

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
VÀ TRUYỀN THÔNG

VŨ THỊ HẰNG

**CÁC PHƯƠNG PHÁP PHÂN ĐOẠN TIẾNG
VIỆT VÀ ỨNG DỤNG**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - Năm 2015

**ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
VÀ TRUYỀN THÔNG**

VŨ THỊ HẰNG

**CÁC PHƯƠNG PHÁP PHÂN ĐOẠN TIẾNG
VIỆT VÀ ỨNG DỤNG**

Chuyên ngành: **KHOA HỌC MÁY TÍNH**

Mã số: **60.48.01**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. BÙI VĂN THANH

Thái Nguyên - Năm 2015

MỤC LỤC

	Trang
LỜI CẢM ƠN	v
LỜI CAM ĐOAN	vi
DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT	vii
DANH MỤC BẢNG.....	viii
DANH MỤC HÌNH	ix
MỞ ĐẦU.....	1
Chương 1. TỔNG QUAN	6
1.1. KHÁI QUÁT VỀ TIẾNG VIỆT	6
1.1.1. Đặc điểm từ tiếng Việt.....	6
1.1.2. Các từ loại tiếng Việt.....	7
1.2. VẤN ĐỀ PHÂN ĐOẠN TIẾNG VIỆT	10
1.2.1. Từ vựng tiếng Việt.....	10
1.2.2. Tiếng – đơn vị cấu tạo lên từ.....	11
1.2.3. Cấu tạo từ.....	13
1.3. PHÂN ĐOẠN TỪ TIẾNG VIỆT BẰNG MÁY TÍNH	17
1.4. TỔNG KẾT CHƯƠNG.....	18
Chương 2. MỘT SỐ PHƯƠNG PHÁP PHÂN ĐOẠN VĂN BẢN TIẾNG VIỆT.....	19
2.1. MÔ HÌNH LRMM	19
2.1.1. Thuật toán Maximum Matching đơn giản.....	19

2.1.2. Thuật toán Maximum Matching phức tạp	19
2.2. PHƯƠNG PHÁP WFST (Weighted Finite-State Transducer)	20
2.3. MÔ HÌNH HỌC MÁY CRF	23
2.3.1. Định nghĩa CRF	23
2.3.2. Hàm tiềm năng của các mô hình CRF	26
2.3.3. Conditional Random Fields	26
2.4. TỔNG KẾT CHƯƠNG	28
Chương 3. BÀI TOÁN PHÂN ĐOẠN TIẾNG VIỆT	29
3.1. PHÁT BIỂU BÀI TOÁN	29
3.1.1. Cấu trúc chương trình	30
3.1.2. Tiền xử lý số liệu	32
3.1.3. Tách câu	34
3.1.4. Tách từ	36
3.1.5. Khử nhập nhằng	36
3.2. CÁC LOẠI NHẬP NHẰNG KHI TÁCH TỪ	36
3.2.1. Nhập nhằng do so khớp cực đại FMM/BMM sinh ra	37
3.2.2. Nhập nhằng theo một số loại khác	37
3.3. CÁCH KHỬ NHẬP NHẰNG	41
3.3.1 Cải tiến phương pháp so khớp cực đại	41
3.3.2 Khử nhập nhằng theo một số loại khác	43
3.4. TỔNG KẾT CHƯƠNG	50
Chương 4. THỬ NGHIỆM VÀ ĐÁNH GIÁ	52

4.1. KHO NGŨ LIỆU THỬ NGHIỆM VÀ CÁCH ĐÁNH GIÁ	52
4.2. QUY TRÌNH THỬ NGHIỆM.....	54
4.3. KẾT QUẢ THỬ NGHIỆM.....	55
4.4. GIAO DIỆN CHƯƠNG TRÌNH ỨNG DỤNG	56
KẾT LUẬN VÀ KIẾN NGHỊ.....	60
DANH MỤC TÀI LIỆU THAM KHẢO.....	62

LỜI CẢM ƠN

Em xin chân thành cảm ơn Ban Giám hiệu, Phòng Đào tạo Sau Đại học, Khoa Công nghệ Thông tin Trường Đại học công nghệ thông tin và truyền thông Thái Nguyên đã tận tình giúp đỡ, tạo mọi điều kiện thuận lợi cho em trong quá trình học tập, nghiên cứu và thực hiện luận văn.

Đặc biệt, em xin gửi lời tri ân sâu sắc đến TS Bùi Văn Thanh – người đã dành nhiều thời gian, công sức và tận tình hướng dẫn khoa học cho em trong suốt quá trình hình thành và hoàn chỉnh luận văn.

Xin chân thành cảm ơn Quý Thầy, Cô đã giảng dạy, truyền đạt cho em những tri thức quý báu, thiết thực trong suốt khóa học.

Cuối cùng xin bày tỏ lòng biết ơn đối với gia đình, người thân, bạn bè, đồng nghiệp đã giúp đỡ, động viên, đóng góp ý kiến quý báu cho em trong việc hoàn thành luận văn này.

Thái Nguyên, ngày tháng năm 2015

Tác giả

Vũ Thị Hằng

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn trực tiếp của TS. Bùi Văn Thanh.

Mọi trích dẫn sử dụng trong báo cáo này đều được ghi rõ nguồn tài liệu tham khảo theo đúng qui định.

Mọi sao chép không hợp lệ, vi phạm quy chế đào tạo, hay gian trá, tôi xin chịu hoàn toàn trách nhiệm.

Thái Nguyên, ngày tháng năm 2015

Tác giả

Vũ Thị Hằng

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

Tiếng Anh

Từ viết tắt	Tên đầy đủ	Diễn giải
BMM	Back Maximum Matching	Phương pháp so khớp cực đại lùi
CRFs	Conditional Random Fields	Trường ngẫu nhiên có điều kiện
FMM	Forward Maximum Matching	Phương pháp so khớp cực đại tiến
LRMM	Left Right Maximum Matching	Phương pháp so khớp cực đại
WEST	Weighted Finite State Transducer	Phương pháp chuyển dịch trạng thái hữu hạn

DANH MỤC BẢNG

	Trang
Bảng 1.1. Hệ thống các từ loại tiếng Việt theo sách giáo khoa Ngữ văn THCS.....	7
Bảng 1.2. Cấu trúc của tiếng trong tiếng Việt	12
Bảng 2.1. Trọng số theo từ.....	22
Bảng 4.1. Bảng số liệu các mục	53
Bảng 4.2. Kết quả thử nghiệm	55
Bảng 4.3. Kết quả phân đoạn	56

DANH MỤC HÌNH

	Trang
Hình 2.1. Đồ thị vô hướng không có chu trình	24
Hình 2.2. Đồ thị vô hướng mô tả cho CRF	25
Hình 2.3. Mô tả các hàm tiềm năng	26
Hình 3.1. Mô hình bài toán phân đoạn tiếng Việt	30
Hình 3.2. Cấu trúc chương trình phân đoạn tiếng Việt	31
Hình 4.1. Chọn chế độ lấy dữ liệu mẫu	52
Hình 4.2. Chương trình phân đoạn văn bản	54
Hình 4.3. Giao diện chính của chương trình	57
Hình 4.4. Chức năng phân đoạn văn bản	58
Hình 4.5. Kết quả sau khi phân đoạn văn bản	59