

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

VŨ CHÍ HIẾU

**NGHIÊN CỨU MÔ HÌNH NGÔN NGỮ N-GRAM CHO
TIẾNG VIỆT VÀ ỨNG DỤNG SỬA LỖI DẤU THANH
TRONG TIẾNG VIỆT**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2016

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

VŨ CHÍ HIẾU

**NGHIÊN CỨU MÔ HÌNH NGÔN NGỮ N-GRAM CHO
TIẾNG VIỆT VÀ ỨNG DỤNG SỬA LỖI DẤU THANH
TRONG TIẾNG VIỆT**

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 0101

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Giáo viên hướng dẫn: TS. VŨ TÁT THẮNG

Thái Nguyên – 2016

LỜI CAM ĐOAN

Tôi xin cam đoan, toàn bộ nội dung liên quan tới đề tài được trình bày trong luận văn là bản thân tôi tự tìm hiểu và nghiên cứu, dưới sự hướng dẫn khoa học của **TS. Vũ Tất Thắng** Viện công nghệ thông tin thuộc Viện Khoa học và Công nghệ Việt Nam.

Các tài liệu, số liệu tham khảo được trích dẫn đầy đủ nguồn gốc.

Thái Nguyên, ngày 20 tháng 3 năm 2016

Học viên

Vũ Chí Hiếu

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn tới trường Đại học CNTT&TT – Đại học Thái Nguyên đã tạo điều kiện và tổ chức khóa học này để tôi có thể có điều kiện tiếp thu kiến thức mới và có thời gian để hoàn thành Luận văn Cao học này.

Tôi xin được cảm ơn TS.Vũ Tất Thắng, người đã tận tình chỉ dẫn tôi trong suốt quá trình xây dựng đề cương và hoàn thành luận văn.

Tôi xin chân thành cảm ơn các thầy cô đã truyền đạt cho em những kiến thức quý báu trong quá trình học Cao học và làm Luận văn.

Tôi chân thành cảm ơn các bạn bè, anh chị em trong lớp cao học K13 đã giúp đỡ, đóng góp ý kiến chia sẻ những kinh nghiệm học tập, nghiên cứu trong suốt khóa học.

Cuối cùng tôi kính gửi thành quả này đến gia đình và người thân của tôi, những người đã hết lòng chăm sóc, dạy bảo và động viên tôi để tôi có kết quả ngày hôm nay.

Mặc dù tôi đã cố gắng hoàn thành Luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn không tránh khỏi những thiếu sót. Xin kính mong nhận được sự cảm thông và tận tình chỉ bảo của quý Thầy Cô và các bạn.

Thái Nguyên, ngày 20 tháng 3 năm 2016

Học viên

Vũ Chí Hiếu

MỤC LỤC

LỜI CAM ĐOAN	i
LỜI CẢM ƠN	ii
LỜI NÓI ĐẦU	1
CHƯƠNG I: MÔ HÌNH NGÔN NGỮ N-GRAM.....	3
1.1. Giới thiệu chung:	3
1.2. Công thức tính “xác suất thô”:	5
1.3. Vấn đề khó khăn khi xây dựng mô hình ngôn ngữ N-gram:.....	5
1.3.1. Phân bố không đều:.....	5
1.3.2. Kích thước bộ nhớ của mô hình ngôn ngữ:	6
1.4. Các phương pháp làm mịn:.....	6
1.4.1. Các thuật toán chiết khấu (discounting):	7
1.4.1.1. Phương pháp làm mịn Add-One:	7
1.4.1.2. Phương pháp làm mịn Witten - Bell:	9
1.4.1.3. Phương pháp làm mịn Good - Turing:	10
1.4.2. Phương pháp truy hồi:	11
1.4.3. Phương pháp nội suy:	12
1.4.4. Phương pháp làm mịn Kneser - Ney:	13
1.4.5. Phương pháp làm mịn Chen - GoodMan:.....	14
1.5. Kỹ thuật làm giảm kích thước dữ liệu:.....	15
1.5.1. Đồng hóa (Quantization):	16
1.5.2. Loại bỏ (pruning):	16
1.5.2.1. Cắt bỏ (cut-off):.....	17
1.5.2.2. Sự khác biệt trọng số (Weighted difference):	18
1.5.3. Nén (Compression):.....	19
1.6. Độ đo trong đánh giá mô hình:.....	19
1.6.1. Entropy - Độ đo thông tin:.....	19
1.6.2. Perplexity - Độ hỗn loạn thông tin:	21
1.6.3. Error rate - Tỷ lệ lỗi:	22
CHƯƠNG II: XÂY DỰNG N-GRAM CHO TIẾNG VIỆT	23
2.1. Giới thiệu:	23

2.2. Công cụ tách từ cho tiếng Việt - vnTokenizer:.....	23
2.3. Bộ công cụ SRILM:.....	27
2.3.1. N-gram-count:	27
2.3.2. N-gram:.....	29
2.4. Bộ công cụ trợ giúp xây dựng tập văn bản huấn luyện:	30
2.5. Phương pháp tách câu, tách từ, gán nhãn từ loại và phân tích cú pháp:.....	31
2.5.1. Tách câu:	31
2.5.2. Tách từ:.....	33
2.5.3. Gán nhãn từ loại:	36
2.5.4. Phân tích cú pháp:	38
2.6. Dữ liệu huấn luyện:.....	39
2.7. Kết quả xây dựng mô hình:.....	39
2.7.1. Số lượng các cụm N-gram với tiếng Việt dựa trên âm tiết:	39
2.7.2. Số lượng các cụm N-gram với tiếng Việt dựa trên từ:	40
2.8. Phân bố thống kê của tần số các cụm N-gram:.....	41
2.8.1. Với âm tiết.....	41
2.8.2. Với từ:.....	42
2.9. Phương pháp loại bỏ (Cut-off):	44
2.9.1. Với âm tiết.....	44
2.9.2. Với từ:.....	44
2.10. Các phương pháp làm mịn:.....	45
2.10.1. Với âm tiết:.....	45
2.10.2. Với từ:.....	45
CHƯƠNG III: ỨNG DỤNG N-GRAM TRONG BÀI TOÁN BÀI TOÁN SỬA	
LỖI DẤU THANH TRONG TIẾNG VIỆT	47
3.1. Tổng quan:	47
3.2. Bài toán sửa lỗi dấu thanh trong tiếng Việt:.....	48
3.2.1. Phát biểu bài toán:	48
3.2.2. Đặc điểm:	48
3.2.3. Hướng giải quyết:.....	49
3.3. Các hệ thống thêm dấu ứng dụng về N-gram đã có:	49
3.3.1. Công cụ AMPad:.....	49

3.3.2. VietPad:.....	50
3.4. Đề xuất hệ thống:.....	51
3.5. Cài đặt thử nghiệm và đánh giá hệ thống	54
KẾT LUẬN	58
HƯỚNG PHÁT TRIỂN CỦA ĐỀ TÀI	59

DANH MỤC ẢNH

Hình 2 - 1: Quy trình tách từ	24
Hình 2 - 2: Số lượng các cụm N-gram với âm tiết khi tăng kích thước dữ liệu	40
Hình 2 - 3: số lượng các cụm N-gram với từ khi tăng kích thước dữ liệu	41
Hình 2 - 4: Số lượng các cụm N-gram (âm tiết) có tần số từ 1 đến 10	42
Hình 2 - 5: Số lượng các cụm Ngram (từ) có tần số từ 1 đến 10	43
Hình 3 - 1: Thêm dấu tiếng Việt tự động bằng AMPad.....	50
Hình 3 - 2: Gõ tiếng Việt không dấu trên VietPad.....	51
Hình 3 - 3: Lưu đồ thực hiện của mô hình đề xuất	52
Hình 3 - 4: Giao diện chương trình	55
Hình 3 - 5: Chương trình thực hiện khi văn bản đầu vào hoàn toàn không có dấu.....	55
Hình 3 - 6: Chương trình thực hiện khi văn bản đầu vào có các từ có dấu xen kẽ	56

DANH MỤC BẢNG BIỂU

Bảng 2- 1: Số lượng các cụm N-gram trong văn bản huấn luyện với âm tiết	39
Bảng 2- 2: Số lượng các cụm N-gram trong văn bản huấn luyện với từ	40
Bảng 2- 3: Tần số của tần số các cụm N-gram áp dụng cho âm tiết.....	42
Bảng 2- 4: Tần số của tần số các cụm Ngram với từ	43
Bảng 2- 5: Bộ nhớ và độ hỗn loạn thông tin khi áp dụng loại bỏ trong âm tiết.....	44
Bảng 2- 6: Bộ nhớ và độ hỗn loạn thông tin khi áp dụng loại bỏ với từ	45
Bảng 2- 7: Độ hỗn loạn thông tin của các phương pháp làm mịn cho âm tiết	45
Bảng 2- 8: Độ hỗn loạn thông tin của các phương pháp làm mịn cho từ	46