

Học viện Công nghệ Bưu chính Viễn thông

Phan Thị Hà

**Nghiên cứu việc xây dựng, chuẩn hóa và khai thác kho
ngữ liệu từ nguồn Internet cho xử lý tiếng Việt**

Chuyên ngành: Truyền dữ liệu và mạng máy tính Mã số: 62.48.15.01

Nghiên cứu sinh: Phan Thị Hà

Cán bộ hướng dẫn: 1. PGS.TS Trần Hồng Quân 2. TS. Nguyễn Thị Minh Huyền

2014

LỜI CAM ĐOAN

Tôi cam đoan rằng nội dung của luận án này là kết quả nghiên cứu của bản thân. Tất cả những tham khảo từ các nghiên cứu liên quan đều được nêu rõ nguồn gốc một cách rõ ràng trong danh mục tài liệu tham khảo được đề cập ở phần sau của luận án. Những đóng góp trong luận án là kết quả nghiên cứu của tác giả đã được công bố trong các bài báo của tác giả ở phần sau của luận án và chưa được công bố trong bất kỳ công trình khoa học nào khác.

Tác giả luận án

Phan Thị Hà

Lời cảm ơn

Trước tiên, tôi muốn gửi lời cảm ơn đến Thầy Cô giáo hướng dẫn của tôi, PGS.TS Trần Hồng Quân, TS Nguyễn Thị Minh Huyền. Thầy, Cô đã tận tình chỉ bảo tôi từ những việc tưởng chừng đơn giản như cách thức thu thập tài liệu tham khảo cho đến phương pháp nghiên cứu khoa học. Sự tận tình hướng dẫn, cộng với sự động viên, khích lệ thường xuyên của Thầy Cô đã giúp tôi tự tin, say mê hơn trong con đường nghiên cứu khoa học. Tôi cảm thấy thực sự trưởng thành sau những năm được học tập và nghiên cứu dưới sự hướng dẫn của Thầy cô, một lần nữa em xin được cảm ơn Thầy Cô và kính chúc Thầy Cô luôn mạnh khỏe, hạnh phúc, thành công trên mọi lĩnh vực, đặc biệt là trên con đường khoa học. Mong rằng sẽ có nhiều lớp nghiên cứu sinh lại tiếp tục được Thầy Cô hướng dẫn trong những năm tiếp theo.

Tôi xin chân thành cảm ơn Tập đoàn Bưu Chính Viễn Thông, Ban lãnh đạo Học viện Công nghệ Bưu Chính Viễn Thông đã động viên và tạo điều kiện thuận lợi cho tôi trong suốt quá trình thực hiện luận án.

Tôi cũng muốn bày tỏ lòng biết ơn đối với tập thể các Thầy Cô khoa Công nghệ Thông tin và các Thầy Cô Khoa Đào tạo Sau Đại học, Học Viện Công nghệ Bưu Chính Viễn Thông, nơi tôi làm việc và học tập trong những năm qua. Các Thầy Cô luôn tạo điều kiện để tôi hoàn thành tốt công việc của mình, và sự dạy dỗ của Quý thầy cô đã giúp tôi trưởng thành.

Xin bày tỏ lời cảm ơn của tôi đến các chuyên gia về xử lý ngôn ngữ tự nhiên, dự án KC01.01/06-10, trung tâm từ điển VietLex đã hỗ trợ việc thu thập tài liệu và các góp ý hữu ích về ý tưởng và kỹ thuật phục vụ cho nghiên cứu của tôi.

Cuối cùng, chân thành cảm ơn người thân, bạn bè luôn bên cạnh động viên, hỗ trợ về mặt tinh thần để tôi vượt qua khó khăn và hoàn thành tốt luận án.

MỤC LỤC

MỤC LỤC	iii
DANH MỤC HÌNH VẼ	vi
DANH MỤC BẢNG	vii
DANH MỤC CỤM TỪ VIẾT TẮT	viii
MỞ ĐẦU	x
Đặt vấn đề	x
Mục tiêu và phạm vi nghiên cứu của luận án	xiii
Kết quả đạt được	xiv
Bố cục của luận án	xv
CHƯƠNG 1. TỔNG QUAN VỀ KHO NGỮ LIỆU.....	1
1.1 Kho ngữ liệu văn bản.....	1
1.2 Xây dựng, chuẩn hóa và khai thác kho ngữ liệu.....	5
1.2.1 Thu thập kho ngữ liệu văn bản	5
1.2.2 Chú giải ngôn ngữ và vấn đề chuẩn hóa.....	7
1.2.3 Khai thác kho ngữ liệu.....	9
1.3 Kho ngữ liệu tiếng Việt	13
1.3.1 Hiện trạng	13
1.3.2 Các vấn đề được nghiên cứu trong luận án.....	13
1.4 Kết chương.....	17
CHƯƠNG 2. XÂY DỰNG KHO NGỮ LIỆU THÔ TỪ INTERNET.....	18
2.1 Giới thiệu	18
2.2 Xây dựng kho ngữ liệu thô tiếng Việt	18
2.2.1 Lựa chọn danh sách từ hạt giống	19
2.2.2 Thu thập địa chỉ URL	21
2.2.3 Lọc nội dung chính của các trang web (URLs)	23
2.2.4 Phát hiện sự trùng lặp gần nhau.....	28
2.2.5 Xây dựng công cụ và kết quả thu thập kho ngữ liệu	32
2.3 Kết chương.....	32

CHƯƠNG 3. CHUẨN HÓA MÔ HÌNH CHÚ GIẢI TIẾNG VIỆT	34
3.1 Giới thiệu	34
3.2 Mô hình MAF của ISO/TC 37/SC 4.....	34
3.3 Mô hình SynAF của ISO/TC 37/SC 4	36
3.4 Chuẩn hóa theo mô hình MAF cho tiếng Việt.....	38
3.4.1 Xác định đơn vị cơ sở (segment)	41
3.4.2 Hình thái từ (Wordform).....	41
3.4.3 <i>Nội dung hình thái cú pháp</i>	42
3.5 Chuẩn hóa theo mô hình SynAF cho tiếng Việt	42
3.6 Kết chương.....	50
CHƯƠNG 4. KHAI THÁC KHO NGỮ LIỆU THÔ CHO NGHIÊN CỨU TỪ VỰNG TIẾNG VIỆT	51
4.1 Giới thiệu	51
4.1.1 Nghiên cứu từ vựng	51
4.1.2 Sketch Engine	52
4.1.3 Ngữ liệu trong Sketch Engine.....	53
4.2 Xây dựng ngữ liệu tiếng Việt cho Sketch Engine.....	56
4.2.1 Tách từ và gán nhãn từ loại.....	56
4.2.2 Xây dựng bộ quan hệ ngữ pháp tiếng Việt	57
4.2.3 Triển khai hệ thống Sketch Engine cho tiếng Việt.....	64
4.2.4 Đánh giá bộ quan hệ ngữ pháp tiếng Việt	67
4.3 Kết chương.....	67
CHƯƠNG 5. KHAI THÁC KHO NGỮ LIỆU CÓ CHÚ GIẢI CHO PHÂN TÍCH CÚ PHÁP TIẾNG VIỆT	69
5.1 Giới thiệu	69
5.2 Văn phạm hình thức.....	70
5.2.1 Khái niệm chung về văn phạm	70
5.2.2 Văn phạm phi ngữ cảnh (Context Free Grammar - CFG)	72
5.2.3 Văn phạm kết nối cây (Tree Adjoining Grammar – TAG).....	74
5.3 Trích rút tự động văn phạm CFG cho tiếng Việt.....	77

5.3.1	Thuật toán trích rút từ VietTreebank	77
5.3.2	Phân tích cú pháp tiếng Việt với văn phạm PCFG	86
5.3.3	Thử nghiệm và đánh giá	89
5.3.4	Nhược điểm của văn phạm PCFG trong phân tích ngữ pháp	90
5.4	Trích rút tự động văn phạm LTAG cho tiếng Việt	90
5.4.1	Thuật toán trích rút từ VietTreebank	90
5.4.2	Xây dựng thuật toán trích rút từ từ điển tiếng Việt.....	100
5.4.3	So sánh, đánh giá tập cây khởi tạo trích rút từ VietTreebank và từ điển...	105
5.5	Kết chương.....	107
KẾT LUẬN		109
DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA NGHIÊN CỨU SINH LIÊN		
QUAN ĐẾN LUẬN ÁN.....		112
TÀI LIỆU THAM KHẢO		113
PHỤ LỤC		125

DANH MỤC HÌNH VẼ

Hình 3. 1 Mô hình tổng quan của MAF [59]	35
Hình 3. 2. Mô hình SynAF [60].....	38
Hình 4. 1 Danh sách tần suất và tính trội của các từ lân cận với tính từ “đẹp”	65
Hình 4. 2. Phác thảo thông tin của 2 từ ”đẹp”, ”xinh”	66
Hình 4. 3. Một số danh sách các từ có quan hệ ngữ pháp với tính từ “đẹp”	67
Hình 5. 1 Biểu diễn văn phạm G dưới dạng cây	73
Hình 5. 2. Phép thay thế.....	75
Hình 5. 3. Phép kết nối	75
Hình 5. 4. Ví dụ về dẫn xuất với phép kết nối và phép thế trong văn phạm TAG	76
Hình 5. 5. Xử lý các cụm từ bằng thuật toán 5.5	94
Hình 5. 6. Ví dụ minh họa việc xây dựng cây phân tích	94
Hình 5. 7. Cây phân tích cú pháp.....	95
Hình 5. 8. Cây phân tích của cây cú pháp trong hình 5.7 sau khi chèn thêm nút.....	95
Hình 5. 9. Các mẫu cây cơ sở spine (ứng với quan hệ đối-vị từ) và phụ trợ (ứng với quan hệ phụ trợ hoặc đẳng lập)	96
Hình 5. 10. Các cây cơ bản.....	97
Hình 5. 11. Ghép các nút liên kết, đường đi trung tâm được đánh dấu bởi nét đôi.....	98
Hình 5. 12. Số mẫu cây tăng dần theo kích thước của Treebank:	100
Hình 5. 13. Sơ đồ so sánh tập cây cơ bản	105
Hình 5. 14. Một cây cơ bản không hợp lệ.....	106

DANH MỤC BẢNG

Bảng 1. 1. Thống kê các kho ngữ liệu đơn ngữ tiếng Việt	13
Bảng 2. 1. Thống kê số URL thu được của thuật toán 2.1.....	21
Bảng 2. 2. Tỷ lệ văn bản và thẻ xuất hiện trong phần nội dung chính của một số trang web tin tức Việt Nam	25
Bảng 2. 3. So sánh tỷ lệ “nội dung chính văn bản cần lấy/ toàn bộ nội dung văn bản trích rút được”	27
Bảng 2. 4. Kết quả thống kê thu thập tự động kho ngữ liệu từ web	32
Bảng 3. 1. Tập từ loại được đối sánh với danh mục phân loại dữ liệu chuẩn ISO 12620	40
Bảng 3. 2. Tập nhãn cú pháp thành phần, nhãn phân loại câu được đối sánh với danh mục phân loại dữ liệu chuẩn ISO 12620	44
Bảng 3. 3. Tập nhãn chức năng cú pháp đối sánh với danh mục phân loại dữ liệu chuẩn ISO 12620	45
Bảng 5. 1. Quá trình trích rút luật theo thuật toán 5.1	81
Bảng 5. 2.. Số các luật thu được	89
Bảng 5. 3. Bảng thành phần trung tâm cho treebank tiếng Việt	92
Bảng 5. 4.. Danh sách các đối.....	93
Bảng 5. 5. .Ghép một số nhãn cú pháp của VietTreebank thành một.....	98
Bảng 5. 6. Hai văn phạm G1, G2 được trích rút từ VietTreebank.....	100
Bảng 5. 7. Thống kê bộ cây cơ bản Spin từ từ điển so sánh với cây cơ bản của VietTreebank	105

DANH MỤC CỤM TỪ VIẾT TẮT

Cụm từ viết tắt	Cụm từ đầy đủ tiếng Anh	Dịch tiếng Việt
ANC	The American National Corpus	Kho ngữ liệu Quốc gia Mỹ
API	Application Programming Interface	Giao diện lập trình ứng dụng
BNC	The British National Corpus	Kho ngữ liệu Anh ngữ
BTE	Body Text Extraction	Trích văn bản phần thân
CES	Copus Encoding Standard	Tiêu chuẩn mã hóa kho ngữ liệu
COCA	The Copus of Contemporary American English	Kho ngữ liệu Anh Mỹ hiện đại
CRF	Conditional Random Field	Trường ngẫu nhiên có điều kiện
HMM	Hidden Markov Model	Mô hình Markov ẩn
HTML	HyperText Markup Language	Ngôn ngữ đánh dấu siêu văn bản
I/O	Input/Output	Đầu vào/ đầu ra
ISO	International Organization for Standardization	Tổ chức tiêu chuẩn hóa Quốc tế
LAF	Linguistic Annotation Framework	Khung chú giải ngôn ngữ học
LDC	Linguistic Data Consortium	Tổ chức dữ liệu ngôn ngữ học
MAF	Morphosyntactic Annotation Framewor	Khung chú giải hình thái cú pháp
MD5	Message Digest 5	Tóm tắt thông điệp
MDFA	Minimal deterministic finite state automata	Otomat hữu hạn trạng thái tối thiểu
MEM	Maximum Entropy Model	Mô hình Entropy cực đại
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên

POS	Part-Of-Speech	Từ loại
SGML	Standard Generalized Markup Language	Ngôn ngữ đánh dấu tổng quát hóa chuẩn
SynAF	Syntactic Annotation Framework	Mô hình chú giải cú pháp
URL	Uniform Resource Locator	Định vị tài nguyên đồng nhất
WFST	Weighted Finit State Transducer	Máy chuyển hữu hạn trạng thái có trọng số
WWW	World Wide Web	Mạng toàn cầu
XML	eXtensible Markup Language	Ngôn ngữ đánh dấu mở rộng
CFG	Context Free Grammar	Văn phạm phi ngữ cảnh
PCFG	Probability Context Free Grammar	Văn phạm phi ngữ cảnh kết hợp xác suất
TAG	Tree Adjoining Grammar	Văn phạm kết nối cây
LTAG	Lexicalized Tree Adjoining Grammar	Văn phạm kết nối cây từ vựng hóa
CYK	Cocke – Younger – Kasami algorithm	Thuật toán CYK
SSL	Semi-supervised learning	Học bán giám sát