MỞ ĐẦU

Ngày nay, ở nước ta, việc soạn thảo văn bản bằng máy tính đã trở nên phổ biến. Tính thẩm mỹ và tốc độ cao đã quyết định nên điều đó. Nhưng các lỗi chính tả vẫn thường xuyên xuất hiện trong các văn bản. Điều đó làm cho các văn bản mất giá trị và cũng như sẽ gây ra những hiểu lầm rất nguy hiểm. Các lỗi có thể là do gõ nhầm hay do phát âm sai (một lỗi truyền thống hay gặp). Do đó, việc bắt lỗi chính tả tiếng Việt trên máy tính trở nên rất cấp thiết.

Bắt lỗi chính tả tiếng Việt cũng như xử lý ngôn ngữ tự nhiên lâu nay đã được quan tâm nghiên cứu ở nước ta. Nhiều sản phẩm bắt lỗi chính tả đã ra đời như: Vietkey Office, VietSpell… Nhưng vẫn còn nhiều lỗi bị bỏ qua trong quá trình phân tích mà đặc biệt là lỗi do cách phát âm. Những phần mềm này chỉ mới đáp ứng được phần nào sự mong đợi của người sử dụng.

Hiện nay, một hướng mới để bắt lỗi chính tả tiếng Việt đã ra đời và đang được nghiên cứu hoàn thiện. Đó là, bắt lỗi chính tả dựa trên mức cú pháp hay có thể nói là dựa trên cấu trúc cú pháp của các thành phần trong câu tiếng Việt để bắt các lỗi chính tả. Đây là một hướng giải quyết vấn đề có tính đột phá, một cơ hội để bắt được tất cả các lỗi chính tả đã mở ra. Dựa trên hướng giải quyết mới này, TS Phan Thị Tươi đã nghiên cứu và hoàn thành đề tài "Bắt lỗi chính tả tự động cho tiếng Việt bằng máy tính". Theo [8, 9], đề tài đã bắt được 95% các loại lỗi chính tả trong văn bản nhà nước phạm vi hành chính, xã hội. Để đạt được điều đó, phần mềm có 56.300 từ và khoảng 96 từ loại với 31.030 luật sinh hỗ trợ. Điều này cho thấy rõ khả năng bắt lỗi của phương pháp mới là cao hơn rất nhiều. Tuy nhiên, phương pháp này có những khó khăn như: vấn đề nhập nhằng ngữ nghĩa trong phân tích cú pháp và hiện tượng bùng nổ tổ hợp. Để giải quyết triệt để vấn đề lỗi chính tả, ta phải vượt qua những khó khăn trên.

Dựa trên phương pháp bắt lỗi chính tả tiếng Việt mới, lấy đề tài của TS Phan Thị Tươi làm cơ sở, khoá luận thực hiện nghiên cứu và phát triển phương pháp bắt lỗi chính tả tiếng Việt. Khoá luận sẽ đề cập đến một số vấn đề, như: từ điển tiếng Việt, quá trình thực thi bắt lỗi, phương pháp gợi ý, giải thuật Earley cho phân tích cú pháp, ngữ pháp tiếng Việt,… Đồng thời, khoá luận còn thực hiện cài đặt một số chương trình thử nghiệm, một số chương trình hỗ trợ quá trình nghiên cứu.

Khoá luận được trình bày trong 52 trang, ngoài phần mở đầu, kết luận và tài liệu tham khảo, khoá luận được chia thành ba chương:

Chương 1: **Bắt lỗi chính tả tiếng Việt**

- Trình bày tổng quan về các loại lỗi chính tả và phương pháp bắt lỗi chính tả.
- Trình bày những phân tích và thiết kế từ điển.
- Trình bày hai phương pháp bắt lỗi chính tả dựa trên âm tiết là: phương pháp dựa trên nguyên lý cấu tạo âm tiết và phương pháp dựa trên từ điển. Đồng thời còn trình bày phương pháp tạo gợi ý sửa lỗi ở mức âm tiết.
- Trình bày giải thuật Earley và những cải tiến
- Trình bày phương pháp bắt lỗi mức cú pháp: phân tích từ, phương pháp bắt lỗi chính tả dựa trên mức cú pháp, phương pháp gợi ý sửa lỗi.

Chương 2: **Ngữ pháp tiếng Việt với phân tích cú pháp**

- Trình bày khái quát về ngữ pháp tiếng Việt, bao gồm: từ loại, cụm từ và câu. Đồng thời, chỉ ra cách mã hoá từ loại và lập luật sinh cho chương trình bắt lỗi.

Chương 3: **Cài đặt chương trình**

- Trình bày hệ thống chương trình bắt lỗi chính tả và minh hoạ chương trình bắt lỗi.

Dù đã có nhiều cố gắng trong quá trình làm khoá luận, nhưng do hạn chế về mặt thời gian cũng như khả năng của bản thân nên không thể tránh khỏi những sơ suất. Mong nhận được sự chỉ giáo của quý thầy cô và bạn bè gần xa.

Nhân đây, em xin chân thành cám ơn thầy Nguyễn Gia Định đã tận tình giúp đỡ và hướng dẫn em trong suốt thời gian làm khoá luận. Thầy đã chỉ bảo và truyền đạt cho em nhiều kinh nghiệm quý báu.

Em cũng xin tỏ lòng biết ơn đến quí thầy cô khoa Công Nghệ Thông Tin - Trường Đại Học Khoa Học - Huế đã giảng dạy và truyền đạt những kiến thức cần thiết và bổ ích trong suốt thời gian học tập tại trường. Đồng thời, cũng xin cám ơn tất cả các bạn bè đã động viên, giúp đỡ tôi hoàn thành khoá luận này.

Gia đình và người thân luôn là nguồn động viên, cổ vũ hết sức to lớn, con xin ghi nhớ công ơn này.

Huế, tháng 5 năm 2004

Sinh viên thực hiện khoá luận

Lê Viết Mẫn