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 hoa 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 Quí 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

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

DANH MỤC CỤM TỪ VIẾT TẮT

Cụm từ viết tắt	Cụm từ đầy đủ tiếng Anh	Dich 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	Worl 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 nội cây từ vựng hóa
CYK	Cocke – Younger – Kasami algorithm	Thuật toán CYK
SSL	Semi-supervised learning	Học bán giám sát