NGHIÊN CỨU XỬ LÝ NGÔN NGỮ TỰ NHIÊN, ỨNG DỤNG VÀO DỊCH TỰ ĐỘNG ANH – VIỆT, VIỆT – ANH

Mã số đề tài: 221304

Tên chủ nhiệm đề tài: **PGS. TS PHAN THỊ TƯƠI**

Cơ quan công tác: Ttrường Đại học Bách Khoa – ĐHQG tp.HCM

Địa chỉ liên lạc: 268 Lý Thường Kiệt, Quận 10, TP.HCM

Điện thoại: 08-8650161 Email: tuoi@dit.hcmut.edu.vn

Thành viên tham gia:

1. Kết quả nghiên cứu của đề tài

- Tìm hiểu các phương pháp phân tích cú pháp cho ngôn ngữ tự nhiên và cho tiếng Việt.
- Chọn lọc nghĩa trong quá trình phân tích cú pháp cho tiếng Việt để áp dụng vào dịch máy song ngữ Anh – Việt, Việt – Anh.
- Xây dựng mô hình dịch máy Việt – Anh dùng phương pháp phân tích cú pháp có xác suất.
- Xây dựng chương trình xử lý tính hợp nhất trong văn phạm có hệ thống nét cho tiếng Việt.
- Chuẩn bị ngữ liệu để xây dựng từ điển song ngữ Anh – Việt, Việt – Anh phục vụ cho dịch máy Việt – Anh.
- Hiện thực mô hình dịch máy có xác suất từ Anh sang Việt trên cơ sở cú pháp.

2. Ý nghĩa thực tiễn và hiệu quả của việc ứng dụng kết quả nghiên cứu

Dịch máy song ngữ đã được nghiên cứu nhiều năm nay ở các nước. Dịch máy Anh – Việt cũng đã được nghiên cứu hơn 10 năm nay ở Việt Nam, song chưa có dịch máy từ Việt sang Anh. Thậm chí dịch máy Anh – Việt hiện nay đều chưa hoàn thiện. Nhóm đề tài đã xây dựng mô hình dịch máy trên cơ sở xác suất từ Anh sang Việt và ngược lại. Đây cũng là một đóng góp cho vấn đề nghiên cứu xử lý ngôn ngữ tiếng Việt cho dịch máy. Thông tin hiện nay rất nhiều, chúng ta cần dịch từ Anh sang Việt và từ Việt sang Anh, do đó nếu các kết quả nghiên cứu được áp dụng thì đề tài không chỉ có ý nghĩa khoa học mà còn có ý nghĩa thực tiễn rất lớn. Kết quả bước đầu nếu xây dựng được từ điển điện tử song ngữ Anh – Việt, Việt – Anh (Lexicon) cũng là một đóng góp rất lớn cho lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt.

3. Kết quả đào tạo sau đại học

- Số học viên cao học đang hướng dẫn: 01
- Số nghiên cứu sinh đang hướng dẫn: 04
- Số học viên cao học đã bảo vệ : 06
- Số nghiên cứu sinh đã bảo vệ: 0

4. Danh mục các sản phẩm khoa học của đề tài

4.1.Các công trình đã công bố trên các tạp chí khoa học

Bài báo "Phân tích cụm danh từ tiếng Việt sử dụng văn phạm hợp nhất" đăng ở Tạp chí Bưu chính Viễn thông và Công nghệ thông tin – chuyên san các công trình nghiên cứu – triển khai viễn thông và công nghệ thông tin, tác giả: Trần Ngọc Tuấn, Phan Thị Tươi, số 13 – tháng 12/2004.

4.2. Các báo cáo khoa học tại các hội nghị Quốc gia

[1]. Báo cáo "Vietnamese-to-English statistical machine translation model" tại hội thảo quốc gia lần thứ VII "Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông" từ ngày 18 – 20/8/2004 tại Đà Nẵng, tác giả Trần Ngọc Tuấn, Phan Thị Tươi.

[2]. Báo cáo "Feature-based Grammar in Adaption to Vietnamese Natural Language Processing" tại hội thảo khoa học công nghệ thông tin của chương trình quốc gia KC.01 (ICT.RDA) 2004 tại Hà Nội từ ngày 24 – 25/9/2004, tác giả: Trần Ngọc Tuấn, Phan Thị Tươi.

[3]. Báo cáo "English-Vietnamese dictionary with lexical conceptual structure for machine translation" tại hội thảo khoa học Quốc gia lần thứ II "Nghiên cứu cơ bản và Ứng dụng công nghệ thông tin" (FAIR'2005) tại Trường Đại học Bách khoa từ ngày 23 – 24/9/2005, tác giả: Lê Mạnh Hải, Phan Thị Tươi, Nguyễn Chí Hiếu.

[4]. Báo cáo "Hệ thống truy xuất thông tin hỗ trợ tiếng Việt: cơ chế hoạt động và hiện thực", tại hội thảo khoa học Quốc gia lần thứ II "Nghiên cứu cơ bản và Ứng dụng công nghệ thông tin" (FAIR'2005) tại Trường Đại học Bách khoa từ ngày 23 – 24/9/2005tác giả: Nguyễn Chánh Thành, Phan Thị Tươi.

[5]. Báo cáo "Tự động rút trích các cụm danh từ Anh – Việt từ kho ngữ liệu song ngữ", tại hội thảo khoa học Quốc gia lần thứ II "Nghiên cứu cơ bản và Ứng dụng công nghệ thông tin" (FAIR'2005) tại Trường Đại học Bách khoa từ ngày 23 – 24/9/2005, tác giả: Nguyễn Chí Hiếu, Phan Thị Tươi, Nguyễn Xuân Dũng.

[6]. Báo cáo "Gán nhãn từ loại cho tiếng Việt dựa trên văn phong", tại hội thảo khoa học Quốc gia lần thứ II "Nghiên cứu cơ bản và Ứng dụng công nghệ thông tin" (FAIR'2005) tại Trường Đại học Bách khoa từ ngày 23 – 24/9/2005, tác giả: Nguyễn Quang Châu, Phan Thị Tươi, Cao Hoàng Trụ.

4.3. Các công trình đã hoàn thành sẽ công bố

[1]. Báo cáo "Applying Natural Language Processing to Machine Translation" tại hội thảo quốc tế về khoa học công nghệ thông tin (RIVF'06) từ ngày 12 – 16/02/2006, tác giả: Nguyễn Chí Hiếu, Phan Thị Tươi, Nguyễn Xuân Dũng, Lê Mạnh Hải (được đăng kỷ yếu hội nghị ở dạng poster).

[2]. Báo cáo "Vietnamese Proper Noun Recognition" tại hội thảo quốc tế về khoa học công nghệ thông tin (RIVF'06) từ ngày 12 – 16/02/2006, tác giả:

Nguyễn Quang Châu, Phan Thị Tươi, Cao Hoàng Trụ (được đăng kỷ yếu hội nghị ở dạng full paper).

[3]. Báo cáo "Syntax-based SMT Model in Adaption to Vietnamese-English Translation" tại hội thảo quốc tế về khoa học công nghệ thông tin (RIVF'06) từ ngày 12 – 16/02/2006, tác giả: Trần Ngọc Tuấn, Phan Thị Tươi (được đăng kỷ yếu hội nghị ở dạng poster).

[4]. Bài báo "Unification Grammar in a Semantic Approach for Vietnamese Compound Noun Parsing" đăng trên tạp chí Tin học và Điều khiển học, tác giả: Trần Ngọc Tuấn, Phan Thị Tươi (đã được chấp nhận của tạp chí).

[5]. Bài báo "Sử dụng kỹ thuật Pruning vào bài toán xác định từ loại", đăng trên tạp chí Phát triển Khoa học và Công nghệ ĐHQG TP.HCM, tác giả: Nguyễn Chí Hiếu, Phan Thị Tươi, Nguyễn Xuân Dũng, Nguyễn Quang Châu (đã được chấp nhận của tạp chí).

[6]. Bài báo "Dịch máy Anh – Việt trên cơ sở cụm từ", gửi đăng trên tạp chí Tin học và Điều khiển học, tác giả: Nguyễn Chí Hiếu, Phan Thị Tươi, Nguyễn Xuân Dũng.