

**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 hướng 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

| | |
|---|----|
| LỜI CAM ĐOAN | i |
| MỤC LỤC | ii |
| DANH SÁCH CÁC HÌNH..... | v |
| MỞ ĐẦU | vi |
| CHƯƠNG 1. TỔNG QUAN VỀ DỊCH MÁY | 1 |
| 1.1. Định nghĩa dịch máy | 1 |
| 1.2. Vai trò của dịch máy | 2 |
| 1.3. Lịch sử của dịch máy | 3 |
| 1.3.1. Giai đoạn 1930 - 1940..... | 3 |
| 1.3.2. Giai đoạn 1940 - 1970..... | 4 |
| 1.3.3. Giai đoạn 1970 – 1990 | 5 |
| 1.3.4. Giai đoạn 1990 - hiện nay | 6 |
| 1.4. Phân loại dịch máy..... | 7 |
| 1.5. Phạm vi của luận văn | 8 |
| 1.6. Kết chương..... | 8 |
| CHƯƠNG 2. CÁC PHƯƠNG PHÁP DỊCH MÁY | 9 |
| 2.1. Các chiến lược dịch cơ bản..... | 9 |
| 2.1.1. Dịch trực tiếp (Direct MT) | 9 |
| 2.1.2. Dịch chuyển đổi cú pháp (Syntactic-transfer MT)..... | 10 |
| 2.1.3. Dịch qua ngôn ngữ trung gian (Interlingual MT) | 11 |
| 2.1.4. Dịch chuyển đổi cú pháp + phân giải ngữ nghĩa..... | 13 |
| 2.2. Các cách tiếp cận của dịch máy hiện nay..... | 14 |
| 2.2.1. Dịch máy dựa trên luật (RBMT: Rule-Based MT) | 14 |

| | | |
|---|---|-----------|
| 2.2.2. | Dịch máy dựa trên thống kê (SMT: Statistics-Based MT)..... | 18 |
| 2.2.3. | Dịch máy dựa trên cơ sở tri thức (KBMT: Knowlegde-Based MT).... | 20 |
| 2.2.4. | Dịch máy dựa trên ví dụ (EBMT: Example-Based MT)..... | 21 |
| 2.2.5. | Dịch máy dựa trên ngữ liệu (CBMT: Corpus-Based MT) | 21 |
| 2.2.6. | Các cách tiếp cận lai (hybrid MTs) | 22 |
| 2.3. | Nhận xét các chiến lược và các cách tiếp cận | 23 |
| 2.3.1. | Nhận xét về các chiến lược | 23 |
| 2.3.2. | Nhận xét về các cách tiếp cận | 25 |
| 2.4. | Kết chương | 26 |
| CHƯƠNG 3. CÁC ĐẶC TRƯNG CỦA DỊCH TRUNG (HÁN) – VIỆT..... | | 27 |
| 3.1. | Ngữ nghĩa đất nước học trong hai ngôn ngữ Hán-Việt..... | 27 |
| 3.1.1. | Văn hóa dân tộc và nội hàm ngữ nghĩa đất nước học | 27 |
| 3.1.2. | Thông tin ngữ nghĩa đất nước học trong từ vựng | 28 |
| 3.2. | Đặc điểm tương đồng và khác biệt của ngôn ngữ Trung (Hán)-Việt | 31 |
| 3.2.1. | Đặc điểm tương đồng và khác biệt về mặt ngữ âm..... | 31 |
| 3.2.2. | Đặc điểm tương đồng và khác biệt về mặt văn tự..... | 36 |
| 3.2.3. | Đặc điểm tương đồng và khác biệt về mặt từ vựng | 36 |
| 3.2.4. | Đặc điểm tương đồng và khác biệt về mặt ngữ pháp | 37 |
| 3.2.5. | Đặc điểm tương đồng và khác biệt về mặt tu từ | 38 |
| 3.3. | Nguyên nhân sự tương đồng và khác biệt ngữ nghĩa đất nước học giữa hai ngôn ngữ Hán và Việt | 38 |
| 3.3.1. | Phong tục tập quán dân tộc..... | 39 |
| 3.3.2. | Bối cảnh văn hoá lịch sử | 40 |
| 3.3.3. | Tín ngưỡng tôn giáo | 41 |
| 3.3.4. | Hoàn cảnh địa lý..... | 42 |

| | |
|---|----|
| 3.4. Kết chương | 43 |
| CHƯƠNG 4. XÂY DỰNG KHỐI LIỆU VÀ ĐÁNH GIÁ..... | 45 |
| MỘT SỐ CÔNG CỤ XỬ LÝ TIẾNG TRUNG | 45 |
| 4.1. Xây dựng kho ngữ liệu Trung-Việt (Corpus Trung-Việt) | 45 |
| 4.1.1. Khái niệm về Corpus..... | 45 |
| 4.1.2. Quy trình xây dựng Corpus..... | 46 |
| 4.1.3. Xây dựng Corpus Trung Việt..... | 48 |
| 4.2. Đánh giá một số công cụ xử lý tiếng Trung..... | 50 |
| 4.2.1. Công cụ phân tích cú pháp (Parser) | 50 |
| 4.2.2. POS Tagger (Part-Of-Speech Tagger) | 53 |
| 4.3. Kết chương | 54 |
| KẾT LUẬN | 55 |
| TÀI LIỆU THAM KHẢO | 56 |

DANH SÁCH CÁC HÌNH

| | |
|---|----|
| Hình 1.1. Quá trình xử lý tài liệu dịch máy | 1 |
| Hình 2.1. Mô hình dịch trực tiếp | 10 |
| Hình 2.2. Mô hình dịch kiểu chuyển đổi cú pháp..... | 10 |
| Hình 2.3. Chuyển đổi cây cú pháp ngôn ngữ nguồn sang cây của ngôn ngữ đích... | 11 |
| Hình 2.4. Mô hình dịch liên ngôn ngữ..... | 12 |
| Hình 2.5. Các chiến lược dịch trong máy dịch | 13 |
| Hình 2.6. Mức độ phân tích, chuyển đổi và tổng hợp trong các chiến lược dịch..... | 14 |
| Hình 2.7. Kết quả phân tích cú pháp câu “I see the man in the car” | 17 |
| Hình 2.8. Kết quả phân tích cú pháp câu “I saw the man in a day” | 17 |
| Hình 4.1. Bộ gõ tiếng Trung Sougou pinyin | 48 |
| Hình 4.2. Giao diện phần mềm Text & Word joiner | 49 |
| Hình 4.3. Giao diện phần mềm Stanford-parser | 50 |
| Hình 4.4. Chọn file đầu vào..... | 51 |
| Hình 4.5. Chọn parser | 51 |
| Hình 4.6. Kết quả phân tích cú pháp | 52 |
| Hình 4.7. Giao diện phần mềm Stanford postagger | 53 |
| Hình 4.8. Nhập dữ liệu đầu vào | 54 |
| Hình 4.9. Kết quả thu được khi gán thẻ..... | 54 |

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 cả việc 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

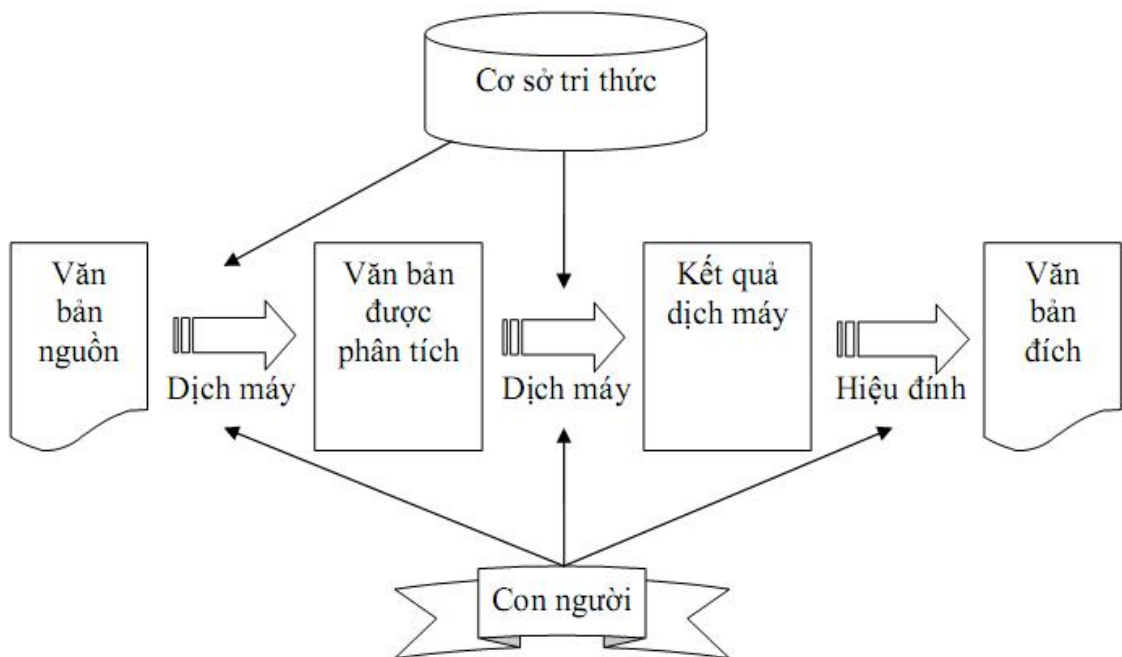
CHƯƠNG 1. TỔNG QUAN VỀ DỊCH MÁY

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 trình dịch có thể chia thành hai giai đoạn: Đầu tiên, văn bản được phân