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

DƯƠNG THỊ HÀ

XỬ LÍ BÀI TOÁN THÊM DẤU CHO TIẾNG VIỆT KHÔNG DẤU DỰA TRÊN NGHIÊN CỨU MÔ HÌNH NGÔN NGỮ N_GRAM

CHUYÊN NGHÀNH: KHOA HỌC MÁY TÍNH

Học viên thực hiện: Dương Thị Hà

Lớp: K9B

Giáo viên hướng dẫn: TS. Vũ Tất Thắng

2012

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 9 năm 2012

Học viên

Dương Thị Hà

LỜI CẢM ƠN

Trước tiên, 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 thày cô đã truyền đạt cho chúng tôi 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 K9 đã giúp đỡ, đóng góp ý kiến chia sẽ 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 9 năm 2012

Học viên

Dương Thị Hà

DANH MỤC HÌNH

Trang

DANH MỤC BẢNG

Trang

MỤC LỤC

Trang

MỞ ĐẦU

1. Lí do chọn đề tài

Ngôn ngữ tự nhiên là ngôn ngữ được con người sử dụng trong giao tiếp hàng ngày, nó khác hẳn với ngôn ngữ nhân tạo (ngôn ngữ lập trình, ngôn ngữ máy…). Việc làm cho máy tính hiểu được ngôn ngữ tự nhiên không phải dễ dàng. Để hiểu đúng nội dung của một văn bản viết bằng ngôn ngữ tự nhiên, trong quá trình đọc hay nghe thì thực tế là ta đã nhận thức được ngữ cảnh của văn bản đó. Mặt khác, ngôn ngữ tự nhiên có các bộ luật, cấu trúc ngữ pháp phong phú hơn nhiều so với các ngôn ngữ máy tính, để có thể xây dựng một bộ luật về ngữ pháp, từ vựng…, thật hoàn chỉnh để máy có thể hiểu ngôn ngữ tự nhiên là một việc rất tốn công sức và đòi hỏi người thực hiện phải có hiểu biết sâu sắc về ngôn ngữ học.

Mô hình ngôn ngữ (Language Model – LM) có thể cho biết xác suất một câu (hoặc cụm từ) thuộc một ngôn ngữ có xác suất sinh ra là bao nhiêu hay nói cách khác thì LM phản ánh một phân bố xác suất của các từ, cụm từ trên các tập văn bản.

Đòi hỏi tiên quyết, để máy tính xử lí ngôn ngữ tự nhiên chính là việc xây dựng mô hình ngôn ngữ, mà ngày nay mô hình thống kê thường được sử dụng bởi nó dựa trên các lí thuyết tường minh của xác suất thống kê để mô hình hóa ngôn ngữ, và thường đạt được độ chính xác cao trong các hệ thống thực tế. Xử lí ngôn ngữ tự nhiên dựa trên thống kê, không nhắm tới việc con người xây dựng mô hình ngữ pháp mà lập chương trình cho máy tính có thể "học" , nhờ vào việc thống kê các từ và cụm từ có trong văn bản. Trong các mô hình ngôn ngữ tiếng nói thì N-gram là một trong số những mô hình được sử dụng rộng rãi nhất.

Mô hình ngôn ngữ là một bộ phận quan trọng của lĩnh vực xử lí ngôn ngữ tự nhiên. Có nhiều lĩnh vực trong xử lí ngôn ngữ tự nhiên sử dụng LM như: kiểm lỗi chính tả, phát sinh câu ngẫu nhiên, dịch máy hay phân đoạn từ… Trên thế giới, đã có rất nhiều nước công bố nghiên cứu về LM áp dụng cho ngôn ngữ của họ nhưng ở Việt Nam, việc nghiên cứu và xây dựng một LM chuẩn cho tiếng Việt vẫn còn mới mẻ và gặp nhiều khó khăn.

Trong thực tế, sử dụng tiếng Việt không dấu đang trở thành thói quen không tốt của nhiều người Việt Nam trên Internet. Vì để gõ tiếng Việt có dấu đòi hỏi phải mất công sức, phải có font chữ, bộ gõ. Việc tự động thêm dấu và phân tích các từ này là vấn đề cần thiết và thú vị.

Chính điều này đã thúc đẩy tôi lựa chọn và tập trung "**Nghiên cứu mô hình ngôn ngữ N-gram và ứng dụng thêm dấu cho tiếng Việt không dấu**", để có thể tạo ra một trong những kết quả cơ bản nhất về xử lí ngôn ngữ nói chung, và có ích cho việc xử lí ngôn ngữ tiếng Việt vốn vô cùng phong phú của chúng ta nói riêng.

Ứng dụng của phương pháp thêm dấu là khá nhiều như: Thêm dấu cho các mail; cho các quản trị web, các trang web yêu cầu viết tiếng Việt nhưng người dùng không có sẵn bộ gõ; thêm dấu cho tin nhắn điện thoại…

2. Mục tiêu và nhiệm vụ

a) Mục tiêu: Do phạm vi bài toán khá lớn và thời gian làm luận văn là có giới hạn nên mục tiêu nghiên cứu của luận văn tập trung ở các điểm sau:

Về học thuật:

Đề tài này tập trung vào việc ứng dụng một số phương pháp tách từ, tiếng, phương pháp làm mịn trong mô hình ngôn ngữ N-gram nhằm tăng hiệu quả thêm dấu cho tiếng Việt không dấu.

Về phát triển và triển khai ứng dụng:

Kết quả của đề tài sẽ ứng dụng trong việc hỗ trợ trong việc thêm dấu cho tiếng Việt không dấu.