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

BÙI THANH THUỶ

NGHIÊN CỨU VỀ DỊCH MÁY THỐNG KÊ DỰA VÀO CỤM TỪ VÀ ỨNG DỤNG DỊCH TỪ TIẾNG VIỆT SANG TIẾNG ANH

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan toàn bộ nội dung trong luận văn này do tôi tự nghiên cứu, đọc, dịch tài liệu, tổng hợp và thực hiện. Trong luận văn tôi có sử dụng một số tài liệu tham khảo như đã trình bày trong phần tài liệu tham khảo.

Người viết luận văn

Bùi Thanh Thủy

LỜI CẢM ƠN

Đầu tiên tôi xin gửi lời cảm ơn chân thành đến TS. **Nguyễn Văn Vinh** đã tận tình hướng dẫn, chỉ bảo cho tôi trong suốt quá trình làm luận văn. Em cũng xin cam ơn anh **Trần Hồng Việt**, nghiên cứu sinh Trường đại học công nghệ, giảng viên Trường Đại học Kinh tế kỹ thuật công nghiệp đã giúp đỡ em trong quá trình làm luận văn

Tôi cũng xin gửi lời cảm ơn đến các thầy cô trường Đại học Công nghệ thông tin và Truyền thông – Đại học Thái Nguyên, các thầy cô Viện Công nghệ thông tin đã truyền đạt những kiến thức và giúp đỡ tôi trong suốt quá trình học của mình.

Tôi cũng xin gửi lời cảm ơn tới Ban giám hiệu, Phòng Đào tạo, các đồng nghiệp trường Cao đẳng nghề Phú Thọ, gia đình và bạn bè những người đã động viên tạo mọi điều kiện giúp đỡ tôi để hoàn thành luận văn.

MỤC LỤC

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1.Lý do chon đề tài

Trong quá trình phát triển và hội nhập văn hóa, kinh tế thế giới. Quá trình giao lưu giữa người Việt Nam và người nước ngoài ngày càng nhiều dẫn đến khó khăn trong quá trình giao tiếp và sử dụng văn bản tài liệu tiếng Anh. Hiện nay có nhiều hệ thống tự động dịch miễn phí trên mạng như: google translate, vietgle, vdict, lạc việt,… Những hệ thống này cho phép dịch tự động các văn bản với một cặp ngôn ngữ chọn trước (ví dụ dịch từ tiếng Anh sang tiếng Việt). Điều ấy cho thấy sự phát triển của dịch máy càng ngày càng tiến gần hơn đến ngôn ngữ tự nhiên của con người.

Vào những năm gần đây, dịch máy nói chung, dịch máy thống kê nói riêng được phát triển mạnh và ứng dụng rộng rãi. Kết quả thực tế của hệ thống dịch này rất tốt. Ngôn ngữ của máy dịch ngày càng gần với ngôn ngữ của người. Ngoài ra cùng với hệ thống dịch máy thống kê, các sản phẩm ứng dụng ngày càng nhiều giúp con người trao đổi thông tin dễ dàng hơn, tốc độ nhanh hơn và cùng với nhiều ngôn ngữ hơn.

Hiện nay, phương pháp dịch thống kê dựa trên cụm từ là phương pháp cho kết quả dịch tốt nhất hiện nay. Điều này được thể hiện của qua các hệ dịch máy của Google, Vietgle. Hơn nữa việc dịch giữa tiếng Việt sang tiếng Anh là rất cần thiết khi khối lượng văn bản tiếng Anh ngày càng lớn trong thời kỳ Việt Nam hội nhập sâu rộng với quốc tế.

Chính vì lý do đó, tôi lựa chọn và thực hiện đề tài "Nghiên cứu về dịch thống kê dựa vào cụm từ và áp dụng cho dịch từ tiếng Việt sang tiếng Anh".

2. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Nghiên cứu về các phương pháp, mô hình dịch máy thống kê

- Thử nghiệm và đánh giá kết quả dịch từ tiếng Việt sang tiếng Anh

Phạm vi nghiên cứu:

Đề tài tập trung vào nghiên cứu phương pháp dịch thống kê dựa vào cụm từ và ứng dụng dịch tài liệu, văn bản tiếng Việt, tiếng Anh.

3. Hướng nghiên cứu của đề tài

- Nghiên cứu, tìm hiểu, phân tích về dịch máy thống kê trên cơ sở cụm từ.

- Cài đặt thử nghiệm tối ưu hóa cụm từ bằng hệ dịch máy thống kế Moses

4. Phương pháp nghiên cứu

- Tìm hiểu các hệ dịch tự động đã có để tìm ra các phương pháp dịch máy mà các hệ dịch đang sử dụng.

- Nghiên cứu và đánh giá các phương pháp dịch máy, những ưu điểm và hạn chế, sau đó tìm ra phương pháp có hiệu quả và đề xuất áp dụng cho bài toán đề tài đặt ra.

- Nghiên cứu các phương pháp đánh giá chất lượng dịch máy để đánh giá hiệu quả dịch cho hệ thống đề tài đã xây dựng.

5. Ý nghĩa khoa học của đề tài

Ý nghĩa khoa học:

Dịch máy dựa vào cụm từ là một trong những phương pháp dịch máy hiệu quả nhất hiện nay. Hơn nữa dữ liệu văn bản ngày càng lớn và đa dạng. chính vì vậy nghiên cứu về hệ dịch dựa vào cụm từ và ứng dụng cho dịch Việt – Anh có ý nghĩa khoa học cũng như thực tiễn

6. Cấu trúc luận văn

+ Chương 1: Tổng quan về dịch máy

+ Chương 2: Dịch máy thống kê dựa vào cụm từ và áp dụng cho ngôn ngữ Việt _ Anh

+ Chương 3: Thực nghiêm, đánh giá

+ Kết luận

CHƯƠNG 1 – TỔNG QUAN VỀ DỊCH MÁY

1.1. Khái niệm về hệ dịch máy

1.1.1. Định nghĩa

Các hệ dịch máy (machine translation system-MT) là các hệ thống sử dụng máy tính để dịch từ một thứ tiếng (trong ngôn ngữ tự nhiên) sang một hoặc vài thứ tiếng khác.

Ngôn ngữ của văn bản cần dịch được gọi là ngôn ngữ nguồn, ngôn ngữ của văn bản đã dịch ra được gọi là ngôn ngữ đích.

1.1.2. Vai trò của dịch máy

Hiện nay trên thế giới có khoảng hơn 5000 ngôn ngữ khác nhau, với một số lượng ngôn ngữ lớn như vậy đã gây ra rất nhiều khó khăn trong việc trao đổi thông tin, trong giao tiếp, đồng thời ngăn cản sự phát triển của thương mại và mậu dịch quốc tế.

Với những khó khăn như vậy con người đã phải dùng đến một đội ngũ phiên dịch khổng lồ, để dịch các văn bản, tài liệu, lời nói, ngôn ngữ từ tiếng nước này sang tiếng nước khác. Những công việc đó mang tính chất thủ công, tỉ mỉ đòi hỏi người dịch phải làm mất rất nhiều thời gian và công sức, trong khi khối lượng văn bản cần dịch ngày càng nhiều.

Để khắc phục được những nhược điểm trên con người đã nghĩ đến việc thiết kế một mô hình tự động trong công việc dịch ngôn ngữ, do đó ngay từ khi xuất hiện chiếc máy tính điện tử đầu tiên (năm 1946) người ta đã tiến hành nghiên cứu về dịch máy. Việc đưa ra mô hình tự động cho việc dịch đã và đang được phát triển, mặc dù chưa giải quyết được triệt để lớp ngôn ngữ tự nhiên. Nhưng sự ra đời của chúng đã khẳng định được lợi ích to lớn về mặt chiến lược và phát triển kinh tế, đồng thời các vấn đề liên quan đến dịch máy