ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRẦN HỒNG VIỆT

CẢI TIẾN CHẤT LƯỢNG DỊCH MÁY THỐNG KÊ CHO CẶP NGÔN NGỮ ANH-VIỆT DỰA VÀO CÂY PHÂN TÍCH CÚ PHÁP PHỤ THUỘC

Chuyên ngành: Khoa học máy tính
Mã số: 62 48 01 01

TÓM TẮT LUẬN ÁN

Hà Nội - 2018

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

Người hướng dẫn khoa học:

1. TS.Nguyễn Văn Vinh
2. PGS.TS. Nguyễn Lê Minh

Mở đầu

1. Tính cấp thiết của luận án

Vấn đề quan trọng của dịch máy liên quan đến việc làm thế nào để sinh ra thứ tự các từ (cụm) chính xác trong ngôn ngữ đích. Trong hệ dịch máy thống kê dựa trên cụm từ (PBSMT), việc đảo cụm từ vẫn còn đơn giản và chất lượng chưa cao. Bên cạnh đó, do các ngôn ngữ có nhiều đặc điểm khác nhau dẫn tới không thể mô hình hóa chính xác trong quá trình dịch.

Phương pháp tiền xử lý với cách tiếp cận tổ hợp có ưu điểm là giữ được điểm mạnh của hệ thống dịch máy dựa trên cụm từ, giảm thiểu thời gian giải mã, cũng như giữ điểm mạnh của dịch máy theo cú pháp trong bài toán đảo trật tự từ. Những vấn đề thách thức đặt ra:

- Một số nghiên cứu đã áp dụng đảo trật tự từ dựa trên cây cú pháp phụ thuộc cho chiều Anh-Việt. Tuy nhiên những nghiên cứu này chủ yếu dùng các luật bằng tay, chưa áp dụng các luật tự động trong bài toán dịch.

- Ít nghiên cứu sử dụng tiền xử lý dựa vào cây cú pháp phụ thuộc, tồn tại nhiều hạn chế cần cải tiến để nâng cao chất lượng.

Với ưu điểm của cấu trúc cây phân tích phụ thuộc trong việc thể hiện quan hệ phụ thuộc từ, tốc độ nhanh, phù hợp với vấn đề sắp xếp lại trật tự từ, luận án tập trung nghiên cứu đề tài: *"Cải tiến chất lượng dịch máy thống kê cho cặp ngôn ngữ Anh-Việt dựa vào cây phân tích cú pháp phụ thuộc"*.

2. Mục tiêu của luận án

- Nghiên cứu các phương pháp giải quyết bài toán đảo cụm từ trong dịch máy thống kê dựa vào cụm theo hướng tiếp cận tiền xử lý.

- Xây dựng, mở rộng các luật thủ công và phát triển các luật tự động áp dụng để cải thiện chất lượng dịch máy thống kê.

- Nghiên cứu hệ thống dịch thống kê Moses, tích hợp tri thức ngôn ngữ, đề xuất phương pháp mới, thực nghiệm.

3. Đóng góp của luận án

- Nghiên cứu các hiện tượng ngôn ngữ, đề xuất các luật đảo trật tự từ thủ công từ việc lựa chọn đặc trưng về ngôn ngữ trên cây cú pháp phụ thuộc.
- Đề xuất phương pháp sử dụng đa phân lớp trong học máy để giải quyết bài toán sắp xếp lại trật tự từ. Các luật được học tự động từ ngữ liệu.
- Đề xuất phương pháp sử dụng mạng nơ-ron để giải quyết bài toán sắp xếp lại câu nguồn theo thứ tự từ câu đích.
- Phân tích ảnh hưởng của các lỗi phân tích cú pháp đến chất lượng dịch qua việc áp dụng các luật sắp xếp lại trật tự từ phía câu nguồn.

Kết quả nghiên cứu được công bố trong 10 công trình: 08 báo cáo trong kỷ yếu của hội nghị quốc tế có phản biện; 01 báo cáo trong kỷ yếu của hội thảo quốc gia có phản biện; 01 bài báo ở tạp chí trong nước có phản biện.

4. Bố cục của luận án

- **Chương 1** Tổng quan các vấn đề liên quan luận án.
- **Chương 2** Phương pháp dựa vào luật thủ công cho bài toán đảo trật tự từ trong dịch máy thống kê.
- **Chương 3** Phương pháp sử dụng các luật tự động bằng học máy với đa phân lớp.
- **Chương 4** Phương pháp sử dụng mạng nơ-ron kết hợp các thông tin ngữ cảnh.
- **Chương 5** Ảnh hưởng của cây phân tích cú pháp phụ thuộc và xây dựng hệ thống thử nghiệm.

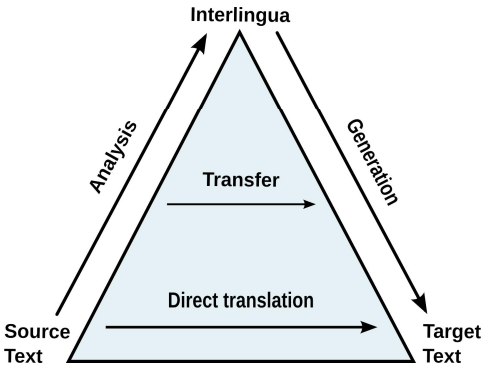
Chương 1

Tổng quan các vấn đề liên quan luận án

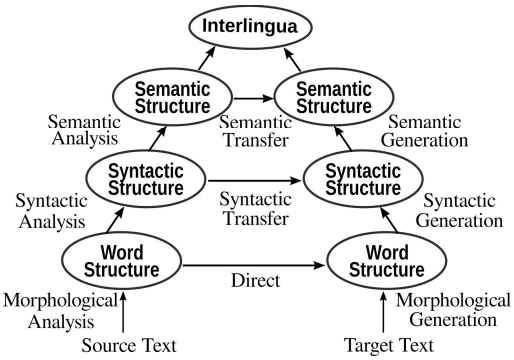
Chương này trình bày tổng quan về các vấn đề nghiên cứu trong luận án, bao gồm: dịch máy (Machine Translation - MT), dịch máy thống kê (Statistical Machine Translation - SMT), mô hình dịch máy dựa trên cụm từ, phân tích cú pháp, cú pháp phụ thuộc, các nghiên cứu liên quan, đưa ra vấn đề còn tồn tại mà luận án sẽ tập trung giải quyết.

1.1 Lịch sử dịch máy

Dịch là một quá trình chuyển nghĩa của các từ hay văn bản sang ngôn ngữ khác, liên quan đến việc giải mã nghĩa của ngôn ngữ nguồn và sau đó mã hóa lại theo nghĩa vào ngôn ngữ đích. Quá trình đòi hỏi kiến thức đầy đủ về ngôn ngữ bao gồm: hình thái học, cú pháp, ngữ nghĩa...



(a) Tháp chuyển đổi thể hiện quá trình dịch theo các phương pháp khác nhau



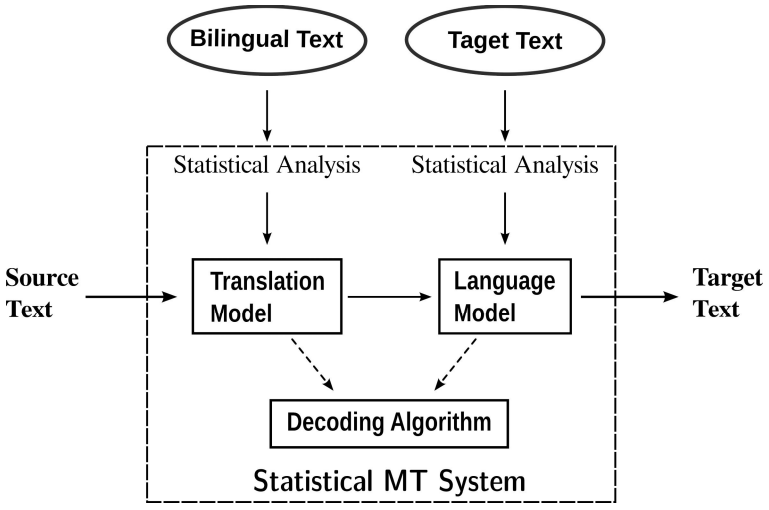
(b) Tháp chuyển đổi thể hiện các kiểu phân tích trong sơ đồ hình tháp

Hình 1.1: Sơ đồ hình tháp thể hiện các hệ thống dịch máy khác nhau.

1.2 Tổng quan về dịch máy

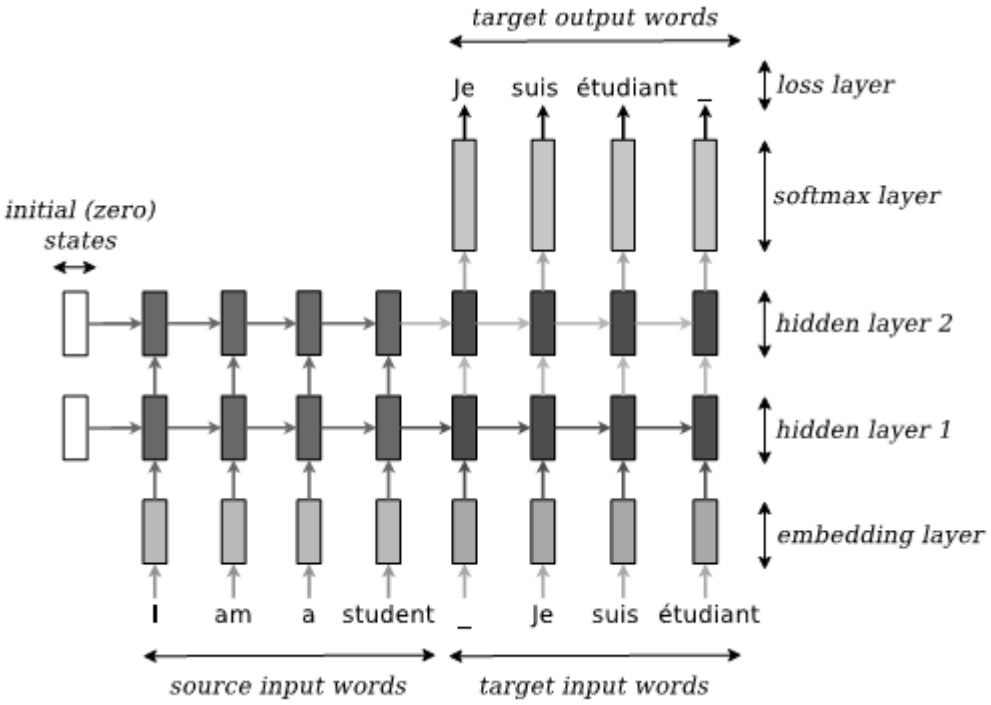
1.3 Dịch máy thống kê

Dịch máy thống kê (SMT) là một phương pháp tiếp cận của dịch máy dựa trên phân tích thống kê tập dữ liệu các cặp câu từ hai ngôn ngữ, ngữ liệu song ngữ.



Hình 1.2: Kiến trúc cơ bản của hệ thống dịch máy thống kê

1.4 Dịch máy mạng nơ-ron



Hình 1.3: Hệ thống dịch máy dựa trên mạng nơ-ron

1.5 Phân tích cú pháp phụ thuộc

1.6 Vấn đề đảo trật tự từ trong dịch máy

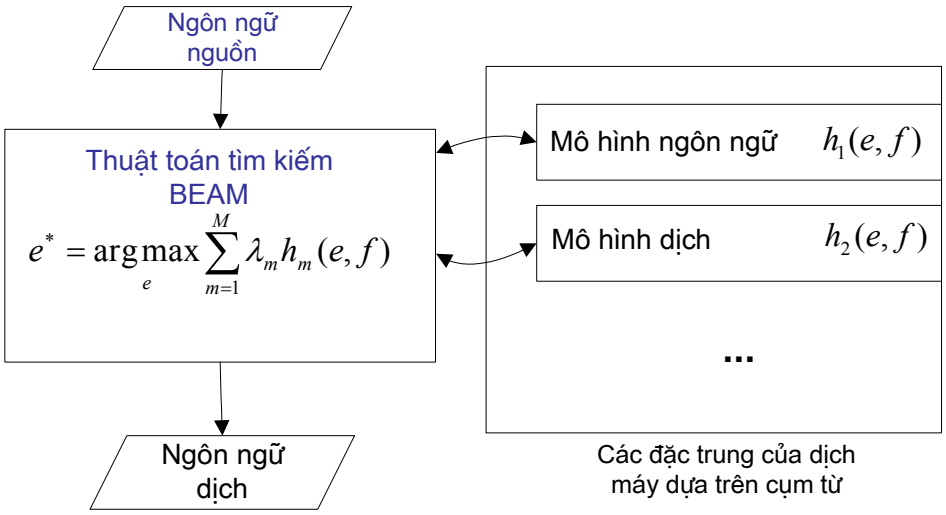
1.6.1 Sự khác nhau về thứ tự từ giữa các ngôn ngữ

1.6.2 Bài toán sắp xếp lại trật tự từ

Bài toán dịch máy thống kê gồm hai bài toán con: đoán định tập hợp từ trong bản dịch và xác định thứ tự của các từ dịch (bài toán sắp xếp lại).

1.7 Mô hình dịch máy dựa trên cụm từ

Kiến trúc của mô hình dịch dựa trên cụm từ trong hình 1.4



Hình 1.4: Kiến trúc của mô hình dịch dựa trên cụm từ

1.8 Các nghiên cứu liên quan

1.8.1 Sử dụng các luật thủ công cho vấn đề tiền xử lý

1.8.2 Sử dụng các luật tự động cho vấn đề tiền xử lý

1.9 Kết luận chương

Chương 2

Phương pháp dựa vào luật thủ công cho bài toán đảo trật tự từ trong dịch máy thống kê

Trình bày cách giải quyết vấn đề sắp xếp lại trật tự từ (đảo trật tự từ) dựa trên tiền xử lý cho bài toán dịch với kho ngữ liệu song ngữ Anh – Việt. Từ phân tích các thông tin trên cây cú pháp phụ thuộc và các hiện tượng ngôn ngữ, sử dụng các luật thủ công để giải quyết vấn đề đảo trật tự từ như bước tiền xử lý hệ thống dịch máy.

2.1 Vấn đề đảo trật tự từ trong dịch máy

Việc đảo trật từ từ dựa vào cây phân tích phụ thuộc và áp dụng các luật sắp xếp lại để tiến hành thay đổi thứ tự các từ.

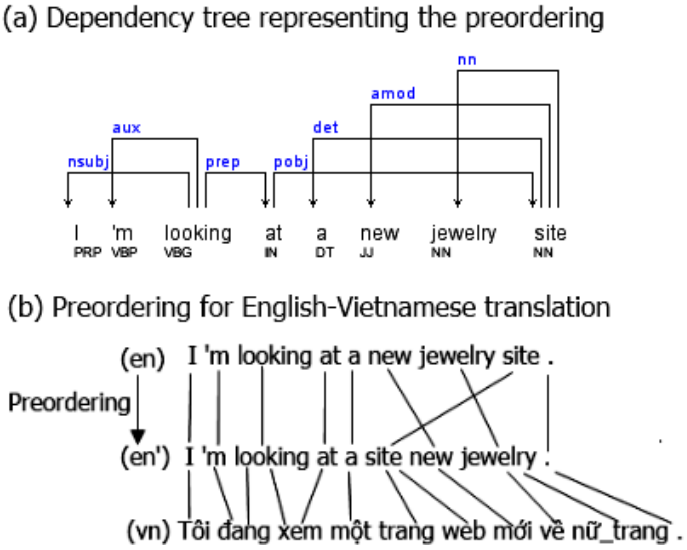
2.2 Các nghiên cứu liên quan

2.3 Dịch máy thống kê dựa trên cụm từ

Thực hiện dịch câu nguồn sang câu đích bằng cách chia câu nguồn thành các chuỗi cụm từ, mỗi cụm được dịch sang ngôn ngữ đích. Biểu diễn của quá trình qua công thức:

$$\hat{t} = \underset{t,a}{argmax} \sum_{i=1}^{n} \lambda_i f_j(s,t,a) \tag{2.3.1}$$

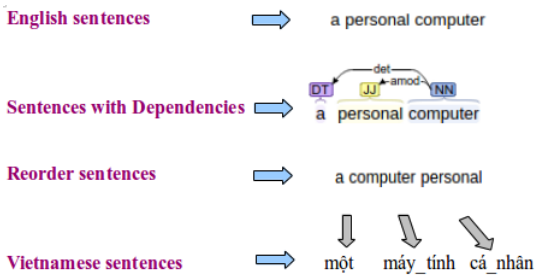
2.4 Tiền xử lý cú pháp phụ thuộc cho dịch máy thống kê



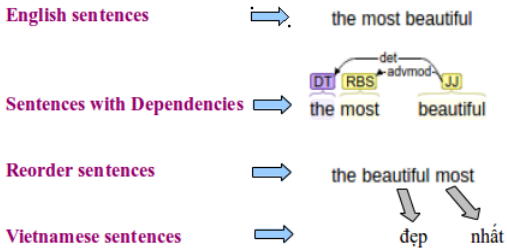
Hình 2.1: Ví dụ về tiền xử lý cho dịch Anh-Việt.

2.4.1 Phân tích hiện tượng ngôn ngữ và vấn đề sắp xếp lại

Tập trung vào việc phân tích các cấu trúc thông dụng nhất của tiếng Anh khi dịch sang tiếng Việt như trong hình 2.2 và hình 2.3.



Hình 2.2: Ví dụ về hiện tượng ngôn ngữ trong cụm danh từ với amod và det. Trong ví dụ này, danh từ "computer" được đảo với tính từ "personal"



Hình 2.3: Ví dụ về hiện tượng ngôn ngữ trong cụm tính từ với advmod và det

2.4.2 Luật chuyển đổi trật tự từ

Ánh xạ: $T \rightarrow (L, W, O)$

- T là từ loại của từ chính (nút cha) trong cụm trên cây cú pháp phụ thuộc.
- L là nhãn phụ thuộc (hay quan hệ phụ thuộc) của các nút con.
- W là trọng số để xác định thứ tự của nút con.
- O là dạng đảo (Normal: không đảo, Reverse: đảo).

T	(L, W, O)
VB*	(advcl, 2, Normal)
	(nsubj, 2, Reverse)
	(prep, -2, Normal)
	(dobj, -2, Normal)
	(prt, -2, Normal)
	(aux, 2, Reverse)
	(auxpass, 2, Normal)
	(neg, 2, Normal)
	(self, 2, Normal)
JJS, JJR, JJ	(advcl, -2, Normal)
	(aux, 2, Normal)
	(auxpass, 2, Normal)
	(neg, 2, Normal)
	(cop, 2, Normal)
	(self, -2, Normal)
NN, NNS, NNP	(prep, -2, Normal)
	(rcmod, -2, Normal)
	(amod, -2, Reverse)
	(self, 2, Normal)
IN, TO	(pobj, -2, Normal)
	(self, 2, Normal)

Hình 2.4: Các luật bằng tay cho việc sắp xếp lại từ tiếng Anh sang tiếng Việt sử dụng tiền xử lý cú pháp phụ thuộc.

2.4.3 Tập các luật đảo trật tự từ thủ công

2.5 Thực nghiệm về sử dụng các luật thủ công dựa trên tiền xử lý trong dịch máy

2.5.1 Tập dữ liệu và cài đặt thực nghiệm

2.5.2 Kết quả thực nghiệm

2.6 Kết luận chương

Sử dụng các luật thủ công để giải quyết vấn đề đảo trật tự từ. Áp dụng phương pháp tiền xử lý đem lại cân bằng giữa tốc độ, thời gian thực hiện và độ chính xác trong quá trình giải mã, nâng cao chất lượng dịch.

Pos-tags nút cha	Nhãn phụ thuộc nút con	Vị trí
VB	aux, nsubj, mark, self, advmod, advcl, discourse, neg, csubj, dep, preconj, expl, cop, appos	2
	rcmod, amod	1
	nsubjpass	0
	possessive	-1
	Tmod, iobj, acomp, prt, parataxis, xcomp, ccomp, vmod, cc, conj, prep, punct, dobj	-2
NN	det, self, nsubj, cop, advmod, mark, aux, neg, num, csubj, predet, discourse, advcl, expl, nsubjpass, tmod, preconj, auxpass, dobj, quantmod	2
	npadvmod	1
	prt	0
	number	-1
	xcomp, ccomp, parataxis, possessive, poss, dep, cc, nn, appos, vmod, rcmod, conj, prep, punct, amod	-2
JJS	self, nsubj, cop, advmod, mark, aux, advcl, csubj, poss, neg	2
	predet	1
	discourse	0
	dobj	-1
	possessive, xcomp, ccomp, amod, parataxis, appos, dep, cc, det, rcmod, conj, vmod, punct, prep	-2
TO	self, advmod, punct, preconj, neg, nsubj, mark, mwe, ccomp, discourse, dobj	2
	parataxis	1
	advcl	0
	xcomp	-1
	aux, tmod, prep, dep, cc, conj, pcomp, pobj	-2

Hình 2.5: Một khảo sát về vị trí từ loại và các nhãn trong việc sắp xếp lại thứ tự từ

Bảng 2.1: Thực nghiệm sử dụng các luật thủ công cho kho ngữ liệu song ngữ Anh-Việt

Hệ thống	**BLEU(%)**	**Mô tả**
System I	26.95	Áp dụng các luật với nhóm danh từ
System II	26.71	Áp dụng các luật với nhóm động từ
System III	27.15	Áp dụng các luật với nhóm tính từ và giới từ
System IV	27.26	Áp dụng các luật thủ công với toàn bộ các nhóm
Baseline	26.52	Hệ thống dịch trên cụm từ trong công cụ Moses

Chương 3

Phương pháp sử dụng các luật tự động bằng học máy với đa phân lớp

Trong chương này, trình bày cách giải quyết bài toán đảo trật tự từ như bước tiền xử lý cho bài toán dịch bằng cách mô hình hóa bài toán đảo trật tự từ với các phân lớp quan hệ thứ tự (vấn đề tiền xử lý dựa trên phân lớp): các luật đảo trật tự từ được sinh tự động từ dữ liệu, được nén thông tin tri thức, các đặc trưng ngôn ngữ vào mô hình học máy.

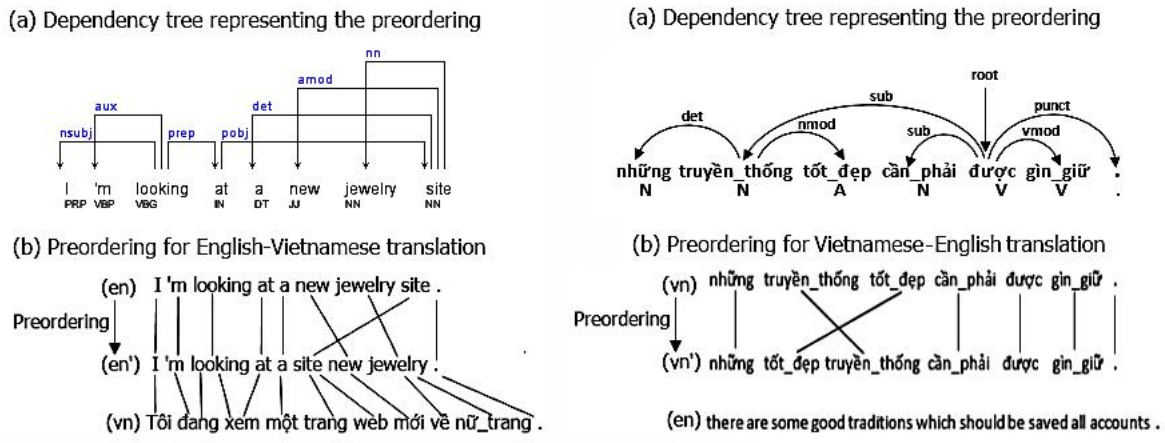
3.1 Tiền xử lý dựa trên phân lớp cho dịch máy dựa theo cụm

Từ những ưu điểm của học máy, chúng tôi đề xuất sử dụng kỹ thuật học máy trong việc giải quyết vấn đề đảo trật tự từ và áp dụng như quá trình tiền xử lý cho hệ thống dịch máy.

3.1.1 Vấn đề tiền xử lý dựa trên phân lớp

Xây dựng mô hình học máy có thể tự động thay đổi thứ tự các từ trong câu ngôn ngữ nguồn sang thứ tự tương ứng với câu ngôn ngữ đích.

3.1.2 Đặc trưng



Hình 3.1: Ví dụ về tiền xử lý cho ngữ liệu song ngữ Anh-Việt.

3.1.3 Mô hình phân lớp

- *Thuật toán 2.1*: Trích xuất tự động các luật với đầu vào bao gồm các cây phụ thuộc của các câu nguồn và cặp gióng hàng từ.

- *Thuật toán 2.2*: Tiến hành bằng cách xét tất cả các luật sau khi hoàn thành theo thuật toán 1 và các cây phụ thuộc phía nguồn để sinh câu mới.

Pos-tag nút cha	Pos-tag nút con	Nhãn phụ thuộc nút con	Thứ tự đảo	Tỉ lệ đảo
NNS	JJR	amod	1_0-0_1	361/612
NNS	NN	nn	1_0-0_1	2137/3116
NNS	PRP$	poss	1_0-0_1	4001/5763
NNS	JJ	amod	1_0-0_1	7485/10606
NN	JJR	amod	1_0-0_1	337/547
NN	PRP$	poss	1_0-0_1	9518/12877
NN	NN	nn	1_0-0_1	2325/3799
NN	DT	det	1_0-0_1	28682/53057
NN	JJ	amod	1_0-0_1	6591/8908
NNP	NNP	nn	1_0-0_1	3798/7099
NNP	DT	det	1_0-0_1	1025/2013
JJ	WRB	advmod	1_0-0_1	455/783
JJ	RBR	advmod	1_0-0_1	429/638
JJ	RBS	advmod	1_0-0_1	707/829
JJ	DT	det	1_0-0_1	865/1683
RB	DT	det	1_0-0_1	318/550

Hình 3.2: Thống kê về quan hệ giữa nút cha với nút con trên ngữ liệu song ngữ.

3.2 Thực nghiệm về phương pháp sử dụng phân lớp cho việc tiền xử lý trong dịch máy

3.2.1 Tập dữ liệu và cài đặt thực nghiệm

3.2.2 Kết quả thực nghiệm

Pos-tag nút cha	Pos-tag nút con thứ nhất	Nhãn phụ thuộc nút con thứ nhất	Pos-tag nút con thứ hai	Nhãn phụ thuộc nút con thứ hai	Thứ tự đảo	Tỉ lệ đảo
VBZ	EX	expl	NNS	nsubj	1_0_2-0_1_2	144/229
NN	DT	det	VB	dep	1_0_2-0_1_2	63/119
NNS	DT	det	JJS	amod	1_2_0-1_0_2	66/115
VBD	EX	expl	NN	nsubj	1_0_2-0_1_2	209/339
NNS	CD	num	JJ	amod	1_2_0-1_0_2	364/489
RB	RB	advmod	IN	mwe	1_0_2-0_1_2	81/121
NN	DT	det	VBP	rcmod	1_0_2-0_1_2	312/605
VBZ	EX	expl	NN	nsubj	1_0_2-0_1_2	872/1350
NN	CD	num	JJ	amod	1_2_0-1_0_2	117/171
NNS	JJ	amod	VBP	rcmod	1_0_2-0_1_2	67/125
NN	PRP$	poss	VB	vmod	1_0_2-0_1_2	171/292
NNS	DT	det	NNP	nn	1_2_0-1_0_2	96/170
NNS	CD	num	NN	nn	1_2_0-1_0_2	69/111
NNS	JJ	amod	VBG	vmod	1_0_2-0_1_2	80/128
NN	PRP$	poss	IN	prep	1_0_2-0_1_2	495/920
NNS	NN	nn	IN	prep	1_0_2-0_1_2	141/260
VBP	NN	nsubj	VB	xcomp	1_0_2-0_1_2	220/386
NN	DT	det	VBD	dep	1_0_2-0_1_2	62/119
IN	RB	advmod	VBG	pcomp	1_0_2-0_1_2	214/341
NN	JJ	amod	IN	prep	1_0_2-0_1_2	494/926
NN	NN	nn	IN	prep	1_0_2-0_1_2	121/229
NN	NN	poss	JJ	amod	1_2_0-0_2_1	64/110
NNS	PRP$	poss	IN	prep	1_0_2-0_1_2	112/210

Hình 3.3: Thống kê về quan hệ giữa nút cha với hai nút con trên ngữ liệu song ngữ.

Bảng 3.1: Hiệu năng cho tác vụ dịch Anh- Việt

Hệ thống	BLEU (%)
Baseline	26.52
Manual Rules	27.26
Auto Rules	27.09
Auto Rules + Manual Rules	27.34

3.3 Kết luận chương

Chương 4

Phương pháp sử dụng mạng nơ-ron kết hợp các thông tin ngữ cảnh

Trong chương này, trình bày nội dung, kết quả nghiên cứu về tiền xử lý cú pháp phụ thuộc cho bài toán dịch máy thống kê Anh-Việt sử dụng phương pháp học máy trong đó mạng nơ-ron dùng các thông tin ngữ cảnh từ word embedding.

4.1 Mô hình đảo dựa trên mạng nơ-ron sử dụng cây cú pháp phụ thuộc cho dịch máy thống kê

Hình 4.1 mô tả kiến trúc và các dữ liệu huấn luyện, trích xuất đặc trưng trong mô hình.

4.1.1 Đặc trưng cho phân lớp và huấn luyện mô hình

Phân lớp head-child

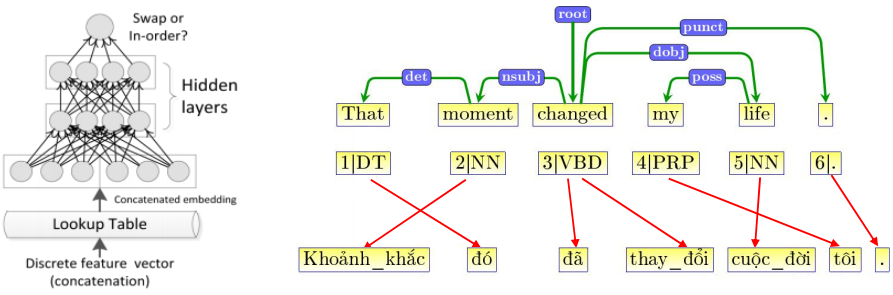
Phân lớp sibling

Các đặc trưng cho hai phân lớp như trong hình 4.2 và hình 4.3.

Lớp truyền thẳng

Mỗi đặc trưng được ánh xạ bởi việc tham chiếu bảng với biểu diễn véc tơ và các véc tơ kết quả được nối và đưa vào mỗi chuỗi các lớp ẩn (các ma trận trọng số) dùng hàm kích hoạt *sigmoid*:

$$\sigma(z) = \frac{1}{1 + e^{-x}} \tag{4.1.1}$$



(a) (b)

Pair	Head			child			Punctuation	Label
	x_h	$T(x_h)$	$L(x_h)$	x_c	$T(x_c)$	$L(x_c)$	$w(x_h,x_c)$	
(3,2)	changed	VBD	root	moment	NN	Nsubj	0	(-1; 0)
(3,5)	changed	VBD	root	Null	Null	Null	0	(0 ; 1)
(2,1)	moment	NN	nsubj	That	DT	det	0	(0 ; 1)
(5,4)	life	NN	dobj	my	PRP	poss	0	(-1; 0)

(c)

Pair	Left child			Right child			Head		Punctuation	Label
	x_l	$T(x_l)$	$L(x_l)$	x_r	$T(x_r)$	$L(x_r)$	x_h	$T(x_h)$	$w(x_l,x_r)$	
(2,5)	moment	NN	nsubj	life	V	dobj	changed	VBD	0	(0 ; 1)

(d)

Hình 4.1: Mô hình đảo cho dịch máy thống kê Anh-Việt sử dụng mạng nơ-ron với cây phân tích phụ thuộc: (a) Kiến trúc phân lớp mạng nơ-ron (b) Một gióng hàng câu từ ngữ liệu song ngữ Anh-Việt với các dữ liệu huấn luyện và đặc trưng được trích xuất cho: (c) phân lớp *cha-con* và (d) phân lớp *anh-em*.

Lớp ẩn đã cho chuyển đổi với véc tơ embedding x, véc tơ trọng số W và một giá trị bias b, đầu ra dự đoán δ xác định bởi:

$$z = W.x + b \tag{4.1.2}$$

$$\delta = tanh(z) \tag{4.1.3}$$

Huấn luyện mạng nơ-ron

$$L = -\frac{1}{T}\sum_{i=1}^{T} y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \tag{4.1.4}$$

Đặc trưng	Mô tả
Pair	Cặp từ với quan hệ nút cha-con
x_h	Từ nút cha x_h
$T(x_h)$	Part-of-speech (POS) tag của nút cha x_h
$L(x_h)$	Nhãn phụ thuộc $L(x_h)$ giữa x_h với nút cha của x_h
x_c	Từ của nút con x_c
$T(x_c)$	Part-of-speech (POS) tag của nút con x_c
$L(x_c)$	Nhãn phụ thuộc $L(x_h)$ giữa x_h với nút con x_c
$\omega\ (x_h, x_c)$	Giá trị logic $\omega(x_h, x_c)$ để chỉ nếu có dấu câu là con của nút cha x_h, tồn tại giữa nút cha x_h và nút con x_c
Label	Nhãn có giá trị trong khoảng -1 đến 1 để cho biết nút con ở bên trái hay bên phải hoặc giữ nguyên vị trí với nút cha.

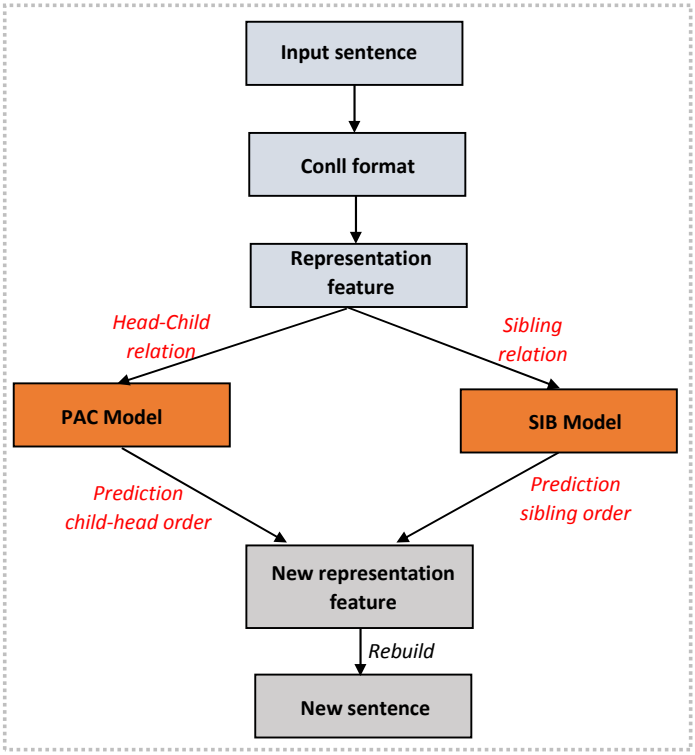
Hình 4.2: Các đặc trưng cho quan hệ *head-chlid* trong mô hình phân lớp

Đặc trưng	Mô tả
Pair	Cặp từ với quan hệ anh-em
x_l	Từ của nút bên trái x_l
$T(x_l)$	Part-of-speech (POS) tag của nút x_l
$L(x_l)$	Nhãn phụ thuộc $L(x_l)$ giữa nút x_l và x_h
x_r	Từ của nút bên phải x_r
$T(x_r)$	Part-of-speech (POS) tag của nút x_r
$L(x_r)$	Nhãn phụ thuộc $L(x_r)$ giữa nút x_r và x_h
x_h	Từ của nút cha x_h
$T(x_h)$	Part-of-speech (POS) tag của x_h
$\omega(x_l, x_r)$	Giá trị logic $\omega(x_l, x_r)$ để chỉ nếu có dấu câu là con của nút cha x_h, tồn tại giữa nút x_l và nút x_r
Label	Nhãn có giá trị trong khoảng -1 đến 1 cho biết nút con phải ở bên trái hay bên phải hoặc giữ nguyên vị trí so với nút con trái.

Hình 4.3: Các đặc trưng cho quan hệ sibling trong mô hình phân lớp

4.1.2 Khung làm việc cho đảo trật tự từ

Khung làm việc mô tả trong hình 4.4. Chúng tôi áp dụng *thuật toán 4.1* (Xây dựng mô hình huấn luyện) và *thuật toán 4.2* (Sắp xếp lại) trong khung làm việc của chúng tôi.



Hình 4.4: Khung làm việc cho quá trình tiền xử lý câu nguồn từ dữ liệu song ngữ Anh-Việt.

4.2 Thực nghiệm về phương pháp sử dụng mạng nơ-ron kết hợp thông tin ngữ cảnh

4.2.1 Tập dữ liệu và cài đặt thực nghiệm

4.2.2 Điểm BLEU

4.3 Phân tích và thảo luận

4.4 Kết luận chương

Bảng 4.1: Thống kê ngữ liệu

Corpus	Sentence pairs	Training Set	Development Set	Test Set
General	133403	131019	1304	1080
			Vietnamese	English
Training	Sentences		131019	
	Average Length		18.91	17.98
	Word		2481762	2360727
	Vocabulary		39071	54086
Development	Sentences		1304	
	Average Length		22.73	21.41
	Word		9092	8567
	Vocabulary		1537	1920
Test	Sentences		1080	
	Average Length		22.70	21.42
	Word		22707	21428
	Vocabulary		2882	3816

Bảng 4.2: Hiệu năng cho tác vụ dịch Anh- Việt

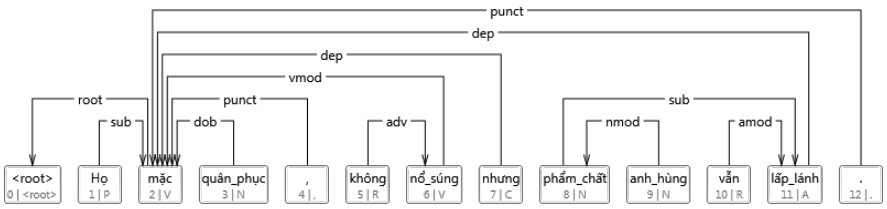
Hệ thống	BLEU (%)
Baseline	26.5
Manual Rules	27.12
Auto Rules	27.07
DPNN Classifier	27.16

Chương 5

Ảnh hưởng của cây phân tích cú pháp phụ thuộc và xây dựng hệ thống thử nghiệm

Trong chương này, thực hiện phân tích so sánh để quan sát hiệu quả của các lỗi phân tích cú pháp khác nhau đối với việc sắp xếp lại bằng cách kết hợp các phương pháp thực nghiệm và mô tả.

5.1 Phân tích cú pháp phụ thuộc

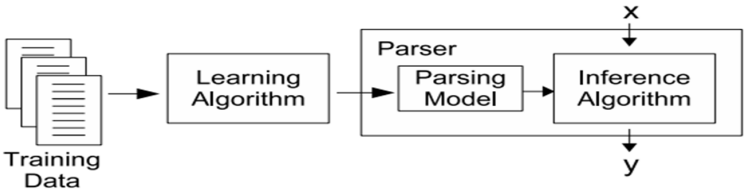


Hình 5.1: Biểu diễn đồ thị cây phân tích phụ thuộc với các nhãn quan hệ.

Theo quy ước phổ biến trong các tài liệu về cú pháp phụ thuộc thì mục từ nằm ở gốc của mũi tên là từ chính – gọi là head, mục từ nằm ở đầu mũi tên là từ phụ - gọi là dependent.

5.1.1 Bài toán phân tích cú pháp phụ thuộc

Bài toán tổng quát: Cho một câu, phân tích cú pháp đưa ra mô tả về quan hệ và vai trò ngữ pháp của các từ, cụm từ và hình thái của câu đó.



Hình 5.2: Mô hình bài toán tổng quát về phân tích cú pháp phụ thuộc

5.1.2 Định dạng dữ liệu theo chuẩn CoNLL

5.1.3 Sử dụng tập nhãn cho cú pháp phụ thuộc

5.2 Ảnh hưởng của lỗi phân tích cú pháp phụ thuộc tới chất lượng dịch máy

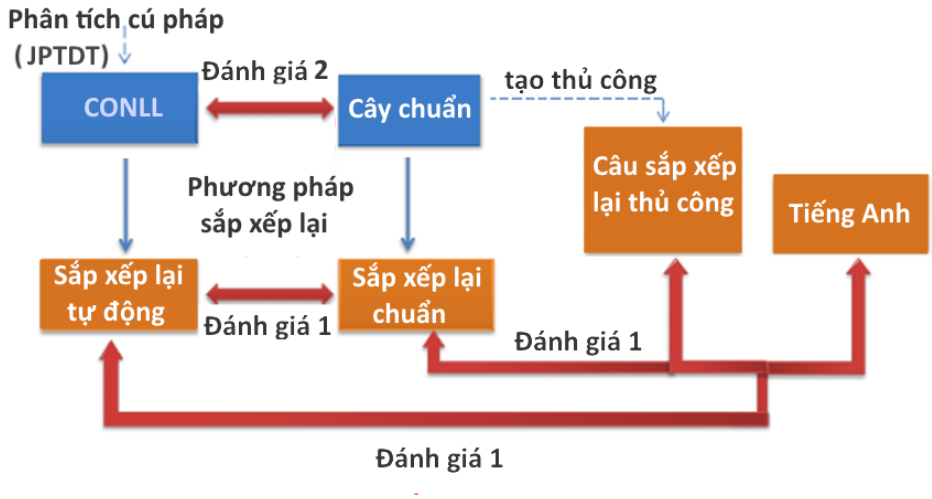
5.2.1 Phương pháp phân tích lỗi

- Do sự tương tự từ mốc chuẩn và câu được sắp xếp lại dựa trên Gold-Tree, cũng như giữa mốc chuẩn và câu được sắp xếp lại dựa trên từng cây cú pháp.
- Xác định mức độ lỗi phân tích cú pháp ảnh hưởng đến sắp xếp lại.

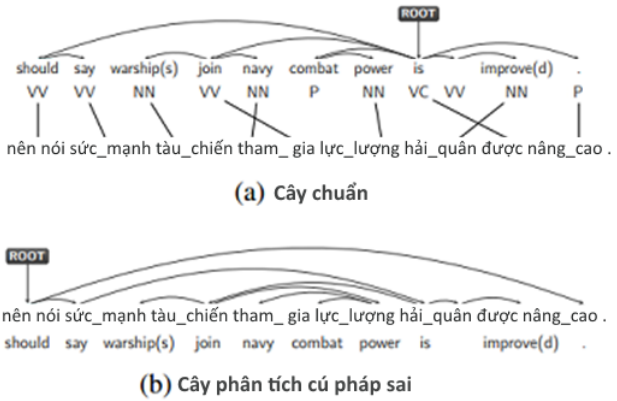
5.2.2 Đánh giá

Sử dụng độ đo Kendall's tau (τ) xếp hạng độ tương quan để đo độ tương tự thứ tự từ trong các cặp câu gồm dữ liệu chuẩn và dữ liệu được sắp xếp lại.

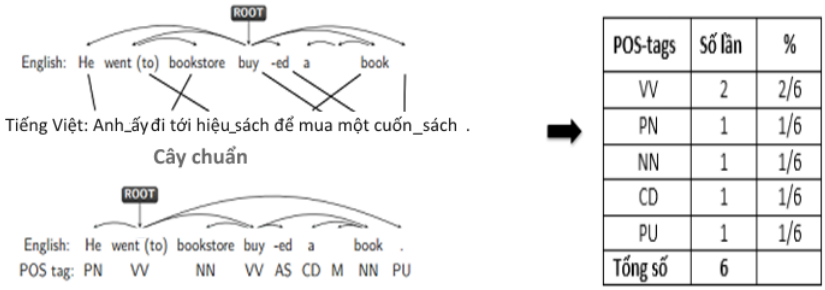
$$\tau = \frac{\#of\ concordant\ pairs}{\#of\ all\ pairs} \times 2 - 1 \tag{5.2.1}$$



Hình 5.3: Mô tả phương pháp phân tích lỗi.



Hình 5.4: Ví dụ về lỗi do xác định sai loại phụ thuộc nút gốc khi so sánh dữ liệu thống kê giữa cây được sinh ra với cây được sinh từ dữ liệu chuẩn.



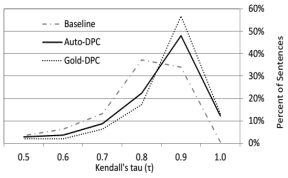
POS-tags	Số lần	%
VV	2	2/6
PN	1	1/6
NN	1	1/6
CD	1	1/6
PU	1	1/6
Tổng số	6	

Hình 5.5: Ví dụ về lỗi từ loại khi so sánh dữ liệu thống kê giữa cây được sinh ra với cây được sinh từ dữ liệu chuẩn.

Đánh giá 1: sử dụng tập các tiếng Anh được sắp xếp lại thủ công như điểm chuẩn và so sánh nó với tập các câu tiếng Anh được sắp xếp lại tự động.

Câu tiếng Anh được sắp xếp thủ công: That moment changed my life .

Câu tiếng Anh được sắp xếp tự động: moment that changed life my .



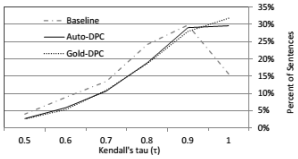
τ	Baseline	Auto-DPC	Gold-DPC
1	3	60	65
$1 \sim 0.9$	167	236	279
$0.9 \sim 0.8$	183	110	85
$0.8 \sim 0.7$	65	43	31
$0.7 \sim 0.6$	31	18	10
$0.6 \sim 0.5$	17	14	10
$0.5 \sim 0.4$	14	4	3
$0.4 \sim 0.3$	1	2	5
$0.3 \sim 0.2$	2	2	1
$0.2 \sim 0.1$	4	1	1
$0.1 \sim 0.0$	1	0	0
$-0.0 \sim -0.1$	0	0	0
$-0.1 \sim -0.2$	0	0	0
$-0.2 \sim -0.3$	0	1	1
$-0.3 \sim -0.4$	0	0	0

(a)

Đánh giá 2: sử dụng tập các câu tham chiếu tiếng Việt đóng vai trò điểm chuẩn và so sánh với tập các câu tiếng Anh được sắp xếp lại tự động.

Tiếng Anh: That moment changed my life .

Tiếng Việt: Khoảnh_khắc đó đã thay_đổi cuộc_đời tôi .



τ	Baseline	Auto-DPC	Gold-DPC
1	339	641	687
$1 \sim 0.9$	645	629	608
$0.9 \sim 0.8$	523	408	403
$0.8 \sim 0.7$	292	232	236
$0.7 \sim 0.6$	192	127	114
$0.6 \sim 0.5$	87	59	55
$0.5 \sim 0.4$	42	39	35
$0.4 \sim 0.3$	11	18	15
$0.3 \sim 0.2$	16	4	4
$0.2 \sim 0.1$	6	3	3
$0.1 \sim 0.0$	4	1	1
$-0.0 \sim -0.1$	4	2	2
$-0.1 \sim -0.2$	2	1	1
$-0.2 \sim -0.3$	0	0	0
$-0.3 \sim -0.4$	1	0	0

(b)

5.2.3 Phân tích nguyên nhân gây lỗi đảo trật tự từ

- Lỗi phụ thuộc: từ loại không phải là một phụ thuộc độc lập với nút cha.
- Lỗi nút cha: từ loại sai khi được nhận biết như nút cha.

5.3 Kết luận chương

Kết luận

Sắp xếp lại trật tự từ trong bước tiền xử lý như một phương pháp bổ sung có hiệu quả đối với các hệ thống dịch máy truyền thống, đóng vai trò quan trọng trong bản dịch.

1. Tóm lược các kết quả và đóng góp của luận án

Các kết quả và đóng góp bao gồm:

- Đề xuất các luật đảo trật tự từ thủ công bằng việc lựa chọn các đặc trưng về ngôn ngữ trên cây phân tích cú pháp phụ thuộc.
- Chúng tôi đề xuất luật đảo trật tự từ tự động. Với hai đề xuất gồm:
  - Khai thác các đặc trưng về ngôn ngữ và đề xuất phương pháp sử dụng đa phân lớp trong kỹ thuật học máy để giải quyết bài toán đảo trật tự từ như việc đoán nhận thứ tự đúng của ngôn ngữ của câu đầu vào tương ứng với thứ tự trong ngôn ngữ đích.
  - Đề xuất phương pháp sử dụng mạng nơ-ron để giải quyết bài toán sắp xếp lại câu nguồn theo thứ tự từ câu đích trước khi đưa vào hệ dịch để nâng cao chất lượng bản dịch.
- Đề xuất phân tích ảnh hưởng của các lỗi phân tích cú pháp đến chất lượng dịch qua việc áp dụng các luật sắp xếp lại trật tự từ phía câu nguồn.

2. Hạn chế và hướng phát triển của luận án

Mở rộng nghiên cứu của chúng tôi đến các cặp ngôn ngữ hoặc từng ngôn ngữ khác. Thử nghiệm phương pháp học tự động với kho ngữ liệu lớn, có độ phủ tốt để có thể xây dựng các luật bằng tay có chất lượng tốt cũng như học tự động để có các luật sắp xếp lại trật tự từ tốt hơn. Ngoài ra chúng tôi sẽ tiến hành sử dụng cách tiếp cận tích hợp vào hệ dịch máy mạng nơ-ron để có thể xây dựng hệ thống dịch tốt cho cả hai chiều dịch Anh-Việt, Việt-Anh.

Danh mục công trình khoa học của tác giả liên quan đến luận án

[1] Viet Hong Tran, Huyen Vu Thuong, Vinh Van Nguyen and Minh Le Nguyen, *"Dependency-based Pre-ordering For English-Vietnamese Statistical Machine Translation"*, In VNU Journal of Science: Computer Science and Communication Engineering, 2017, pages 175-179.

[2] Viet Hong Tran, Quan Hoang Nguyen and Vinh Van Nguyen *"A Neural Network Classifier Based on Dependency Tree English-Vietnamese Statistical Machine Translation"*, In Proceedings of the 19th International Conference on Intelligent Text Processing and Computational Linguistics, 2018. Available: http://site.cicling.org/2018/accepted.html

[3] Viet Hong Tran, Huyen Vu Thuong, Vinh Van Nguyen and Minh Le Nguyen, *"A Classifier-based Preordering Approach for English-Vietnamese Statistical Machine Translation"*, In Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics.

[4] Viet Hong Tran, Huyen Vu Thuong, Vinh Van Nguyen and Minh Le Nguyen, *"A Reordering Model For Vietnamese-English Statistical Machine Translation Using Dependency Information"*, In Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2016 IEEE RIVF International Conference on, pages 175-179.

[5] Viet Hong Tran, Vinh Van Nguyen and Minh Le Nguyen, *"Improving English-Vietnamese Statistical Machine Translation Using Pre-processing Dependency*

Syntactic", In Proceedings of the Pacific Association for Computational Linguistics 2015, pages 115-121.

[6] Viet Hong Tran, Huyen Vu Thuong, Vinh Nguyen Van and Trung Le Tien, *"The English-Vietnamese Machine Translation System for IWSLT 2015"*, In Proceeding of the 12th International Workshop on Spoken Language Translation, 2015, pages 80-84. Available: http://workshop2015.iwslt.org.

[7] Viet Hong Tran, Anh Tuan Pham, Vinh Van Nguyen, Hoai Xuan Nguyen, Huy Quang Nguyen, *"Parameter Learning for Statistical Machine Translation using CMA-ES"*, In Proceedings of the Sixth International Conference KSE 2014, Series: Advances in Intelligent Systems and Computing, Vol. 326, pages 251-259.

[8] Luan Nghia Pham, Viet Hong Tran, Vinh Van Nguyen, *"Vietnamese Text Accent Restoration with Statistical Machine Translation "*, Proceeding of 27th Pacific Asia Conference on Language, Information and Computation. Available: http://aclweb.org/anthology/Y13-1044

[9] Hoai Thu Vuong, Vinh Van Nguyen, Viet Hong Tran and Akira Shimazu, *"Improving Statistical Machine Translation with Processing Shallow Parsing"*, Proceeding of 26th Pacific Asia Conference on Language, Information and Computation. Available: http://www.aclweb.org/anthology/Y/Y12/Y12-1043.pdf

[10] Trần Hồng Việt, Vương Hoài Thu, Nguyễn Văn Vinh,Trần Lâm Quân, *"Áp dụng tiền xử lý cú pháp nông trong dịch máy thống kê"*, Kỷ yếu hội thảo Quốc gia lần thứ XV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông", trang 410-416.