)	**15.1**(+6.9)	**14.4**(+6.4)

với $e1_i = (x_1, x_2, ..., x_l)$ và $e2_j = (y_1, y_2, ..., y_l)$.

Giá trị $e^{\cos(e1_i, e2_j)}$ chủ yếu được sử dụng để chuẩn hóa kết quả tính toán của điểm $score_i$ trong công thức 3.1, trong đó, $e = 2.718$ được sử dụng để chuẩn hóa kết quả và $\cos(e1_i, e2_j)$ là cosin của góc giữa hai vectơ $e1_i$ và $e2_j$ và được tính như công thức 3.3.

$$\cos(e1_i, e2_j) = \frac{\Sigma_{l=1}^{n} x_l . y_l}{\sqrt{\Sigma_{l=1}^{n} x_l^2} \cdot \sqrt{\Sigma_{l=1}^{n} y_l^2}} \tag{3.3}$$

Bước 3: Tại mỗi bước huấn luyện, thay thế $t1_i$ bởi một trong các đơn vị dịch tương tự với nó (có số điểm $score_i$ nhỏ nhất). Quá trình thay thế được thực hiện tự động sau mỗi bước đi qua tập dữ liệu huấn luyện.

3.3.1.2 Phương pháp biến đổi vectơ từ phía ngôn ngữ nguồn

Dựa trên ý tưởng về khoảng cách vectơ, cách tiếp cận này giả sử vectơ embedding $e1_i$ của đơn vị dịch $t1_i$, $\forall\, t1_i \notin \{A \cup B\}$ được biến đổi đến một giá trị xấp xỉ như trong công thức 3.4.

$$e1_i = e1_i + d_i \tag{3.4}$$

với d là khoảng cách giữa vectơ $e1_i$ với vectơ trung bình của tất cả các vectơ $e2_j$ của đơn vị dịch $t2_j$, $\forall\, t2_j \in \{A \cup B\}$ như công thức 3.5.

$$d_i = e1_i - \frac{\sum_{j=1}^{j=M} e2_j}{M} \tag{3.5}$$

với M là tổng số đơn vị dịch trong $\{A \cup B\}$.

3.3.1.3 Áp dụng phương pháp tăng cường dữ liệu tổng hợp

Dữ liệu đơn ngữ tiếng Anh từ tập dữ liệu của Nghị viện Châu Âu được sử dụng để sinh dữ liệu tổng hợp.

3.3.2 Thử nghiệm và kết quả

3.3.2.1 Tập dữ liệu và tiền xử lý

Tập dữ liệu song ngữ Anh-Việt và Pháp-Việt được trích xuất từ miền TED Talks và được thống kê như trong Bảng 3.10 của luận án.

Trong quá trình tiền xử lý, tất cả các văn bản tiếng Anh, tiếng Pháp và tiếng Việt được tách từ và chuẩn hóa sử dụng các kịch bản từ công cụ Moses [2].

3.3.2.2 Thiết lập hệ thống và huấn luyện

Các thực nghiệm được tiến hành trên kiến trúc dịch Transformer với các thiết lập cơ bản như trong mục 1.6.3. Để áp dụng các phương pháp đề xuất, kiến trúc Transformer được chỉnh sửa để tích hợp các phương pháp trong mục 3.3.1.

3.3.2.3 Kết quả và phân tích

Độ đo `multi-BLEU` [3] được sử dụng để đánh giá chất lượng các hệ dịch. Các kết quả được trình bày trong Bảng 3.11.

3.4 Kết luận chương

Trong chương này tác giả đề xuất hai hệ dịch đa ngữ từ tiếng Trung, Nhật sang tiếng Việt và tiếng Anh, Pháp sang tiếng Việt.

2. https://github.com/moses-smt/mosesdecoder/tree/master/scripts
3. https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Bảng 3.11. Kết quả thực nghiệm trên hệ thống dịch đa ngữ từ tiếng Anh, Pháp sang tiếng Việt so với hệ dịch chỉ dựa trên song ngữ.

Chiều	Hệ thống	Tập phát triển	Tập đánh giá
Anh → Việt	Hệ thống cơ sở (chỉ sử dụng song ngữ))	31.74	35.13
	Đa ngữ	31.66 (-0.08)	**36.18 (+1.05)**
	Đa ngữ + làm mịn thông thường	**31.88 (+0.14)**	**36.56 (+1.43)**
	Đa ngữ + học từ tương tự	**31.93 (+0.19)**	**36.75 (+1.62)**
	Đa ngữ + biến đổi vectơ từ	**32.11 (+0.37)**	**36.74 (+1.61)**
	Đa ngữ + dữ liệu tổng hợp	30.86 (-0.88)	35.09 (-0.04)
Pháp → Việt	Hệ thống cơ sở (chỉ sử dụng song ngữ))	23.07	23.03
	Đa ngữ	**24.49 (+1.42)**	**24.22 (+1.19)**
	Đa ngữ + làm mịn thông thường	**24.51 (+1.44)**	**24.86 (+1.83)**
	Đa ngữ + học từ tương tự	**24.37 (+1.30)**	**24.70 (+1.63)**
	Đa ngữ + biến đổi vectơ từ	**24.60 (+1.53)**	**24.96 (+1.93)**
	Đa ngữ + dữ liệu tổng hợp	**25.59 (+2.52)**	**25.57 (+2.54)**
	Chỉ sử dụng dữ liệu tổng hợp	19.00	18.71

Với hệ dịch từ tiếng Trung, Nhật sang tiếng Việt tác giả đề xuất sử dụng các phương pháp phân đoạn khác nhau cho văn bản tiếng Nhật và chỉ ra các lợi thế đối với từng tác vụ dịch.

Với hệ dịch từ tiếng Anh, Pháp sang tiếng Việt, tác giả đề xuất hai kỹ thuật có thể áp dụng trong quá trình làm mịn hệ thống dịch đa ngữ.

Các tập dữ liệu sử dụng trong các thực nghiệm của chương này do tác giả và các cộng sự thu thập và được công bố cho mục đích nghiên cứu [4] .

4. https://github.com/ngovinhtn/Low-resource-Machine-Translation.git

Chương 4

CẢI TIẾN CHẤT LƯỢNG DỊCH CÁC TỪ HIẾM

Trong chương này, tác giả tập trung trình bày các đề xuất khác nhau để cải tiến việc dịch các từ hiếm trong hệ thống NMT.

4.1 Đặt vấn đề

Dịch các từ hiếm là một trong những thách thức lớn của dịch máy. Các nghiên cứu đã có chưa giải quyết triệt để vấn đề này.

4.2 Cải tiến quá trình giải mã thông qua việc chú thích các từ hiếm

4.2.1 Ý tưởng

Bước 1. Sinh từ điển dóng hàng giữa ngôn ngữ nguồn và ngôn ngữ đích. Tác giả sử dụng công cụ MGiza++ [1] cho mục đích này.

Bước 2: Gán nhãn cho các từ hiếm với một ký tự đặc biệt bắt đầu và kết thúc.

Bước 3. Quá trình giải mã thực hiện tính điểm $score_{ij}$ để ước lượng dóng hàng giữa từ hiếm thứ i trong câu đích với từ hiếm thứ j trong cầu nguồn như Công thức 4.2.

$$\text{score}_{\text{ij}} = \beta * \sum \text{P}_{\text{sw}_{\text{kj}}} + \gamma * \sum \text{P}_{\#_{\text{mj}}} + \epsilon * \sum \text{P}_{\text{a}_{\text{lj}}} \qquad (4.2)$$

Giải thuật 8 mô hình hoá các bước trong phương pháp đề xuất.

4.2.2 Thực nghiệm và kết quả

Tác giả tiến hành thực nghiệm phương pháp đề xuất trên hệ dịch song ngữ và đa ngữ. Các kết quả được đánh giá qua độ do SacreBLEU.

4.2.2.1 Tập dữ liệu và tiền xử lý

Chi tiết về các tập dữ liệu được thống kê như trong Bảng 4.3 của luận án. Đối với hệ dịch đa ngữ, tác giả trộn tất cả các tập dữ liệu như trong Bảng 4.3 để tạo thành tập huấn luyện 5.5 triệu cặp câu.

4.2.2.2 Hệ thống và huấn luyện

Trong thực nghiệm này, tác giả sử dụng công cụ mã nguồn `ViNMT` [2] cài đặt kiến trúc Transformer được phát triển từ một đề tài KC 4.0 với các thiết lập hệ thống cơ bản giống như mục 1.6.3.

4.2.2.3 Kết quả và phân tích

Các kết quả thực nghiệm được trình bày trong Bảng 4.4 và Bảng 4.5 trên điểm SacreBLEU.

1. https://github.com/moses-smt/giza-pp
2. https://github.com/KCDichDaNgu/KC4.0_MultilingualNMT

GIẢI THUẬT 8: Giải thuật tìm kiếm vị trí từ hiếm thứ j trong câu nguồn phù hợp với từ hiếm thứ i trong câu đích.

Input : *attn_cache* mảng chứa trọng số dóng hàng chéo từ mô hình dịch,
marked_src_sent: mảng α = [0 0 ... 1 1 1 ...] đánh dấu vị trí của từ hiếm như Hình 4.1,
word_query: mảng các sub-word của từ hiếm thứ i trong câu đích, bao gồm cả ký hiệu "#".

Output: Từ hiếm thứ j tốt nhất tương ứng với từ hiếm thứ i

$P_{sw_{kj}}[size]$ ={0}; $P_{\#_{mj}}[size]$ ={0}; $P_{a_{lj}}[size]$ = {0} // size số lượng từ hiếm trong câu nguồn
$\beta = 0.4$; $\gamma = 0.4$; $\epsilon = 0.2$
p_words = $spm(word\_query)$ // mảng các đơn vị dịch con của i
/* Tính $P_{sw_{kj}}$: */
for $i \leftarrow 0$ **to** $n - 1$ **do**
    piece = p_words[i]
    attn_pos = arg_max(attn_cache[piece])
    rare_word = marked_src_sent[attn_pos]
    if $rare\_word \neq 0$ **then**
        $P_{sw_{kj}}[rare\_word - 1] += 1$
/* Tính $P_{\#_{mj}}$: Hai ký hiệu "#" gần nhất với i */
for $i \leftarrow 0$ **to** 2 **do**
    attn_pos = arg_max(attn_cache[$\#_i$])
    rare_word = marked_src_sent[attn_pos]
    if $rare\_word \neq 0$ **then**
        $P_{\#_{mj}}$[rare_word-1] += 1
/* Tính $P_{a_{lj}}$: */
neighbor_word = get_neighbor(word_query) // Lấy hai từ lân cận gần nhất
while $neighbor\_word \neq None$ **do**
    $pos = attn\_cache[neighbor\_word_i]$
    $attn\_pos = arg\_max(pos)$
    $rare\_word = find\_nearest(attn\_pos)$ // Tìm từ hiếm gần nhất từ vị trí đang xem xét
    if $rare\_wordis \neq 0$ **then**
        $P_{a_{lj}}[rare\_word - 1] += 1$
$score_{ij} = \beta * P_{sw_{kj}} + \gamma * P_{\#_{mj}} + \epsilon * P_{a_{lj}}$
$rare\_word = arg\_max(score_{ij})$
return *rare_word*

Bảng 4.1. Điểm BLEU trên tập đánh giá ALT cho chiều dịch từ tiếng Trung sang tiếng Việt khi cải tiến quá trình giả mã.

STT	Hệ thống	150 ngàn cặp câu	2.5 triệu cặp câu
1	Hệ thống cơ sở	18.1	28.1
2	Mạng con trỏ theo Song và cộng sự (2019)	17.5	28.0
3	Phương pháp đề xuất	**18.3**	**29.1**

Bảng 4.2. Điểm BLEU trên hệ thống dịch đa ngữ khi được áp dụng phương pháp cải tiến quá trình giải.

STT	Chiều dịch	Hệ thống đa ngữ cơ sở	Phương pháp đề xuất
1	Anh → Việt	32.4	**34.0** (+1.6)
2	Trung → Việt	28.0	**29.8** (+1.8)
3	Lào → Việt	24.4	**25.1** (+0.7)
4	Khmer → Việt	28.9	**29.2** (+0.3)

4.3 Kết hợp vectơ từ, tách hình thái từ có giám sát và sử dụng cơ sở dữ liệu WordNet khi dịch từ hiếm

4.3.1 Kết hợp vectơ embediing trong câu nguồn

Tác giả đề xuất cải tiến công thức công thức kết hợp vectơ embedding trong câu nguồn tới xác suất đầu ra như Công thức 4.6.

$$p(y_j \mid y_{<j}, x) = \text{softmax}(W(z_j + l_j) + b) \qquad (4.6)$$

Các thực nghiệm và kết quả sẽ được trình bày trong mục 4.3.4.

4.3.2 Tách hình thái theo cách tiếp cận học có giám sát cho văn bản tiếng Anh

Ý tưởng chính của phương pháp là thu thập các phụ tố (gồm tiền tố và hậu tố) của ngôn ngữ đang xem xét, sau đó tách các phụ tố của các từ hiếm trong tập dữ liệu, trong khi đảm bảo thành phần của từ sau khi tách phụ tố phải tồn tại trong tập dữ liệu. Các kết quả thực nghiệm của phương pháp được trình bày trong mục 4.3.4.

4.3.3 Sử dụng quan hệ đồng nghĩa trong cơ sở dữ liệu WordNet

Ý tưởng của tác giả là thay thế một từ hiếm trong tập huấn luyện hoặc tập đánh giá bởi các từ đồng nghĩa với nó khi từ này xuất hiện trong tập dữ liệu. Các kết quả thực nghiệm được trình bày trong mục 4.3.4.

4.3.4 Thực nghiệm và kết quả

Độ đo `multi-BLEU` được sử dụng để đánh giá các thực nghiệm.

4.3.4.1 Tập dữ liệu và tiền xử lý

Cặp ngôn ngữ Nhật-Việt gồm 106758 cặp câu từ miền TED Talks, cặp Anh-Việt, gồm 133317 cặp câu từ hội nghị IWSLT 2015. Tác giả tiến hành thu thập các từ đồng nghĩa từ

Bảng 4.8. Kết quả thực nghiệm các hệ dịch trên cặp ngôn ngữ Nhật-Việt

STT	Hệ thống	Nhật → Việt	
		dev2010	tst2010
(1)	Hệ thống cơ sở	7.91	9.42
(2)	+ Vectơ từ nguồn	7.77	**9.96**
(3)	+ WordNet	**8.37**	**10.34**
STT	**Hệ thống**	**Việt → Nhật**	
		dev2010	tst2010
(1)	Hệ thống cơ sở	9.53 (9.53)	10.95 (10.99)
(2)	+ Vectơ từ nguồn	**10.51 (10.51)**	**11.37 (11.39)**

cơ sở dữ liệu WordNet song ngữ Anh-Nhật [3] và thu được 315850 cho tiếng Anh và 1419948 cho tiếng Nhật.

4.3.4.2 Hệ thống và huấn luyện

Các thực nghiệm được tiến hành trên kiến trúc RNN với các thiết lập cơ bản như trong mục 1.6.3. Các hệ thống được huấn luyện sau 16 lần đi qua tập huấn luyện với kích thước các mini-batch là 32.

4.3.4.3 Kết quả và phân tích

Bảng 4.8 và Bảng 4.9 trình bày các kết quả trên các hệ dịch cho hai cặp ngôn ngữ nêu trên.

Bảng 4.9. Kết quả thực nghiệm các hệ dịch trên cặp ngôn ngữ Anh-Việt

STT	Hệ thống	Anh → Việt	
		tst2012	tst2013
(1)	Hệ thống cơ sở	26.91 (24.39)	29.86 (27.52)
(2)	+ Vectơ từ nguồn	**27.41 (24.92)**	**30.41 (28.05)**
(3)	+ Sennrich's BPE	**26.96 (24.46)**	**30.10 (27.84)**
(4)	+ Tách hình thái	**27.16 (24.67)**	**30.60 (28.34)**
(5)	+ WordNet	**27.46 (24.99)**	**30.85 (28.54)**
(6)	Toan và Chiang (2017)	-	26.7
(7)	Huang và cộng sự (2018)	-	28.07
STT	**Hệ thống**	**Việt → Nhật**	
		tst2012	tst2013
(1)	Hệ thống cơ sở	27.97 (28.52)	30.07 (29.89)
(2)	+ Vectơ từ nguồn	**28.42 (29.04)**	**30.12 (29.93)**

4.4 Tóm tắt chương

Tác giả trình bày các phương pháp pháp khác nhau để cải tiến chất lượng dịch các từ hiếm như cải tiến quá trình giả mã, kết hợp vectơ embedding, tác hình thái từ có giám sát và sử dụng WordNet.

3. http://compling.hss.ntu.edu.sg/wnja/

Chương 5

ẢNH HƯỞNG CỦA PHÂN ĐOẠN TỪ LÊN CÁC HỆ THỐNG DỊCH DỰA TRÊN MẠNG NƠRON VÀ DỊCH MÁY THEO MIỀN

5.1 Đặt vấn đề

Việc sử dụng các phương pháp phân đoạn khác nhau ít nhiều cũng ảnh hưởng đến chất lượng dịch máy do làm thay đổi khả năng chia sẻ các đơn vị dịch cũng như độ dài câu. Trong phần này, tác giả đề xuất hệ dịch dựa trên ký tự và phương pháp phân đoạn văn bản tiếng Việt dựa trên cách tiếp cận dịch không giám sát.

5.2 Hệ thống dịch máy dựa trên ký tự

5.2.1 Ý tưởng

Tác giả đề xuất giả pháp dịch dựa vào ký tự. Giải pháp này không yêu cầu các công cụ phân đoạn từ hay các tài nguyên khác ngoài tập dữ liệu song ngữ sẵn có.

5.2.2 Sự khác biệt giữa kiến trúc Transformer và kiến trúc RNN khi dịch dựa trên các ký tự

Sự khác biệt giữa hai kiến trúc thể hiện trên ba khía cạnh là khả năng học đồng thời các biểu diễn và phân đoạn từ, mô hình hóa ngữ cảnh dài, khả năng song song hóa các tính toán.

5.2.3 Thực nghiệm và kết quả

Tác giả đánh giá các giả thuyết đã nêu trên tác vụ dịch từ Nhật-Việt và tiến hành so sánh chúng trên kiến trúc Transformer và RNN với các thiết lập cụ thể như sau:

5.2.3.1 Tập dữ liệu

Tác giả sử dụng bốn tập dữ liệu song ngữ Nhật-Việt: (1) TED Talks gồm 106758 cặp câu, (2) Tập ALT gồm 18088 cặp câu, (3) Tatoeba 2000 cặp câu và (4) Glosbe gồm 210 cặp câu, tập phát triển `dev2010` và tập đánh giá `tst2010`.

5.2.3.2 Tiền xử lý dữ liệu

Trong thực nghiệm này, `kytea` sử dụng cho việc tách từ các văn bản tiếng Nhật với 50000 thao tác BPE. Các văn bản tiếng Việt được tách từ sử dụng công cụ Moses hoặc `pyvi`.

5.2.3.3 Kiến trúc hệ thống

Các thực nghiệm được tiến hành trên hai kiến trúc dịch RNN và Transformer với các thiết lập cơ bản như trong mục 1.6.3.

5.2.3.4 Kết quả

Tác giả đánh giá chất lượng các hệ thống dịch sử dụng độ đo `multi-BLEU`, các kết quả như trong Bảng 5.3.

Bảng 5.3. So sánh kết quả thực nghiệm trên các hệ thống dịch từ tiếng Nhật sang tiếng Việt và ngược lại.

Nhật⇒Việt			
	Hệ thống	BLEU	Δ BLEU
(1)	Word2WordRecurrent	11.05	-2.29
(2)	Word2WordTransformer	11.72	-1.62
(3)	Char2CharRecurrent	10.06	-3.28
(4)	Char2CharTransformer	**13.34**	-
Việt⇒Nhật			
	Hệ thống	BLEU	Δ BLEU
(1)	Word2WordRecurrent	11.13	-3.92
(2)	Word2WordTransformer	13.07	-1.98
(3)	Char2CharRecurrent	9.61	-5.44
(4)	Char2CharTransformer	**15.05**	-

Các hệ thống dịch dựa trên ký tự đạt được kết quả tốt nhất trên tất cả các chiều dịch khi sử dụng kiến trúc Transformer trong khi không đòi hỏi sử dụng các công cụ tách từ hoặc tri thức bên ngoài.

5.3 Phân đoạn từ cho văn bản tiếng Việt sử dụng học không giám sát

Trong nghiên cứu này, tác giả đề xuất phương pháp tách từ cho văn bản tiếng Việt theo cách tiếp cận học không giám sát và thử nghiệm cho bài toán dịch máy.

5.3.1 Ý tưởng đề xuất

Phương pháp đề xuất được minh hoạ như trong Giải thuật 11.

5.3.2 Thực nghiệm và kết quả

5.3.2.1 Tập dữ liệu và tiền xử lý

Tác giả tiến hành thu thập dữ liệu từ miền TED Talks được trích xuất bởi WIT3 gồm 106758 cặp câu cho tập huấn luyện. Các tập đánh giá và tập phát triển đã được công bố trong hội nghị IWSLT năm 2010, trong đó tập phát triển dev2010 gồm 568 cặp câu, tập đánh giá tst2010 gồm 1220 cặp câu.

5.3.2.2 Hệ thống và huấn luyện

Tác giả huấn luyện hệ thống dịch máy sử dụng kiến trúc RNN với các thiết lập cơ bản như trong mục 1.6.3 với các tham số thay đổi như kích thước mỗi mini-batch là 64 cặp câu và độ rộng beam là 16.

5.3.2.3 Kết quả và phân tích

Tác giả đánh giá chất lượng của các hệ thống dịch sử dụng độ đo `multi-BLEU` từ Moses. Các kết quả được trình bày trong Bảng 5.4.

GIẢI THUẬT 11: Phương pháp tách từ sử dụng học không giám sát cho văn bản tiếng Việt.

```
Input: - T1: văn bản đầu vào để học các luật ghép cặp
- min_freq: ngưỡng tần số cho việc ghép cặp hai âm tiết
Output: T2: văn bản được tách từ
Function get_most_freq_pair (T1, min_freq)
dict =
dumpWord = Tập các ký tự đặc biệt, dấu câu
for line in text: do
    w1 = the_first_word
    for each word in line: do
        w2 = next_word
        if w1 or w2 not in dumpWord and w1 không phải là số: then
            dict[w1,w2] += 1
        w1 = w2
code_file=all_pairs_has_freq > min_freq
return code_file
Function do_BPE (T2, codes_file):
for each pair in codes_file: do
    original_word = pair[0] + " " + pair[1]
    replaced_word = pair[0] + "_" + pair[1]
    replace(T2, original_word,replaced_word)
    return T2
```

Bảng 5.4. Kết quả thực nghiệm trên hệ thống dịch NMT giữa tiếng Việt và tiếng Nhật trong phương pháp phân đoạn từ không giám sát cho văn bản tiếng Việt.

Tiếng Việt→Tiếng Nhật			
	Hệ thống	dev2010	tst2010
(1)	Hệ thống SMT cơ sở	-	8.73
(2)	Hệ thống NMT cơ sở	8.68	9.39
(3)	+ VNBPE	9.12	9.89
(4)	+ JPBPE	**9.74**	**11.13**
Tiếng Nhật →Tiếng Việt			
	System	dev2010	tst2010
(1)	Hệ thống SMT cơ sở	-	7.73
(2)	Hệ thống NMT cơ sở	6.85	8.18
(3)	+ VNBPE	7.36	8.47
(4)	+ JPBPE	7.77	9.04
(5)	+ Dịch ngược	8.25	9.39
(6)	+ Dịch trộn lẫn	**8.56**	**9.64**

5.4 Cải thiện chất lượng dịch máy theo miền

Trong phần này tác giả đề xuất hệ thống dịch kết hợp một số phương pháp để nâng cao chất lượng dịch từ tiếng Anh sang tiếng Việt trên miền tin tức của hội nghị VLSP năm 2020.

5.4.1 Tập dữ liệu đánh giá

Tác giả sử dụng các tập dữ liệu đơn ngữ và song ngữ được cung cấp bởi hội nghị VLSP năm 2020.

Bảng 5.6. So sánh kết quả thực nghiệm trong luận án với các hệ thống dịch khác trong VLSP năm 2020 khi sử dụng đánh giá tự động.

Hạng	**Đội thi**	**Điểm BLEU**	**Điểm TER**
1	Hệ thống của luận án	38.39	0.45
2	RD-VAIS	33.89	0.53
3	Bluesky	32.38	0.56
4	Lab-914	32.10	0.50
5	NLP-HUST	23.72	0.62
6	THORLab	2.53	-

Bảng 5.7. So sánh kết quả thực nghiệm trong luận án với các hệ thống dịch khác trong VLSP năm 2020 khi đánh giá bằng chuyên gia.

Hạng	Đội thi	Điểm
1	Lab-914	1.554
2	Hệ thống của luận án	1.327
3	RD-VAIS	0.864
4	Bluesky	0.536
5	NLP-HUST	-0.043
6	THORLab	-4.239

5.4.2 Thực nghiệm và kết quả

5.4.2.1 Tiền xử lý dữ liệu

Trong thực nghiệm này, các văn bản tiếng Anh và tiếng Việt được tách dấu câu và chuẩn hoá sử dụng kịch bản Moses.

5.4.2.2 Thiết lập hệ thống

Các thực nghiệm được tiến hành trên kiến trúc Transformer với các thiết lập cơ bản như trong mục 1.6.3.

5.4.2.3 Kết quả và phân tích

Các kết quả thực nghiệm được đánh giá thông quá các độ đo tự động Sacrebleu và TER và chuyên gia. Bảng 5.6 và Bảng 5.7 trình bày các kết quả so sánh. Hệ thống dịch của tác giả xếp thứ hai so với hệ thống dịch của "Lab-914" với điểm số khác biệt không nhiều nhưng tác giả chỉ sử dụng lượng dữ liệu rất nhỏ (hơn 800 ngàn cặp câu) so với "Lab-914" sử dụng (24 triệu cặp câu).

5.5 Tóm tắt chương

Trong chương này, tác giả đề xuất hệ thống dịch dựa trên ký tự, phương pháp tách từ theo cách tiếp cận học không giám sát cho văn bản tiếng Việt, phương pháp kết hợp để tăng cường chất lượng dịch theo miền trong tác vụ dịch tin tức của hội nghị VLSP năm 2020.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong luận án, tác giả đề xuất một số cách tiếp cận khác nhau để cải tiến chất lượng dịch máy trong điều kiện hạn chế tài nguyên song ngữ trên các tác vụ dịch có liên quan đến tiếng Việt. Các đóng góp chính có thể được tóm tắt như sau:

Thứ nhất, tác giả đề xuất phương pháp sinh dữ liệu tổng hợp đơn giản và hiệu quả trong điều kiện hạn chế tài nguyên song ngữ. Phương pháp đề xuất có nhiều lợi thế so với các cách tiếp trước đó khi không đòi hỏi thêm các tài nguyên bên ngoài như mô hình dịch, mô hình ngôn ngữ được huấn luyện trước, từ điển song ngữ, bộ phân tích cú pháp, ... Các kết quả nghiên cứu được công bố trong công trình 1.

Thứ hai, tác giả đề xuất các hệ dịch đa ngữ cho các cặp ngôn ngữ có nhiều điểm tương đồng như hệ thống dịch từ tiếng Trung, Nhật sang tiếng Việt và từ tiếng Anh, tiếng Pháp sang tiếng Việt. Đối với hệ dịch từ tiếng Trung, Nhật sang tiếng Việt, tác giả đề xuất sử dụng các phương pháp phân đoạn từ khác nhau cho văn bản tiếng Nhật và chỉ ra các lợi của các tác vụ dịch với từng phương pháp. Đối với hệ dịch từ tiếng Anh, Pháp sang tiếng Việt, tác giả đề xuất hai phương pháp tăng cường dịch các từ hiếm trong không gian đa ngữ. Các thực nghiệm cho thấy các hệ dịch đa ngữ đạt được sự cải thiện đáng kể so với các hệ thống dịch dựa trên song ngữ trong điều kiện hạn chế tài nguyên. Các kết quả nghiên cứu được công bố trong các công trình 1 và 4.

Thứ ba, tác giả đề xuất các kỹ thuật khác nhau để nâng cao chất lượng dịch các từ hiếm trong điều kiện hạn chế tài nguyên bao gồm: cải tiến quá trình giải mã dựa trên các việc gán nhãn từ hiếm, kết hợp vectơ embedding trong câu nguồn, sử dụng tách từ có giám sát cho văn bản tiếng Anh và khai thác quan hệ đồng nghĩa từ cơ sở dữ liệu từ vựng WordNet. Các thực nghiệm được tiến hành trên các cặp ngôn ngữ hạn chế tài nguyên và đạt được những sự cải thiện đáng kể. Các kết quả nghiên cứu được công bố trong các công trình 2, 3 và 7.

Thứ tư, tác giả đề xuất hệ dịch dựa trên ký tự cho cặp ngôn ngữ Nhật-Việt và so sánh hiệu quả với các hệ dịch ở dựa vào mức từ và ký tự trên hai kiến trúc dịch phổ biến. Các thực nghiệm cho thấy kiến trúc Transformer đạt được những kết quả dịch tốt hơn kiến trúc RNN.khi dịch dựa trên ký tự so với dịch dựa từ hoặc sub-word. Bên cạnh đó, tác giả đề xuất phương pháp phân đoạn từ cho văn bản tiếng Việt sử dụng cách tiếp cận học không giám sát và đạt được hiệu quả dịch tương đương so với công cụ tách từ `pyvi`. Phương pháp của tác giả tuy có độ chính xác không cao nhưng có thể phù hợp với bài toán dịch máy, đồng thời có thể áp dụng rộng rãi trong các miền khác khi không đòi hỏi dữ liệu được gán nhãn trước như các phương pháp đã có. Các kết quả nghiên cứu được công bố trong các công trình 6, 8 và 9.

Thứ năm, tác giả đề xuất kết hợp các kỹ thuật đơn giản để cải thiện dịch máy theo miền trong tác vụ dịch tin tức của hội nghị VLSP 2020. Các thực nghiệm cho thấy, tác giả chỉ sử dụng một tập dữ liệu rất nhỏ (hơn 800 ngàn cặp câu) nhưng cho hiệu quả dịch gần tương

đương với hệ thống dịch sử dụng một lượng dữ liệu lớn rất nhiều lần (24 triệu cặp câu). Điều này giúp giảm đáng kể thời gian và chi phí huấn luyện hệ dịch. Các kết quả nghiên cứu được công bố trong công trình 5.

Thứ sau, tác giả đóng góp thêm một số tập dữ liệu song ngữ liên quan đến tiếng Việt cho mục đích nghiên cứu như Anh-Việt, Pháp-Việt, Trung-Việt và Nhật-Việt.

Hướng phát triển:

Trong các nghiên cứu tiếp theo, tác giả sẽ xem xét kỹ lưỡng hơn các tình huống cụ thể trong các phương pháp đề xuất như sử dụng lượng dữ liệu tổng hợp gấp nhiều lần hơn so với tập dữ liệu song ngữ sẵn có trong công trình số 1, áp dụng các phương pháp học các từ tương tự trong không gian đa ngữ cho nhiều cặp ngôn ngữ, cải tiến công thức tính toán độ đo khoảng cách và cosin giữa các vectơ như trong công trình số 4, đánh giá mức độ ảnh hưởng của nhiều từ xung quanh tới từ hiếm đang xem xét thay vì chỉ sử dụng hai từ liền kề gần nhất trong phương pháp cải tiến quá trình giả mã ở công trình số 2, xem xét việc thay thế các từ hiếm bởi nhiều từ đồng nghĩa dựa theo ngữ cảnh cụ thể thay vì chỉ sử dụng một từ có tần số cao nhất nhất như trong công trình 7.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC

1. **Thi-Vinh Ngo**, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha, and Le-Minh Nguyen. *An Efficient Method for Generating Synthetic Data For Low-Resource Machine Translation: An empirical study of Chinese, Japanese to Vietnamese Neural Machine Translation.* Applied Artificial Intelligence, Volume 36, Issue 1, 2022, Open Access, DOI: 10.1080/08839514.2022.2101755, Taylor & Francis, **SCI-E**.
2. Minh-Cong Nguyen-Hoang, **Thi-Vinh Ngo**, Van-Vinh Nguyen. *A Simple and Fast Strategy for Handling Rare Words in Neural Machine Translation.* In Proceedings of the 2st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop (SRW AACL-IJCNLP 2022). November 21-23, 2022, Taiwan, China, Online. Association for Computational Lingustics.
3. **Ngô Thị Vinh**, Nguyễn Phương Thái. *Nâng cao hiệu quả dịch từ hiếm cho cặp ngôn ngữ Trung-Việt và Nhật-Việt.* Hội thảo quốc gia lần thứ XXIV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông (VNICT 2021), Trang 325-330, Thái Nguyên, 13-14/12/2021. Nhà xuất bản khoa học kỹ thuật.
4. **Thi-Vinh Ngo**, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, and Le-Minh Nguyen. *Improving Multilingual Neural Machine Translation For Low-Resource Languages: French, English - Vietnamese.* In Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020), pages 55–61, Suzhou, China, Online. Association for Computational Linguistics.
5. **Thi-Vinh Ngo**, Minh-Thuan Nguyen, Minh Cong Nguyen Hoang, Hoang-Quan Nguyen, Phuong-Thai Nguyen, Van-Vinh Nguyen. *The UET-ICTU Submissions to the VLSP 2020 News Translation Task.* In Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing (VLSP 2020), pages 71–76, December 18, 2020, Hanoi, Vietnam. Association for Computational Lingustics.
6. **Thi-Vinh Ngo**, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. *How Transformer Revitalizes Character-based Neural Machine Translation: An Investigation on Japanese-Vietnamese Translation Systems.* In Proceedings of the 16th International Conference on Spoken Language Translation (IWSLT 2019), November 2-3, 2019, Hong Kong, China. Association for Computational Linguistics.
7. **Thi-Vinh Ngo**, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. *Overcoming the Rare Word Problem for Low-Resource Language Pairs in Neural Machine Translation.* In Proceedings of the 6th Workshop on Asian Translation (WAT 2019), pages 207–214, November 4, 2019, Hong Kong, China. Association for Computational Linguistics.
8. **Thi-Vinh Ngo**, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. *Combining Advanced Methods in Japanese-Vietnamese Neural Machine Translation.* 10th International Conference on Knowledge and Systems Engineering (KSE 2018), pages 318–322, November 1-3, 2018, Ho Chi Minh, Vietnam. Springer.
9. **Ngô Thị Vinh**, Nguyễn Phương Thái. *Dịch máy Nhật - Việt sử dụng mô hình mạng nơron học sâu.* Tạp chí khoa học và công nghệ Đại học Thái Nguyên, pages 9-14, Tập 178, số 02, 2018. Nhà xuất bản Đại học Thái Nguyên.