ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG

Ngô Hữu Huy

NGHIÊN CỨU MỘT SỐ CÔNG CỤ PHỤC VỤ CHO VIỆC PHÁT TRIỂN HỆ THỐNG HỖ TRỢ DỊCH TRUNG - VIỆT

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG

Ngô Hữu Huy

NGHIÊN CỨU MỘT SỐ CÔNG CỤ PHỤC VỤ CHO VIỆC PHÁT TRIỂN HỆ THỐNG HỖ TRỢ DỊCH TRUNG - VIỆT

Chuyên ngành : Khoa học máy tính
Mã số : 60.48.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS Nguyễn Ái Việt

Thái Nguyên - 2012

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn là kết quả của sự tìm hiểu, nghiên cứu các tài liệu một cách nghiêm túc dưới sự hường dẫn của TS Nguyễn Ái Việt.

Nội dung của luận văn được phát triển từ ý tưởng, sự sáng tạo của bản thân và kết quả hoàn toàn trung thực.

Học viên

Ngô Hữu Huy

MỤC LỤC

DANH SÁCH CÁC HÌNH

MỞ ĐẦU

Chế tạo ra một loại máy có khả năng dịch tự động để giúp cho con người vượt qua rào cản ngôn ngữ là một mơ ước của loài người đã có từ thế kỷ XVII, rất lâu trước khi máy tính điện tử ra đời. Khi khoa học công nghệ phát triển mạnh, yêu cầu nắm bắt thông tin về kỹ thuật nhanh và chính xác trở nên cần thiết.

Chẳng bao lâu sau khi máy tính điện tử đầu tiên ra đời, bên cạnh những ứng dụng tính toán trong lĩnh vực toán học và vật lý, con người nghĩ ngay đến việc sử dụng "não bộ của máy tính" cho những ứng dụng thực tiễn, trong đó có vấn đề dịch máy. Lần đầu tiên, việc sử dụng máy tính điện tử để dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác được đề cập đến trong những cuộc thảo luận giữa Andrew D. Booth và Warren Weaver vào năm 1946. Vượt qua nhiều trở ngại về lý thuyết và công nghệ, Booth đã cho ra mắt "hệ dịch dựa trên từ điển" đầu tiên tại hội nghị của MIT vào năm 1952 [4] [15] [16].

Trong sự phát triển nhanh chóng của mạng máy tính và công nghệ truyền thông, con người ngày càng có điều kiện tiếp xúc với nguồn tri thức rất phong phú ở nhiều dạng khác nhau (chữ viết, hình ảnh, âm thanh, .v.v.), được thể hiện ở nhiều ngôn ngữ khác nhau. Nhu cầu đọc hiểu và trao đổi thông tin trở nên cần thiết hơn bao giờ, thế nhưng, nhu cầu này đã gặp phải một rào cản - sự khác biệt về mặt ngôn ngữ. Và, ngôn ngữ, tự hân nó đã vốn rất phức tạp.

Con người đã tìm cách vượt qua rào cản ngôn ngữ theo nhiều cách khác nhau, từ việc xây dựng các bộ từ điển song ngữ, các nghiên cứu về dịch thuật liên ngữ, phương pháp học ngoại ngữ nhanh chóng, cho đến cảviệc tạo ra một ngôn ngữ chung cho loài người - quốc tế ngữ Esperanto. Vào thời điểm sức mạnh của máy tính đã được khẳng định, bài toán sử dụng máy tính để chuyển đổi tri thức được viết bằng ngôn ngữ này sang một ngôn ngữ khác được đặt ra. Trong khoảng 50 năm, có rất nhiều phương pháp dịch máy đã được giới thiệu và triển khai. Hiện nay, đã có nhiều hệ dịch tự động được thương mại hóa ở dạng các chương trình máy tính hoặc các dịch vụ web.

Sự nhìn nhận về vấn đề dịch máy (Machine Translation) đã nhiều lần thay đổi trong hơn năm mươi năm qua, từ chỗ hình dung rằng dịch thuật là công việc đơn giản, máy sẽ dịch mọi loại văn bản không kém gì con người, như một bộ máy vạn năng, cho đến chỗ khẳng định rằng dịch máy tự động, chất lượng cao là hoàn toàn không khả thi. Ngày hôm nay, hầu hết các chuyên gia về dịch máy đều có chung quan điểm rằng máy tính chỉ có thể biên dịch văn bản chất lượng chấp nhận được trong một lĩnh vực chuyên môn hẹp, hoặc chỉ có thể hỗ trợ dịch thô để đọc hiểu. Phương pháp dịch máy dựa hoàn toàn vào kho ngữ liệu như Dịch máy dựa trên Thống kê (Statistical Machine Translation) hay Dịch máy dựa trên mẫu ví dụ (Example-based Machine Translation) được xem là chỉ có ích để dịch với chất lượng tương đối thấp cho mọi loại văn bản [4] [15] [16].

Hiện nay số người nói tiếng Trung trên thế giới là nhiều nhất. Tiếng Trung Quốc chiếm một vị trí quan trọng trên trường quốc tế, đồng thời nó có ảnh hưởng rất lớn đến sự phát triển của nền văn hóa và kinh tế trên toàn thế giới.

Với mong muốn học hỏi, tôi mạnh dạn chọn đề tài “Nghiên cứu một số công cụ phục vụ cho việc phát triển hệ thống hỗ trợ dịch Trung-Việt” cho luận văn của mình. Luận văn được trình bày trong 4 chương, khái quát như sau:

Chương 1: Tổng quan

Chương 2: Các phương pháp dịch máy

Chương 3: Các đặc trưng của dịch Trung (Hán)-Việt

Chương 4: Xây dựng khối liệu và đánh giá một số công cụ xử lý tiếng Trung

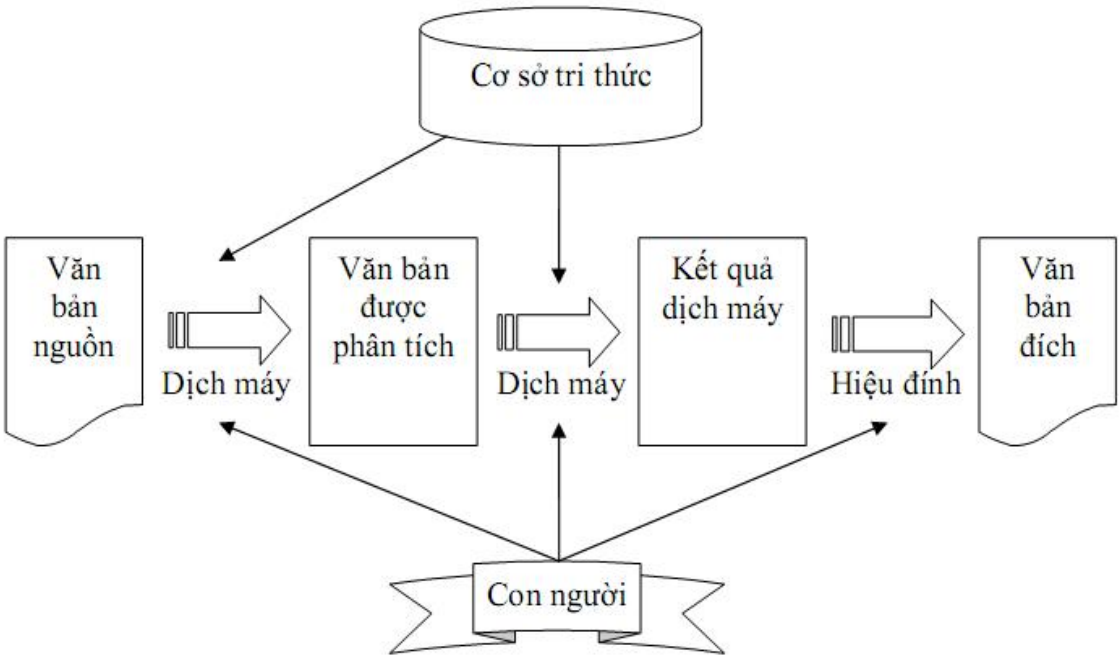
CHƯƠNG 1. TỔNG QUAN VỀ DỊCH MÁY

1.1. Định nghĩa dịch máy

Khái niệm dịch máy đã được nhiều tác giả trong lĩnh vực xử lý ngôn ngữ tự nhiên định nghĩa, tuy có một vài điểm khác biệt nhưng, hầu hết đều tương đương với định nghĩa sau:

Một hệ dịch máy (Machine Translation System) là một hệ thống sử dụng máy tính để chuyển đổi văn bản được viết trong ngôn ngữ tự nhiên này thành bản dịch tương đương trong ngôn ngữ khác [15] [16].

Ngôn ngữ của văn bản cần dịch còn gọi là ngôn ngữ nguồn, ngôn ngữ của bản dịch được gọi là ngôn ngữ đích. Sơ đồ sau thể hiện vị trí của hệ dịch máy trong tiến trình dịch tài liệu.



Hình 1.1. Quá trình xử lý tài liệu dịch máy

Đầu vào của một hệ dịch máy thường là một văn bản viết trong ngôn ngữ nguồn. Quá trinh dịch có thể chia thành hai giai đoạn: Đầu tiên, văn bản được phân