

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN  
BỘ MÔN CÔNG NGHỆ TRI THỨC**

**VĂN CHÍ NAM**

**XỬ LÝ NGỮ NGHĨA  
TRONG HỆ DỊCH TỰ ĐỘNG ANH – VIỆT  
CHO CÁC TÀI LIỆU TIN HỌC**

**LUẬN VĂN CỬ NHÂN TIN HỌC**

**TP. Hồ Chí Minh – Năm 2003**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN CÔNG NGHỆ TRI THỨC**

**VĂN CHÍ NAM - 9912618**

**XỬ LÝ NGỮ NGHĨA  
TRONG HỆ DỊCH TỰ ĐỘNG ANH – VIỆT  
CHO CÁC TÀI LIỆU TIN HỌC**

**LUẬN VĂN CỬ NHÂN TIN HỌC**

**GIÁO VIÊN HƯỚNG DẪN**

**TS. ĐINH ĐIỀN**

**NIÊN KHOÁ 1999 - 2003**





## NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....  
.....

Khoa CNTT - ĐH KHTN TP HCM

Tp. Hồ Chí Minh, ngày tháng 07 năm 2003



## Lời Cảm Ơn

Sau một thời gian thực hiện luận văn tốt nghiệp, đến nay, mọi công việc liên quan đến luận văn đã hoàn tất. Trong suốt thời gian này, tôi đã nhận được rất nhiều sự giúp đỡ. Ở phần đầu tiên của luận văn, cho phép tôi có đôi điều gửi đến những người tôi vô cùng biết ơn.

Xin gửi lời cảm ơn chân thành nhất đến Thầy Đinh Điền, người đã tận tình hướng dẫn, động viên, và giúp đỡ em trong suốt thời gian qua. Nếu không có những lời chỉ dẫn, những tài liệu, ngữ liệu, những lời động viên khích lệ của Thầy thì luận văn này khó lòng hoàn thiện được.

Cũng xin gửi lời biết ơn đến cả nhà, đến pá, đến má, đến mẹ, đến dương ba, đến chế Hiền, đến chế Nghi, những người đã luôn dành những tình thương yêu nhất cho Năm, những người đã luôn hỗ trợ, dõi theo những bước đi của Năm trong tất cả các năm học vừa qua.

Xin tri ân tất cả các Thầy Cô, những người dày công dạy dỗ, truyền cho em rất nhiều tri thức quý báu.

Cảm ơn các bạn, các anh trong nhóm VCL vì những đóng góp của các bạn, các anh cho luận văn này. Đặc biệt xin gửi lời cảm ơn đến với anh Ngô Quốc Hưng, anh Phạm Phú Hội, bạn Nguyễn Thái Ngọc Duy cho những công cụ phục vụ luận văn và những góp ý cho chương trình.

Cảm ơn tất cả bạn bè tôi, những người đã sát cánh cùng vui những niềm vui, cùng chia sẻ những khó khăn của tôi.

Còn rất nhiều điều không thể diễn tả hết bằng lời, xin luôn ghi nhớ mãi trong tim.

**Văn Chí Nam**

## **Lời Nói Đầu**

Những năm gần đây, với sự phát triển nhanh chóng trong lĩnh vực công nghệ thông tin, việc sử dụng các tài liệu để có thể nắm bắt được các tri thức mới vô cùng phổ biến. Song một khó khăn lớn đối với nhiều người Việt chúng ta hiện nay là việc hiểu ngôn ngữ được thể hiện trong các tài liệu (mà chủ yếu là tiếng Anh). Do đó, tạo lập một hệ thống chỉ dịch các tài liệu tin học từ tiếng Anh sang tiếng Việt có ý nghĩa to lớn. Chắc chắn nó sẽ giúp nhiều người Việt có điều kiện tiếp cận tốt các nội dung, kiến thức mới của tin học trên thế giới.

Nhưng vấn đề khó khăn nhất gặp phải trong việc thiết lập một hệ dịch tự động là tính nhập nhằng vốn có của ngôn ngữ tự nhiên, trong đó nhập nhằng lớn nhất là nhập nhằng ngữ nghĩa. Việc chọn ra một nghĩa thích hợp cho từ là một công việc không dễ dàng nhưng cực kỳ lý thú. Giải quyết tốt vấn đề ngữ nghĩa sẽ nâng cao chất lượng cho hệ dịch tự động Anh – Việt.

Đề tài này hướng đến việc giải quyết tốt những nhập nhằng nghĩa của từ trong các tài liệu tin học nhờ vào việc huấn luyện trên ngữ liệu song ngữ để rút ra các luật chuyển đổi. Thông qua việc kết hợp các khối khác của dịch tự động, tạo ra các câu dịch tiếng Việt có thể hiểu được. Sự thay đổi lĩnh vực xem xét không ảnh hưởng nhiều đến cấu trúc của mô hình. Chúng tôi thực hiện việc giới hạn lĩnh vực ngoài ý nghĩa nêu phía trên còn có lý do thử nghiệm mô hình xử lý ngữ nghĩa mới, xem xét tính tương hỗ từ các thông tin trong ngữ liệu song ngữ và đảm bảo chất lượng câu dịch.

Luận văn được tổ chức thành 5 chương và các phụ lục.

- Chương 1 giới thiệu tổng quan về dịch máy nói chung và xử lý ngữ nghĩa nói riêng.
- Chương 2 giới thiệu các cơ sở lý thuyết cần sử dụng, trong đó có đề cập đến thuật toán huấn luyện.
- Chương 3 đưa ra mô hình cài đặt cho khối xử lý ngữ nghĩa
- Chương 4 cụ thể hoá mô hình cài đặt
- Chương 5 tổng kết luận văn và đề ra hướng phát triển.

## Mục Lục

Lời Nói Đầu .....	i
Mục Lục .....	ii
Danh Sách Hình .....	vii
Danh Sách Bảng Biểu .....	viii
<b>Chương 1 TỔNG QUAN .....</b>	<b>1</b>
1.1. SƠ LƯỢC VỀ DỊCH MÁY .....	2
1.1.1. Lịch sử của Dịch Máy .....	2
1.1.2. Khái niệm về Dịch Máy .....	6
1.1.3. Các bước xử lý trong một hệ Dịch Máy .....	7
1.2. XỬ LÝ NGỮ NGHĨA TRONG DỊCH MÁY .....	10
1.2.1. Vai trò và chức năng của xử lý ngữ nghĩa .....	10
1.2.2. Các mức độ nhập nhằng trong tầng xử lý ngữ nghĩa .....	12
1.2.2.1. Nhập nhằng ở mức từ vựng .....	12
1.2.2.2. Mức độ nhập nhằng cấu trúc .....	12
1.2.2.3. Mức độ nhập nhằng liên câu .....	13
1.2.2.4. Mức độ nhập nhằng theo thể loại văn bản .....	14
1.2.3. Các khó khăn trong xử lý ngữ nghĩa .....	15
1.2.3.1. Nhập nhằng nghĩa .....	15
1.2.3.2. Phụ thuộc vào ngữ cảnh .....	15
1.2.3.3. Phụ thuộc vào tri thức .....	15
1.2.3.4. Sự khác biệt giữa tiếng Anh và Việt .....	16
1.2.3.5. Yếu tố khác .....	16
1.3. CÁC CÁCH TIẾP CẬN TRONG XỬ LÝ NGỮ NGHĨA VÀ CÁC CÔNG TRÌNH TRƯỚC ĐÂY .....	17
1.3.1. Xử lý ngữ nghĩa trong thời gian đầu .....	17

1.3.2. Dựa trên trí tuệ nhân tạo .....	18
1.3.3. Dựa trên cơ sở tri thức .....	20
1.3.3.1. Từ điển máy .....	20
1.3.3.2. Từ điển đồng nghĩa .....	22
1.3.3.3. Từ điển điện toán .....	23
1.3.4. Dựa trên ngữ liệu .....	24
<b>Chương 2 CƠ SỞ LÝ THUYẾT .....</b>	<b>27</b>
2.1. CƠ SỞ LÝ THUYẾT VỀ NGÔN NGỮ HỌC .....	28
2.1.1. Nghĩa của từ .....	28
2.1.1.1. Cơ cấu nghĩa của từ .....	29
2.1.1.2. Phân tích nghĩa của từ .....	29
2.1.1.3. Nghĩa của từ trong hoạt động ngôn ngữ .....	30
2.1.2. Quan hệ đồng nghĩa và trái nghĩa trong từ vựng .....	30
2.1.2.1. Từ đồng nghĩa .....	30
2.1.2.2. Từ trái nghĩa .....	31
2.1.3. Biến đổi trong từ vựng .....	31
2.1.3.1. Những biến đổi bề mặt .....	31
2.1.3.2. Những biến đổi trong chiều sâu của từ vựng .....	32
2.2. HỌC DỰA TRÊN CHUYỂN ĐỔI .....	32
2.2.1. Học dựa trên chuyển đổi là gì ? .....	32
2.2.2. Giải thuật học dựa trên chuyển đổi tổng quát .....	33
2.2.3. Mô tả về trình tự tạo luật chuyển đổi .....	35
2.2.4. Yêu cầu trong việc áp dụng thuật toán học dựa trên chuyển đổi vào xử lý ngữ nghĩa .....	37
2.2.5. Nhận xét .....	38
2.3. MỘT SỐ GIẢI THUẬT HỌC DỰA TRÊN CHUYỂN ĐỔI CẢI TIẾN .....	39
2.3.1. Lazy TBL .....	39



2.3.2. TBL đa chiều.....	40
2.3.3. TBL nhanh .....	40
2.4. THUẬT TOÁN FAST-TBL.....	41
2.4.1. Quy ước.....	41
2.4.2. Phát sinh luật.....	42
2.4.2.1. Trường hợp 1 .....	43
2.4.2.2. Trường hợp 2 .....	44
2.5. VĂN PHẠM PHỤ THUỘC.....	46
2.5.1. Giới thiệu .....	46
2.5.2. Vận dụng văn phạm phụ thuộc vào xử lý ngữ nghĩa.....	49
2.5.3. Các loại quan hệ trong bộ phân tích cú pháp dựa trên văn phạm phụ thuộc.....	50
<b>Chương 3 MÔ HÌNH CÀI ĐẶT .....</b>	<b>53</b>
3.1. CÁC NGUỒN TRI THỨC ĐỂ XỬ LÝ NGỮ NGHĨA .....	54
3.1.1. Tri thức về từ loại và hình thái.....	54
3.1.2. Tri thức về ngôn từ.....	56
3.1.3. Tri thức về quan hệ cú pháp và ràng buộc ngữ nghĩa.....	57
3.1.4. Tri thức về chủ đề .....	58
3.1.5. Tri thức về tần suất nghĩa của từ.....	59
3.2. CÁC BƯỚC THỰC HIỆN.....	59
3.3. MÔ HÌNH HUẤN LUYỆN CHO BỘ GÁN NHÃN NGỮ NGHĨA.....	61
3.4. HỆ THỐNG NHÃN NGỮ NGHĨA .....	62
3.4.1. Yêu cầu đối với hệ thống nhãn ngữ nghĩa .....	62
3.4.2. Cơ sở của việc phân lớp ngữ nghĩa.....	63
3.4.3. Nhận xét các hệ thống nhãn ngữ nghĩa có liên quan .....	64
3.5. CHUẨN BỊ NGỮ LIỆU HUẤN LUYỆN.....	66
3.5.1. Giới thiệu kho ngữ liệu song ngữ Anh-Việt VCLEVC .....	66

3.5.2. Rút trích thống kê từ ngữ liệu song ngữ .....	68
3.5.2.1. Thống kê các nghĩa tiếng Việt .....	68
3.5.2.2. Thống kê tần số xuất hiện một nghĩa của từ tiếng Anh .....	69
3.5.2.3. Ý nghĩa.....	70
3.5.3. Xây dựng ngữ liệu huấn luyện.....	70
3.5.3.1. Gán nhãn ngữ nghĩa bán tự động cho ngữ liệu.....	71
3.5.3.2. Xây dựng “ngữ liệu vàng” .....	72
<b>Chương 4 CÀI ĐẶT THỬ NGHIỆM.....</b>	<b>75</b>
4.1. GÁN NHÃN CƠ SỞ .....	76
4.1.1. Mô hình gán nhãn cơ sở.....	76
4.1.2. Xử lý ngôn từ, thành ngữ .....	78
4.1.3. Xử lý ràng buộc lựa chọn.....	79
4.1.3.1. Cơ sở tri thức.....	79
4.1.3.2. Thuật toán .....	79
4.1.4. Xử lý dựa trên lĩnh vực xem xét .....	81
4.1.5. Xử lý dựa trên tần số xuất hiện.....	82
4.2. MẪU LUẬT.....	82
4.2.1. Các từ trong ngữ cảnh .....	83
4.2.2. Từ gốc trong ngữ cảnh .....	83
4.2.3. Từ loại trong ngữ cảnh.....	83
4.2.4. Nhãn ngữ nghĩa trong ngữ cảnh.....	83
4.2.5. Từ có quan hệ ngữ pháp trong ngữ cảnh .....	84
4.2.6. Các nhãn trong ngữ cảnh có quan hệ ngữ pháp .....	84
4.3. GẮN NGHĨA TIẾNG VIỆT .....	84
4.3.1. Các từ không cần gán nghĩa tiếng Việt.....	85
4.3.2. Gắn thêm lượng từ Những .....	86
4.3.2.1. Mô tả .....	86

4.3.2.2. Ngữ liệu và mẫu luật.....	87
4.3.3. Quan hệ giữa động từ “to be” và các trường hợp khác.....	88
4.3.4. Các trường hợp đi kèm với giới từ.....	90
4.3.5. Các trường hợp liên quan đến thành ngữ.....	91
4.4. KẾT QUẢ THỰC HIỆN.....	92
4.4.1. Dãy luật tối ưu.....	92
4.4.2. Dãy luật rút ra để giải quyết việc thêm từ trong tiếng Việt.....	93
4.4.3. Thử nghiệm.....	93
<b>Chương 5 KẾT LUẬN – HƯỚNG PHÁT TRIỂN.....</b>	<b>98</b>
5.1. HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN.....	99
5.2. KẾT LUẬN.....	100
Danh Mục Tài Liệu Tham Khảo.....	101
Phụ Lục 1. Danh Sách Nhãn Ngữ Nghĩa Cơ Bản.....	103
Phụ Lục 2. Danh Sách Các Nhãn Từ Loại.....	106
Phụ Lục 3. Trích Một Số Luật.....	108
Phụ Lục 4. Các Kết Quả Dịch Đạt Được.....	111
Phụ Lục 5. Một Số Kết Quả Dịch Thử Nghiệm.....	123
Phụ Lục 6. Một Số Ví Dụ So Sánh.....	138

## Danh Sách Hình

Hình 1-1 : Các chiến lược trong dịch máy (do nhóm GETA đề xuất).....	3
Hình 1-2 : Một hệ dịch trực tiếp.....	4
Hình 1-3 : Mô hình dịch dựa trên chuyển đổi cú pháp và hình ảnh của chuyển đổi cú pháp trên cây cú pháp tiếng Anh sang tiếng Việt .....	4
Hình 1-4 : Một hệ dịch liên ngôn ngữ cho $n$ ngôn ngữ khác nhau .....	5
Hình 1-5 Các bước xử lý trong hệ dịch máy dựa trên chuyển đổi cú pháp.....	9
Hình 1-6 : Cây phân cấp mã ngữ nghĩa trong LDOCE.....	22
Hình 2-1 : Lưu đồ giải thuật học dựa trên chuyển đổi.....	33
Hình 2-2: Minh hoạ của Samuel về trình tự tạo luật chuyển đổi.....	35
Hình 2-3 : Minh hoạ một cây cú pháp thông thường.....	47
Hình 2-4 : Kết quả khi phân tích câu sử dụng văn phạm phụ thuộc.....	48
Hình 2-5 : Hình ảnh một cây quan hệ phụ thuộc .....	48
Hình 2-6 : Các quan hệ phụ thuộc trong câu <i>She is punished by her parents.</i> ....	51
Hình 2-7 : Các quan hệ phụ thuộc trong câu <i>I installed that old driver into my computer.</i> .....	52
Hình 3-1: Mô hình huấn luyện cho bộ gán nhãn ngữ nghĩa .....	61
Hình 3-2 : Minh hoạ các cặp được liên kết trong ngữ liệu song ngữ .....	66
Hình 3-3 : Thể hiện các mối liên kết của một cặp câu.....	67
Hình 3-4 : Công cụ WordAlignEditor.....	67
Hình 3-5 : Công cụ SenseTaggerEditor .....	71
Hình 4-1 : Mô hình cho phương pháp gán nhãn cơ sở.....	78

## Danh Sách Bảng Biểu

Bảng 2-1 : Một số quan hệ khi phân tích bằng văn phạm phụ thuộc.....	51
Bảng 3-1 : Trích thống kê các nghĩa tiếng Việt dựa vào ngữ liệu song ngữ .....	68
Bảng 3-2 : Trích thống kê tần số xuất hiện của nghĩa tiếng Việt của một từ tiếng Anh dựa vào ngữ liệu song ngữ. ....	69
Bảng 4-1 : Trích mẫu luật để thêm từ <i>những</i> .....	88
Bảng 4-2 : Tóm tắt một số trường hợp giải quyết cho động từ <i>be</i> .....	90
Bảng 4-3 : Một số tri thức được áp dụng để giải quyết giới từ.....	91
Bảng 4-4 : Kết quả một số luật chuyển đổi trong xử lý ngữ nghĩa.....	93
Bảng 4-5 : Kết quả một số luật chuyển đổi dùng để thêm từ tiếng Việt.....	93
Bảng 4-6 : Kết quả thử nghiệm.....	93

Khoa CNTT - ĐH KHTN TP.HCM

Chương 1

# TỔNG QUAN

*Chương này nhằm giới thiệu tổng quan về dịch máy nói chung, và xử lý ngữ nghĩa nói riêng. Chúng tôi đề cập các cách tiếp cận và các công trình trước đây trong xử lý ngữ nghĩa. Trong chương này, chúng tôi còn đề cập đến các mức độ nhập nhằng cũng như các khó khăn trong xử lý ngữ nghĩa.*

## **1.1. SƠ LƯỢC VỀ DỊCH MÁY**

### **1.1.1. Lịch sử của Dịch Máy**

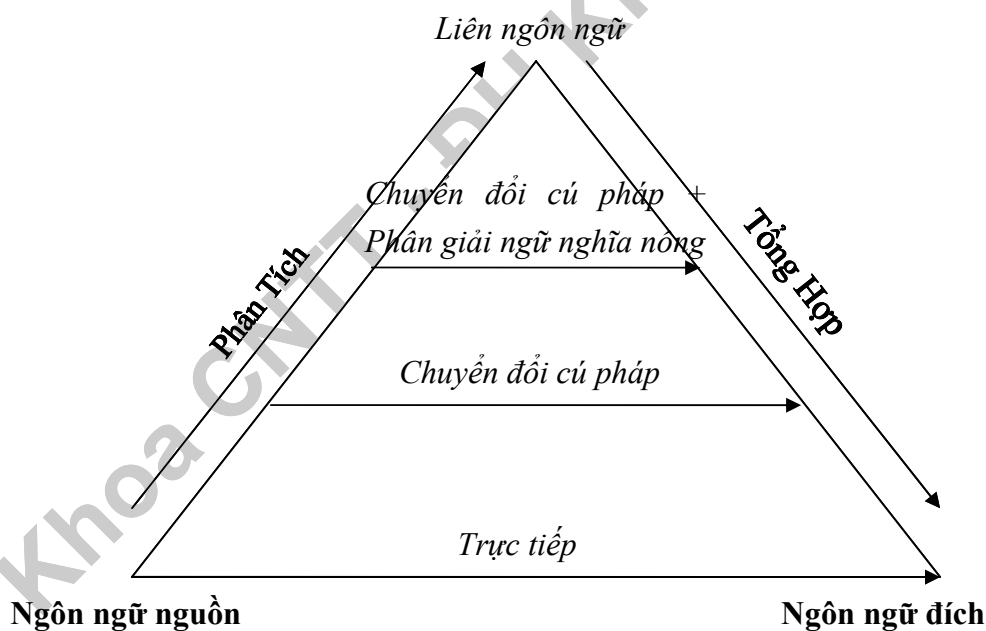
Sau đại chiến thế giới thứ hai, nhờ sự phát triển của máy tính điện tử và do nhu cầu cần nắm bắt những tin tức kịp thời và chính xác trước sự bùng nổ thông tin khoa học - kỹ thuật ngày càng lớn, người ta thấy cần phải trao cho máy tính điện tử nhiệm vụ dịch các văn bản từ ngôn ngữ này sang ngôn ngữ khác, đặc biệt là dịch các tài liệu khoa học - kỹ thuật.

Việc dịch ngôn ngữ tự nhiên hay còn gọi là *Dịch Máy (Machine Translation)* được bắt đầu nghiên cứu từ đầu thập niên 1950. Đây là vấn đề khó khăn nhất trong việc ứng dụng của trí tuệ nhân tạo vào thực tế và cũng là đề tài thời sự gây tranh cãi, và bàn tán sôi nổi từ trước đến nay, lúc hy vọng, lúc thất vọng, lúc phát triển, lúc lu mờ và cũng bị khen và chê nhiều nhất.

Khởi đầu, Dịch Máy cố gắng nhấn mạnh sự quan trọng của việc dịch từng từ dựa trên sự tra tự điển song ngữ và dựa trên thông tin thống kê, tần số từ và những mẫu tuân tự. Trong thời kỳ thập niên 1960, việc Dịch Máy gặp phải nhiều khó khăn và bị chỉ trích. Có trường phái kết luận rằng việc Dịch Máy là không thể thực hiện được và không đáng để bỏ công sức để thực hiện, dẫn đến việc Dịch Máy đã lắng xuống. Những người chống đối lý luận rằng: "... việc dịch ngôn ngữ không những chỉ cần những kiến thức về ngôn ngữ mà còn phải những kiến thức ngoài ngôn ngữ (*extra-linguistic*)...". Trong thời kỳ này (1975) các chính phủ đã không còn trợ cấp cho các chương trình nghiên cứu về Dịch Máy nữa và các chương trình này cũng chấm dứt.

Nhưng may mắn thay, từ cuối thập niên 1980 và nhất là gần đây có một sự trỗi dậy mạnh mẽ việc quan tâm tới việc Dịch Máy và đã đạt được nhiều kết quả đáng khích lệ. Sự hồi sinh này là do kết quả nghiên cứu mới về lý thuyết về ngôn ngữ học, về ngữ pháp học, từ vựng học... và ngoài ra là có sự ra đời những thế hệ máy tính mới có khả năng mạnh hơn nhiều. Tuy nhiên việc Dịch Máy đến nay cũng còn nhiều hạn chế và chỉ dùng chủ yếu phiên dịch các tài liệu kỹ thuật hơn là tác phẩm văn học.

Có nhiều hướng tiếp cận, các chiến lược dịch khác theo cấp độ từ đơn giản đến phức tạp, bao gồm : dịch trực tiếp, dịch theo chuyển đổi cú pháp, chuyển đổi cú pháp + phân giải ngữ nghĩa, dịch qua ngôn ngữ trung gian, dịch dựa trên luật, dịch dựa trên thống kê, dịch dựa trên cơ sở tri thức, dịch dựa trên ngữ liệu... Dưới đây chúng tôi sẽ mô tả một số cách tiếp cận, và chiến lược đó (Xem thêm trong [7]).



**Hình 1-1 : Các chiến lược trong dịch máy (do nhóm GETA đề xuất)**

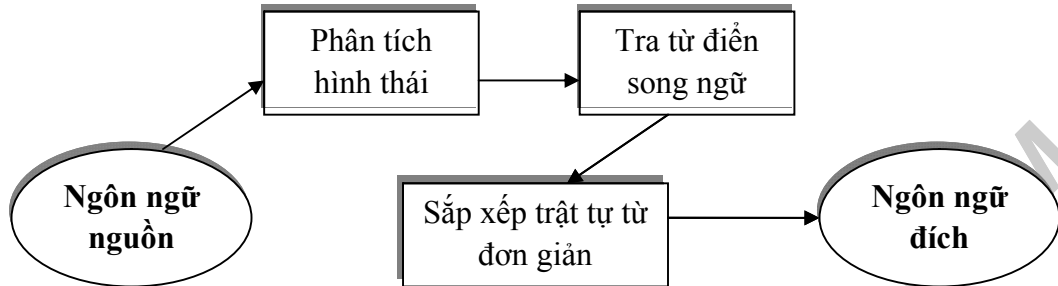
❑ **Dịch trực tiếp :**

Dịch ngôn ngữ bằng cách thay thế những từ trong ngôn ngữ nguồn với những từ trong ngôn ngữ đích một cách máy móc. Những hệ dịch trực tiếp phù hợp

---

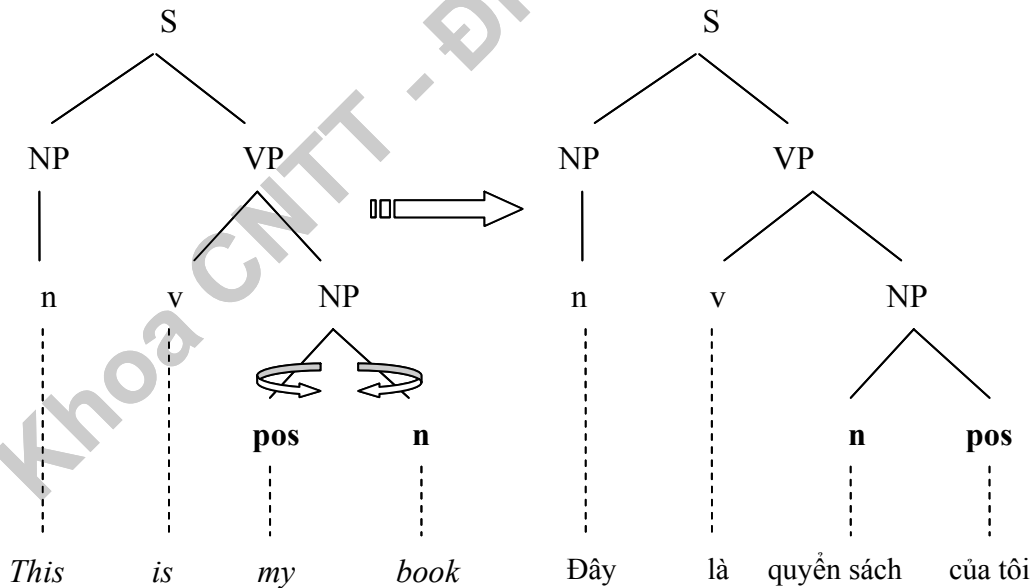
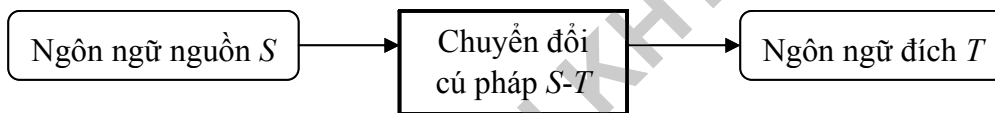


cho những ứng dụng nơi mà văn bản dịch có khối lượng từ nhỏ và số lượng câu giới hạn. Các hệ dịch trực tiếp hoạt động tương đối tốt khi dịch giữa các ngôn ngữ có cùng loại hình.



**Hình 1-2 : Một hệ dịch trực tiếp**

❑ Dịch theo chuyển đổi cú pháp :



**Hình 1-3 : Mô hình dịch dựa trên chuyển đổi cú pháp và hình ảnh của chuyển đổi cú pháp trên cây cú pháp tiếng Anh sang tiếng Việt**

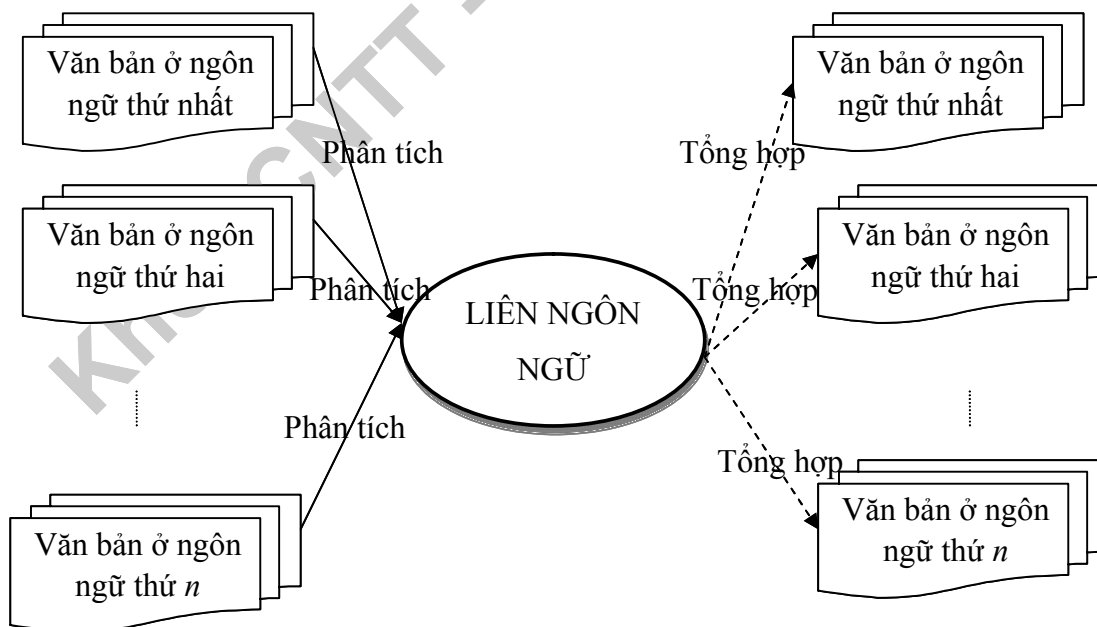
Phân tích cú pháp câu được nhập vào và sau đó áp dụng những luật ngôn ngữ và từ vựng (hay còn được gọi là những luật chuyển đổi) để ánh xạ thông tin văn phạm từ ngôn ngữ này sang ngôn ngữ khác. Theo đó, không thể giải quyết các trường hợp nhập những ngữ nghĩa của câu có cùng cấu trúc nhưng khác nghĩa nhau.

□ **Dịch chuyển đổi cú pháp + cộng phân giải ngữ nghĩa :**

Dung hoà giữa mức độ phân tích cú pháp và phân giải ngữ nghĩa. Hệ chủ yếu dựa vào phân tích cú pháp, và chỉ phân giải ngữ nghĩa ở mức cần thiết để khử nhập những nghĩa thôi.

□ **Dịch qua ngôn ngữ trung gian :**

Xây dựng một ngôn ngữ trung gian biểu diễn độc lập với mọi ngôn ngữ tự nhiên và biểu diễn được mọi sự khác biệt về ý nghĩa đến mức tinh tế nhất của mọi ngôn ngữ có trong hệ dịch đó. Khi dịch một ngôn ngữ nguồn A sang ngôn ngữ đích B thì thực hiện việc chuyển từ ngôn ngữ nguồn A sang ngôn ngữ trung gian, sau đó chuyển từ ngôn ngữ trung gian dịch sang ngôn ngữ đích B. Ưu điểm của hệ liên ngôn ngữ là số lượng bộ dịch được dùng bởi hệ dịch liên ngôn ngữ không nhiều. Song, khó khăn lớn nhất là không dễ xây dựng một ngôn ngữ trung gian !



**Hình 1-4 : Một hệ dịch liên ngôn ngữ cho  $n$  ngôn ngữ khác nhau**

❑ **Dịch dựa trên luật :**

Đây là cách tiếp cận truyền thống xuất phát từ cách làm của các hệ luật dẫn trong hệ chuyên gia trong lĩnh vực trí tuệ nhân tạo. Các luật dẫn được các nhà ngôn ngữ học xây dựng bằng tay. Ưu điểm là dựa được vào lý thuyết ngôn ngữ học. Còn khuyết điểm của các hệ dịch loại này là : tốn công sức xây dựng hệ luật ; các luật không bao quát ; có hiện tượng luật thừa và luật mâu thuẫn...

❑ **Dịch dựa trên thống kê :**

Thay vì xây dựng các từ điển, các quy luật chuyển đổi bằng tay, hệ dịch này tự động xây dựng các từ điển, các quy luật dựa trên thống kê. Cách tiếp cận này không đòi hỏi sự phân tích sâu về ngôn ngữ, chúng thực hiện hoàn toàn tự động các quá trình phân tích, chuyển đổi, tạo câu dựa trên kết quả thống kê có được từ kho ngữ liệu.

❑ **Dịch dựa trên cơ sở tri thức :**

Dựa trên lập luận “*muốn dịch được trước hết phải hiểu được*”, máy tính phải được trang bị tri thức ngôn ngữ và tri thức về thế giới thực y như con người. Đây là một công việc cực kỳ khó khăn. Vì vậy, chất lượng các hệ dịch dựa trên cách tiếp cận này còn rất hạn chế.

❑ **Dịch dựa trên ngữ liệu :**

Đặc điểm của các hệ dịch theo cách tiếp cận này là thay vì xây dựng bộ luật bằng tay, hay dựa trên thống kê thì xây dựng các bộ luật dựa trên các công nghệ máy học để có được các bộ luật chuyển đổi nhờ vào kho ngữ liệu. Các bộ luật này hoàn toàn tuân thủ các lý thuyết ngôn ngữ và dễ đọc hơn các luật rút ra từ thống kê. Các bộ luật này còn có ưu điểm đầy đủ hơn, dễ kiểm soát hơn so với các luật do các nhà ngôn ngữ học đưa ra.

### ***1.1.2. Khái niệm về Dịch Máy***

Khi dùng máy tính điện tử để dịch một văn bản ở ngôn ngữ A, gọi là ngôn ngữ nguồn, sang ngôn ngữ B, gọi là ngôn ngữ đích, người ta cần chuyển văn bản đó vào máy, rồi từ máy, nhờ các qui tắc dịch đã cung cấp sẵn cho nó, chuyển ra văn

---

bản ở ngôn ngữ B. Muốn thế, cần phân tích văn bản A về các mặt từ vựng, cú pháp, ngữ nghĩa rồi chuyển những kết quả đó vào máy. Qua một bộ từ điển máy, ở đó cho sẵn sự tương ứng về từ vựng - ngữ nghĩa, về kết cấu cú pháp giữa 2 ngôn ngữ A và B, chính máy có thể tổng hợp những kết quả đã đưa vào và chuyển ra ngôn ngữ B.

Quá trình dịch máy các văn bản văn học nghệ thuật gặp rất nhiều khó khăn chưa khắc phục được. Cho đến nay chỉ có thể dịch các văn bản khoa học kỹ thuật, loại văn bản có phong cách đơn giản. Nhưng chính hướng nghiên cứu dịch tự động này để thúc đẩy lý thuyết ngôn ngữ học phát triển rất mạnh. Người ta phải chính xác hóa, hình thức hóa các khái niệm ngôn ngữ, phải phát hiện được những sự kiện bản chất trong quan hệ giữa nội dung và hình thức ngôn ngữ, nghiên cứu các điểm giống nhau giữa các ngôn ngữ ([5]).

### **1.1.3. Các bước xử lý trong một hệ Dịch Máy**

Dưới đây mô tả các bước xử lý trong một hệ Dịch Máy được cài đặt bằng phương pháp chuyển đổi cú pháp (*Syntactic Transfer System*) với ngôn ngữ nguồn là tiếng Anh và ngôn ngữ đích là tiếng Việt.

#### **□ Tiền xử lý (*pre-processing*) :**

Văn bản tiếng Anh sau khi được đưa vào hệ Dịch máy được tiền xử lý. Nhiệm vụ của khối này là xử lý sơ bộ văn bản đầu vào, rồi phân tách nó thành các đơn vị rõ ràng để giảm bớt những bước nhập nhằng không đáng có. Bước tiền xử lý bao gồm : tách bỏ những dấu hiệu, những ký tự lạ (những ký tự đồ họa chẳng hạn) ; tách đoạn ; tách câu (nhận dạng được đâu là dấu ngắt câu đúng) ; các danh hiệu, các từ viết tắt...

#### **□ Phân tích hình thái tiếng Anh (*morphological analysis*) :**

Kể từ giai đoạn này, đơn vị xử lý của hệ Dịch Máy là câu. Các câu này lấy được nhờ vào phần *Tiền xử lý*. Mục đích của bước này là xác định đúng từ loại (Part-Of-Speech) của từ tiếng Anh và từ gốc của nó ; nhận dạng những tên riêng (tên địa danh, tên người, địa chỉ email, địa chỉ website).

❑ **Phân tích cú pháp tiếng Anh (*syntactic analysis*) :**

Nhờ vào từ loại của các từ có được từ bước xử lý trước, bước này sẽ xác định được các ngữ trong câu tiếng Anh (ngữ động từ, ngữ danh từ, ngữ giới từ...), chủ ngữ, vị ngữ, tạo cây cú pháp cho câu tiếng Anh. Những thông tin này sẽ được chuyển sang cho bộ phận xử lý ngữ nghĩa và bộ phận chuyển đổi sang cây cú pháp tiếng Việt.

❑ **Xử lý ngữ nghĩa dựa trên tiếng Anh (*semantic processing*) :**

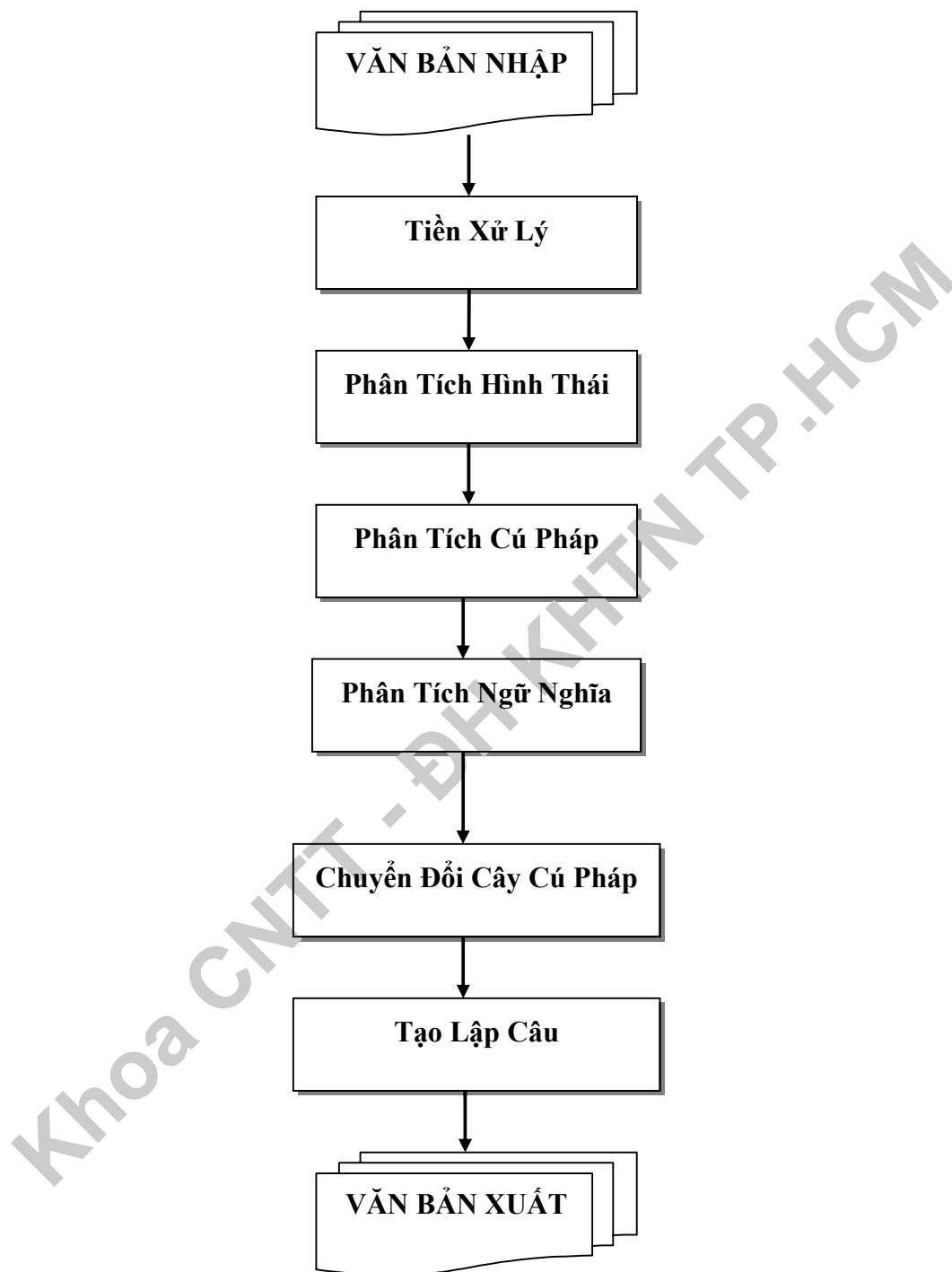
Mục đích của giai đoạn này là từ những thông tin có được của các bước trước (từ loại, cây cú pháp) kết hợp với các thông tin về ngữ cảnh để chọn ra được một *nghĩa thích hợp nhất* cho từ trong câu tiếng Anh.

❑ **Chuyển đổi cây cú pháp tiếng Anh sang tiếng Việt (*syntactic tree transfer*) :**

Bộ phận này nhận cây cú pháp tiếng Anh (từ bộ phận phân tích cú pháp), sau đó chuyển đổi cây cú pháp đó sang cây cú pháp tiếng Việt. Bộ phận này sử dụng các thông tin có được từ bộ phận xử lý ngữ nghĩa để đạt được hiệu quả chuyển đổi cao nhất.

❑ **Tạo câu tiếng Việt nhờ cây cú pháp tiếng Anh đã được chuyển đổi:**

Với cây cú pháp tiếng Anh đã được chuyển sang cây cú pháp tiếng Việt, lúc này hệ thống đạt được trật tự của các từ theo câu tiếng Việt. Gắn kết với kết quả của bước xử lý ngữ nghĩa để tạo thành một câu tiếng Việt cho câu tiếng Anh. Bước này còn phải thực hiện một công việc khác là hoàn chỉnh câu tiếng Việt, điều đó có nghĩa là phải thêm những *hư từ* vào câu tiếng Việt sao cho giúp người đọc càng dễ hiểu càng tốt.



Hình 1-5 Các bước xử lý trong hệ dịch máy dựa trên chuyển đổi cú pháp

## 1.2. XỬ LÝ NGỮ NGHĨA TRONG DỊCH MÁY

### 1.2.1. Vai trò và chức năng của xử lý ngữ nghĩa

Có thể xem việc xác định đúng nghĩa của từ (*xử lý ngữ nghĩa*) là một vấn đề trung tâm của mọi hệ xử lý ngôn ngữ tự nhiên. Hiệu quả làm việc của bộ phận xác định nghĩa của từ có một ảnh hưởng rất lớn đến chất lượng thực hiện của một hệ xử lý ngôn ngữ tự nhiên. Trong một hệ dịch máy, vấn đề xử lý ngữ nghĩa đóng vai trò cốt lõi và hết sức quan trọng. Nó quyết định tính đúng đắn và hiệu quả của một hệ dịch. Một hệ dịch không xử lý tốt ở bộ phận này sẽ dẫn đến kết quả dịch sai nghĩa hoàn toàn thậm chí có thể dẫn đến một câu kết quả hết sức ngớ ngẩn, không thể hiểu nổi.

- Dưới đây là một số ví dụ về trường hợp nhập nhằng gây lỗi cho một hệ dịch máy :

#### Ví dụ 1-1 : I can can a can.

Một câu nhìn vào tưởng chừng rất đơn giản nhưng không dễ giải quyết vì từ *can* có đến 3 nghĩa : (1) *có thể* (động từ hình thái) ; (2) *đóng hộp* (động từ) ; (3) *cái hộp* (danh từ). Trong câu ví dụ trên, cả 3 nghĩa của từ *can* đều xuất hiện. Chỉ cần chọn sai một nghĩa của từ *can* trong câu trên sẽ dẫn đến không hiểu được ý của câu trên, câu tiếng Việt sẽ trở nên ngớ ngếch. Một kết quả thường gặp của câu dịch trên là : Tôi có thể có thể một có thể, trong khi câu trên đáng lẽ phải được dịch là : Tôi có thể đóng hộp một cái hộp.

Người ta nhận thấy rằng muốn giải quyết nhập nhằng tốt cho câu trên cần phải có được một bộ gán nhãn từ loại thật tốt. Lý do là mặc dù có đến 3 nghĩa khác nhau nhưng các nghĩa của từ *can* đã có thể phân biệt được thông qua từ loại của chúng<sup>1</sup>.

---

<sup>1</sup> Nhờ bộ phân tích hình thái tốt, ta có kết quả như sau : I/PRP can/MD can/VB a/DT can/NN.

---

**Ví dụ 1-2 : I enter the new bank(1) near the bank(2) of SaiGon river.**

Nhìn vào trong câu ví dụ trên, người ta dễ dàng nhận được câu dịch chính xác của nó : Tôi **đi vào ngân hàng mới gần bờ** của sông Sài Gòn. Nhưng đối với một hệ dịch máy, đây là một câu chứa nhập nhằng. Nhập nhằng được phát hiện ở 2 từ trong câu trên. Thứ nhất là từ **enter** và thứ hai là từ **bank**.

Áp dụng cách giải quyết của ví dụ trên, tức là có bộ phân tích hình thái thật tốt, ta được : **I/PRP enter/VBP the/DT new/JJ bank/NN of /IN SaiGon/NNP river/NN**. Tuy nhiên, vẫn không thể nào giải quyết được nhập nhằng được cho 2 từ nêu trên.

Từ **enter** có 2 nghĩa động từ (VB): (1) *đi vào*; (2) *nhập* (như trong câu *I enter data into new computer*). Và từ **bank** cũng có 2 nghĩa danh từ (NN) : (1) *ngân hàng* ; (2) *bờ sông*.

Để giải quyết nhập nhằng cho trường hợp này phải sử dụng đến một thông tin khác về các quan hệ trong câu. Ở đây, một quan hệ được tìm thấy giữa **enter** và **bank(1)**, và một quan hệ có được giữa **bank(2)** và **river**. Đầu tiên, vận dụng các ý niệm của ngôn ngữ học tri nhận để biết rằng *enter* là hành động *đi vào không gian kín*. Trong khi với nghĩa *bờ sông* từ *bank* chỉ một *không gian hở*, còn với nghĩa *ngân hàng mới* chỉ một *không gian kín*. Thông qua mối quan hệ giữa từ *enter* và từ *bank* mà chúng ta có thể xác định được nghĩa của cả hai từ. Kế đến, quan hệ giữa **bank(2)** và **river** cho biết nghĩa của từ *bank* phải có thuộc tính tự nhiên, từ đó chọn được nghĩa thích hợp của từ *bank(2)* là *bờ sông*.

Tóm lại, vấn đề giải quyết nhập nhằng ngữ nghĩa là hết sức then chốt và quyết định trong mọi hệ dịch. Một bộ phận giải quyết nhập nhằng ngữ nghĩa hiệu quả sẽ góp phần cải thiện khả năng dịch và độ chính xác của hệ dịch máy một cách đáng kể.



### 1.2.2. Các mức độ nhập nhằng trong tầng xử lý ngữ nghĩa

#### 1.2.2.1. Nhập nhằng ở mức từ vựng

Như câu ví dụ *I enter the bank* ở trên, sau khi phân tích cú pháp, máy tính đã xác định được mối quan hệ giữa động từ *enter* (đi vào) và tân ngữ của nó là *bank* nhưng để chọn nghĩa thích hợp cho từ *bank* (nghĩa *ngân hàng* hay *bờ sông*) thì phải phân tích ngữ nghĩa của động từ *enter* và danh từ *bank*. Trong trường hợp này, vận dụng các ý niệm của ngôn ngữ học tri nhận để biết rằng *enter* là hành động *đi vào không gian kín* trong khi với nghĩa *bờ sông* từ *bank* chỉ một *không gian hở*, còn với nghĩa *ngân hàng* mới chỉ một *không gian kín*. Thông qua mối quan hệ giữa từ *enter* và từ *bank* mà chúng ta có thể xác định được nghĩa của cả hai từ.

Một vài ví dụ cụ thể cho trường hợp này :

Làm sao xác định được nghĩa (tiếng Việt) của từ *old* trong các cụm từ sau : *old man* và *old book*. Các nghĩa của từ *old* đều chỉ một tính chất cũ kỹ, nhưng với con người thì từ *old* có thể có nghĩa *già* hoặc *cũ* trong khi đối với vật chất thì từ *old* chỉ có thể là *cũ* mà thôi. Nhờ đâu mà ta có thể dịch cụm *old man* là *ông già*, còn *old book* là *quyển sách cũ* ? Chúng ta sẽ tìm thấy câu trả lời ở các phần sau.

Một ví dụ thêm nữa rơi vào động từ *enter* (đi vào, nhập) trong hai câu : *I enter the new bank*; và *I enter data into computer*. Câu đầu tiên phải được dịch là *Tôi đi vào ngân hàng* còn câu thứ hai phải được dịch là *Tôi nhập dữ liệu vào máy tính*.

#### 1.2.2.2. Mức độ nhập nhằng cấu trúc

Xét ngữ *Old man and woman*, ta có 2 phân tích : *[Old man] and [woman]* và *Old [man and woman]*. Mỗi phân tích, khi áp dụng vào trong một hệ xử lý ngôn ngữ tự nhiên, sẽ có một cách hiểu khác nhau. Ví dụ trong hệ dịch tự động, cụm từ trên có thể được dịch là *Ông già và người đàn bà* đối với cách phân tích thứ nhất ; và có thể được dịch thành *Ông già và người đàn bà già* đối với cách phân tích thứ hai. Tuy nhiên, chọn cách dịch nào sẽ được quyết định trong bộ phận xác định nghĩa của từ. Trong trường hợp này, bộ xác định sẽ thiên về (chọn) cách phân tích thứ hai

---

do tri thức nhận được về cấu trúc song song *parallel structure* trong ngôn ngữ thông qua liên từ *and*.

Song không phải lúc nào bộ xác định nghĩa cũng chọn một cách (phân tích thứ hai). Hãy xét thêm một ví dụ : *Old man and child*. Cụm từ này cũng được phân tích theo hai cách : *[Old man] and child*, và *Old [man and child]*. Trước khi nói cách xử lý nhập nhằng của bộ xác định nghĩa, chúng ta hãy dịch hai cách phân tích này sang tiếng Việt để dễ hình dung. Đối với cách phân tích thứ nhất, ta có câu dịch *Ông già và đứa trẻ*, trong khi đối với cách phân tích thứ hai ta lại có *Ông già và đứa trẻ già*. Không cần phải nói thêm thì chúng ta cũng có thể biết được cần chọn cách dịch nào ! Tại sao phân tích thứ nhất lại hợp lý hơn phân tích thứ hai ? Như chúng ta đều biết, từ *child* bản thân đã mang tính *trẻ*. Nếu theo cách phân tích thứ hai thì chúng ta đã tạo ra một mâu thuẫn giữa *già* và *trẻ*. Đó là lý do vì sao cách phân tích thứ nhất đã được chọn.

### 1.2.2.3. Mức độ nhập nhằng liên câu

Có một cặp câu ví dụ khá điển hình cho mức độ nhập nhằng liên câu. Hãy xét cặp câu ví dụ sau :

#### Ví dụ 1-3 :

*The monkey ate the banana because it was hungry*

và *The monkey ate the banana because it was ripe.*

Cặp câu này có vẻ rất đơn giản vì chúng ta sẽ không thấy rõ được sự nhập nhằng nếu chỉ đơn thuần dịch câu này (sang tiếng Việt). Với câu thứ nhất, câu dịch là *Con khỉ ăn chuối vì nó đói* và câu thứ hai được dịch là *Con khỉ ăn chuối vì nó chín*. Tới đây, chắc chắn chúng ta còn thắc mắc : nói nhập nhằng nhưng nhập nhằng ở điểm nào. Quá dễ hiểu và dễ thực hiện trong việc xác định nghĩa (!?). Nhưng hãy chú ý đến đại từ *it*. *It* trong câu thứ nhất chỉ về *monkey*; trong khi *it* trong câu thứ hai lại chỉ về *banana*. Có thể nó sẽ không rõ ràng vì *it* nào cũng được dịch là *nó*. Nhưng điều đó lại thực sự quan trọng trong hệ hiểu văn bản. Muốn hiểu được thì phải biết *it* nào chỉ cái nào (*it* – *monkey* hay *it* – *banana*). Một trong các cách hiểu

---

được ghi nhận là xác định đại từ nhân xưng có thể đại diện cho những (cụm) từ nào. Dựa vào các quan hệ đã có để giải quyết nhập nhằng. Ví dụ trong cặp câu trên, *it* có thể đại diện cho *monkey* hoặc *banana*. Ở câu thứ nhất, với *it = monkey*, thì quan hệ *monkey – hungry* mới hợp lý (vì động vật mới đói bụng !), còn *it = banana* thì quan hệ *banana – hungry* là không hợp lý! Còn ở câu thứ hai, với *it = monkey* thì quan hệ *monkey – ripe* là không hợp lý, chỉ có quan hệ *banana – ripe* mới hợp lý.

#### 1.2.2.4. Mức độ nhập nhằng theo thể loại văn bản

Ở mức độ nhập nhằng này, một từ hay một ngữ có thể mang nhiều hơn một nghĩa đúng. Cụ thể sẽ có nhiều kết quả đúng đồng thời, dẫn đến việc chọn lựa nghĩa của chúng phải được kết hợp thêm thông tin về thể loại văn bản.

#### Ví dụ 1-4 : an old driver

Ta có nghĩa các từ *an* : một, *old* : già (đối với người), cũ (đối với đồ vật), *driver* : người tài xế, trình điều khiển (máy tính). Với các nghĩa của từ ta có thể nhận được các câu dịch sau:

- Một tài xế cũ (1)
- Một trình điều khiển già (2)
- Một tài xế già (3)
- Một trình điều khiển cũ (4)

Đối với nghĩa (1), (2) ta có thể thấy đây là hai nghĩa hoàn toàn sai. Nghĩa (3) và (4) đưa ra là những nghĩa đúng. Đối với ngữ cảnh thông thường thì nghĩa (3) sẽ được ưu tiên hơn. Tuy nhiên, nếu văn bản đang dịch ở thể loại tin học thì nghĩa (4) sẽ ưu tiên được chọn. Như vậy, vấn đề khử nhập nhằng nghĩa cũng rất cần thông tin về thể loại văn bản trong quá trình xử lý.

### 1.2.3. Các khó khăn trong xử lý ngữ nghĩa

Từ những phân tích ở các phần trên, có thể rút ra các điểm khó khăn chính trong xử lý ngữ nghĩa như sau :

#### 1.2.3.1. Nhập nhằng nghĩa

Đây là một vấn đề hết sức phức tạp trong xử lý ngữ nghĩa bởi tính đa nghĩa của một từ. Một từ với một chức năng ngữ pháp có thể có nhiều nghĩa khác nhau.

Từ *line* có các nghĩa như sau:

- Hàng (*line of people* : hàng người)
- Dòng (*line printing device* : thiết bị in dòng)
- Đường kẻ (*a thin line* : một đường kẻ mỏng)
- Đường dây (*telephone line* : đường dây điện thoại)
- Tuyến xe (*bus line* : tuyến xe buýt)

Việc chọn lựa nghĩa phù hợp trong câu là một vấn đề khó khăn vì cần phải hiểu được mối quan hệ của từ với ngữ cảnh xung quanh để nhận biết nghĩa chính xác của từ.

#### 1.2.3.2. Phụ thuộc vào ngữ cảnh

Một ý nghĩa của một từ có nghĩa khác nhau nếu nằm trong những ngữ cảnh khác nhau. Ngữ cảnh ở đây có thể được xem như là nội dung của văn bản đang đề cập, ý nghĩa của các câu trước hoặc sau có liên quan đến nó trong đoạn văn, hoặc các từ có liên quan với nó trong câu. Chúng ta sẽ thấy yếu tố ngữ cảnh sẽ tác động như thế nào đến ngữ nghĩa của cụm từ *an old driver*. Nếu ta viết *An old driver drives the car*. thì nghĩa ở đây của *an old driver* là *một người tài xế già* và nếu ta viết *I installed that old driver into this computer*. thì cụm đó lại mang nghĩa là *trình điều khiển cũ*.

#### 1.2.3.3. Phụ thuộc vào tri thức

Ngôn ngữ là phương tiện giao tiếp của con người. Con người sử dụng ngôn ngữ để thể hiện những điều mình nhận thức được trong thế giới xung quanh. Những

---

nhận thức đó chính là tri thức. Do vậy, khi thể hiện những điều mình muốn bằng ngôn ngữ, thì bản thân những điều đó phải phù hợp với tri thức đang có. Ví dụ chúng ta không thể nói *Chiếc xe ăn hết thức ăn* hay *Cái điện thoại đi ngủ*. Xử lý ngữ nghĩa cũng không thể tránh khỏi những vấn đề đó, cần phải biết phân biệt những vấn đề không hợp lý trong ngôn ngữ. Tuy nhiên, để thể hiện được tất cả tri thức không phải là một vấn đề dễ dàng và đang là bài toán hóc búa đối với các nhà khoa học.

#### 1.2.3.4. Sự khác biệt giữa tiếng Anh và Việt

Tiếng Anh và tiếng Việt là ngôn ngữ của hai dân tộc khác nhau, có nền văn hóa khác nhau. Vì vậy, yếu tố khác nhau giữa tiếng Anh và tiếng Việt là một khó khăn trong vấn đề xử lý ngữ nghĩa. Có những khái niệm trong tiếng Anh có thể sử dụng cho tất cả sự vật với cùng một nghĩa nhưng trong tiếng Việt thì không phải như vậy. Ví dụ cho phần này là cụm từ *old book* và *old man* đã được nêu ở trên.

#### 1.2.3.5. Yếu tố khác

Như đã đề cập ở trên, khối xử lý ngữ nghĩa là bước tiếp theo của khối phân tích cú pháp. Do đó kết quả của xử lý ngữ nghĩa chịu ảnh hưởng của khối phân tích cú pháp. Cây cú pháp do khối phân tích có thể đưa ra sai, hoặc quá phức tạp, hoặc thiếu những cấu trúc cú pháp mà khối xử lý ngữ nghĩa cần. Bên cạnh đó, ngoài kiến thức Tin học, công việc xử lý ngữ nghĩa trong hệ dịch máy cần phải có những kiến thức về ngôn ngữ học, tiếng Anh cũng như tiếng Việt. Những kiến thức này hỗ trợ cho việc tìm mối quan hệ giữa cú pháp và ngữ nghĩa, mối quan hệ giữa các nghĩa, sự phân loại...

### 1.3. CÁC CÁCH TIẾP CẬN TRONG XỬ LÝ NGỮ NGHĨA VÀ CÁC CÔNG TRÌNH TRƯỚC ĐÂY

#### 1.3.1. Xử lý ngữ nghĩa trong thời gian đầu

Trong một công trình có từ năm 1949, Weaver thảo luận sự cần thiết phải xác định nghĩa đúng của từ trong dịch máy và định ra những bước cơ bản trong khâu nhập nhằng nghĩa (Xem thêm trong [13]). Ông cho rằng : Nếu một ai đó xem xét từng từ một trong một quyển sách thì rõ ràng người đó không thể xác định được nghĩa của tất cả các từ. Ông cũng cho rằng, nếu như mở rộng vùng xem xét xung quanh từ đó thì không những xác định được nghĩa của từ đó mà còn có thể xác định thêm được nghĩa của những từ xung quanh nữa. Nhưng vùng xung quanh đó có kích thước của sổ xem xét là bao nhiêu ? Năm 1950, một thí nghiệm nổi tiếng do Kaplan thực hiện nhằm tìm câu trả lời cho câu hỏi nêu trên. Kaplan dùng 7 từ để xem xét, và vùng cửa sổ xung quanh xem xét được thay đổi từ một đến hai từ mỗi bên của từ cần xem xét. Kaplan quan sát rằng độ phân giải nghĩa được đưa 2 từ trên mỗi bên của từ xem xét không tốt hơn cũng như không tệ hơn khi đưa toàn bộ câu.

“Sự trùng khớp ngữ nghĩa” (semantic coincidence) (do Reifler đưa ra năm 1955) giữa một từ và ngữ cảnh của nó (xét trên độ phức tạp của ngữ cảnh và vai trò của quan hệ cú pháp) nhanh chóng trở thành một yếu tố quyết định trong việc xác định đúng nghĩa của từ. Reifler cho rằng : *Cấu trúc ngữ pháp có thể giúp khử nhập nhằng nghĩa cho từ.* Ví dụ, với từ *keep*, có thể xác định nghĩa đúng cho nó dựa trên việc xác định túc từ của nó : túc từ của nó là một danh động từ (gerund) (He **kept** eating - Anh ấy **tiếp tục** ăn) hay ngữ tính từ và ngữ danh từ (He **kept** calm – Anh ấy **giữ** bình tĩnh ; He **kept** a record – Anh ấy **giữ** một kỷ lục).

Trong giai đoạn này, dịch máy chủ yếu tập trung vào việc dịch các tài liệu kỹ thuật. Do đó đã có những nghiên cứu về vai trò của lĩnh vực (domain) trong việc khử nhập nhằng cho nghĩa mà sau đó vài thập kỷ (năm 1992) được Gale, Church và Yarowsky lặp lại. Cũng liên quan đến việc sử dụng lĩnh vực của tài liệu cần dịch, có những nghiên cứu nhằm tạo ra các từ điển chuyên dụng. Các từ điển này chỉ chứa

---

những nghĩa thích hợp của một từ nào đó trong các văn bản chỉ của lĩnh vực đó. Ví dụ, một từ điển cho dịch máy về lĩnh vực toán học, không hề chứa nghĩa *kẽng ba góc* (một dụng cụ âm nhạc) của từ *triangle*, mà chỉ chứa nghĩa *hình tam giác* của từ này.

Một điều khá lý thú là ngay trong giai đoạn này cách tiếp cận thống kê đã được đề cập đến (trong công trình của Weaver). Nhiều tác giả đã thực hiện theo công trình này (như Richards năm 1953; Yngve năm 1955, Parker-Rhodes năm 1958). Các ước lượng về mức độ nhập nhằng trong văn bản và trong từ điển được thực hiện bao gồm : Harper xác định số lượng từ nhập nhằng trong một tài liệu vật lý là 30% ; hay Bel'skaja đưa ra con số 500 trong tổng số 2000 từ của từ điển điện toán tiếng Nga đầu tiên là từ đa nghĩa... Còn với Pimsleur, trong năm 1957, ông đề nghị hai mức độ sâu trong dịch : mức đầu tiên là dùng nghĩa thường gặp nhất (ông đưa ra kết quả 80% giải quyết đúng), mức thứ hai, phân biệt các nghĩa thêm (giải quyết được 90% trường hợp đúng). Cách này khá giống với các phương pháp gán nhãn baseline được thực hiện trong những năm gần đây.

### **1.3.2. Dựa trên trí tuệ nhân tạo**

Đây là cách tiếp cận với những lý thuyết rất hay về mạng ngữ nghĩa, khung ngữ nghĩa, và các ý niệm nguyên thủy (như : THING, DO, CAUSE...) và các quan hệ như IS-A, PART-OF... Tuy nhiên, do hầu hết các tri thức về ngữ nghĩa trong cách tiếp cận này đều được xây dựng bằng tay (nên không thể xây dựng được nhiều tri thức về thế giới thực), vì vậy các mô hình này đều dừng lại ở mức độ biểu diễn trên một vài câu. Chẳng hạn các mô hình dùng mạng suy diễn tri thức ngữ nghĩa, dùng logic hình thức, logic – ngôn ngữ, ngữ nghĩa hình thức mà trong đó đều chứa tri thức là “người là động vật, có khả năng suy nghĩ, nói năng, học tập...”. Nhưng trong thực tế thì “*trẻ sơ sinh chưa có thể nói được*” và ngược lại có trường hợp “*người bay được*” như chúng ta thấy trong câu “*Tôi sẽ bay vào sáng mai*”.

Mạng ngữ nghĩa (Semantic Network) được phát triển vào cuối những năm 1950 và nhanh chóng được áp dụng vào trong bài toán biểu diễn nghĩa cho từ. Năm 1962, Masterman sử dụng một mạng ngữ nghĩa để thu được biểu diễn câu trong một

---

liên ngôn ngữ gồm những khái niệm ngôn ngữ cơ sở. Sự phân biệt nghĩa được thực hiện bằng cách chọn các biểu diễn phản ánh được các nhóm nút có liên quan gần gũi trong mạng. Masterman phát triển một tập gồm 100 loại ý niệm cơ sở (THING, DO). Nhóm của bà xây dựng một từ điển ý niệm gồm 15.000 mục dựa trên tập đó. Trong từ điển đó, các loại ý niệm được tổ chức trong một lattice với sự kế thừa thuộc tính từ cao đến thấp. Quilian xây dựng một mạng bao gồm các liên kết giữa các từ và ý niệm. Các liên kết được gán nhãn với các quan hệ ngữ nghĩa khác nhau để chỉ mối liên kết giữa các từ. Mạng đó được tạo ra từ từ điển nhưng tri thức thể giới thực được mã hoá bằng tay. Khi hai từ được đưa vào mạng, chương trình giả lập kích hoạt tuần tự các nút ý niệm dọc theo đường chứa các liên kết bắt đầu từ phía mỗi từ. Việc khử nhập nhằng được thực hiện do chỉ có một nút ý niệm của mỗi từ nằm trên đường ngắn nhất nối giữa hai từ.

Các cách tiếp cận dựa trên trí tuệ nhân tạo tiếp theo sử dụng các frame chứa thông tin về từ, vai trò và quan hệ của nó với các từ khác trong một câu. Ví dụ, Hayes kết hợp mạng ngữ nghĩa và các frame vai (case frame). Mạng bao gồm các nút thể hiện các nghĩa danh từ và các liên kết do các nghĩa động từ thể hiện. Các case frame sử dụng quan hệ IS-A (là một) và PART-OF (bộ phận của) trên mạng. Về mặt bản chất, cách tiếp cận *preference semantics* (ngữ nghĩa ưu tiên) của Wilk sử dụng các ý niệm nguyên thủy của Masterman là một cách tiếp cận dựa trên vai (case-based) trong hiểu ngôn ngữ tự nhiên và là một trong những cách tiếp cận đầu tiên được thiết kế đặc biệt cho bài toán khử nhập nhằng nghĩa của từ. *Preference semantics* xác định các ràng buộc lựa chọn<sup>2</sup> (selectional restriction) cho các kết hợp giữa các từ trong câu. Trong khi đó, Boguraev cho rằng *preference semantics* không thích hợp cho các động từ đa nghĩa và đã cố gắng cải tiến phương pháp của Wilk bằng cách sử dụng các thông tin về ràng buộc lựa chọn, ràng buộc ưu tiên và case frame. Giống như nhiều hệ thống khác, các hệ thống nêu trên dựa vào đơn vị câu,

---

<sup>2</sup> Một ví dụ về ràng buộc lựa chọn là : *My car drinks gasoline* – Xe tôi uống xăng. Có ràng buộc trong câu trên vì động từ *drink* chỉ thích hợp với chủ thể là một vật thể sống chứ không thể có chủ thể là một vật thể không sống như *car* trong câu trên

---



chính vì vậy đã không sử dụng được các mức độ về thông tin lĩnh vực dịch, thông tin về đề tài đang được đề cập. Kết quả là vài loại nhập nhằng rất khó và không thể giải quyết.

Bộ phận xác định nghĩa của hệ hiểu ngôn ngữ của Dahlgren sử dụng các loại thông tin khác nhau như các ngữ cố định, thông tin cú pháp (để tạo các ràng buộc lựa chọn) và khối lập luận nghĩa thông thường. Chỉ khi hai thông tin đầu tiên không tạo được kết quả thì mới áp dụng khối thứ ba. Khối này thông qua một bản thể học (ontology) để tìm các ancestor thông thường của từ trong ngữ cảnh nhằm xác định độ tương tự bản thể học (ontological similarity). Độ tương tự này là một thành phần khử nhập nhằng khá mạnh. Bà Dahlgren cũng lưu ý rằng ràng buộc lựa chọn của động từ là một nguồn thông tin quan trọng cho việc khử nhập nhằng cho danh từ.

### **1.3.3. Dựa trên cơ sở tri thức**

Các công trình dựa trên trí tuệ nhân tạo của những năm 1970, 1980 rất hay về mặt lý thuyết nhưng không thực tế tí nào vì việc tạo ra một lượng lớn tri thức cần thiết cho khử nhập nhằng của từ rất tốn công sức (còn được gọi là “cổ chai tiếp nhận tri thức” – “knowledge acquisition bottleneck”). Các nghiên cứu trên lĩnh vực này đã chuyển sang một hướng mới vào những năm 1980 khi các tài nguyên như từ điển máy, từ điển đồng nghĩa và ngữ liệu trở nên phổ biến rộng rãi. Người ta cố gắng khai thác tự động tri thức từ những nguồn này, và gần đây là xây dựng các cơ sở tri thức khổng lồ hoàn toàn bằng tay.

#### **1.3.3.1. Từ điển máy**

Các từ điển máy (Machine-Readable Dictionary) ngày càng trở thành một nguồn tri thức phổ biến trong các công việc xử lý ngôn ngữ. Có nhiều công trình liên quan đến việc cố gắng rút trích tự động cơ sở tri thức từ từ điển như của : Michiels, Mullenders, và Noël ; Calzolari ; Chodorow, Byrd, và Heidon ; Markowitz, Ahlswede, và Evens ; Byrd và các cộng sự ; Nakamura và Nagao ; Klavans, Chodorow, và Wacholder ; Wilk và các cộng sự... Các công việc này có những đóng góp đáng kể cho việc nghiên cứu ngữ nghĩa từ vựng, nhưng nó cũng

cho thấy rằng mục tiêu ban đầu – tự động rút trích cơ sở tri thức – khó đạt được hoàn toàn. Hiện nay, cơ sở tri thức cỡ lớn về từ vựng duy nhất được sử dụng rộng rãi là WordNet lại được xây dựng bằng tay. Khó khăn trong công việc này là do sự không đồng nhất trong các từ điển cũng như do các từ điển đó được xây dựng dành cho con người sử dụng chứ không phải để dành cho máy khai thác.

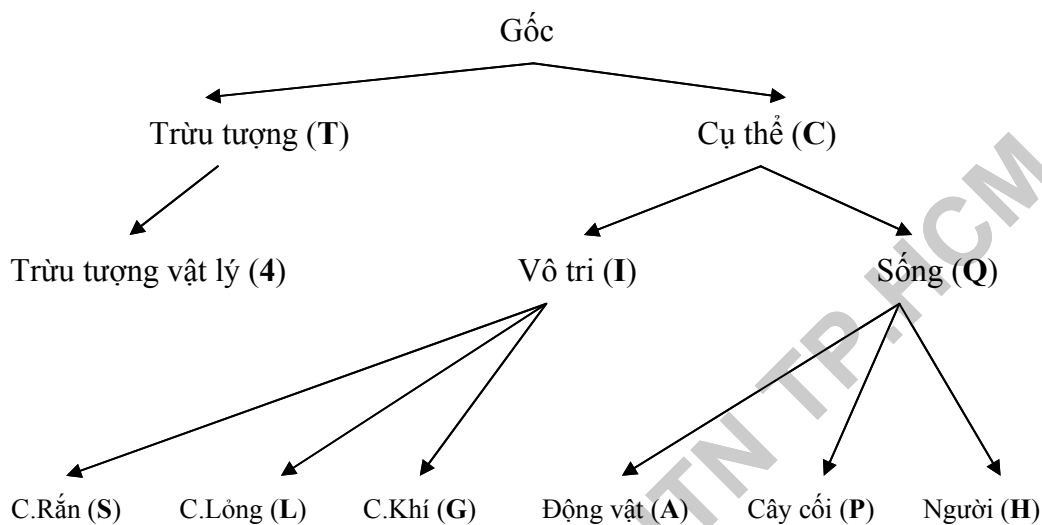
Mặc dù còn có những thiếu sót, song các từ điển máy cung cấp một nguồn thông tin có sẵn cho các nghĩa của từ và vì thế nhanh chóng trở thành nguồn gốc chung cho các nghiên cứu về xử lý ngữ nghĩa. Các phương pháp tiếp theo cố tránh các khó khăn nêu trên thông qua việc sử dụng trực tiếp các định nghĩa, cùng với các cách hiệu quả làm giảm hoặc loại trừ các ảnh hưởng từ tính chất không đồng nhất của từ điển. Tất cả các phương pháp này dựa trên quan điểm : nghĩa hợp lý nhất gán cho những từ xuất hiện đồng thời là nghĩa làm cực đại độ tương quan giữa các nghĩa được chọn.

Năm 1986, Lesk tạo ra một cơ sở tri thức gán mỗi nghĩa trong từ điển với một “*chữ ký*” (thể hiện bằng danh sách các từ xuất hiện trong định nghĩa của nghĩa đó). Việc xác định nghĩa được thực hiện bằng cách chọn nghĩa của từ có “chữ ký” chứa số lượng trùng lặp lớn nhất với các “chữ ký” của các từ trong ngữ cảnh của nó. Phương pháp này chọn nghĩa đúng từ 50% đến 70%. Cách này sẽ dễ bị ảnh hưởng bởi các từ trong các định nghĩa. Tuy nhiên, phương pháp này lại là cơ sở cho hầu hết các công trình xử lý nhập nhằng tiếp theo dựa trên từ điển máy. Wilk và các cộng sự thì tính tần số xuất hiện đồng thời của các từ trong định nghĩa nhằm tạo ra nhiều độ đo độ liên quan giữa các từ để cải tiến tri thức kèm theo mỗi nghĩa. Độ đo này sau đó được dùng với phương pháp vector liên kết mỗi từ và ngữ cảnh của nó.

Về sau, nhiều tác giả (như Krovetz và Croft ; Guthrie và các đồng tác giả ; Janssen ; Braden-Harder ; Liddy và Paik) sử dụng các trường thông tin phụ trong bản điện tử của Từ điển hiện đại tiếng Anh Longman (Longman Dictionary of Contemporary English - LDOCE) (như mã ngữ nghĩa, mã chủ đề của mỗi nghĩa) để cải tiến kết quả. Mã ngữ nghĩa gồm có các ý niệm nguyên thủy (như *Trừu tượng (T)*, *Vật có sự sống (Q)*, *Con người (H)*,...), mã hoá các ràng buộc của danh từ, tính

---

từ và các tham số của động từ. Mã chủ đề phân chia chủ đề cho từ (chẳng hạn về kinh tế, kỹ thuật...).



**Hình 1-6 : Cây phân cấp mã ngữ nghĩa trong LDOCE**

Tuy nhiên, việc dùng các mã ngữ nghĩa của LDOCE lại gặp phải vấn đề do các mã này không có hệ thống. Braden-Harder chỉ ra rằng nếu chỉ đơn giản tìm sự phù hợp giữa mã ngữ nghĩa hay mã chủ đề thì khả năng hiểu nghĩa không hiệu quả. Chẳng hạn, với câu *I tipped the driver*, xét quan hệ giữa từ *tipped* và từ *driver*, có nhiều nghĩa của hai từ này thoả ràng buộc : từ *tip* (với nghĩa liên quan đến tiền – *cho tiền quà*) cần một túc từ chỉ người thì *driver* với nghĩa *tài xế* là phù hợp ; từ *tip* (với nghĩa *đánh gậy*) cần túc từ chỉ một vật thể đặc có thể di chuyển được (movable solid object) thì *driver* với nghĩa *cái bạt đánh gôn* là phù hợp. Do đó câu *I tipped the driver* nếu đơn thuần sử dụng mã ngữ nghĩa thì chưa thể biết được nghĩa chính xác của cả từ *tipped* lẫn từ *driver*.

### 1.3.3.2. Từ điển đồng nghĩa

Từ điển đồng nghĩa (thesaurus) cung cấp thông tin về các mối quan hệ giữa các từ, đáng lưu ý nhất là quan hệ đồng nghĩa. Thông thường, mỗi thể hiện của một từ trong các phạm trù khác nhau biểu diễn các nghĩa khác nhau của từ đó, điều đó

---

có nghĩa là các phạm trù có mối tương ứng mạnh mẽ với các nghĩa của từ. Tập các từ trong cùng một phạm trù có quan hệ ngữ nghĩa.

Giống từ điển máy, từ điển đồng nghĩa là tài nguyên dành cho con người, và vì vậy, không phải là một nguồn thông tin hoàn hảo về các mối quan hệ trong thế giới thực. Người ta đã nhận thấy rằng các tầng phía trên của cây phân cấp ý niệm quá rộng nên rất khó sử dụng để thiết lập các phạm trù ngữ nghĩa đầy đủ. Song, các từ điển đồng nghĩa cung cấp một mạng rộng lớn các mối liên kết của từ và tập các phạm trù ngữ nghĩa nên có tiềm năng cho việc xử lý ngữ nghĩa.

### 1.3.3.3. Từ điển điện toán

Vào giữa những năm 1980, nhiều cơ sở tri thức khổng lồ bắt đầu được xây dựng bằng tay (như WordNet, CyC, ACQUILEX, COMLEX). Có 2 cách tiếp cận cơ bản liên quan đến việc xây dựng các cơ sở tri thức này : cách tiếp cận liệt kê (enumerative approach) và cách tiếp cận sản sinh (generative approach). Trong cách tiếp cận liệt kê, các nghĩa được cung cấp đầy đủ, rõ ràng. Còn trong cách tiếp cận sản sinh, các thông tin ngữ nghĩa liên quan đến một từ không được xác định rõ ràng, thay vào đó các luật sinh được dùng để tạo ra các thông tin nghĩa chính xác.

Trong số các từ điển điện toán được thực hiện theo cách tiếp cận liệt kê, WordNet là từ điển nổi tiếng nhất và được sử dụng nhiều nhất trong xử lý ngữ nghĩa cho từ trong tiếng Anh. WordNet kết hợp được các đặc tính của nhiều loại tài nguyên khác được khai thác thường xuyên trong xử lý ngữ nghĩa. Nó gồm các định nghĩa của các nghĩa riêng biệt như trong từ điển. Nó tổ chức các nghĩa thành cách *tập đồng nghĩa* (synset), tổ chức thành cây ý niệm phân cấp giống như trong từ điển đồng nghĩa (thesaurus). Ngoài ra nó còn bao gồm các mối liên kết giữa các từ theo các quan hệ ngữ nghĩa như hyponymy/hyperonymy, antonymy, và meronymy. Tuy nhiên, WordNet cũng không phải là một nguồn thông tin đầy đủ để xử lý ngữ nghĩa của từ. Lý do thường được đề cập đến là do sự phân biệt nghĩa quá chi tiết của WordNet. Sự phân biệt này đôi khi không cần thiết lắm trong nhiều ứng dụng xử lý ngôn ngữ tự nhiên, trong đó có dịch máy. (Nhưng thật sự là

không dễ gì xác định được phân biệt nghĩa đến mức độ nào thì phù hợp cho công việc xử lý ngữ nghĩa).

Có nhiều nghiên cứu dựa trên WordNet để khử nhập nhằng nghĩa cho từ. Chẳng hạn, Richardson và Smeaton tạo ra một cơ sở tri thức từ cây phân cấp của WordNet và áp dụng hàm tính độ tương tự ngữ nghĩa để giải quyết nhập nhằng ngữ nghĩa trong truy xuất thông tin (information retrieval). Sussna tính độ đo khoảng cách ngữ nghĩa cho mỗi tập các thuật ngữ (danh từ) đưa vào để khử nhập nhằng. Ông ấy gán trọng số dựa trên các loại quan hệ. Điểm hay của nghiên cứu của ông ấy nằm ở chỗ ông ấy không chỉ sử dụng quan hệ IS-A mà còn sử dụng các loại quan hệ khác nữa (quan hệ đồng nghĩa chẳng hạn).

Hầu hết các công trình khử nhập nhằng ngữ nghĩa cho đến nay vẫn dựa chủ yếu vào sự phân biệt nghĩa đã được liệt kê sẵn. Tuy nhiên, gần đây cũng có công trình khử nhập nhằng nghĩa khai thác các từ điển tự sinh như của Pustejovsky, ..

#### ***1.3.4. Dựa trên ngữ liệu***

Ngữ liệu đã được sử dụng trong ngôn ngữ học từ nửa đầu thế kỷ 20. Một vài công trình có liên quan đến nghĩa của từ như : Palmer nghiên cứu về ngôn từ (collocation) trong tiếng Anh ; Lorge tính tần số của nghĩa cho 570 từ tiếng Anh thông dụng nhất ; Eaton so sánh tần số nghĩa trong 4 ngôn ngữ ; Thorndike ; và Zipf xác định rằng có mối tương hỗ giữa tần số và số lượng từ đồng nghĩa của một từ (dấu hiệu cho thấy sự phong phú của ngữ nghĩa, một từ càng đa nghĩa thì nó càng có nhiều từ đồng nghĩa).

Ngữ liệu cung cấp một lượng lớn các mẫu, cho phép phát triển nhiều mô hình ngôn ngữ số, nên việc sử dụng ngữ liệu đi liền với các phương pháp theo kinh nghiệm (empirical method). Mặc dù các phương pháp định lượng/thống kê được quan tâm, theo đuổi trong thời gian đầu của Dịch Máy, nhưng vào giữa những năm 1960, các quan tâm theo hướng thống kê có sút giảm do xu hướng hướng về các luật ngôn ngữ học hình thức từ các lý thuyết của Zellig Harris và lý thuyết chuyển đổi của Noam Chomsky. Trong suốt 10 đến 15 năm sau đó, chỉ có một lượng nhỏ

các nhà ngôn ngữ học theo đuổi các nghiên cứu trên ngữ liệu, hầu hết là cho các mục đích giáo dục và tạo từ điển. Trong hoàn cảnh có ít các nghiên cứu dựa trên ngữ liệu trong thời gian này, một số nhà nghiên cứu như Weiss, Kelley, và Stone vẫn chú ý đến hướng này. Weiss chứng tỏ rằng các luật khử nhập nặng có thể được học từ các ngữ liệu gán nhãn ngữ nghĩa bằng tay. Dù cho kích thước thực nghiệm không lớn lắm (5 từ, mỗi từ có 20 câu huấn luyện và 30 câu dùng để kiểm tra) nhưng kết quả đạt được thì đáng khích lệ (90%). Còn Kelly và Stone thì sử dụng các thông tin về ngôn từ, quan hệ cú pháp, phạm trù ngữ nghĩa để khử nhập nặng cho 1800 từ trong ngữ liệu nửa triệu từ.

Trong những năm 1980, mối quan tâm về ngôn ngữ học ngữ liệu đã được hồi sinh. Các tiến bộ trong công nghệ cho phép tạo ra, lưu trữ ngữ liệu lớn hơn bao giờ hết, và cho phép phát triển các mô hình mới sử dụng các phương pháp thống kê.

Black đã phát triển một mô hình dựa trên cây quyết định sử dụng một ngữ liệu gồm 22 triệu lượt từ, sau khi gán nhãn ngữ nghĩa bằng tay có xấp xỉ 2000 dòng cho 5 từ dùng để thử. Kể từ đó, các phương pháp học có giám sát từ các ngữ liệu được gán nhãn ngữ nghĩa được nhiều nhà nghiên cứu sử dụng như : Zernik ; Hearst; Leacock, Towell, và Voorhees ; Gale, Church, và Yarowsky ; Bruce và Wiebe ; Miller và các cộng sự ; Niwa và Nitta ; Lehman... Mặc dù số lượng các ngữ liệu khổng lồ ngày càng tăng, song hai trở ngại chính trong việc rút trích tri thức từ vựa từ ngữ liệu là : khó khăn của việc gán nhãn ngữ nghĩa bằng tay, và sự thừa thớt dữ liệu.

Gán nhãn ngữ nghĩa bằng tay cho một ngữ liệu là một công việc cực kỳ tốn kém. Hiện tại rất hiếm các ngữ liệu đã được gán nhãn ngữ nghĩa sẵn. Có thể kể ra vài ngữ liệu đã được gán nhãn sẵn : ngữ liệu của Linguistic Data Consortium khoảng 200.000 câu cho tất cả các nghĩa của 191 từ (sử dụng nghĩa của WordNet) ; ngữ liệu của Cognitive Science Laboratory của đại học Princeton. Tuy nhiên, các ngữ liệu còn quá nhỏ hơn nhiều so với các ngữ liệu cần dùng với các phương pháp thống kê.

Nhiều nghiên cứu hướng đến việc tự động gán nhãn ngữ nghĩa cho một ngữ liệu thông qua phương pháp tăng cường. Hearst đề nghị một thuật toán (CatchWord) gồm một pha huấn luyện trong đó các từ đã được gán nhãn ngữ nghĩa bằng tay. Các số liệu thống kê rút ra được từ ngữ cảnh của các từ này được dùng để khử nhập nhằng cho các ngữ cảnh khác. Trong quá trình sử dụng, nếu có trường hợp mới đảm bảo khử nhập nhằng được, hệ thống tự động tiếp nhận các thông tin thống kê từ trường hợp này để cải tiến tri thức của chương trình. Gần đây, lại có đề nghị dùng phương pháp tăng cường dựa trên lớp (class-based bootstrapping) để gán nhãn ngữ nghĩa trong những lĩnh vực xác định.

Khoa CNTT - ĐH KHTN TP. HCM

Khoa CNTT - ĐHQG KHTN TP.HCM

Chương 2

# CƠ SỞ LÝ THUYẾT



*Chương này dẫn giải các cơ sở lý thuyết cần thiết cho xử lý ngữ nghĩa. Các cơ sở lý thuyết đó bao gồm : cơ sở lý thuyết trong ngôn ngữ học, giải thuật học dựa trên chuyển đổi, và văn phạm phụ thuộc. Đối với giải thuật học dựa trên chuyển đổi, chúng tôi thảo luận chi tiết về fnTBL.*

## 2.1. CƠ SỞ LÝ THUYẾT VỀ NGÔN NGỮ HỌC

### 2.1.1. Nghĩa của từ

Là những liên hệ được xác lập trong nhận thức của chúng ta giữa từ với những cái mà nó chỉ ra. Nghĩa của từ tồn tại trong từ, nói rộng ra là trong hệ thống ngôn ngữ. Trong ý thức, trong bộ óc trí tuệ của con người chỉ tồn tại *sự hiểu biết về nghĩa của từ* chứ không phải là *nghĩa của từ*. Nghĩa của từ bao gồm :

- Nghĩa biểu vật (denotative meaning) : liên hệ giữa từ và sự vật (hiện tượng, thuộc tính, hành động,...).
- Nghĩa biểu niệm (significative meaning) : liên hệ giữa từ và ý (ý nghĩa, ý niệm, biểu niệm,...).
- Nghĩa ngữ dụng (pragmatical meaning), còn gọi là nghĩa biểu thái, nghĩa hàm chỉ, là mối liên hệ giữa từ với thái độ chủ quan, cảm xúc của người nói.
- Nghĩa cấu trúc (structural meaning) là mối quan hệ giữa từ với các từ khác trong hệ thống từ vựng. Quan hệ giữa từ này với từ khác thể hiện trên hai trục : trục đối vị (paradigmatic axis) và trục ngữ đoạn (syntagmatic axis).

Nghĩa và khái niệm của từ gắn bó rất chặt chẽ với nhau nhưng chúng không phải là trùng nhau. Ví dụ : khái niệm từ “*nước cứng*” và nghĩa “*nước*” của nó.

### 2.1.1.1. Cơ cấu nghĩa của từ

Một từ có thể có một hay nhiều nghĩa có quan hệ với nhau, ngay trong từng nghĩa cũng bao gồm các nghĩa tố (seme) có quan hệ với nhau. Cách phân loại nghĩa thường như sau :

- Nghĩa gốc - nghĩa phát sinh.
- Nghĩa tự do - nghĩa hạn chế.
- Nghĩa trực tiếp - nghĩa chuyển tiếp.
- Nghĩa thường trực - không thường trực.

Để xây dựng, phát triển thêm nghĩa của các từ, trong ngôn ngữ người ta dùng 2 phương pháp:

- Chuyển nghĩa ẩn dụ (metaphor). Ví dụ, *cánh* trong *cánh chim*, *cánh máy bay*, *cánh quạt*,...
- Chuyển nghĩa hoán dụ (metonymy). Ví dụ, *Vụng vá vai (áo) tài và rách (áo)*.

### 2.1.1.2. Phân tích nghĩa của từ

#### ❑ Theo ngữ cảnh :

*Ngữ cảnh của một từ* là chuỗi từ kết hợp với nó hoặc bao xung quanh nó, đủ làm cho nó được cụ thể hóa và hoàn toàn xác định về nghĩa.

#### Ví dụ 2-1 :

Từ *chắc* trong các ngữ cảnh sau : “*lúa đã chắc hạt*”, “*ông này chắc đã có con lớn*”,... Sở dĩ chúng ta xác định được một nghĩa cụ thể là vì trong mỗi ngữ cảnh, từ thể hiện khả năng kết hợp ngữ pháp và từ vựng của mình.

#### ❑ Kết hợp ngữ pháp :

Khả năng đứng vào một vị trí nhất định trong những cấu trúc nhất định nào đó.

**Ví dụ 2-2 :**

Động từ kết hợp với các từ : *đã, đang, xong, mãi,..* tạo thành : *đã làm, làm xong, đang đi,..*

□ **Kết hợp từ vựng :**

Khả năng kết hợp giữa một nghĩa của từ này với một nghĩa của từ khác, sao cho tổ hợp được tạo thành phản ánh đúng với thực tại, phù hợp với logic và thói quen sử dụng ngôn ngữ của người bản ngữ.

**Ví dụ 2-3 :**

*ăn cơm, học bài* chứ không thể *ăn bài, học cơm* được.

**2.1.1.3. Nghĩa của từ trong hoạt động ngôn ngữ**

Khi đi vào hoạt động ngôn ngữ, nghĩa của từ giảm tính trừu tượng, tăng tính xác định, cụ thể. Đồng thời, nó cũng gia tăng những sắc thái mới.

**Ví dụ 2-4 :**

- Số từ *Một trăm* trong : “Yêu nhau vạn sự chẳng nề ; Một trăm chỗ lệch cũng kê cho bằng”.

- Câu *Những tư tưởng xanh lục không màu đang ngủ một cách giận dữ* (N.Chomsky) đúng về mặt ngữ pháp nhưng vô lý.

**2.1.2. Quan hệ đồng nghĩa và trái nghĩa trong từ vựng**

**2.1.2.1. Từ đồng nghĩa**

Từ đồng nghĩa là những từ tương đồng với nhau về nghĩa, khác nhau về âm thanh, và phân biệt với nhau về một vài sắc thái ngữ nghĩa hoặc sắc thái phong cách... nào đó, hoặc đồng thời cả hai.

**Ví dụ 2-5 : Các nhóm đồng nghĩa :**

- Trong tiếng Anh : start, begin, commence (bắt đầu).
- Trong tiếng Việt : cố, gắng, cố gắng.

Những từ đồng nghĩa với nhau không nhất thiết phải tương đương nhau về số lượng nghĩa, các từ đồng nghĩa thường chỉ đồng nghĩa ở một nghĩa nào đó, vì vậy các từ đa nghĩa có thể tham gia vào nhiều nhóm đồng nghĩa khác nhau. Trong mỗi nhóm đồng nghĩa, thường có một từ trung tâm.

**Ví dụ 2-6 :**

Nhóm : “yếu, yếu ớt, yếu đuối,..” có từ “yếu” là từ trung tâm.

**2.1.2.2. Từ trái nghĩa**

Từ trái nghĩa là những từ có nghĩa đối lập nhau trong mối quan hệ tương liên. Chúng khác nhau về ngữ âm và phản ánh những khái niệm tương phản về logic. Ví dụ : “chân cứng, đá mềm”.

Để xác định cặp trái nghĩa phải dựa trên nhiều tiêu chí như :

- Cùng có khả năng kết hợp với một từ bất kỳ nào đó mà qui tắc ngôn ngữ cho phép. Ví dụ : người khôn - người dại.
- Đảm bảo mối quan hệ liên tưởng đối lập với nhau một cách thường xuyên và mạnh. Ví dụ : cứng - mềm > cứng - dẻo.
- Riêng đối với tiếng Việt, chúng ta thấy số lượng âm tiết thường bằng nhau.

Ngoài ra chúng ta cũng có các cặp tuy không trái nghĩa nhưng lại được dùng như trái nghĩa. Ví dụ : Đầu voi - đuôi chuột.

**2.1.3. Biến đổi trong từ vựng**

**2.1.3.1. Những biến đổi bề mặt**

Hiện tượng rơi rụng bớt từ ngữ do những nguyên nhân sau đây gây ra:

- Tranh chấp về giá trị sử dụng, như : *chiến* biến đổi thành *chùa*
- Sự biến đổi ngữ âm, như : *mời* biến đổi thành *lời*
- Sự rút gọn từ, như : *omnibus* biến đổi thành *bus*
- Lịch sử và xã hội.

Sự xuất hiện những từ ngữ mới :

- Phát sinh từ mới bằng cách tạo, vay mượn, rút gọn, cải tiến,...

### **2.1.3.2. Những biến đổi trong chiều sâu của từ vựng**

Biến đổi về phương diện ngữ nghĩa của từ, theo hướng :

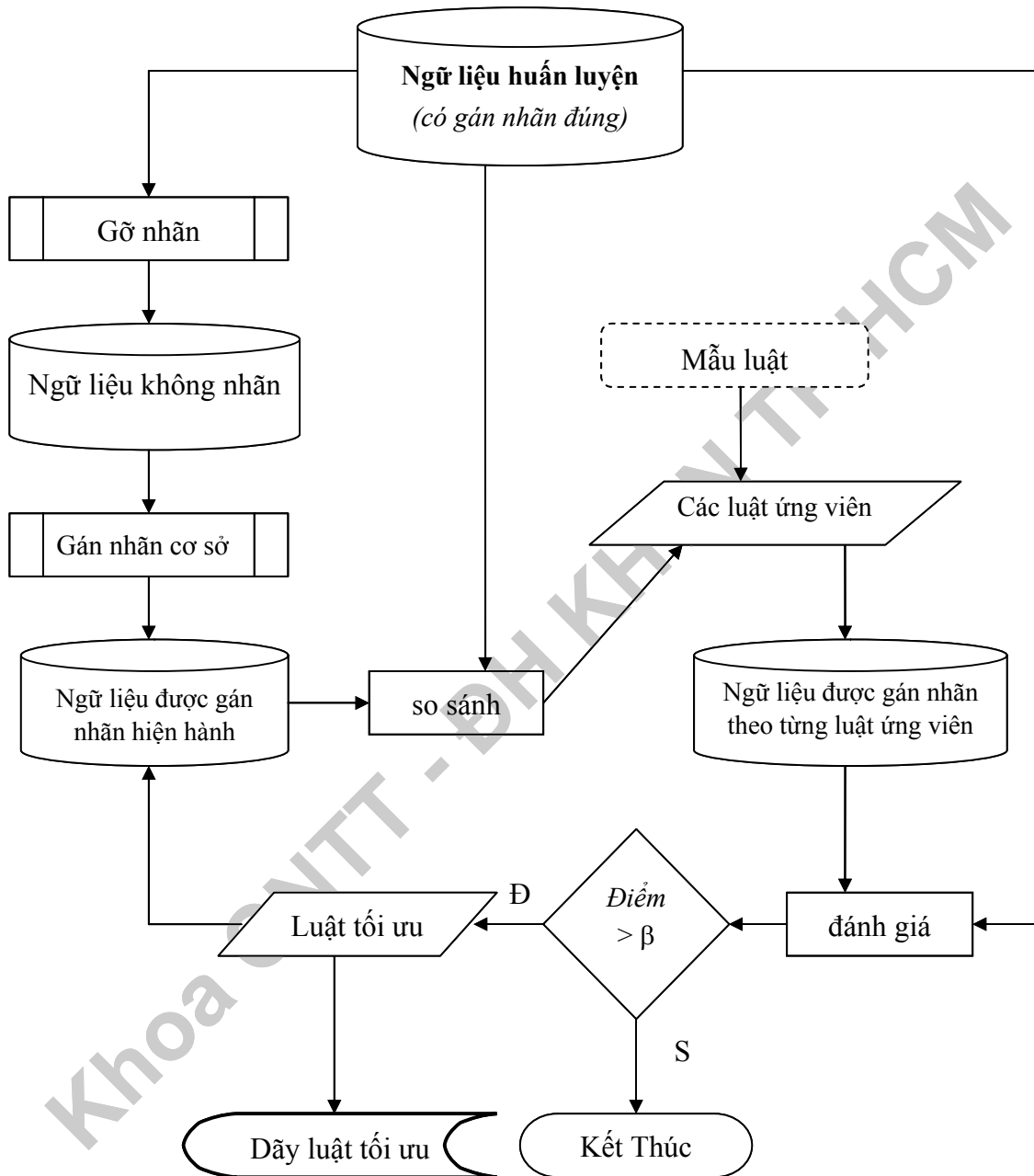
- Thu hẹp nghĩa của từ.
- Mở rộng nghĩa của từ.

## **2.2. HỌC DỰA TRÊN CHUYỂN ĐỔI**

### **2.2.1. Học dựa trên chuyển đổi là gì ?**

Học dựa trên chuyển đổi (transformation-based learning – TBL) hay còn được gọi là học hướng lỗi (error driven) là một giải thuật học giám sát được Eric Brill đề xuất năm 1993 trong luận án tiến sĩ của ông [11]. Giải thuật này dựa trên cơ sở ngôn ngữ học cấu trúc của Z.S.Harris. Bộ học phát sinh một tập các luật chuyển đổi theo thứ tự dựa trên ngữ liệu huấn luyện đã được gán nhãn và mẫu luật định dạng các kiểu hình thành luật. Định dạng của bộ luật, mẫu luật, số lượng luật phát sinh được tùy thuộc vào từng công việc cụ thể.

### 2.2.2. Giải thuật học dựa trên chuyển đổi tổng quát



**Hình 2-1 : Lưu đồ giải thuật học dựa trên chuyển đổi**

Cách dễ nhất để hiểu ý tưởng chính của việc học dựa trên chuyển đổi là xem qua một ví dụ.

**Ví dụ 2-7 :**

- a) Some students, there is no denying, are more charismatic and powerful on stage than others. (stage : sân khấu)
- b) Clearly, there is a difference in scale and dimension between the stage, the television screen and the cinema screen. (stage : sân khấu)
- c) At different stages of development. (stage : giai đoạn).
- d) At an early stage the Roberts decided to do away with the lawn which sloped towards the house. (stage : giai đoạn).

Chúng ta có thể nhận thấy rằng từ *stage* được sử dụng trong những ví dụ trên theo những nghĩa khác nhau. Trong hai câu đầu tiên, từ *stage* được dịch là *sân khấu* và trong ngữ liệu huấn luyện nó được gán nhãn là SANKHAU. Nhưng nhãn ban đầu của từ *stage* được chọn là GIAIDOAN, có nghĩa *một thời kỳ phát triển*. Hệ học sẽ so sánh nhãn ban đầu với nhãn được gán đúng trong từng trường hợp và phát hiện rằng nhãn ban đầu (GIAIDOAN) khác với nhãn đúng (SANKHAU) trong ví dụ (a) và (b). Hệ thống bắt đầu tìm những manh mối trong ngữ cảnh mà nhãn SANKHAU được chọn : nó phát sinh tất cả các luật ứng viên có thể dựa trên mẫu luật cho trước. Theo các mẫu luật được đưa trong Ví dụ 2-8, có thể có các luật ứng viên sau :

- Thay thế nhãn GIAIDOAN bằng nhãn SANKHAU nếu từ thứ hai bên trái của từ gây nhầm lẫn (*stage*) là từ *powerful*.
- Thay thế nhãn GIAIDOAN bằng nhãn SANKHAU nếu theo sau từ gây nhầm lẫn (*stage*) là dấu phẩy (,).
- Thay thế nhãn GIAIDOAN bằng nhãn SANKHAU nếu trước từ gây nhầm lẫn (*stage*) là từ *the*.

Hệ học áp dụng những luật ứng viên này vào trong các trường hợp của từ *stage* trong toàn bộ ngữ liệu huấn luyện và chọn luật đúng nhất, và luật này được đưa vào trong bộ luật. Luật ứng viên thứ ba có thể được chấp nhận, SANKHAU là nhãn gán đúng trong các trường hợp từ trước từ *stage* là một mạo từ xác định. Tuy

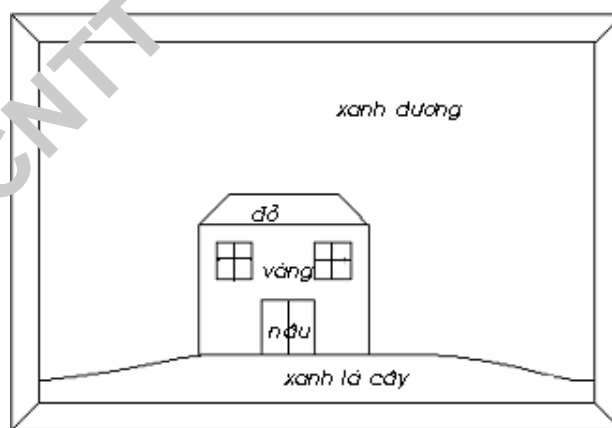
nhiên, không phải luật này lúc nào cũng đúng. Trong ngữ liệu huấn luyện có thể có những câu như sau :

- In the early stages of his political career
- At the planning stage
- Ready for the final stage

Mặc dù luật không phải lúc nào cũng gán nhãn đúng cho từ *stage*, nhưng nó vẫn được đưa vào trong bộ luật nếu số trường hợp nó tìm thấy các nhãn chính xác là đủ lớn. Các lỗi do luật phát sinh có thể được chỉnh sửa bằng những luật đặc biệt hơn ở phía sau của chuỗi luật. Ví dụ, có thể tìm thấy một luật đặc biệt hơn như sau : nếu theo sau từ *stage* là giới từ *of* thì từ *stage* có thể chỉ một *giai đoạn/thời kỳ*. Những luật như vậy thay thế những nhãn sai bằng những nhãn chính xác hơn.

### 2.2.3. Mô tả về trình tự tạo luật chuyển đổi

Theo Samuel và các cộng sự [18], trình tự tạo bộ luật chuyển đổi được so sánh với trình tự công việc của một họa sĩ vẽ tranh. Một bức tranh được mô tả như sau : một căn nhà màu vàng có mái nhà màu đỏ, cửa cái màu nâu, hai cửa sổ, cỏ xanh và bầu trời xanh dương.



**Hình 2-2: Minh họa của Samuel về trình tự tạo luật chuyển đổi**

Để vẽ bức tranh trên, trước tiên, họa sĩ sẽ sơn toàn bộ với màu xanh dương (màu chiếm nhiều nhất trên bức vẽ) – màu của bầu trời, không chừa chỗ cho căn nhà hay bãi cỏ. Sau đó ông ta sẽ chia mặt xanh dương thành hai phần bằng một



đường ngang và sơn phần dưới với màu xanh lá cây. Ông ta sẽ phủ lên phần màu xanh dương và xanh lá cây khi ông ta sơn bức tường màu vàng và mái đỏ của ngôi nhà. Cuối cùng, chi tiết nhỏ nhất, cửa cái và cửa sổ, sẽ được sơn lên trên phần tường màu vàng bằng các cái cọ nhỏ theo đúng màu tương ứng.

Trình tự tạo tập luật chuyên đổi cũng tương tự. Luật đầu tiên thường tổng quát và có thể còn rất nhiều lỗi, nhưng đầu ra của luật này sẽ là đầu vào của những luật được áp dụng sau đó. Những luật sau thường đặc biệt hơn và có thể chỉnh sửa các lỗi do những luật tổng quát tạo ra lúc đầu.

Các luật chuyển đổi có thể có nhiều loại khác nhau : thêm, thay thế, hoặc xóa các nhãn. Chúng tôi chỉ tập trung vào loại luật được áp dụng trong công việc của chúng tôi, *luật thay thế (replacement rule)*.

Những luật chuyển đổi phát sinh bởi thuật toán này đã được áp dụng thành công vào nhiều bài toán khác nhau trong lĩnh vực xử lý ngôn ngữ tự nhiên như : tách từ, tách câu, phân tích hình thái học, bắt lỗi chính tả, nhận diện tên riêng, gán nhãn từ loại, phân tích cú pháp,... Trong các bài toán kể trên, đa số các kết quả mà TBL đạt được đều rất cao và tương đương với những phương pháp học tiên tiến khác. TBL được các nhà ngôn ngữ học - máy tính đánh giá là trực quan, dễ hiểu, gần với công việc của các nhà ngôn ngữ học. Một đặc điểm nổi bật của TBL so với các giải thuật học khác là tính trực quan, tính biểu trưng và tính kế thừa. Các nhà ngôn ngữ học hoàn toàn có thể theo dõi, can thiệp vào suốt quá trình học, quá trình thực thi cũng như các kết quả trung gian và cuối cùng, TBL cho phép sửa sai trên đầu ra của một hệ khác.

Việc học luật được thực hiện như sau. Mỗi từ sắp được gán nhãn được gán cho một nhãn ban đầu. Nhãn ban đầu của một từ có thể là nhãn thường gặp nhất của từ đó trong ngữ liệu huấn luyện, hay là một nhãn nào đó theo qui ước lúc đầu. Giai đoạn này được gọi là giai đoạn gán nhãn cơ sở. Việc gán nhãn này có thể rất ngây ngô. Tuy nhiên, nó sẽ được chỉnh sửa trong quá trình áp dụng chuỗi luật.

Mỗi từ trong ngữ liệu huấn luyện có thể được gán nhãn khác với nhãn ban đầu của từ đó. Khi đó bộ học sẽ phát sinh các luật dựa trên tập mẫu luật. Đó là tất cả

---

các luật thích hợp khớp với mẫu luật, thay thế những nhãn sai thành nhãn đúng. Dưới đây là ví dụ của vài mẫu luật.

**Ví dụ 2-8 :**

- Thay thế nhãn **X** bằng nhãn **Y** nếu từ đứng trước có nhãn **Z**.
- Thay thế nhãn **X** bằng nhãn **Y** nếu từ đứng sau là từ **Z**.
- Thay thế nhãn **X** bằng nhãn **Y** nếu từ kế tiếp hoặc từ liền sau là từ **Z**.
- Thay thế nhãn **X** bằng nhãn **Y** nếu từ trước đó là **Z** và từ trước trước đó là từ **W**.

Trong ví dụ trên của các mẫu luật, có thể hiểu **X, Y, Z, W** như các *biến*. Giá trị của các *biến* này như thế nào sẽ được xác định nhờ vào quá trình huấn luyện. Phân trên chúng ta đã thấy cụ thể các luật được tạo ra nhờ vào các mẫu luật của này.

Luật tốt nhất trong những luật ứng viên vừa được phát sinh sẽ được chọn. Ở đây, luật tốt nhất được hiểu là luật làm tăng độ chính xác nhiều nhất khi áp dụng cho toàn bộ ngữ liệu huấn luyện. Hệ thống huấn luyện sẽ ghi nhớ trường hợp nhãn sai được chỉnh thành đúng và số lỗi tạo ra khi áp dụng luật đó.

Hệ thống sẽ ngừng phát sinh luật khi điểm của luật (số lần áp dụng thành công trừ cho số lỗi do luật tạo ra) nhỏ hơn ngưỡng (do người sử dụng chọn). Có thể ngăn không cho bộ học phát sinh những luật quá đặc biệt bằng cách đặt giá trị ngưỡng cao.

**2.2.4. Yêu cầu trong việc áp dụng thuật toán học dựa trên chuyển đổi vào xử lý ngữ nghĩa**

Bên cạnh xem xét những từ trong ngữ cảnh, các luật có thể sử dụng những thông tin khác về những từ đó, ví dụ các đặc điểm về hình thái, cú pháp và ngữ nghĩa. Để lấy được những luật như vậy chúng ta cần đánh dấu những từ trong ngữ liệu huấn luyện những nhãn có thể cho bộ học những thông tin cần thiết về những từ đó. Chúng ta cần cung cấp cho bộ học một tập mẫu luật biểu diễn tất cả các loại luật mà chúng ta mong muốn phát sinh. Ví dụ, chúng ta có thể dùng các luật theo định dạng sau :

---

- Thay thế nhãn **X** bằng nhãn **Y** nếu *tân ngữ trực tiếp* (direct object) của từ gây nhầm lẫn là từ **Z**.
- Thay thế nhãn **X** bằng nhãn **Y** nếu từ đứng trước từ gây nhầm lẫn có *từ loại* là **Z**.
- Thay thế nhãn **X** bằng nhãn **Y** nếu từ đứng trước từ gây nhầm lẫn *đồng nghĩa* với từ **Z**.
- Thay thế nhãn **X** bằng nhãn **Y** nếu từ đứng sau từ gây nhầm lẫn là *hyponymy*<sup>3</sup> của từ **Z**.

Trình tự của các luật chuyển đổi phát sinh bởi bộ học có thể được dùng để gán nhãn ngữ nghĩa cho từ trong các văn bản mới. Trước khi áp dụng luật, mỗi từ sắp được gán nhãn được gán cùng nhãn ban đầu như trong giai đoạn huấn luyện. Nếu không áp dụng được luật nào, nhãn ban đầu không được thay thế. Vì vậy, việc chọn nhãn ban đầu thực chất là chọn một nhãn mặc định áp dụng cho từ trong trường hợp không có luật thích hợp.

Việc chọn lựa nhãn ban đầu, tập mẫu luật, ngưỡng, và số lượng các luật phải phát sinh phụ thuộc vào công việc và kích thước của ngữ liệu huấn luyện. Độ chính xác đạt đến khi áp dụng luật có thể khác nhau cho từng nhiệm vụ và dữ liệu khác nhau, đôi khi dùng nhiều bộ dữ liệu kiểm tra là hữu ích để đạt kết quả tốt nhất.

### **2.2.5. Nhận xét**

Các luật chuyển đổi dễ hiểu. Chúng ta có thể xoá những luật có thể gây hậu quả khi sử dụng, thêm những luật mới hoặc chỉnh sửa những luật đã có. Tuy nhiên, khi thêm, chỉnh sửa hay xoá bỏ các luật, chúng ta phải chú ý rằng trình tự các luật trong bộ luật là quan trọng. Thay đổi hoặc xoá những luật ở phần đầu có thể làm cho những luật ở phần giữa hoặc phần cuối không áp dụng được.

---

<sup>3</sup> Quan hệ hyponymy (ký hiệu  $\rightsquigarrow$ ) là quan hệ cụ thể hoá. Ví dụ, một quan hệ hyponymy : *written symbol*  $\rightsquigarrow$  *character*  $\rightsquigarrow$  *letter* (theo WordNet).

---

Việc học dựa trên chuyển đổi đường như là một cách thích hợp để có được những từ tiếng Việt tương ứng với những từ tiếng Anh bị nhập nhằng nghĩa trong các ngữ cảnh khác nhau. Định dạng của tập mẫu luật uyển chuyển cho phép dùng nhiều nguồn tri thức khác nhau để có thể tìm được các nghĩa tương ứng với những từ gây nhầm lẫn. Như đã đề cập, trình tự các luật chuyển đổi có thể được thay thế bằng tay. Ví dụ, chúng ta có thể thay thế những luật đặc biệt bằng những luật cụ thể hơn.

### **2.3. MỘT SỐ GIẢI THUẬT HỌC DỰA TRÊN CHUYỂN ĐỔI CẢI TIẾN**

Bên cạnh những ưu điểm của giải thuật TBL như đã trình bày ở trên, TBL có một số khuyết điểm như : số luật rút ra quá lớn, chỉ cho ra một kết quả, thao tác chỉ trên một công việc, và quan trọng nhất là thời gian huấn luyện quá lâu. Để khắc phục các khuyết điểm này người ta đã đưa ra một số giải thuật cải tiến như sau.

#### **2.3.1. Lazy TBL**

Giải thuật Lazy TBL (LTBL) được K. Samuel ([17]) đưa ra vào năm 1998 nhằm khắc phục số lượng luật phát sinh quá lớn khi số các mẫu luật tăng lên. Việc xác định một tập mẫu luật hiệu quả là một điều không đơn giản. Nếu chúng ta bỏ sót một mẫu luật nào đó thì TBL sẽ không rút ra được những luật hiệu quả từ mẫu luật đó. Nếu chúng ta đưa vào quá nhiều mẫu luật thì sẽ dẫn đến việc TBL phải kiểm tra vô vàn các luật ứng viên của tất cả các mẫu luật đó. Điều này khiến cho giai đoạn huấn luyện của TBL sẽ không còn khả thi. Chính tác giả của giải thuật TBL (Eric Brill) đã tránh tình trạng xấu này bằng cách chỉ đưa vào 30 mẫu luật và mỗi mẫu luật chỉ kiểm tra một hoặc hai điều kiện.

Giải thuật LTBL khắc phục hạn chế bằng cách vẫn cho phép một tập mẫu luật đầy đủ nhưng trong quá trình huấn luyện, tại mỗi bước lặp, LTBL chỉ cho phép một số giới hạn  $R$  các luật ứng viên được xem xét. Việc lựa chọn luật ứng viên nào là dựa vào phương pháp lấy mẫu ngẫu nhiên (phương pháp Monte Carlo). LTBL dựa trên giả thiết “*các luật hiệu quả sẽ sửa được nhiều lỗi trong ngữ liệu, có nghĩa*

là sẽ xuất hiện trong không gian thử nghiệm nhiều hơn và sẽ dễ được chọn nhiều hơn”. Kết quả thực nghiệm cho thấy với  $R$  càng nhỏ, LTBL sẽ giảm đáng kể chi phí huấn luyện còn độ chính xác giảm không đáng kể (có thể xem thêm kết quả thống kê so sánh giữa LTBL và TBL nguyên thủy trong [17]).

### **2.3.2. TBL đa chiều**

Trong quá trình xử lý ngôn ngữ, có nhiều công việc được thực hiện đồng thời không nhất thiết phải nối tiếp nhau. Việc thực hiện đồng thời có ưu điểm là tận dụng cùng thông tin ngữ cảnh (dễ tổ chức lưu trữ trong bộ nhớ) và nhờ sự tương tác lẫn nhau giữa các công đoạn có thể làm tăng độ chính xác của mỗi công đoạn. Đối với phương pháp học bằng mạng nơron, yêu cầu thực hiện nhiều công việc đã được thực hiện dễ dàng bằng cách tăng cường thêm số nút của tầng xuất. Để thực hiện yêu cầu song song nói trên đối với TBL, Radu Floarian và Grace Ngai đã đưa ra giải thuật TBL đa chiều (Multi-dimension TBL) mà trong đó các tác giả đã thay thế hàm đánh giá (chấm điểm) của TBL gốc (chỉ cho một công việc) bằng hàm đánh giá trên nhiều công việc đồng thời.

### **2.3.3. TBL nhanh**

Bước cải tiến đáng kể về TBL có lẽ là giải thuật Fast TBL([16]). Giải thuật này được Radu Florian, và Grace Ngai đưa ra vào năm 2001 nhằm khắc phục khuyết điểm lớn nhất của TBL, đó là thời gian huấn luyện quá lâu<sup>4</sup> (nhất là khi kích thước huấn luyện tăng lên). Để khắc phục nhược điểm này, trước đó cũng đã có một số giải thuật được đề nghị như : kiểu thống kê của Ramshaw và Marcus, ICA của Hepple, Lazy TBL của Samuel... nhưng các giải thuật này đều ít nhiều làm giảm độ chính xác của TBL hoặc chi phí bộ nhớ quá lớn.

Thay vì phải phát sinh từng luật ứng viên ở mỗi bước lặp như trong TBL, fnTBL lưu lại các luật này trong bộ nhớ cùng với điểm của nó. Ngoài ra, việc tính

---

<sup>4</sup> Nguyên nhân khiến TBL huấn luyện quá lâu là TBL phải lần lượt thử từng luật ứng viên của mỗi mẫu luật bằng cách cho luật này tác động lên toàn bộ ngữ liệu, rồi sau đó đánh giá (tính điểm) dựa trên ngữ liệu vàng.

điểm không cần phải so sánh trên toàn bộ dữ liệu như trong TBL, fnTBL chỉ so sánh trong vùng lân cận vị trí mà luật ứng viên tác động mà thôi (vì các vùng khác không thay đổi). Kết quả là fnTBL làm giảm thời gian huấn luyện từ 10 đến 130 lần, bộ nhớ tăng không đáng kể, và quan trọng nhất là độ chính xác hoàn toàn không thay đổi.

## 2.4. THUẬT TOÁN FAST-TBL

### 2.4.1. Quy ước

- $S$  : Không gian mẫu
- $C$  : tập hợp các nhãn ngôn ngữ dùng cho việc gán nhãn (trong luận văn này, đây chính là hệ thống nhãn ngữ nghĩa)
- $C[s]$  là nhãn được gán cho mẫu  $s$ , và  $T[s]$  là nhãn đúng của  $s$  (nhãn của  $s$  trong ngữ liệu huấn luyện).
- $p$  : vị từ được định nghĩa trên không gian  $S$ .
- Một luật  $r$  là một cặp gồm vị từ, nhãn  $(p,t)$  và nhãn  $t \in C$ . Có nghĩa là mẫu  $s \in S$  được gán nhãn  $t$  nếu vị từ  $p$  thoả trên  $s$ .
- Với một luật  $r = (p,t)$ ,  $p_r$  dùng để chỉ vị từ  $p$ , còn  $t_r$  dùng để chỉ thành phần  $t$  trong  $r$ .
- Một luật  $r$  được áp dụng trên mẫu  $s$  khi và chỉ khi  $p_r(s) = true$  và  $t_r \neq C[s]$ . Kết quả của việc áp dụng luật này trên mẫu  $s$  là  $r(s)$ .
- $G(r)$  : tập các mẫu được luật  $r$  sửa từ sai thành đúng.

$$G(r) = \{s \mid C[s] \neq T[s] \wedge C[r(s)] = T[s]\}$$

- Khi đó điểm tốt của luật  $r$  là :  $good(r) = |G(r)|$
- $B(r)$  : tập các mẫu bị luật  $r$  sửa từ đúng thành sai.

$$B(r) = \{s \mid C[s] = T[s] \wedge C[r(s)] \neq T[s]\}$$

- Khi đó điểm xấu của luật  $r$  là :  $bad(r) = |B(r)|$

➤ Sử dụng hàm đánh giá  $f(r) = good(r) - bad(r)$ .

Về căn bản thuật toán FastTBL (fnTBL) giống thuật toán TBL, nó chỉ khác biệt ở phần phát sinh điểm – phát sinh các luật ứng viên.

#### 2.4.2. Phát sinh luật

Cho một luật  $b$  mới được học (được đưa vào chuỗi luật), mục tiêu là đi xác định được luật  $r$  (đã được phát sinh trước đó) bị thay đổi (do tác động của luật  $b$ ). Chúng ta có thể thấy rằng luật  $r$  bị thay đổi điểm (cần xác định lại) khi một trong hai giá trị điểm (xấu hoặc tốt,  $bad(r)$  hoặc  $good(r)$ ) của nó bị thay đổi. Rõ ràng rằng, nếu cả tập  $G(r)$  lẫn tập  $B(r)$  không bị tác động khi áp dụng luật  $b$  thì giá trị của hàm đánh giá của luật  $r$  vẫn như cũ. Ưu điểm của thuật toán fnTBL nằm ở chỗ : không cần phải xác định (tính toán lại) toàn bộ điểm của tất cả các luật mà chỉ cần xác định lại (cập nhật) điểm của những luật bị tác động mà thôi.

Khi xem xét ảnh hưởng của  $b$  lên một mẫu  $s$ , chúng ta phải tính đến tác động gián tiếp của các nhãn lân cận mẫu  $s$ . Gọi vùng lân cận của một mẫu  $s$  là  $V(s)$ . Nếu các mẫu độc lập với nhau thì  $V(s) = \{s\}$ .

Cho luật tốt nhất  $b$  tác động lên mẫu  $s \in S$  (nghĩa là  $b(s) \neq C[s]$ ). Chúng ta cần phải xác định được luật  $r$  bị ảnh hưởng khi  $s$  chuyển thành  $b(s)$ .  $f(r)$  cần phải được cập nhật nếu và chỉ nếu tồn tại ít nhất một mẫu  $s'$  sao cho :

$$(s' \in G(r)) \wedge (b(s') \notin G(r))$$

$$(s' \in B(r)) \wedge (b(s') \notin B(r))$$

$$(s' \notin G(r)) \wedge (b(s') \in G(r))$$

$$(s' \notin B(r)) \wedge (b(s') \in B(r))$$

Mỗi điều kiện nêu trên tương ứng với một trường hợp cập nhật lại giá trị của  $good(r)$  hay  $bad(r)$ . Khi xem xét ảnh hưởng của việc áp dụng luật  $b$  vào mẫu  $s$ , chỉ những mẫu  $s'$  thuộc về tập  $V(s)$  mới cần được kiểm tra.

Cho  $s' \in V(s)$ . Có 2 trường hợp cần phải xem xét : (1)  $b$  áp dụng lên  $s'$  được, (2)  $b$  không áp dụng lên  $s'$  được.

### 2.4.2.1. Trường hợp 1

□  $C[s'] = C[b(s')]$  (b không ảnh hưởng tới s').

Lần lượt biến đổi các biểu thức :

$$(1) (s' \in G(r)) \wedge (b(s') \notin G(r))$$

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s']) \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee t_r \neq T[b(s')])$$

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s']) \wedge (p_r(b(s')) = false \vee C[s'] = t_r \vee t_r \neq T[s'])$$

(do  $C[s'] = C[b(s')]$  (điều kiện) và  $T[s'] = T[b(s')]$  (hiển nhiên) )

$$\Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge p_r(b(s')) = false$$

$$(2) (s' \notin G(r)) \wedge (b(s') \in G(r))$$

$$\Leftrightarrow p_r(b(s')) = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge p_r(s') = false$$

$$(3) (s' \in B(r)) \wedge (b(s') \notin B(r))$$

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s']) \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee C[b(s')] \neq T[b(s')])$$

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s']) \wedge (p_r(b(s')) = false \vee C[s'] = t_r \vee C[s'] \neq T[s'])$$

(do  $C[s'] = C[b(s')]$  (điều kiện) và  $T[s'] = T[b(s')]$  (hiển nhiên) )

$$\Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s'] \wedge p_r(b(s')) = false$$

$$(4) (s' \notin B(r)) \wedge (b(s') \in B(r))$$

$$\Leftrightarrow p_r(b(s')) = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s'] \wedge p_r(s') = false$$

□ **Tóm lại :**

$$(1a) (s' \in G(r)) \wedge (b(s') \notin G(r)) \Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge p_r(b(s')) = false$$

$$(2a) (s' \notin G(r)) \wedge (b(s') \in G(r)) \Leftrightarrow p_r(b(s')) = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge p_r(s') = false$$



$$(3a) (s' \in B(r)) \wedge (b(s') \notin B(r)) \Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s'] \\ \wedge p_r(b(s')) = false$$

$$(4a) (s' \notin B(r)) \wedge (b(s') \in B(r)) \Leftrightarrow p_r(b(s')) = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s'] \\ \wedge p_r(s') = false$$

Nhận xét rằng nếu các điều kiện trong biểu thức (1a) và (3a) xảy ra thì tương ứng các điểm  $good(r)$  và  $bad(r)$  bị giảm. Do đó, ta có thuật toán cập nhật điểm thứ nhất như sau :

- Tạo ra tất cả vị từ  $p$  (dựa vào các mẫu luật) thoả mẫu  $s'$ .
- **If**  $C[s'] \neq T[s']$  **then**
  - **If**  $p(b(s')) = false$  **then** giảm  $good(r)$  trong đó  $r = (p, T[s'])$ .
- **else**
  - **If**  $p(b(s')) = false$  **then** giảm  $bad(r)$  với tất cả các luật  $r$  có vị từ là  $p$  và  $t_r \neq C[s']$ .

Nhận xét rằng nếu các điều kiện trong biểu thức (2a) và (4a) xảy ra thì tương ứng các điểm  $good(r)$  và  $bad(r)$  được tăng lên. Do đó, ta có thuật toán cập nhật điểm thứ hai như sau :

- Tạo ra tất cả vị từ  $p$  (dựa vào các mẫu luật) thoả mẫu  $b(s')$ .
- **If**  $C[s'] \neq T[s']$  **then**
  - **If**  $p(s') = false$  **then** tăng  $good(r)$  trong đó  $r = (p, T[s'])$ .
- **else**
  - **If**  $p(s') = false$  **then** tăng  $bad(r)$  với tất cả các luật  $r$  có vị từ là  $p$  và  $t_r \neq C[s']$ .

#### 2.4.2.2. Trường hợp 2

□  $C[s'] \neq C[b(s')]$  (b có ảnh hưởng tới s')

Lần lượt biến đổi các biểu thức :

$$(1) (s' \in G(r)) \wedge (b(s') \notin G(r))$$


---

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s']) \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee t_r \neq T[b(s')])$$

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s']) \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee t_r \neq T[s'])$$

$$(do T[s'] = T[b(s')] \text{ (hiên nhiên) })$$

$$\Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r)$$

$$(2) (s' \notin G(r)) \wedge (b(s') \in G(r))$$

$$\Leftrightarrow p_r(b(s')) = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge (p_r(s') = false \vee C[s'] = t_r)$$

$$(3) (s' \in B(r)) \wedge (b(s') \notin B(r))$$

$$\Leftrightarrow (p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s']) \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee C[b(s')] \neq T[b(s')])$$

$$\Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s'] \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee C[b(s')] \neq T[s'])$$

$$(do T[s'] = T[b(s')] \text{ (hiên nhiên) })$$

$$(4) (s' \notin B(r)) \wedge (b(s') \in B(r))$$

$$\Leftrightarrow p_r(b(s')) = true \wedge C[b(s')] \neq t_r \wedge C[b(s')] = T[s'] \wedge (p_r(s') = false \vee C[s'] = t_r \vee C[s'] \neq T[s'])$$

□ **Tóm lại :**

$$(1b) (s' \in G(r)) \wedge (b(s') \notin G(r)) \Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r)$$

$$(2b) (s' \notin G(r)) \wedge (b(s') \in G(r)) \Leftrightarrow p_r(b(s')) = true \wedge C[s'] \neq t_r \wedge t_r = T[s'] \wedge (p_r(s') = false \vee C[s'] = t_r)$$

$$(3b) (s' \in B(r)) \wedge (b(s') \notin B(r)) \Leftrightarrow p_r(s') = true \wedge C[s'] \neq t_r \wedge C[s'] = T[s'] \wedge (p_r(b(s')) = false \vee C[b(s')] = t_r \vee C[b(s')] \neq T[s'])$$

$$(4b) (s' \notin B(r)) \wedge (b(s') \in B(r)) \Leftrightarrow p_r(b(s')) = true \wedge C[b(s')] \neq t_r \wedge C[b(s')] = T[s'] \wedge (p_r(s') = false \vee C[s'] = t_r \vee C[s'] \neq T[s'])$$

Nhận xét rằng nếu các điều kiện trong biểu thức (1b) và (3b) xảy ra thì tương ứng các điểm  $good(r)$  và  $bad(r)$  bị giảm. Do đó, ta có thuật toán cập nhật điểm thứ ba như sau :

- Tạo ra tất cả vị từ  $p$  (dựa vào các mẫu luật) thoả mẫu  $s'$ .
- **If**  $C[s'] \neq T[s']$  **then**
  - **If**  $p(b(s')) = false$  **or**  $C[b(s')] = t_r$  **then** giảm  $good(r)$  trong đó  $r = (p, T[s'])$ .
- **else**
  - Giảm  $bad(r)$  với tất cả các luật  $r$  có vị từ là  $p$  và  $t_r \neq C[s']$ .

Nhận xét rằng nếu các điều kiện trong biểu thức (2b) và (4b) xảy ra thì tương ứng các điểm  $good(r)$  và  $bad(r)$  được tăng lên. Do đó, ta có thuật toán cập nhật điểm thứ tư như sau :

- Tạo ra tất cả vị từ  $p$  (dựa vào các mẫu luật) thoả mẫu  $b(s')$ .
- **If**  $C[b(s')] \neq T[s']$  **then**
  - **If**  $p(s') = false$  **or**  $C[s'] = t_r$  **then** tăng  $good(r)$  trong đó  $r = (p, T[s'])$ .
- **else**
  - Tăng  $bad(r)$  với tất cả các luật  $r$  có vị từ là  $p$  và  $t_r \neq C[b(s')]$ .

## 2.5. VĂN PHẠM PHỤ THUỘC

### 2.5.1. Giới thiệu

Văn phạm phụ thuộc là loại văn phạm biểu diễn cấu trúc cú pháp dưới dạng các liên kết giữa các từ riêng lẻ thay vì dưới dạng cây cú pháp thông thường. Văn phạm phụ thuộc bắt nguồn từ văn phạm truyền thống Latin và Ả rập. Quan hệ cơ bản trong văn phạm phụ thuộc là quan hệ giữa một *head* (cha) và một *dependent* (phụ thuộc) của nó. Một từ (thường thường là động từ chính của câu) sẽ làm *head* cho toàn câu, mỗi từ khác sẽ phụ thuộc vào một *head* nào đó và cũng có thể làm

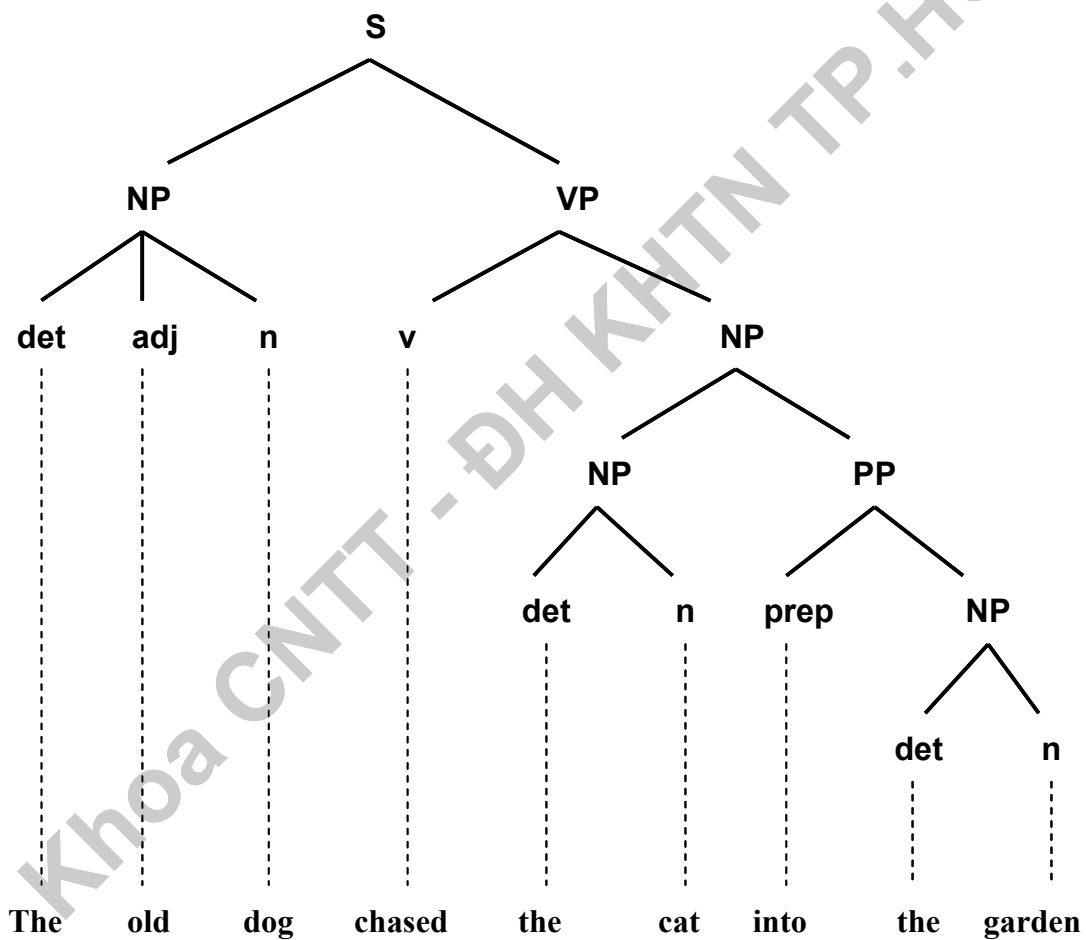
---

*head* cho một số các phụ thuộc khác. Luật văn phạm sẽ xác định các *head* cần phải lấy *dependent* như thế nào (ví dụ tính từ phụ thuộc vào danh từ, chứ không phụ thuộc vào động từ).

**Ví dụ 2-9 :**

Với câu “*The old dog chased the cat into the garden*”

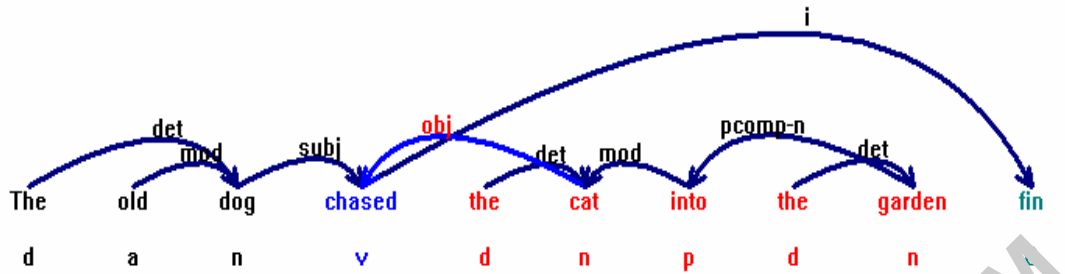
Theo cách phân tích cú pháp thông thường, ta có cây cú pháp như sau:



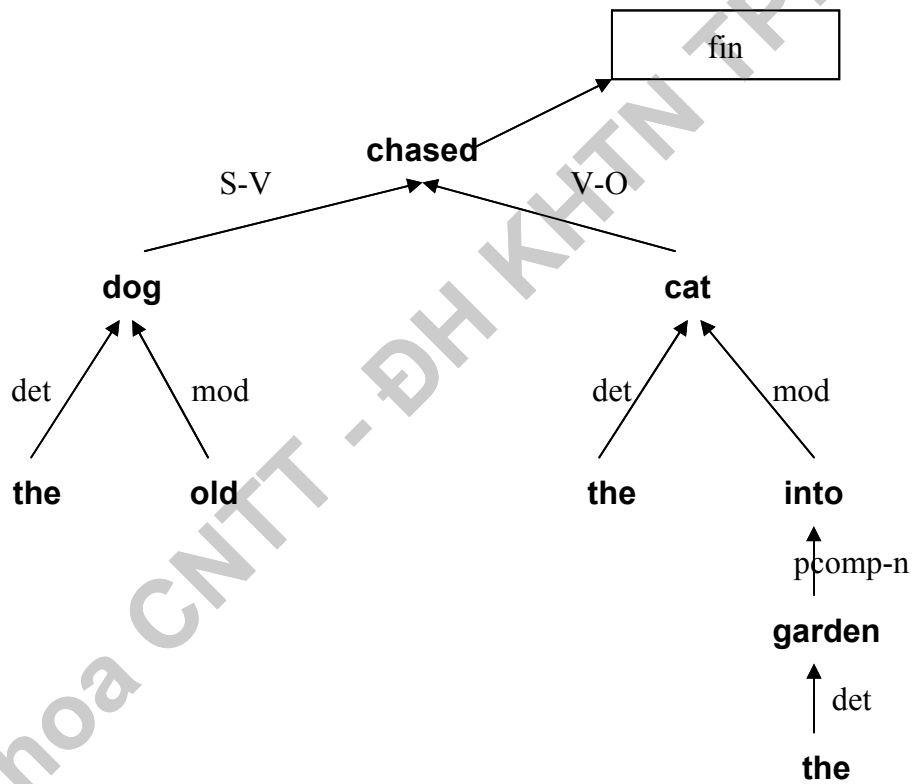
**Hình 2-3 : Minh họa một cây cú pháp thông thường**

Trong khi đó, khi phân tích bằng văn phạm phụ thuộc, ta được hình ảnh của các các mối liên kết giữa các các từ.

---



Hình 2-4 : Kết quả khi phân tích câu sử dụng văn phạm phụ thuộc



Hình 2-5 : Hình ảnh một cây quan hệ phụ thuộc

Trong ví dụ trên, ta có một số mối liên kết (quan hệ<sup>5</sup>) giữa các từ như sau :

- Động từ *chased* là head cho cả câu.
- Với cụm từ (ngữ danh từ) *the old dog*, từ *dog* là head của từ *the* và *old* (*the* là det của *dog*, còn từ *old* là mod của từ *dog*).
- Cụm từ (ngữ giới từ) *into the garden* có từ *into* là head (*garden* là pcomp-n của từ *into*).
- Còn trong cụm từ *the garden*, từ *garden* là head.
- Ngoài ra, còn có một số quan hệ : *dog* là subj của *chased*, còn *cat* là obj của *chased*.

### 2.5.2. Vận dụng văn phạm phụ thuộc vào xử lý ngữ nghĩa

Phân tích cú pháp dựa trên văn phạm phụ thuộc trong xử lý ngữ nghĩa có nhiều đặc điểm thuận lợi so với các phương pháp phân tích cú pháp thông thường khác.

Thứ nhất, trong xử lý ngữ nghĩa (của hệ dịch máy) nói riêng và trong dịch máy nói chung, yêu cầu đặt ra là xác định đúng nghĩa (dịch đúng) các tài liệu được đưa vào. Do đó, phân tích cú pháp không cần phải đưa ra kết quả quá sâu, chỉ cần phân tích nông đến một mức cần thiết để giải quyết công việc cần thực hiện. Vấn đề đặt ra ở đây là nông đến mức nào, sâu đến mức nào ? Trong xử lý ngữ nghĩa, phân tích cú pháp với mục đích xác định được mối quan hệ ngữ pháp giữa các thành phần trong câu : thành phần nào đóng vai trò là chủ ngữ của câu, thành phần nào đóng vai trò là động từ chính, thành phần nào sẽ giữ chức năng của một tân ngữ của động từ, có các ngữ nào, các mệnh đề thuộc loại gì (chính/phụ...)... Cái đặc biệt (cũng là cái hay) khi áp dụng văn phạm phụ thuộc để phân tích cú pháp là phân tích đủ mức (nông/sâu) cần thiết để có thể đưa ra được đầy đủ các thành phần ngữ pháp trong câu.

---

<sup>5</sup> Chúng tôi sẽ đề cập sau đây các khái niệm : *det*, *mod*, *subj*, *obj*, *pcomp-n*... Đó là tên của các quan hệ được tìm thấy khi phân tích một câu sử dụng văn phạm phụ thuộc. Chúng tôi sẽ đề cập đến chúng, diễn giải chi tiết trong phần sau.

---

Thứ hai, văn phạm phụ thuộc cần tập văn phạm nhỏ hơn nhiều so với các cách phân tích cú pháp dựa trên luật. Văn phạm phụ thuộc đòi hỏi phải có một tập luật văn phạm (quy định xem một *head* cần một *dependent* như thế nào). Song tập luật văn phạm này mang tính khái quát rất cao<sup>6</sup>, chúng là những tri thức mang tính thống nhất của các ngôn ngữ. Còn phân tích cú pháp dựa trên luật, tập luật dẫn cần thiết để có thể phân tích một câu tiếng Anh là rất nhiều, hơn nữa, tập luật này cũng không dễ gì kiểm soát được. Đặc điểm này còn hàm chứa một đặc điểm liên quan đến tốc độ thực thi của bộ phân tích cú pháp. Bộ phân tích cú pháp dựa trên văn phạm phụ thuộc mà chúng tôi sử dụng có thể phân tích 500 từ trên một giây trên máy tính có tốc độ 700 Mhz và 500 MB bộ nhớ.

### ***2.5.3. Các loại quan hệ trong bộ phân tích cú pháp dựa trên văn phạm phụ thuộc***

Trong phần này chúng tôi giới thiệu các loại quan hệ do bộ phân tích cú pháp (dựa trên văn phạm phụ thuộc) mà chúng tôi sử dụng rút ra được khi phân tích một câu tiếng Anh.

<b>Tên quan hệ</b>	<b>Diễn giải</b>
appo	Quan hệ đồng vị
s	Chủ ngữ bề mặt
subj	Chủ ngữ của động từ
obj	Tân ngữ của động từ
obj2	Tân ngữ thứ hai của động từ
pred	Vị ngữ
rel	Mệnh đề quan hệ

---

<sup>6</sup> Chúng chỉ cần đưa ra các luật văn phạm như : tính từ bổ nghĩa cho danh từ, chứ không phải là động từ ; trong cấu trúc song song (parallel structure), các thành phần song song có quan hệ ngữ pháp giống nhau ...

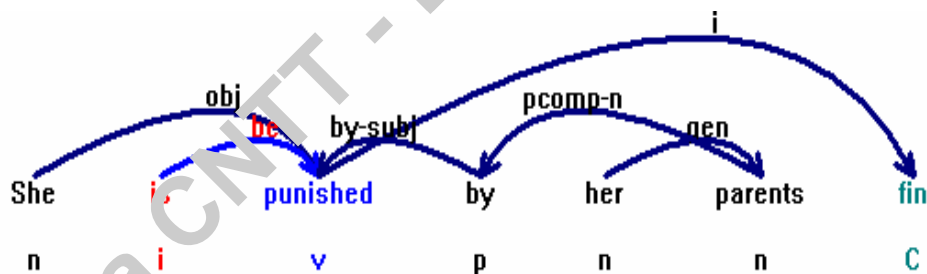
---

mod	Quan hệ bổ nghĩa giữa từ và định ngữ của nó
vrel	Động từ ở dạng bị động
wha, whn	Các từ dạng bắt đầu bằng WH
be	Quan hệ của động từ be và động từ khác (tiếp diễn, bị động)
det	Định từ
pcomp-n	Danh từ bổ nghĩa cho giới từ (trong ngữ giới từ)
gen	Sở hữu cách
by-subj	Từ <i>by</i> trong câu bị động. (sau <i>by</i> sẽ là tác nhân của hành động trong câu bị động).

**Bảng 2-1 : Một số quan hệ khi phân tích bằng văn phạm phụ thuộc**

**Ví dụ 2-10 :**

Cho câu *She is punished by her parents*. Các quan hệ được tìm thấy trong câu này được cho trong hình sau :



**Hình 2-6 : Các quan hệ phụ thuộc trong câu *She is punished by her parents*.**

Head <sup>7</sup> (cha)	Dependent (từ phụ thuộc)	Loại quan hệ
punished	she	obj (tân ngữ)

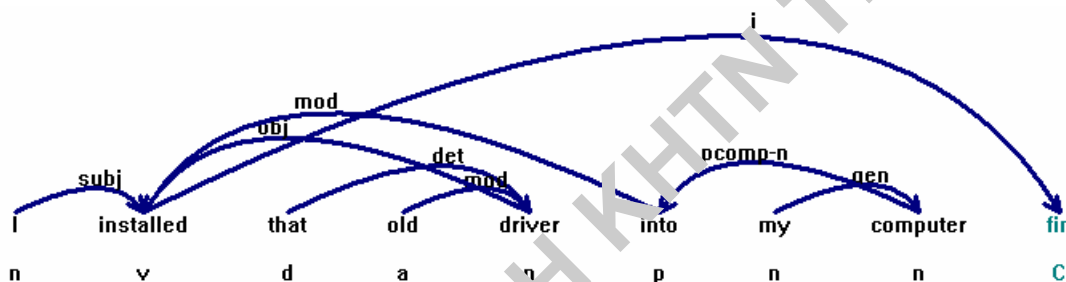
<sup>7</sup> Cho một quan hệ, chúng ta có thể đọc một cách dễ hiểu như sau : *head* là **loaiquanhe** của *dependent*. Chẳng hạn, *she* là **obj** của *punished*.



punished	is	be (bị động)
punished	by	by-subj (bị động)
parents	her	gen (sở hữu)
by	parents	pcomp-n (quan hệ giữa danh từ và giới từ trong ngữ danh từ)

**Ví dụ 2-11 :**

Với câu *I installed that old driver into my computer*, chúng ta có thể thấy được các mối quan hệ như sau :



**Hình 2-7 :** Các quan hệ phụ thuộc trong câu *I installed that old driver into my computer*.

Head (cha)	Dependent (từ phụ thuộc)	Loại quan hệ
installed	I	subj (chủ ngữ)
installed	driver	obj (tân ngữ)
driver	that	det (determiner)
driver	old	mod (bổ nghĩa)
computer	into	pcomp-n
computer	my	gen (sở hữu)

Khoa CNTT - ĐHQG KHTN TP.HCM

Chương 3

# MÔ HÌNH CÀI ĐẶT

*Chương 3 này đưa ra mô hình cài đặt cho khối xử lý ngữ nghĩa. Chương này nêu ra hai công đoạn để có thể đưa từ một câu tiếng Anh sang một câu tiếng Việt có gắn nghĩa : gắn nhãn ngữ nghĩa, và gắn nghĩa tiếng Việt. Trong chương này, chúng tôi còn thảo luận về các nguồn tri thức dùng cho xử lý ngữ nghĩa, hệ thống nhãn ngữ nghĩa, và cách xây dựng ngữ liệu huấn luyện.*

### 3.1. CÁC NGUỒN TRI THỨC ĐỂ XỬ LÝ NGỮ NGHĨA

Để xử lý ngữ nghĩa, người ta phải kết hợp nhiều nguồn tri thức khác nhau : thông tin về từ loại, hình thái, ngôn từ, quan hệ ngữ pháp, quan hệ ngữ nghĩa, và lĩnh vực xem xét. Dưới đây là miêu tả về các sử dụng các nguồn tri thức ấy trong xử lý ngữ nghĩa.

#### 3.1.1. Tri thức về từ loại và hình thái

Như đã trình bày ở Chương 1, thông tin từ loại của từ là một nguồn tri thức đáng kể để khử nhập nhằng nghĩa cho từ dù rằng thông tin này chưa đủ để khử nhập nhằng toàn bộ ngữ nghĩa. Thông tin từ loại có thể được dùng làm một bộ lọc để hạn chế số nghĩa cần xem xét. Trong câu *My/POS bank/NN is/AUX on/IN the/DT corner/NN*, nhờ thông tin từ loại, khối xử lý ngữ nghĩa đã loại bỏ được các nghĩa *gửi ngân hàng*, hay *đắp bờ* (các nghĩa động từ của từ *bank*), mà chỉ xét đến các nghĩa *bờ sông*, hay *ngân hàng* của từ này (các nghĩa có từ loại danh từ).

Trong một ví dụ (Ví dụ 1-1) được nêu ở phần *Vai trò và chức năng của xử lý ngữ nghĩa* (phần 1.2.1), câu *I can can a can* hoàn toàn có thể khử nhập nhằng tốt với điều kiện có được một bộ gán nhãn từ loại tốt. Từ *can* có 3 nghĩa khác nhau, mỗi nghĩa lại có một từ. Do đó, thông qua bộ gán nhãn từ loại, câu này được gán nhãn thành *I/PRP can/MD can/VB a/DT can/NN*. Khi đó, ứng với mỗi từ loại, ta dễ dàng chọn được một nghĩa thích hợp cho từ *can* (có thể (MD – động từ hình thái), đóng hộp (VB – động từ), cái hộp (NN - danh từ)).

Tương tự, chúng ta cũng có câu *I want to book two books*. Nhờ vào bộ gán nhãn từ loại mà chúng ta có từ loại của các từ trong ví dụ như sau : *I/PRP want/VB*

*to/AUX book/VB two/CD books/NNS*. Như vậy, nhờ từ loại khác nhau, *book/VB* (đặt trước) có thể phân biệt được với *books/NNS* (quyển sách).

Tuy nhiên, ở đây chúng ta giả sử rằng công đoạn gán nhãn từ loại đã được làm tốt, và những luật giải quyết nhập nhằng của chúng ta sẽ được áp dụng trên những văn bản đã được chú thích về từ loại. Chúng ta sẽ tập trung vào những từ mà không thể giải quyết nhập nhằng nếu chỉ sử dụng thông tin về từ loại. Chẳng hạn như để xác định nghĩa của từ *boxer* là *võ sĩ quyền anh* (danh từ) hay là *chó boxer* (danh từ) chúng ta phải cần thông tin về ngữ cảnh.

Một số từ có thể được giải quyết nhập nhằng bằng cách xác định từ loại của từ liên quan trong ngữ cảnh. Chẳng hạn, danh từ *way* có thể có hai nghĩa khác nhau: *con đường* (danh từ) hay *phương pháp/cách thức* (danh từ). Trong cả hai trường hợp này danh từ *way* có thể được theo sau bởi từ *to*. Nếu chúng ta biết được từ *to* có từ loại là giới từ (IN - preposition) hay là một phần tạo nên dạng nguyên thể của động từ (AUX - infinitive) thì chúng ta có thể sử dụng thông tin này để xác định nghĩa của danh từ *way*. Trong Ví dụ 3-1, từ *to* là một giới từ và vì thế nghĩa dịch đúng của danh từ *way* sẽ là *con đường, đường*. Trong Ví dụ 3-2, từ *to* là thành phần của một *infinitive* theo sau bởi động từ, cho nên danh từ *way* trong Ví dụ 3-2, sẽ có nghĩa là *phương pháp/cách thức*.

**Ví dụ 3-1 : Từ *way* được khử nhập nhằng nhờ vào giới từ *to* đi sau nó.**

- In a report, for example, the body text lines may reach all the **way** to the left and right margins, but quoted material may be indented 1 inch from each margin.
- Nonetheless, UNIX never really caught on as a consumer operating system, giving **way** to DOS, Windows, and the Mac OS, which generally have been perceived as easier to learn and use.
- Even though Cho and Hermione were on the **way** to becoming friends, they didn't have a relationship anything like Harry had with Cho.

**Ví dụ 3-2 : Từ *way* được khử nhập nhằng nhờ vào *to Inf* đi sau nó.**

- One might think that pen-based systems would be a handy **way** to enter text into the computer for word processing.
- Adding RAM is a relatively inexpensive **way** to boost a system's overall performance.
- The F1 key, for example, became the universal **way** to access online help.

Một ví dụ khác là việc sử dụng danh từ *stage* với các từ có từ loại chỉ số thứ tự (ORD – Ordinal number) hay số đếm (CD – Cardinal number). Khi danh từ này được đi trước bởi một từ chỉ thứ tự (Ví dụ 3-3) hoặc theo sau bởi một từ chỉ số đếm (Ví dụ 3-4) thì dường như nó có nghĩa là *giai đoạn* hơn là *sân khấu*.

**Ví dụ 3-3 : Từ *stage* được khử nhập nhằng nhờ vào một số thứ tự đứng trước :**

- In *next **stage***, this system will connect to the character recognition system in order to translate texts automatically.
- Besides, I also send my sincere thanks to all professors who help me in *last **stage***.

**Ví dụ 3-4 : Từ *stage* được khử nhập nhằng nhờ vào một từ chỉ số đếm đứng sau :**

- in **stage** *two*
- **Stage** *1*

### 3.1.2. Tri thức về ngôn từ

Ngôn từ cũng là một nguồn thông tin đáng chú ý trong khử nhập nhằng ngữ nghĩa của từ. Những hư từ (giới từ, mạo từ, đại từ, liên từ...) xung quanh từ đang được xét có thể giúp xác định cách dịch cụ thể của từ.

Chẳng hạn, với từ *way* được nêu ở trên, nếu phía sau nó có giới từ *of* thì nó được dịch là *cách* thay vì dịch là *đường*. Một ví dụ khác, trong cụm từ *the date and*

*the fruit*, thông qua liên từ *and*, chúng ta biết được từ *date* và từ *fruit* có quan hệ song song *parallel structure* nên hai từ này có nghĩa phải chia sẻ một ý niệm nào đó. Do từ *fruit* được dịch với nghĩa *quả, trái cây* nên từ *date* bắt buộc phải có nghĩa là *quả chà* là thay vì nghĩa *ngày tháng*.

### 3.1.3. Tri thức về quan hệ cú pháp và ràng buộc ngữ nghĩa

Thông tin về một số quan hệ ngữ nghĩa giữa các từ có thể được sử dụng để hình thành những luật giải quyết nhập nhằng tổng quát hơn. Chẳng hạn như động từ *raise* có thể được dịch theo nhiều cách khác nhau. Động từ này có thể được giải quyết nhập nhằng một cách dễ dàng nếu chúng ta biết rằng tân ngữ trực tiếp của nó là một danh từ chỉ động vật (như *pigs, dogs, chickens, ...*)

Thông tin về những đặc trưng ngữ nghĩa của một từ cũng có thể được sử dụng cho việc phát sinh những quy luật trong trường hợp ngữ liệu huấn luyện không chứa đủ số những ví dụ nhập nhằng đối với một vài nghĩa đặc biệt. Ví dụ như từ *date* vừa có nghĩa là *ngày* vừa có nghĩa là *quả chà* là. Tuy nhiên nghĩa *quả chà* là hiếm khi xuất hiện. Để phát sinh các quy luật giải quyết nhập nhằng cho danh từ *date* chúng ta có thể phân tích ngữ cảnh của những từ khác cũng chỉ về *trái cây*. Chúng ta có thể giả sử rằng từ *date* có nghĩa là *trái cây* và những từ như *banana, pineapple* hay *apricot* thường được sử dụng trong ngữ cảnh tương tự.

Trong các ví dụ ở phần trên, chúng ta đã quen với câu (1) *The old man has an old book.* hay câu (2) *I installed that old driver into my computer.* Nhờ đâu (thông tin nào) mà các câu này có những cách dịch chính xác<sup>8</sup>? Câu trả lời là đều nhờ vào thông tin quan hệ ngữ pháp. Ở đây, các quan hệ ngữ pháp được xem xét là: S – V (chủ ngữ và động từ), V – O (động từ và tân ngữ), A – N (tính từ và danh từ). Thứ nhất, trong câu (1) có hai quan hệ ngữ pháp cần lưu ý: A – N của *old* và *man*; A – N của *old* và *book*. *man* có thuộc tính người nên giúp khử được nhập nhằng

---

<sup>8</sup> *old man* được dịch là *người đàn ông già* chứ không dịch là *người đàn ông cũ*. *old book* dịch là *quyển sách cũ* chứ không phải là *quyển sách già*. *old driver* được dịch là *trình điều khiển cũ* chứ không phải là *trình điều khiển già* hay *tài xế già* hay *tài xế cũ*.

---

cho từ *old* (cũ hay già ??? => chọn *già*), *book* có thuộc tính đồ vật nên giúp khử nhập nhằng cho từ *old* (cũ hay già ??? => chọn *cũ*). Thứ hai, câu (2) có các quan hệ ngữ pháp đáng quan tâm : S – V của *I* và *installed*, V – O của *installed* và *driver*, A–N của *old* và *driver*. Như đã biết, tân ngữ của *installed* (cài đặt) phải là một từ có nghĩa thuộc nhóm *phần mềm* hay *phần cứng*, khi đó giúp chọn nghĩa đúng cho từ *driver* (phải là *trình điều khiển* (nhóm phần mềm) chứ không là *tài xế* (người)). Sau khi từ *driver* được xác định nghĩa thì từ *old* sẽ được xác định nghĩa theo.

### 3.1.4. Tri thức về chủ đề

Trong một số trường hợp nhập nhằng, chúng ta có thể xác định được nghĩa đúng của từ nếu ta biết được chủ đề của văn bản. Chẳng hạn, từ *bank*, nếu đang nói về vấn đề về “tài chính” thì nó thường có nghĩa là *ngân hàng* ; từ *driver* có nghĩa là *trình điều khiển* (nếu chủ đề là lĩnh vực tin học) ; *sentence* có nghĩa là *câu* (nếu chủ đề là ngôn ngữ, văn phạm) hoặc *bản án* nếu đang nói về pháp luật ; *element* có nghĩa *nguyên tố* (trong lĩnh vực hoá học) ; và *phần tử* (trong toán/tin học).

Để xác định được chủ đề của văn bản đang cần dịch, ta cần xem xét sự xuất hiện của một số từ chuyên môn trong lĩnh vực đó. Chẳng hạn, nếu trong văn bản ta thấy xuất hiện những từ như : *ellipsis* (tính lược), *bilingual* (song ngữ), *anaphora* (thế đại từ), *phrase* (ngữ) thì ta có thể đoán nhận văn bản này đang nói về chủ đề ngôn ngữ học.

Chúng ta có thể xác định được chủ đề một cách tự động bằng cách xem xét các từ chuyên môn lân cận từ đang cần khử nhập nhằng theo công thức của Yarowsky (lân cận từ xem xét là cửa sổ 50-từ xung quanh từ đang khử nhập nhằng):

$$ARGMAX_{Scat} \sum_{w \in W} \log \frac{\Pr(w | SCat) \Pr(SCat)}{\Pr(w)}$$

#### Công thức 3-1 : Công thức xác định chủ đề văn bản

Trong đó :

- Scat : mã chủ đề
- W : khung cửa sổ chứa từ *w*

Do xác suất  $\Pr(\text{Scat})$  không phụ thuộc vào  $w$  nên công thức trên được viết lại thành :

$$\text{ARGMAX}_{\text{Scat}} \sum_{w \in W} \log \frac{\Pr(w | \text{Scat})}{\Pr(w)}$$

**Công thức 3-2 : Công thức xác định chủ đề văn bản (sau khi biến đổi)**

### 3.1.5. Tri thức về tần suất nghĩa của từ

Không phải từ nào cũng thuộc về một chủ đề nào đó, vì vậy tính thông dụng của một nghĩa nào đó được dựa trên độ đo về tần suất xuất hiện của từ đó với nghĩa cụ thể. Chẳng hạn, danh từ *pen* sẽ có nghĩa thông dụng nhất là *bút/viết* (bên cạnh các nghĩa ít thông dụng hơn như *chuông*, *lông chim*) ; *ball* thường có nghĩa là *quả banh/hòn bi* hơn là *buổi khiêu vũ*.

Độ đo tần suất xuất hiện của mỗi nghĩa của từ được thống kê trên những ngữ liệu rất lớn thuộc nhiều loại văn bản khác nhau. Chính vì vậy, trong WordNet và trong LDOCE, các nghĩa được sắp xếp theo thứ tự giảm dần (nghĩa thông dụng nhất sẽ được liệt kê đầu tiên).

## 3.2. CÁC BƯỚC THỰC HIỆN

Khối xử lý ngữ nghĩa là một bộ phận trong hệ dịch tự động Anh-Việt (Xem Hình 1-5). Khối này kế thừa các kết quả có được từ các khối xử lý trước : tiền xử lý, phân tích hình thái học, phân tích cú pháp.

Khối này gồm có hai công đoạn : (1) gán nhãn ngữ nghĩa cho các từ trong câu ; (2) gán nghĩa tiếng Việt cho các từ với nhãn có sẵn. Công đoạn thứ nhất bắt đầu với câu tiếng Anh đã được phân tích cú pháp, gán nhãn từ loại, rút trích các quan hệ ngữ pháp, ngữ nghĩa, áp dụng tập luật rút ra được trong quá trình huấn luyện, gán nhãn ngữ nghĩa cho các từ trong câu. Sau đó, kết hợp với kết quả rút ra được từ quá trình chuyển đổi cây cú pháp, công đoạn thứ hai sẽ có cách chọn nghĩa tiếng Việt hợp lý cho nhãn, hình thành câu tiếng Việt có thể hiểu được.



Phần gán nhãn ngữ nghĩa cho các từ trong câu tiếng Anh được thực hiện trên cơ sở áp dụng các luật rút ra được trong quá trình huấn luyện sử dụng thuật toán học dựa trên chuyển đổi fnTBL. Phần này đòi hỏi một ngữ liệu lớn để huấn luyện, một phương pháp gán nhãn cơ sở (baseline), hệ thống mẫu luật để tạo luật và quan trọng hơn cả đó là hệ thống nhãn ngữ nghĩa áp dụng trong quá trình gán nhãn ngữ nghĩa. Các công việc cần thực hiện trong phần này bao gồm :

- Xây dựng hệ thống nhãn ngữ nghĩa thích hợp.
- Chuẩn bị ngữ liệu.
- Tạo ngữ liệu vàng<sup>9</sup>.
- Xây dựng mẫu luật.
- Áp dụng thuật toán học rút luật.

Sau khi công đoạn thứ nhất hoàn thành, chúng ta sẽ nhận được đầu ra là câu tiếng Anh trong đó mỗi từ đã được gán tương ứng với một nhãn ngữ nghĩa. Kết quả của công đoạn này sẽ được kết hợp với kết quả của giai đoạn chuyển đổi cây cú pháp cùng với từ điển tiếng Việt có nhãn tương ứng để tạo được câu tiếng Việt. Các công việc cần thực hiện trong phần này bao gồm :

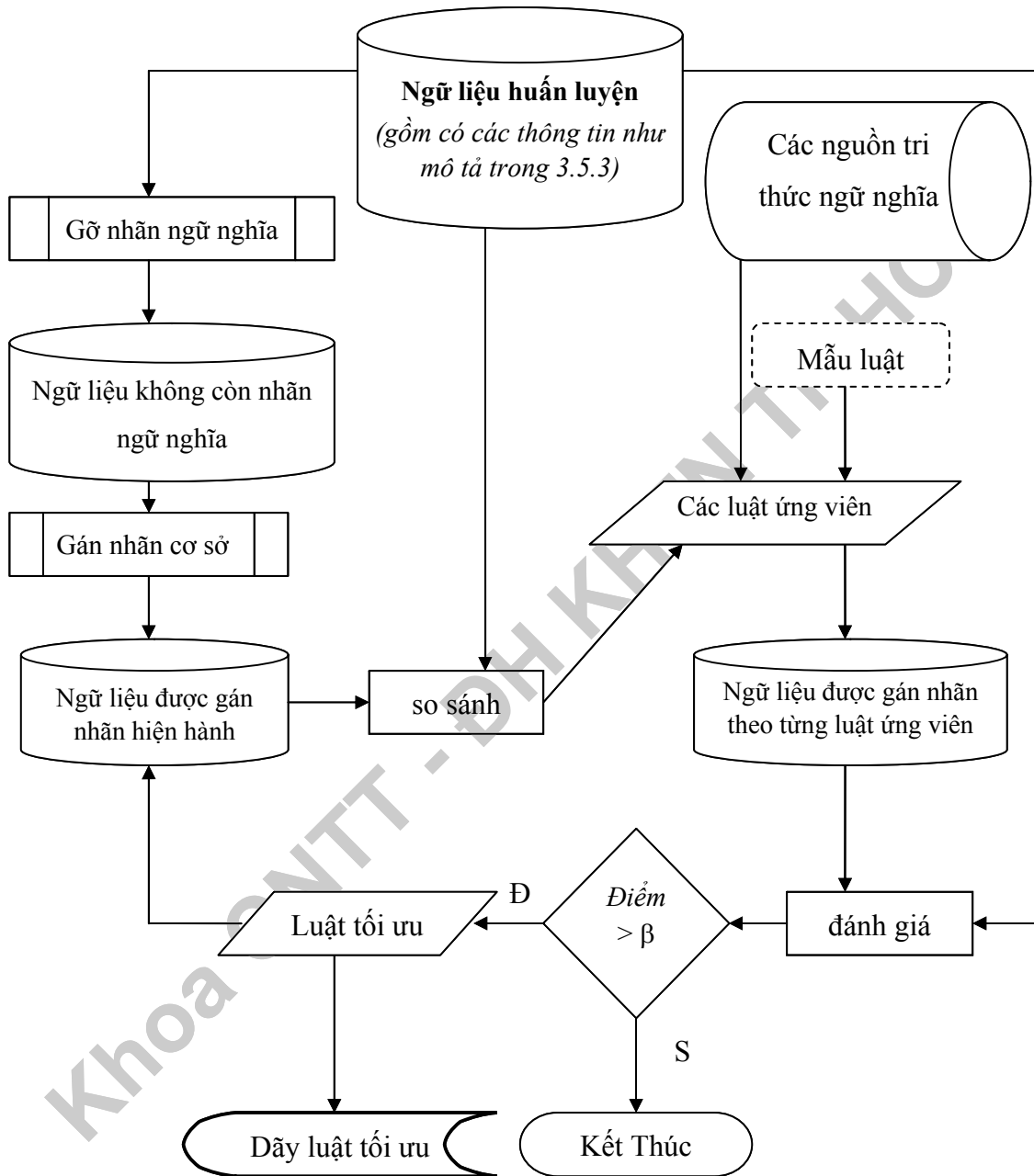
- Gán tiếng Việt vào nhãn.
- Hoàn chỉnh câu tiếng Việt (loại bỏ, hoặc thêm vào các hư từ, lượng từ...)

---

<sup>9</sup> Ngữ liệu vàng (golden corpus). Đây là một loại ngữ liệu mà trong đó các thông tin được đưa vào là hoàn toàn chính xác. Các ngữ liệu này thường phải được xây dựng bằng tay. Các ngữ liệu này rất hiếm do việc tạo lập rất khó khăn. Chính vì vậy người ta gọi chúng là *ngữ liệu vàng*.

---

### 3.3. MÔ HÌNH HUẤN LUYỆN CHO BỘ GÁN NHÃN NGỮ NGHĨA



Hình 3-1: Mô hình huấn luyện cho bộ gán nhãn ngữ nghĩa

### 3.4. HỆ THỐNG NHÃN NGỮ NGHĨA

#### 3.4.1. Yêu cầu đối với hệ thống nhãn ngữ nghĩa

Nếu như các hệ thống nhãn ngữ pháp (gồm có nhãn từ loại, cú pháp) đã được thống nhất và xác định rõ ràng (hệ thống nhãn ngữ pháp của Penn Tree Bank gồm khoảng 100 nhãn) thì ngược lại việc xây dựng hệ thống nhãn ngữ nghĩa thống nhất đến nay vẫn chưa hoàn tất và hiện nay vẫn đang tồn tại nhiều hệ thống nhãn ngữ nghĩa khác nhau. Các hệ thống nhãn được xây dựng tùy thuộc vào yêu cầu của từng công việc và mức độ khừ nhập nhằng của các công trình tương ứng. Chung quy lại, yêu cầu đối với hệ thống nhãn gồm có hai điểm cần lưu ý : không quá mịn cũng không quá thô, nhưng phải đủ để xác định các nghĩa khác nhau cho các từ.

Một hệ thống nhãn phân giải quá chi tiết (quá mịn) làm cho việc xây dựng ngữ liệu cực kỳ khó khăn. Như đã biết, trong quá trình huấn luyện, để đạt được kết quả tốt thì mỗi nhãn cần có vài trăm lần xuất hiện. Với hệ thống nhãn quá mịn (WordNet 1.7.1 có 195817 nghĩa khác nhau) mỗi nhãn cho một nghĩa, ngữ liệu cần đến vài chục triệu câu, vài trăm triệu đến vài tỷ từ. Đây là một việc gần như không thể<sup>10</sup>.

Trong khi đó, một hệ thống nhãn quá thô sẽ làm cho việc khừ nhập nhằng nghĩa không hiệu quả. Chẳng hạn, từ *letter* sẽ có hai nghĩa tiếng Việt tương ứng là (1) *bức thư* ; (2) *chữ cái*. Nếu như chọn chung một thuộc tính TXT (Text) làm nhãn cho cả hai nghĩa của từ *letter* thì không thể nào xác định được khi nào có nghĩa *bức thư*, khi nào có nghĩa *chữ cái*.

Do đó, cái cần thiết của hệ thống nhãn ngữ nghĩa phục vụ cho công việc khừ nhập nhằng nghĩa của từ là phải xác định được những trường hợp nào không cần

---

<sup>10</sup> Xây dựng ngữ liệu dành cho công việc khừ nhập nhằng ngữ nghĩa đòi hỏi rất nhiều công sức và thời gian. Nó đòi hỏi phải cung cấp đầy đủ các thông tin chính xác : từ gốc, từ loại, cú pháp, và nhãn ngữ nghĩa.

phải phân giải nghĩa<sup>11</sup>, trường hợp nào cần phải phân giải nghĩa<sup>12</sup>. Dựa trên lý luận “*dịch để phục vụ người đọc chứ không phải phục vụ cho máy hiểu văn bản*”, chúng tôi đã xây dựng một hệ thống nhãn ngữ nghĩa riêng để phục vụ cho quá trình xử lý ngữ nghĩa.

### 3.4.2. Cơ sở của việc phân lớp ngữ nghĩa

Lâu nay, chúng ta quá quen với các từ điển thông thường (đơn ngữ hay song ngữ) được sắp xếp theo thứ tự abc của mục từ. Chính vì vậy mà hai mục từ *animal* (động vật) và *zoo* (sở thú), hoặc *aunt* (cô/dì) và *uncle* (chú/bác) được đặt ở các vị trí rất xa nhau, chẳng có liên quan gì với nhau về mặt ngữ nghĩa. Từ điển theo trật tự abc thì hợp lý và chặt chẽ về mặt hình thức (hình thái) nhưng lại không hợp lý về mặt hợp lý về mặt nội dung (ngữ nghĩa) và cũng không phù hợp với tư duy ngôn ngữ của con người.

Một thực nghiệm được các nhà ngôn ngữ học - tâm lý thực hiện để kiểm xem ở con người hệ thống ngữ nghĩa (từ điển) được sắp xếp như thế nào. Họ cho một từ kích thích *aunt* cho nhiều người khác nhau và đặt câu hỏi là anh/chị sẽ nghĩ đến từ nào đầu tiên. Kết quả thu được là đa số đều cho biết trong đầu họ nghĩ đến từ *uncle* trước nhất. Điều đó chứng tỏ rằng, ngay “lời nói bên trong” của con người chúng ta, từ *uncle* và *aunt* đã có quan hệ với nhau. Đây cũng chính là nền tảng lý thuyết về ngữ nghĩa từ vựng mà các nhà làm từ điển phân lớp ý niệm đã dựa vào khi xây dựng các hệ thống phân lớp ngữ nghĩa và gán nhãn ngữ nghĩa cho mỗi lớp đó. Đến nay, đã có một số hệ thống phân lớp như trên : từ điển LLOCE/LDOCE, WordNet, CoreLex...

---

<sup>11</sup> Trong WordNet chẳng hạn, từ *coffee* có 4 nghĩa khác nhau (một loại thức uống, một loại cây, một loại hạt, và một loại màu). Song, khi dịch qua tiếng Việt, từ *coffee* chỉ cần một nghĩa duy nhất là *cà phê* thôi. Việc xem xét nó là một loại thức uống, một loại cây, hay một loại màu sắc thì người đọc chắc chắn đủ tri thức để có thể hiểu được.

<sup>12</sup> Tuy nhiên, cũng có những từ cần phải phân giải nghĩa rõ ràng, như từ *letter* được nêu ở trên chẳng hạn.

---

Kết quả nghiên cứu về *phổ quát ngôn ngữ* cho thấy : một số *phổ quát ngôn ngữ* là từ các hiện tượng tâm lý – ngôn ngữ học, vì thế, một cách khái quát, nó phụ thuộc vào mối quan hệ giữa *ngôn ngữ* và *tư duy* của con người ; một số *phổ quát ngôn ngữ* lại là những hiện tượng về dân tộc – ngôn ngữ học, vì thế nó phụ thuộc vào mối quan hệ giữa *ngôn ngữ* và *văn hoá*. Các nhà nghiên cứu chia *phổ quát ngôn ngữ* thành 2 dạng :

❑ **Các phổ quát về thực thể :**

Là những nét chung về sự tổ chức các thực thể ngôn ngữ. Chẳng hạn, mọi ngôn ngữ đều tồn tại các phạm trù danh từ và động từ, nó là cơ sở để biểu hiện cấu trúc chìm của câu trong mọi ngôn ngữ.

❑ **Các phổ quát về dạng thức :**

Chẳng hạn, ngữ pháp tạo sinh coi rằng bộ phận cơ sở của cú pháp trong mọi ngôn ngữ thì giống nhau.

Ngoài các phổ quát ngôn ngữ về ngữ âm, ngữ pháp, ngữ nghĩa (là những phổ quát chỉ đề cập tới một phương diện ký hiệu hoặc tới cái biểu đạt hoặc tới cái được biểu đạt), người ta còn chú ý tới các phổ quát ngôn ngữ về ký hiệu, chúng đề cập tới cái quan hệ giữa cái biểu đạt và cái được biểu đạt.

*Giáo trình ngôn ngữ học đại cương* (Ferdinand de Saussure) đã chỉ ra hai dạng quan hệ : ngang (tuyến tính, hình tuyến, ngữ đoạn) và dọc (hệ hình, trục tuyến). Tương ứng với quan hệ ngang có trường nghĩa tuyến tính và trường nghĩa liên tưởng, còn ứng với quan hệ dọc có trường nghĩa biểu vật và trường nghĩa biểu niệm. Trường nghĩa biểu vật là tập hợp những từ đồng nghĩa về ý nghĩa biểu vật và trường nghĩa biểu niệm là một tập hợp các từ có chung cấu trúc biểu niệm.

### **3.4.3. Nhận xét các hệ thống nhãn ngữ nghĩa có liên quan**

Trong phần này chúng tôi đề cập đến hệ thống nhãn ngữ nghĩa của LLOCE, LDOCE, WordNet và CoreLex.

Cách phân chia các lớp của LLOCE thực chất là dựa trên cơ sở lý thuyết phân chia trường ngữ nghĩa theo trục dọc (trường nghĩa biểu vật và biểu niệm). Đối

với WordNet, ngoài việc dựa trên cơ sở lý thuyết phân chia theo trường biểu vật và biểu niệm, nó còn dựa vào cơ sở phân chia theo trường nghĩa tuyến tính và trường nghĩa liên tưởng (qua các quan hệ chức năng, bộ phận, tính chất...).

Do mục tiêu ban đầu là hệ thống các ý niệm chung nhất cho mọi ngôn ngữ của nhân loại nên việc biểu diễn hệ thống các ý niệm trong WordNet được dựa trên cơ sở lý thuyết về ngôn ngữ học-tri nhận (cognitive linguistics), ngôn ngữ học tâm lý (psycho-linguistics),... nhưng tất cả các lý thuyết này đều hướng tới một mục tiêu chung là nghiên cứu về sự chung nhất của mọi ngôn ngữ trên thế giới hay còn gọi là phổ quát của ngôn ngữ.

Hệ thống nhãn LDOCE chỉ chú trọng đến danh từ, có số lượng từ khá lớn (45.000) nhưng sự phân chia lớp ngữ nghĩa quá thô (chỉ có 32 lớp).

Hệ thống nhãn LLOCE đơn giản, hệ thống phân cấp chỉ gồm 3 cấp (chủ đề - nhóm - lớp), số nhãn không quá lớn (gồm 2441 nhãn). Hệ thống phân lớp 3 cấp nên giữa các lớp khó tìm mối quan hệ với nhau. Số lượng từ trong LLOCE còn khá hạn chế (chỉ gồm 16.000 mục từ).

Hệ thống nhãn của WordNet rất chi tiết, đầy đủ (cho các từ loại chính), vì vậy số lượng rất lớn (hơn 100.000 nhãn). WordNet có ưu điểm là phân cấp chi tiết (hàng chục cấp) và giữa các lớp đồng nghĩa còn có nhiều kiểu quan hệ khác nhau.

Hệ thống nhãn CoreLex (dành cho danh từ) phân biệt được từ đồng nghĩa và từ đồng tự trong khi đó WordNet thì không .

Trọng tâm hệ thống nhãn ngữ nghĩa của chúng tôi là để khử nhập nhằng ngữ nghĩa của từ cho mục đích dịch chứ không phải cho mục đích hiểu nên không cần phải phân giải ngữ nghĩa chi tiết như trong WordNet.

Hệ thống nhãn LDOCE thì quá thô, không đủ sức khử nhập nhằng cho các từ cùng lớp nhưng khác nghĩa.

Hệ thống nhãn CoreLex được xây dựng từ các lớp cơ bản của WordNet và có cá mã số nhãn là các từ viết tắt chữ đầu, dễ nhớ hơn các nhãn của các hệ thống khác (chỉ dùng con số, hay chữ số).

### 3.5. CHUẨN BỊ NGỮ LIỆU HUẤN LUYỆN

#### 3.5.1. Giới thiệu kho ngữ liệu song ngữ Anh-Việt VCLEVC

Kho ngữ liệu song ngữ VCLEVC (VCL English-Vietnamese Corpus) được thu nhập từ nhiều nguồn văn bản song ngữ khác nhau (sách, từ điển, ngữ liệu...) thuộc lĩnh vực Khoa học, Kỹ thuật,.. Việc thu nhập phải tuân theo các tiêu chí nhất quán về mặt ngôn ngữ, về văn phong, lĩnh vực... (xem thêm trong [6])

Sau khi thu thập từ nhiều nguồn khác nhau, các văn bản song ngữ này được tiền xử lý : chuẩn hoá về dạng văn bản (text only), font chữ, chuẩn hoá chính tả... Sau đó, các văn bản song ngữ này được đánh mã số tương ứng với từng cặp câu.

Các cặp câu này được thực hiện liên kết từ<sup>13</sup> (Word Align) tự động nhờ vào một chương trình.

Dưới đây là mẫu của ngữ liệu song ngữ đã được liên kết :

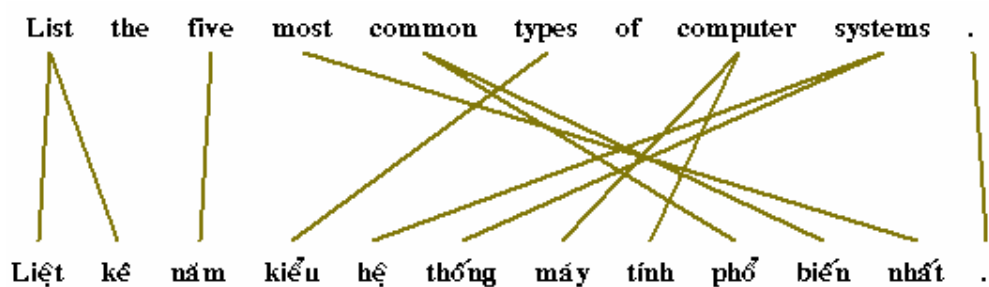
List the five most common types of computer systems .
Liệt kê năm kiểu hệ thống máy tính phổ biến nhất .
1 1 3 6 9 9 8 8 5 5 4 10
Identify two unique features of supercomputers .
Xác định hai đặc trưng duy nhất của siêu máy tính .
[1_1,2] [2_3] [3_6,7] [4_5,4] [5_8] [6_9,10,11] [7_12]
Differentiate workstations from personal computers
Phân biệt trạm làm việc với máy tính cá nhân .
1 1 2 2 2 0 5 5 4 4 6

**Hình 3-2 : Minh hoạ các cặp được liên kết trong ngữ liệu song ngữ**

---

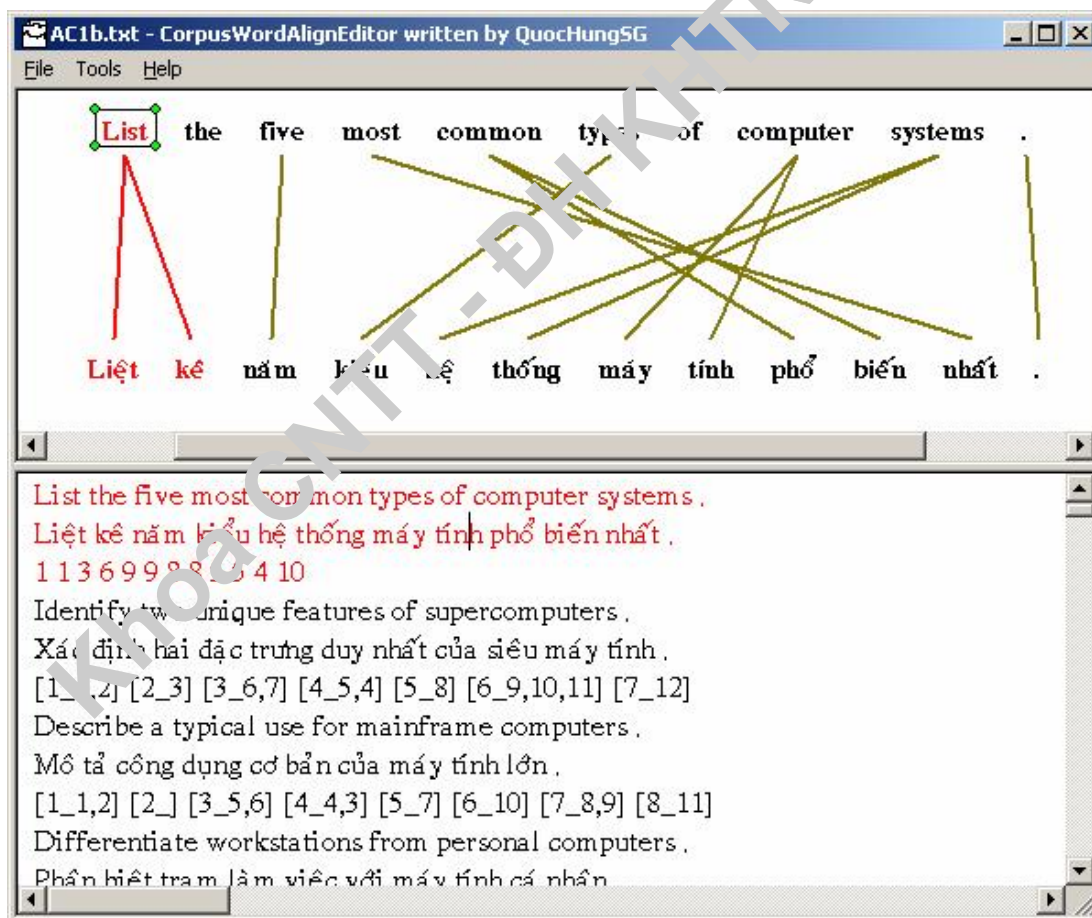
<sup>13</sup> Liên kết từ là xác định mối liên kết giữa một (nhiều) từ tiếng Anh với một (nhiều) từ tiếng Việt tương ứng.

---



Hình 3-3 : Thể hiện các mối liên kết của một cặp câu

Sau khi được chương trình liên kết tự động, các mối liên kết này lại được hiệu chỉnh bằng tay (do người thực hiện) để đảm bảo rằng các mối liên kết thu được là hoàn toàn chính xác. Việc hiệu chỉnh bằng tay được thực hiện nhờ vào công cụ CorpusWordAlignEditor.



Hình 3-4 : Công cụ WordAlignEditor



### 3.5.2. Rút trích thống kê từ ngữ liệu song ngữ

Phần trên vừa mô tả về kho ngữ liệu song ngữ VCLEVC. Tiếp theo, sẽ là bước thực hiện tiếp theo trong xây dựng ngữ liệu huấn luyện cho xử lý ngữ nghĩa : Các cặp câu song ngữ đã liên kết được sử dụng để rút trích các thống kê

#### 3.5.2.1. Thống kê các nghĩa tiếng Việt

Từ tiếng Anh	Các nghĩa tiếng Việt tương ứng
accomplished	hoàn hảo
according	tùy thuộc##theo##dựa theo
according to	theo##tùy theo
account	chiếm##giải thích##tài khoản
accountants	kế toán
accuracy	tính chính xác##độ chính xác
accurate	chính xác##sự chính xác
achieve	đạt tới##đạt được
upgrade	nâng cấp##việc nâng cấp##sự nâng cấp
upgraded	nâng cấp
upgrades	bản nâng cấp
upgrading	nâng cấp##việc nâng cấp
upper	góc trên##viết hoa
uppercase	viết hoa

**Bảng 3-1 : Trích thống kê các nghĩa tiếng Việt dựa vào ngữ liệu song ngữ**

### 3.5.2.2. Thống kê tần số xuất hiện một nghĩa của từ tiếng Anh

Từ tiếng Anh với nghĩa tiếng Việt	Tần số xuất hiện
after##sau khi	25
after all##nói cho cùng	3
computer##máy tính	557
computer##máy vi tính	3
computerized##vi tính hóa	1
computerized##điện toán hóa	1
develop##phát triển	9
developed##cải tiến	3
developed##phát triển	20
knowledge##kiến thức	9
knowledgebase##cơ sở tri thức	3
microprocessors##bộ vi xử lý	14
microprocessors##bộ xử lý	3
microprocessors##vi xử lý	2
types##kiểu	22
types##loại	68
typewriter##máy đánh chữ	7

**Bảng 3-2 : Trích thống kê tần số xuất hiện của nghĩa tiếng Việt của một từ tiếng Anh dựa vào ngữ liệu song ngữ.**

### 3.5.2.3. Ý nghĩa

Các thống kê này có nhiều ý nghĩa trong việc xây từ điển Anh – Việt phục vụ cho dịch máy, từ điển nhãn ngữ nghĩa cho các từ tiếng Anh, từ điển nhãn ngữ nghĩa cho các từ tiếng Việt :

- Dựa trên các thống kê này, chúng tôi bao quát đầy đủ các cách dịch nghĩa gặp trong thực tế của các từ tiếng Anh trong một lĩnh vực trong ứng (bởi vì các tài liệu thu thập được để sử dụng cho việc xây dựng kho ngữ liệu song ngữ là do người dịch nên các cách dịch nghĩa của từ rất phong phú, đa dạng). Từ tập hợp nghĩa tìm thấy được cho mỗi từ, chúng tôi còn dựa vào tần số xuất hiện của chúng và mức độ tương đồng về nghĩa tiếng Việt để xem xét, chọn lựa những cách dịch nghĩa nên đưa vào từ điển, những cách dịch phải loại bỏ đi<sup>14</sup>.
- Việc thống kê này đảm bảo từ điển sau khi xây dựng xong có lượng từ đủ lớn để phục vụ cho hệ dịch.
- Xác định được các nghĩa khác nhau của các từ. Từ đó có cơ sở kiểm chứng việc xây dựng hệ thống nhãn ngữ nghĩa có đủ để xác định được nghĩa cho các từ chưa.

### 3.5.3. Xây dựng ngữ liệu huấn luyện

Bước kế đến trong khối xử lý ngữ nghĩa là xây dựng từ điển Anh-Việt, từ điển nhãn ngữ nghĩa cho từ tiếng Anh, từ điển nhãn ngữ nghĩa cho từ tiếng Việt. Bước này sẽ không được mô tả chi tiết ở đây. Kết quả đầu ra của bước này là các nghĩa tiếng Anh được gắn kèm theo một nhãn ngữ nghĩa, các nghĩa tiếng Việt được

---

<sup>14</sup> Trên bảng thống kê về tần số xuất hiện của nghĩa tiếng Việt đối với từ tiếng Anh (Bảng 3-2), từ *computer* có hai cách dịch tiếng Việt (theo thống kê) là *máy vi tính*, *máy tính*. Hai cách dịch này thực chất thể hiện một nghĩa trong tiếng Việt. Dựa vào tần số của chúng, chúng tôi đưa cách dịch *máy tính* (bỏ cách dịch *máy vi tính*) vào trong từ điển phục vụ cho dịch máy của mình.

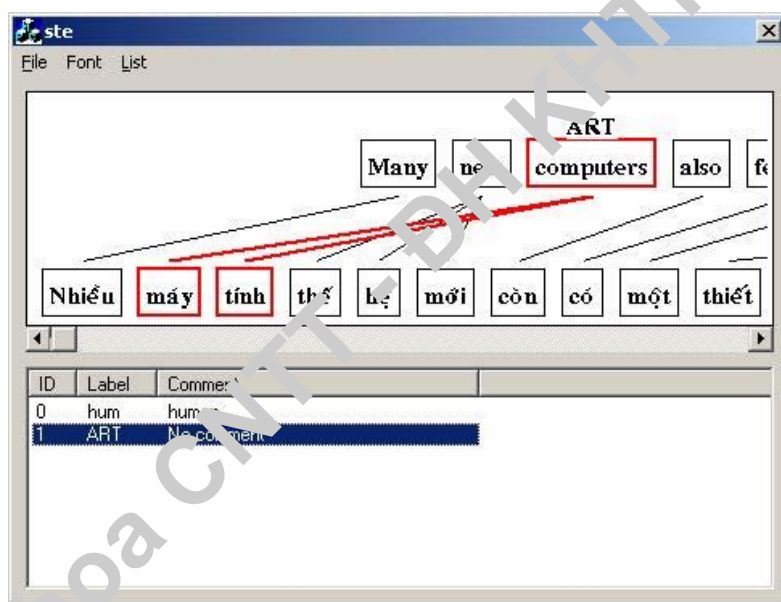
---

gắn kèm theo một nhãn ngữ nghĩa. Cả hai từ điển nhãn ngữ nghĩa này sử dụng chung một hệ thống nhãn ngữ.

### 3.5.3.1. Gán nhãn ngữ nghĩa bán tự động cho ngữ liệu

Lưu ý rằng, cho đến lúc này, kho ngữ liệu chỉ chứa các câu đã được liên kết đúng hoàn toàn, chưa có một thông tin nào để sử dụng cho việc xử lý ngữ nghĩa. Nhưng nhờ vào các mối liên kết của từ tiếng Anh và từ tiếng Việt cộng với lý luận “*các ý niệm giống nhau của thế giới thực thể hiện trong các ngôn ngữ khác nhau sẽ giống nhau*”, chúng tôi có thể gán ngữ nghĩa cho các từ trong câu tiếng Anh một cách bán tự động.

Để gán nhãn bán tự động cho ngữ liệu song ngữ này, chúng tôi đã xây dựng một công cụ có tên là SenseTaggerEditor.



**Hình 3-5 : Công cụ SenseTaggerEditor**

Công cụ này sẽ sử dụng từ điển ngữ nghĩa cho các từ tiếng Anh, từ điển ngữ nghĩa cho các từ tiếng Việt, và ngữ liệu song ngữ để hỗ trợ người thực hiện công việc gán nhãn. Công cụ này sẽ thực hiện như sau :

- Cho phép người gán nhãn xác định hai cặp câu đã được liên kết từ.

- Với mỗi liên kết tìm thấy được, tìm tập hợp nhãn ngữ nghĩa đối với từ tiếng Anh  $E$ , tập hợp nhãn ngữ nghĩa đối với từ tiếng Việt  $\Omega$ .
- Giao<sup>15</sup> hai tập hợp nhãn ngữ nghĩa này lại ( $\Sigma = E \cap \Omega$ ).
- Xây ra các trường hợp sau đối với tập  $\Sigma$  :
  - Trường hợp 1 : Số nhãn trong tập  $\Sigma$  bằng 0 ( $\Sigma = \emptyset$ ), người đánh nhãn sẽ chọn từ danh sách nhãn ngữ nghĩa.
  - Trường hợp 2 : Số nhãn trong tập  $\Sigma$  là 1 ( $|\Sigma| = 1$ ), công cụ sẽ gán nhãn trong tập  $\Sigma$  cho cả hai từ (tiếng Anh và tiếng Việt).
  - Trường hợp 3 : Số nhãn trong tập  $\Sigma$  nhiều hơn 1 ( $|\Sigma| > 1$ ), công cụ cho phép người đánh nhãn chọn từ danh sách nhãn của tập  $\Sigma$ .

Mặc dù công việc gán nhãn ngữ nghĩa cho một ngữ liệu là một công việc khó khăn nhưng nhát nhẽo, nhảm chán và tốn thời gian, song từ các thông tin liên kết từ (các mối liên kết có được giữa các từ tương quan hỗ trợ qua lại với nhau) chúng tôi đã tạo được một ngữ liệu có gán nhãn ngữ nghĩa ít khó khăn, ít tốn thời gian hơn.

### 3.5.3.2. Xây dựng “ngữ liệu vàng”

Sau khi các từ trong ngữ liệu đã được gán nhãn ngữ nghĩa, chúng tôi tách các câu tiếng Việt ra riêng. Lúc này, ngữ liệu song ngữ chỉ còn là ngữ liệu đơn ngữ (chỉ còn tiếng Anh). Các câu tiếng Anh này được đưa qua các khối : bộ phân tích hình thái, phân tích cây cú pháp, rút trích các quan hệ ngữ pháp xuất hiện trong các câu đó.

Các thông tin được chúng tôi sử dụng để tạo thành ngữ liệu vàng bao gồm:

- Các từ xung quanh từ đang xét

---

<sup>15</sup> Lý do để chúng tôi giao hai tập hợp ngữ nghĩa này đã được đề cập phía trên. Khi hai từ có mối liên kết với nhau, chúng sẽ cùng chia sẻ nhau một ý niệm của thế giới thực. Ý niệm này được tìm thấy qua việc giao hai tập ý niệm (mỗi tập ứng với mỗi từ - tiếng Anh, tiếng Việt).

---

- Các từ gốc (lemma) xung quanh từ đang xét
- Từ loại của các từ xung quanh từ đang xét
- Các từ có quan hệ ngữ pháp với từ đang xét
- Nhãn của các từ xung quanh từ đang xét
- Nhãn của các từ có quan hệ ngữ pháp với từ đang xét

❑ **Từ trong câu :**

Thông tin từ trong câu được lấy là những từ xung quanh từ đang xét theo một cửa sổ có độ lớn là [-KICHTHUOC,+KICHTHUOC]. Thông tin này được coi là ngữ cảnh của từ đang xét.

**Ví dụ 3-5:**

Với KICHTHUOC = 3, các từ cần quan tâm chính là các từ TU\_L3, TU\_L2, TU\_L1, TU\_0, TU\_R1, TU\_R2, TU\_R3

TU_L3	TU_L2	TU_L1	<b>TU_0</b>	TU_R1	TU_R2	TU_R3
Which	are	now	<b>becoming</b>	available	should	make

❑ **Từ gốc :**

Đây là kết quả lấy được từ bộ phân tích hình thái. Mỗi từ được đưa về dạng gốc đúng theo từ loại của nó.

**Ví dụ 3-6 :**

Với KICHTHUOC = 3, các từ gốc cần quan tâm chính là các từ gốc GOC\_L3, GOC\_L2, GOC\_L1, GOC\_0, GOC\_R1, GOC\_R2, GOC\_R3

GOC_L3	GOC_L2	GOC_L1	<b>GOC_0</b>	GOC_R1	GOC_R2	GOC_R3
which	be	now	<b>become</b>	available	should	make

❑ **Từ loại :**

Từ loại của các từ xung quanh từ loại đang xét với kích thước cửa sổ xem xét là [-KICHTHUOC, +KICHTHUOC]

---

**Ví dụ 3-7 :**

Với KICHTHUOC = 3, các từ loại cần quan tâm là POS\_L3, POS\_L2, POS\_L1, POS\_0, POS\_R1, POS\_R2, POS\_R3

POS_L3	POS_L2	POS_L1	<b>POS_0</b>	POS_R1	POS_R2	POS_R3
WDT	VBP	RB	<b>VBG</b>	JJ	MD	VB

□ **Từ có quan hệ ngữ pháp với từ đang xét, nhãn của từ có quan hệ ngữ pháp với từ đang xét :**

Ở đây, chúng tôi sử dụng các loại quan hệ (đã được đề cập trong phần 2.5.3. Các loại quan hệ trong bộ phân tích cú pháp dựa trên văn phạm phụ thuộc) như : subj, obj, obj2, mod, pred, pnomp-n.

Các thông tin được tạo lập nhờ các bộ phân tích cú pháp, rút trích quan hệ được tổ chức thành dạng cột. Mỗi hàng gồm đầy đủ các thông tin nêu trên dùng để khởi nhập nhằm cho một từ đứng ở vị trí trung tâm (TU\_0). Các thông tin này là thông tin được thực hiện tự động nên không đảm bảo là chính xác hoàn toàn. Để bảo đảm ngữ liệu đang xây dựng là ngữ liệu vàng, chúng tôi phải tiếp thêm phần hậu xử lý, hiệu chỉnh những thông tin chưa chính xác. Do tỷ lệ đúng của bộ phân tích cú pháp, gán nhãn từ loại, phân tích hình thái học, và rút trích quan hệ ngữ pháp là khá cao<sup>16</sup> nên công việc hậu xử lý cũng đỡ vất vả.

---

<sup>16</sup> Bộ phân tích cú pháp có độ chính xác khoảng 89%, phân tích từ loại, hình thái học có độ chính xác khoảng 98%, rút trích quan hệ ngữ pháp có độ chính xác khoảng 89%.

---

Khoa CNTT - ĐHQG KHTN TP.HCM

Chương 4

# CÀI ĐẶT THỬ NGHIỆM



*Chương 4 sẽ cụ thể hoá mô hình cài đặt qua việc đề cập chi tiết hơn về một cách gán nhãn cơ sở khá đặc biệt và công đoạn gán nghĩa tiếng Việt. Ở phần gán nhãn ngữ nghĩa cơ sở, chúng tôi nêu ra cách sử dụng tất cả các tri thức để hình thành nhãn mặc định cho từ. Chúng tôi cũng nêu những trường hợp thêm/bớt từ cho nghĩa tiếng Việt để đảm bảo đạt được một câu kết quả hoàn chỉnh trong công đoạn gán nghĩa tiếng Việt. Phần cuối của chương này là các kết quả thử nghiệm đạt được.*

## **4.1. GÁN NHÃN CƠ SỞ**

### **4.1.1. Mô hình gán nhãn cơ sở**

Gán nhãn cơ sở đóng một vai trò quan trọng trong phương pháp học dựa trên chuyển đổi. Tất cả các trường hợp trước khi thực hiện chuyển đổi đều phải có một nhãn. Gán nhãn cơ sở là nhằm gán một nhãn ban đầu cho các trường hợp đó. Trong xử lý ngữ nghĩa, gán nhãn cơ sở là nhằm gán một nhãn ngữ nghĩa ban đầu cho một từ cần phải khử nhập nhằng (xác định ngữ nghĩa). Nhãn ban đầu này được xem là nhãn mặc định cho từ đó. Gọi là nhãn mặc định bởi vì trong trường hợp không tìm ra một luật nào để sửa lỗi (chuyển đổi) thì nhãn ban đầu cũng chính là nhãn đầu ra sau khi áp dụng toàn bộ chuỗi luật chuyển đổi.

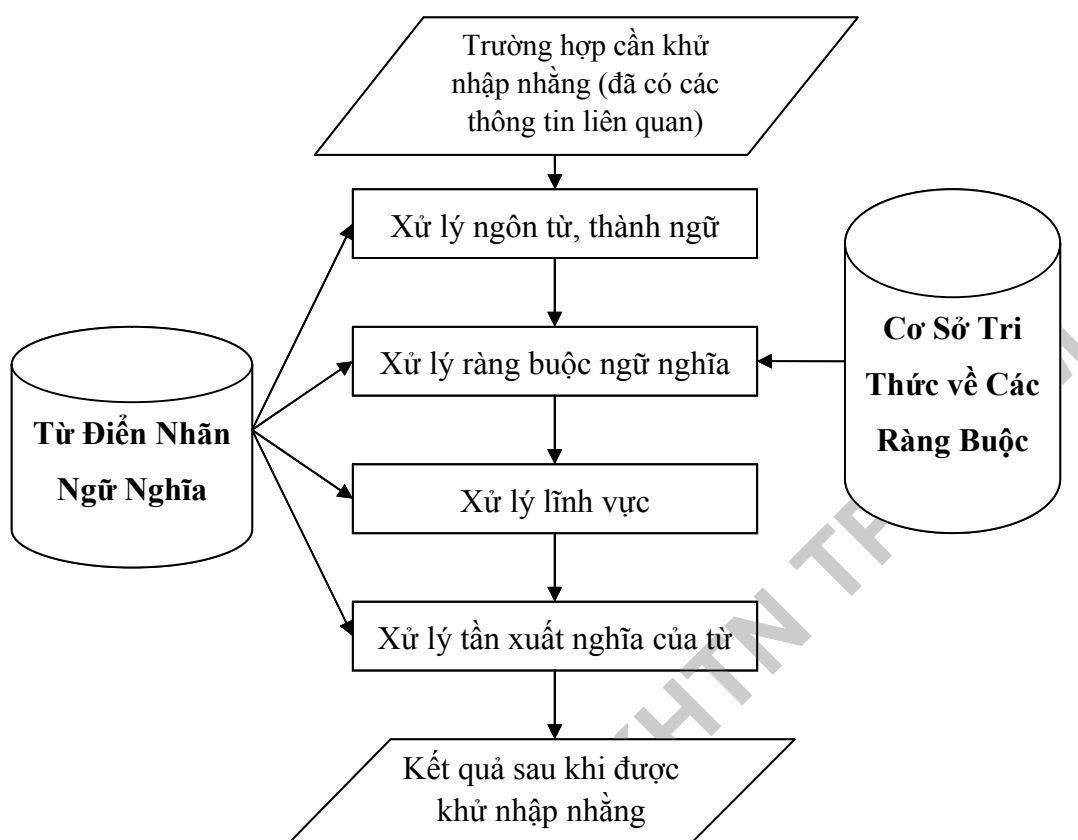
Trong phương pháp học dựa trên chuyển đổi, bộ phận thực hiện gán nhãn cơ sở là bộ phận được sử dụng chung cho cả quá trình huấn luyện lẫn quá trình áp dụng. Trong quá trình huấn luyện, sau khi ngữ liệu được gỡ bỏ nhãn liên quan đến ngữ nghĩa thì được đưa vào bộ phận gán nhãn cơ sở ; còn trong quá trình áp dụng, sau khi một trường hợp cần khử nhập nhằng được đưa vào (đã được xác lập các thông tin cần thiết), nó được đưa qua bộ phận gán nhãn cơ sở.

Có nhiều cách gán nhãn cơ sở khác nhau : (1) có thể gán một cách ngẫu ngô một nhãn bất kỳ làm nhãn cơ sở ; (2) gán mặc định tất cả các trường hợp bằng một nhãn bất kỳ ; (3) gán nhãn cơ sở dựa trên thống kê - nhãn nào có tần số xuất hiện cao nhất đối với một từ nào đó thì nhãn đó sẽ được gán cho từ đó khi gặp lại...

Các cách gán nhãn cơ sở nêu trên đây đều có thể đáp ứng một nhu cầu nào đó. Điều đó có nghĩa là sau khi thực hiện một trong các cách gán nhãn cơ sở nêu trên thì các bộ học dựa trên chuyển đổi đều tạo ra được một chuỗi các luật chuyển đổi, song kết quả áp dụng các dãy luật này có thể sẽ không bằng một cách gán nhãn cơ sở có phương pháp do các cách gán nhãn cơ sở nêu trên chưa sử dụng các thông tin hữu ích (đã được đưa vào ngay từ đầu). Và một cách rõ ràng là bộ huấn luyện sẽ phải làm việc nhiều hơn, lâu hơn, phát sinh nhiều luật sửa lỗi (chuyển đổi) hơn.

Như đã đề cập ở phần trên (phần 2.2. HỌC DỰA TRÊN CHUYỂN ĐỔI), một tính chất rất đặc biệt của phương pháp học này là khả năng sửa lỗi trên đầu ra của một bộ phận khác và kết quả sau khi áp dụng dãy luật từ phương pháp học này sẽ luôn có tỷ lệ chính xác lớn hơn hoặc bằng kết quả có được từ đầu ra của bộ phận trước đó. Nhận thấy tầm quan trọng của gán nhãn cơ sở và tính chất đặc biệt này của phương pháp học này, chúng tôi đã thực hiện một bộ phận gán nhãn ngữ nghĩa cơ sở từ các thông tin được đưa vào. Bộ gán nhãn cơ sở này tận dụng tất cả các thông tin để khử nhập nhằng ngữ nghĩa. Bộ huấn luyện sẽ nhận kết quả của bộ gán nhãn cơ sở này để điều chỉnh và nâng cao chất lượng cho khối xử lý ngữ nghĩa.

Trong mô hình thực hiện dưới đây chúng tôi không đề cập đến bộ lọc từ loại. Thật ra, có thể mặc nhiên hiểu rằng trước khi đưa vào mô hình theo miêu tả dưới đây, chúng tôi đã đưa các thông tin qua bộ lọc từ loại để loại bỏ những nhãn ngữ nghĩa không phù hợp với từ loại của từ trong ngữ cảnh (câu). Dưới đây là mô hình thực hiện.



**Hình 4-1 : Mô hình cho phương pháp gán nhãn cơ sở**

#### **4.1.2. Xử lý ngôn từ, thành ngữ**

Ngôn từ và thành ngữ thường đưa ra các nghĩa khá đặc biệt cho từ. Chẳng hạn, từ *make* trong cụm từ thành ngữ *make up Poss mind* (như *make up my mind*, *make up her mind*) có nghĩa rất đặc biệt, không thể chọn nhãn ngữ nghĩa tương đương với nghĩa *làm*, hay *trang điểm* (riêng cụm từ thành ngữ *make up* có nghĩa là *trang điểm*) mà phải chọn nhãn ngữ nghĩa tương đương với nghĩa *quyết định* của từ *make*.

Trong giai đoạn này, chúng tôi chỉ gán nhãn ngữ nghĩa cho những cụm từ thành ngữ nào có từ loại là danh từ, hay động từ. Những ngôn từ đặc biệt khác không nhập nhằng (chẳng hạn, những cụm từ như *from day to day* – ngày qua ngày, *from A to Z* – từ đầu đến cuối...), chỉ có tác dụng trong việc gán nghĩa tiếng Việt thích hợp sẽ được chuyển sang công đoạn thứ hai của khối này.

### 4.1.3. Xử lý ràng buộc lựa chọn

Đây là bộ phận có vai trò quyết định cho chất lượng của bộ gán nhãn cơ sở này. Tất cả các thông tin ngữ pháp rút trích được nhờ bộ phân tích cú pháp dựa trên văn phạm phụ thuộc được sử dụng để chọn nghĩa. Bộ phận này còn sử dụng thông tin từ cơ sở tri thức các ràng buộc ngữ nghĩa.

#### 4.1.3.1. Cơ sở tri thức

Trong việc xét điều kiện ràng buộc về ngữ nghĩa, chúng ta phải xét đến tính cấp bậc trong hệ thống nhãn ngữ nghĩa, trong đó khái niệm con sẽ kế thừa các nét nghĩa của khái niệm cha và có thêm nét nghĩa mới riêng của chúng. Thông tin đặc điểm ngữ nghĩa của từng từ thực cũng như các ràng buộc đã được xác định trong từ điển LDOCE hoặc là FrameNet.

Cơ sở tri thức về ràng buộc ngữ nghĩa có thể tổ chức cho từng nghĩa của từ. Chẳng hạn, đối với nghĩa *đi vào* của từ *enter* chủ thể của nó phải là người, tân ngữ của nó phải là một không gian kín (Clo-SPA), nên nó được lưu như sau *enter(subj:HUM, obj:clo-SPA)* hay với nghĩa *già* của từ *old* thì được lưu thành *old(modto:HUM)*.

#### 4.1.3.2. Thuật toán

##### □ Thuật toán xử lý ràng buộc lựa chọn

- B1 : Sắp xếp các quan hệ theo thứ tự ưu tiên (các quan hệ theo thứ tự như sau : s, subj, obj, obj2, mod, pcomp-n)
- B2 : Duyệt các quan hệ theo thứ tự
- B3 : Với mỗi quan hệ
  - Chọn các nhãn ngữ nghĩa làm cho quan hệ có điểm cao nhất
  - Gán các nhãn ngữ nghĩa này cho các từ.
- Lặp lại B3

##### □ Thuật toán xác định điểm cao nhất cho một quan hệ

- Vào : danh sách các nhãn ngữ nghĩa của hai từ có trong quan hệ

- Ra : nhãn ngữ nghĩa được chọn tương ứng với các từ được gán nhãn.
- **If** một trong hai từ trong quan hệ đã được gán nhãn ngữ nghĩa (gọi nhãn này là  $Sem_i$ ) **then** //(chúng ta sẽ lấy nhãn ngữ nghĩa này làm gốc để chọn các ngữ nghĩa cho từ còn lại).
  - $MaxPoint = 0$  //điểm cao nhất
  - Lặp, với mỗi nhãn ngữ nghĩa  $S_{jk}$  của từ  $Word_j$  chưa được gán nhãn
    - $Point = PointScore(Sem_i, S_{jk})$
    - **If** ( $Point > MaxPoint$ ) **then**
      - $MaxPoint = Point$
      - $Sem_j = S_{jk}$
  - **If** ( $MaxPoint > 0$ ) **then return**  $Sem_i, Sem_j$
- **else** //cả hai từ ( $Word_i, Word_j$ ) chưa được gán nhãn ngữ nghĩa
  - $MaxPoint = 0$
  - Lặp Với mỗi nhãn  $Sem_{ik}$  của từ  $Word_i$ , với mỗi nhãn  $Sem_{jl}$ 
    - $Point = PointScore(Sem_{ik}, Sem_{jl})$
    - **If** ( $Point > MaxPoint$ ) **then**
      - $MaxPoint = Point$
      - $Sem_j = S_{jl}$
      - $Sem_i = Sem_{ik}$
  - **If** ( $MaxPoint > 0$ ) **then return**  $Sem_i, Sem_j$
- **Thuật toán xác định điểm quan hệ giữa các nhãn ngữ nghĩa**
- Vào : hai nhãn ngữ nghĩa cần tính điểm (có nhãn  $Sem_i$  phụ thuộc vào nhãn  $Sem_j$  do trên văn phạm phụ thuộc  $Word_i$  phụ thuộc vào  $Word_j$ ).
- Ra : điểm quan hệ phụ thuộc giữa hai nhãn ngữ nghĩa

- B1 : đọc từ cơ sở tri thức các đặc điểm ngữ nghĩa, các ràng buộc ngữ nghĩa liên quan đến các nhãn ngữ nghĩa.
- B2 : Point = 0;
- B3 : Lặp, với ràng buộc ngữ nghĩa  $C_{jl}$  của  $Sem_j$ 
  - Lặp, với đặc điểm ngữ nghĩa  $S_{ik}$  của  $Sem_i$ 
    - **If** ( $S_{ik}$  Thỏa  $C_{jl}$ ) **then**
      - Point += FIX\_POINT
    - **else**
    - **If** ( $\exists$  một cha của  $S_{ik}$  Thỏa  $C_{jl}$ ) **then**
      - Point += NOR\_POINT
- B4 : Lặp, với đặc điểm ngữ nghĩa  $S_{jl}$  của  $Sem_j$ 
  - Lặp, với đặc điểm ngữ nghĩa  $S_{ik}$  của  $Sem_i$ 
    - **If** ( $S_{jl} = S_{ik}$ ) **then**
      - Point += FIX\_POINT
    - **else**
    - **If** ( $S_{ik}$  nằm trên cùng một nhánh cây ý niệm với  $S_{jl}$ ) **then**
      - Point += NOR\_POINT
- B5 : **return** Point

(Lưu ý, FIX\_POINT > NOR\_POINT)

#### 4.1.4. *Xử lý dựa trên lĩnh vực xem xét*

Sau khi qua tầng xử lý ràng buộc ngữ nghĩa dựa trên các quan hệ cú pháp, một số từ chưa được gán nhãn ngữ nghĩa. Lúc này chúng tôi sẽ xem xét đến lĩnh vực (mà văn bản đó thuộc về) để chọn một nhãn ngữ nghĩa thích hợp gán cho từ đó. Trong luận văn này chúng tôi ưu tiên xử lý các văn bản là các tài liệu tin học nên lĩnh vực được ưu tiên xem xét theo thứ tự là CPT (tin học), TECH (kỹ thuật), SCI

(khoa học). Thật ra để giải quyết tốt trường hợp lĩnh vực xem xét này đáng lẽ ra phải kèm theo một bộ nhận dạng lĩnh vực<sup>17</sup>, song nếu chỉ sử dụng ngữ cảnh xung quanh của từ cần xem xét để xác định đó là lĩnh vực gì thì bao nhiêu thông tin đó là không đủ. Chính vì vậy, chúng tôi giả thiết rằng đa số sẽ là các từ sẽ chọn nghĩa nằm trong lĩnh vực tin học, tỷ lệ các từ không phải trong lĩnh vực tin học là không đáng kể.

#### **4.1.5. Xử lý dựa trên tần số xuất hiện**

Tầng xử lý cuối cùng này có vai trò không quan trọng lắm. Cách thực hiện của tầng này giống một phương pháp gán nhãn cơ sở ngây ngô thông thường. Nó sẽ xem xét những từ nào chưa được gán nhãn mà từ đó sẽ chọn nhãn ngữ nghĩa của từ đó có tần số xuất hiện cao nhất.

## **4.2. MẪU LUẬT**

Trong phương pháp học dựa trên chuyển đổi, để hình thành được các luật chuyển đổi, chúng ta cần phải thiết lập các mẫu luật. Các mẫu luật chính là giới hạn của luật. Chúng là cơ sở của luật. Luật chuyển đổi có dạng nào, có những yếu tố nào, sử dụng các thông tin gì, tất cả đều được mẫu luật quyết định. Xây dựng mẫu luật không dễ bởi vì :

- Đưa quá nhiều mẫu luật sẽ làm cho bộ huấn luyện huấn luyện lâu do phát sinh quá nhiều luật thoả mẫu luật .
- Đưa quá ít mẫu luật dẫn đến không đủ mẫu luật để phát sinh luật chuyển đổi, điều này sẽ nguy hiểm vì rằng thiếu luật chuyển đổi nên một số trường hợp sai không được sửa thành đúng
- Phải chọn lựa thông tin thích hợp để đưa vào mẫu luật.

Chúng tôi đã chọn lựa và thiết lập các mẫu luật như sau

---

<sup>17</sup> Vì rằng trong các tài liệu tin học không có gì đảm bảo là chỉ toàn nói về tin học, không có nói thêm về các lĩnh vực khác như dịch vụ, thương mại chẳng hạn.

---

#### 4.2.1. Các từ trong ngữ cảnh

Các mẫu luật thuộc nhóm này có dạng là :  $\bigcup_{i \in [-m, +n]} TU\_i$  ( $m, n \in \mathbb{N}$ )

Một số mẫu luật trong trường hợp này được nêu dưới đây.

$TU\_0 \Rightarrow \text{Sense}$

$\$TU[-3,-1] TU\_0 \$TU[1,3] \Rightarrow \text{Sense}$

$TU\_0 \text{ Sense} \Rightarrow \text{Sense}$

$TU\_0 \$TU[-3,-1] \text{ Sense} \Rightarrow \text{Sense}$

#### 4.2.2. Từ gốc trong ngữ cảnh

Các mẫu luật thuộc nhóm này có dạng là :  $\bigcup_{i \in [-m, +n]} GOC\_i$  ( $m, n \in \mathbb{N}$ )

Một số mẫu luật trong trường hợp này được nêu dưới đây.

$GOC\_0 \text{ Sense} \Rightarrow \text{Sense}$

$GOC\_0 \$GOC[-3,-1] \text{ Sense} \Rightarrow \text{Sense}$

$GOC\_0 \$GOC[1,3] \text{ Sense} \Rightarrow \text{Sense}$

$\$GOC[-3,-1] GOC\_0 \$GOC[1,3] \text{ Sense} \Rightarrow \text{Sense}$

#### 4.2.3. Từ loại trong ngữ cảnh

$POS\_0 \text{ Sense} \Rightarrow \text{Sense}$

$POS\_0 \$POS[-3,-1] \text{ Sense} \Rightarrow \text{Sense}$

$POS\_0 \$POS[1,3] \text{ Sense} \Rightarrow \text{Sense}$

$\$POS[-3,-1] POS\_0 \$POS[1,3] \text{ Sense} \Rightarrow \text{Sense}$

#### 4.2.4. Nhân ngữ nghĩa trong ngữ cảnh

$\$SEM[-3,-1] \text{ Sense} \Rightarrow \text{Sense}$

$\$SEM[1,3] \text{ Sense} \Rightarrow \text{Sense}$

$\$SEM[-3,-1] \$SEM[1,3] \text{ Sense} \Rightarrow \text{Sense}$



#### **4.2.5. Từ có quan hệ ngữ pháp trong ngữ cảnh**

TU\_SUBJ Sense => Sense

TU\_OBJ Sense => Sense

TU\_MOD Sense => Sense

TU\_OBJ2 Sense => Sense

TU\_PRED Sense => Sense

TU\_SUBJ TU\_OBJ TU\_OBJ2 Sense => Sense

TU\_SUBJ TU\_OBJ Sense => Sense

#### **4.2.6. Các nhãn trong ngữ cảnh có quan hệ ngữ pháp**

SEM\_SUBJ Sense => Sense

SEM\_OBJ Sense => Sense

SEM\_MOD Sense => Sense

SEM\_OBJ2 Sense => Sense

SEM\_PRED Sense => Sense

SEM\_SUBJ SEM\_OBJ SEM\_OBJ2 Sense => Sense

SEM\_SUBJ SEM\_OBJ Sense => Sense

### **4.3. GẮN NGHĨA TIẾNG VIỆT**

Sau đây chúng tôi sẽ đề cập đến những bước thực hiện để xây dựng một câu tiếng Việt hoàn chỉnh, tức là phải thêm những hư từ, lượng từ... để có cách dịch của câu tiếng Việt *nghe giống tiếng Việt hơn*. Một số trường hợp chúng tôi đã áp dụng thuật toán fnTBL để giải quyết.

#### 4.3.1. Các từ không cần gắn nghĩa tiếng Việt

Chắc chắn rằng không phải từ nào cũng được gắn nghĩa tiếng Việt. Một số từ mặc dù có nhãn ngữ nghĩa nhưng không thể nào có nghĩa tiếng Việt. Đó là trường hợp của các tên riêng. Các tên riêng có nghĩa tiếng Việt chính là nó<sup>18</sup>. Trong bộ phân tích cú pháp, nhãn tên riêng được gán bằng hai nhãn cú pháp NNP và NNPS. Tuy nhiên không phải lúc nào quan điểm về tên riêng cũng hợp lý cho cả tiếng Việt lẫn tiếng Anh. Ví dụ như, trong tiếng Anh, các từ là ngày, tháng (như *Monday, Sunday, Tuesday,...*, *January, February,...*), các từ là tên quốc gia (*England, France,...*) được xem là các tên riêng ; song khi các từ này khi được đưa qua tiếng Việt thì không thể nào không dịch. Lại có những trường hợp không phải là tên riêng<sup>19</sup> (không được viết hoa chữ đầu tiên) nhưng vẫn được gán nhãn cú pháp NNP, hay NNPS.

Do đó, chúng tôi đã phải giải quyết bằng heuristic sau :

- Các từ có nhãn cú pháp NNP, NNPS nhưng không được viết hoa ký tự đầu tiên thì sẽ được dịch.
- Các từ có nhãn cú pháp NNP, NNPS nhưng chỉ quốc gia, ngày tháng sẽ được dịch.
- Các trường hợp có nhãn cú pháp NNP, NNPS còn lại đều không được dịch.

---

<sup>18</sup> Ví dụ, trong câu *I book two books from Mr. Book*, từ *Book* cuối cùng không được dịch mà phải giữ nguyên.

<sup>19</sup> Theo quy ước, các tên riêng phải được viết hoa ít nhất chữ cái đầu tiên.

---

### 4.3.2. Gắn thêm lượng từ Những

#### 4.3.2.1. Mô tả

Trong tiếng Anh, các danh từ số nhiều có thể được nhận biết bằng các tiếp vĩ ngữ (suffix) -s, -es. Còn trong tiếng Việt, các từ *những*, *các*<sup>20</sup> ... thường được gắn thêm vào các danh từ để chỉ số nhiều. Việc gắn các từ (*những*, *các*) này vào trong cách dịch các danh từ số nhiều tiếng Anh là không thể tùy tiện. Điều đó có nghĩa là chúng ta không thể cứ gộp một danh từ số nhiều của tiếng Anh (được gắn nhãn cú pháp là NNS) là thêm các từ này (*những*, *các*...) vào.

#### Ví dụ 4-1 :

- *three pupils/NNS* không thể dịch là *ba những học sinh*.
- *several new pens/NNS* không thể dịch là *vài những cây viết mới*.
- *highspeed data/NNS lines/NNS* không thể dịch là *những dòng những dữ liệu tốc độ cao*.
- *hundreds/NNS of cars/NNS* không thể dịch là *những trăm của những xe hơi*.
- *several categories/NNS of application programs/NNS* không thể dịch là *những nhiều những loại của những chương trình ứng dụng*.

#### Ví dụ 4-2 :

- *All these pupils/NNS* có thể được dịch là *Tất cả những học sinh này*.
- *small businesses/NNS and individual users/NNS* có thể được dịch là *những doanh nghiệp nhỏ và những người dùng cá nhân*.
- *various traditional design skills/NNS* có thể được dịch là *những kỹ năng thiết kế truyền thống khác nhau*.

---

<sup>20</sup> Trong luận văn này, chúng tôi chỉ xem xét việc gắn thêm từ *những* vào danh từ số nhiều mà thôi (các từ khác có nghĩa tương tự).

---

Chúng tôi đã tiến hành xây dựng ngữ liệu huấn luyện, áp dụng thuật toán fnTBL để xác định những danh từ số nhiều nào (NNS) cần phải thêm từ *những*, những từ không thể thêm.

#### 4.3.2.2. Ngữ liệu và mẫu luật

Nhận thấy rằng việc thêm từ *những* cho những danh từ số nhiều phụ thuộc vào ngữ cảnh xung quanh từ đó, và phần lớn là các từ loại của ngữ cảnh đó, không phụ thuộc nhiều vào các quan hệ ngữ pháp, chúng tôi tiến hành xây dựng ngữ liệu huấn luyện ở mức gọn nhẹ (chỉ sử dụng thông tin từ loại, từ trong câu, và chỉ với 3 nhãn<sup>21</sup> mà thôi).

Ngữ liệu huấn luyện gồm những câu được trích từ kho ngữ liệu VCLEVC. Chúng tôi đã để bộ gán nhãn từ loại phân tích hình thái học, và từ loại cho chúng. Nhãn cơ sở trong trường hợp này là EMPTY cho các từ có từ loại không phải NNS, NHUNG cho các từ có từ loại NNS. Sau đó, ngữ liệu huấn luyện này được người chỉnh lại bằng tay để có thể là ngữ liệu vàng.

Mẫu luật được tạo chỉ chứa các thông tin từ loại và hình thái.

Mẫu Luật	Diễn Giải
$\$TU_{[-5,-1]}=X \Rightarrow TAG$	Một trong 5 từ đứng trước từ cần xét có nhãn <b>X</b> .
$\$TU_{[1,5]}=X \Rightarrow TAG$	Một trong 5 từ đứng sau từ cần xét có nhãn <b>X</b> .
$\$TU_{[-5,-1]}=X \ \$TU_{[1,5]}=Y \Rightarrow TAG$	Một trong 5 từ đứng trước có nhãn là <b>X</b> và một trong 5 từ đứng sau có nhãn là <b>Y</b> .
$\$POS_{[-5,-1]}=X \ POS0=Y \Rightarrow TAG$	Một trong 5 từ đứng trước có từ loại là <b>X</b> .

<sup>21</sup> Đó là các nhãn NHUNG (trường hợp gắn thêm *những* vào danh từ số nhiều thì hợp lý), KHONG (gán từ *những* vào danh từ số nhiều thì không hợp lý), EMPTY (gán cho các từ không phải danh từ số nhiều).

	và từ loại của từ đang xét là <b>Y</b>
\$POS_[1,5]=X POS0=Y => TAG	Một trong 5 từ đứng sau có từ loại là <b>X</b> và từ loại của từ đang xét có từ loại là <b>Y</b>

**Bảng 4-1 : Trích mẫu luật để thêm từ *những***

Sau khi thực hiện việc huấn luyện trên ngữ liệu cùng với tập mẫu luật được xây dựng riêng để giải quyết trường hợp này. Chúng tôi đã rút ra được một số luật để áp dụng. Chúng tôi đã thực hiện tối ưu luật và chỉ sử dụng các luật chuyển từ nhãn NHUNG sang nhãn KHONG, nghĩa là sử dụng các luật này chỉ để xác định những vị trí danh từ được gán nhãn từ loại là NNS (danh từ số nhiều) nhưng không được thêm lượng từ *những* vào (xem thêm Phụ Lục 3 trang 108).

#### 4.3.3. Quan hệ giữa động từ “to be” và các trường hợp khác

Động từ *to be* là một động từ đặc biệt. Nó có thể có quan hệ với danh từ, tính từ, các động từ khác.

Ví dụ 4-3 :

He                    is                    a                    teacher

Ví dụ 4-4 :

She                    is                    beautiful

Ví dụ 4-5 :


They                    are                    reading                    books

Ví dụ 4-6 :

He                    is                    hated                    by                    his                    friends

**Ví dụ 4-7:**

These books are written by me



Trong Ví dụ 4-3 và Ví dụ 4-4, *be* có quan hệ với danh từ *teacher* và tính từ *beautiful*. Quan hệ này thông qua văn phạm phụ thuộc có tên là *pred*. Do đó, dựa trên quan hệ này, chúng ta sẽ xác định các dịch của động từ *be* như sau : (1) nếu từ phụ thuộc nó là danh từ thì *be* được gắn nghĩa là ; (2) nếu từ phụ thuộc nó là tính từ thì *be* được gắn nghĩa thì.

Trong Ví dụ 4-5, Ví dụ 4-6, và Ví dụ 4-7, động từ *be* có quan hệ với các động từ. Trong văn phạm phụ thuộc, quan hệ này có chung tên là *be*. Tuy nhiên, trong trường hợp của Ví dụ 4-5, *be* và động từ *read* tạo thành thì tiếp diễn (continuous tense) nên từ *read* cần phải được thêm từ *đang* để chỉ sự tiếp diễn. Còn trong Ví dụ 4-6 và Ví dụ 4-7, *be* và động từ (dạng động từ quá khứ phân từ *hated, written*) tạo thành dạng bị động (passive voice) song đối với động từ *hate* thì *be* được dịch kèm là *bị*, còn với động từ *writte* thì *be* được dịch kèm là *được*. Đây là một trường hợp khác biệt giữa tiếng Anh và tiếng Việt (như đã nêu trong phần 1.2.3.4. *Sự khác biệt giữa tiếng Anh và Việt*). Sự phân biệt *bị* hay *được* đối với động từ trong thể bị động có được nhờ vào đặc tính của động từ (*xấu* hay *tốt*).

Do vậy, chúng tôi đưa ra cách giải quyết cho mỗi quan hệ giữa động từ *be* và một số trường hợp như sau :

Quan hệ	Từ loại của từ phụ thuộc	Đặc tính	Dịch kèm
pred	Danh từ <sup>22</sup>		là
pred	Tính từ		thì

---

<sup>22</sup> Có nhiều nhãn từ loại danh từ : NNS, NN, NNP,... Xem thêm trong Phụ Lục 2 trang 106.

be	Động từ VBG		đang
be	Động từ VBN	Tốt	được
be	Động từ VBN	Xấu	bị

**Bảng 4-2 : Tóm tắt một số trường hợp giải quyết cho động từ *be***

#### **4.3.4. Các trường hợp đi kèm với giới từ**

Trong tiếng Anh, các giới từ (preposition) rất khó sử dụng, muốn sử dụng tốt giới từ thông thường thì phải nhớ. Chúng ta không thể viết là *on July* mà phải viết là *in July*, cũng không thể viết là *in Sunday* mà phải viết là *on Sunday*.

Còn khi dịch sang tiếng Việt, các giới từ không dễ dịch tí nào. Chỉ với giới từ *in* thôi, trong cụm từ *in English*, thì *in* được là *bằng* (như trong câu *I write this speech in English*), hay như với *in Vietnam* thì *in* được dịch là *ở* (như trong câu *There are a lot of beautiful places in Vietnam*), còn trong *in December* thì *in* được dịch là *vào* (như trong câu *SEA Games 22 will be organized in December 2003*).

Giới từ là một loại hư từ, không được gán nhãn ngữ nghĩa nên muốn chọn cách dịch hợp lý cho các giới từ thường phải dựa vào các kinh nghiệm được các nhà ngôn ngữ học xác định. Các tri thức này được chúng tôi tổ chức dưới dạng các ràng buộc ngữ nghĩa. Các ràng buộc đặc biệt này tồn tại giữa giới từ và danh từ (trong ngữ giới từ, ràng buộc này có giữa giới từ và danh từ chính của ngữ). Trong khi gán nghĩa cho giới từ, chúng tôi còn phải xác định một số từ<sup>23</sup> để thêm vào danh từ để cho câu dễ hiểu hơn.

---

<sup>23</sup> Ví dụ, với câu *SEA Games 22 will be organized in 2003*, chúng tôi phải xác định được con số *2003* ở đây là chỉ số năm để có thể dịch là *SEA Games 22 sẽ được tổ chức vào năm 2003*.

---

Dưới đây là một số tri thức được áp dụng trong dịch máy

Giới từ	Từ phụ thuộc có nhân thuộc chỉ	Được dịch	Ví dụ
at	thời gian	lúc	at 6:00
at	địa điểm	tại	at home
for	thời gian	trong	for 5 years
in	ngày tháng, thời gian	vào	in November, in 1999
in	ngôn ngữ	bằng	in English
in	địa điểm	ở	in Vietnam
on	ngày tháng	vào	on Sunday

**Bảng 4-3 : Một số tri thức được áp dụng để giải quyết giới từ**

#### 4.3.5. Các trường hợp liên quan đến thành ngữ

Nói đến thành ngữ chúng ta nghĩ ngay đến những cụm từ có nghĩa không thể nào xác định bằng cách gán các nghĩa của các từ trong nhóm lại với nhau. Những thành ngữ như thế cần phải dùng một từ điển đặc biệt. Từ điển này liệt kê các thành ngữ với nghĩa của nó. Khi đó trong lúc gán nghĩa tiếng Việt cho một câu, chúng tôi sẽ xem xét so khớp những cụm từ với từ điển để xác định nghĩa cho các thành ngữ. Tuy nhiên không có từ điển nào có đầy đủ các thành ngữ. Chúng tôi nhận thấy rằng:

- Với thành ngữ cực kỳ đặc biệt như trong *It is raining cats and dogs*<sup>24</sup> hay trong *to be or not to be*<sup>25</sup>, chắc chắn chúng ta không thể nào dịch được các cụm từ đó nếu không sử dụng từ điển.
- Song cũng có những trường hợp như trong *keep an eye on something*<sup>26</sup>, đây cũng là một thành ngữ, nhưng nó có thể hoàn toàn

---

<sup>24</sup> *It is raining cats and dogs* là Trời đang mưa như trút.

<sup>25</sup> Câu nói nổi tiếng của Hamlet “sống hay là chết”.

<sup>26</sup> *keep an eye on* = để mắt vào

---



hiểu được nếu gán nghĩa theo cách thông thường, tức là từ nhãn ngữ nghĩa suy ra nghĩa tiếng Việt và ghép nối lại.

Từ đó, chúng tôi đã xác định một giải pháp giải quyết các trường hợp này. Đối với những cụm từ quá đặc biệt, chúng tôi phải lưu nó trong từ điển; còn đối với những từ có thể ghép nối các nghĩa trong cụm từ thành ngữ thì chúng tôi không lưu vào trong từ điển. Cách thực hiện này có các ưu điểm như sau : (1) từ điển sẽ nhỏ hơn ; (2) trong trường hợp các thành ngữ xuất hiện trong văn bản có một ít thay đổi, nếu chỉ so khớp thì sẽ không xác định đúng nghĩa, ngược lại sử dụng cách này, nghĩa xuất hiện trong câu có thể không hay lắm (so với nghĩa trong từ điển) nhưng hoàn toàn có thể hiểu được !

#### 4.4. KẾT QUẢ THỰC HIỆN

##### 4.4.1. Dãy luật tối ưu

Chúng tôi đã thực hiện huấn luyện trên ngữ liệu huấn luyện và tập mẫu luật đã được xây dựng. Sau khi thực hiện, chúng tôi đã rút ra được 2537 luật chuyển đổi với ngưỡng là 2. Dưới đây chúng tôi xin trích một vài luật được rút ra (xem thêm trong Phụ Lục 3 trang 108).

\$SEM[-3,-1]=EMPTY Sense=DONTKNOW => Sense=PHM
\$POS[1,3]=NN POS0=VB \$POS[-3,-1]=NN => Sense=vsta
\$SEM[-3,-1]=EMPTY \$POS[1,3]=JJ POS0=VB \$POS[-3,-1]=NN Sense=PHM => Sense=vsta
\$SEM[-3,-1]=EMPTY \$POS[1,3]=PUNC POS0=NNP \$POS[-3,-1]=IN Sense=PHM => Sense=HUM
\$SEM[-3,-1]=DEV Sense=PHM => Sense=DEV
\$POS[1,3]=PUNC POS0=NNP \$POS[-3,-1]=PUNC => Sense=HUM

\$SEM[1,3]=DEV \$POS[-3,-1]=DT POS0=NN \$POS[1,3]=NN Sense=PHM =>  
 Sense=DEV

**Bảng 4-4 : Kết quả một số luật chuyển đổi trong xử lý ngữ nghĩa**

**4.4.2. Dãy luật rút ra để giải quyết việc thêm từ trong tiếng Việt**

Sau khi huấn luyện, chúng tôi thu được 135 luật chuyển đổi với ngưỡng là 0.75. Dưới đây là một số luật chuyển đổi đó :

POS_0=NNS => SENSE=NHUNG
TU_0=data => SENSE=KHONG
TU_[-5,-1]=of => SENSE=KHONG
POS_[-5,-1]=CD POS_0=NNS => SENSE=KHONG
TU_0=people => SENSE=KHONG
\$TU_[-5,-1]=or \$TU_[1,5]=. => SENSE=KHONG

**Bảng 4-5 : Kết quả một số luật chuyển đổi dùng để thêm từ tiếng Việt**

**4.4.3. Thử nghiệm**

Chúng tôi tiến hành thử nghiệm qua hai bước : (1) thử nghiệm bộ gán nhãn cơ sở ; (2) thử nghiệm với dãy luật được rút ra. Kết quả thử nghiệm được tiến hành với tập thử nghiệm gồm 1500 mẫu, được trích trong VCLEVC. Các mẫu này đã không được sử dụng cho huấn luyện.

Dưới đây là bảng kết quả :

Kiểu thử nghiệm	Số trường hợp		Tỷ lệ	
	Đúng	Sai	Đúng	Sai
Gán nhãn cơ sở	1178	322	78,53%	21,47 %
Kết hợp với luật	1216	284	81,07%	18,93%

**Bảng 4-6 : Kết quả thử nghiệm**

CÀI ĐẶT THỬ NGHIỆM

Bảng kết quả thử nghiệm trên đây mang tính chất tham khảo cho tính đúng đắn của mô hình. Kết quả thực tế hơn của hệ dịch tự động đó là những câu dịch tiếng Việt đã được hoàn chỉnh phần gán nghĩa. Dưới đây là kết quả đã được gán tiếng Việt (Các câu dịch hoàn chỉnh, sau khi chuyển đổi có thể xem thêm trong Phụ Lục 4 trang 111).

I can can a can											
I	<i>can</i>	<i>can</i>	a	<i>can</i>							
Tôi	có thể	đóng hộp	một	cái hộp							
I want to book two books											
I	want	to	<i>book</i>	two	<i>books</i>						
Tôi	muốn	để	đặt trước	hai	cuốn sách						
An old man is reading an old book											
An	<i>old</i>	man	<i>is</i>	reading	an	<i>old</i>	book				
một	già	người đàn ông		đang đọc	một	cũ	cuốn sách				
He is not only handsome but also intelligent.											
He	<i>is</i>	not only	handsome	but also	intelligent	.					
anh ấy	(thì)	không những	đẹp trai	mà còn	thông minh	.					
He is at his home alone and he is reading a book.											
He	<i>is</i>	<i>at</i>	his	home	alone	and	he	is reading	a	book	.
anh ấy	(thì)	tại	của anh ấy	nhà	một mình	và	anh ấy	đang đọc	một	cuốn sách	.

CÀI ĐẶT THỬ NGHIỆM

He was scolded by his wife but he will be appreciated by his lover					
He	<i>was scolded</i>	by	his	wife	but
anh ấy	bị mắng	bởi	của anh ấy	người vợ	nhưng
he	will	<i>be appreciated</i>	by	his	lover
anh ấy	sẽ	được đánh giá cao	bởi	của anh ấy	người yêu
He can program many subtle programs with new and interesting features.					
He	can	<i>program</i>	many	subtle	<i>programs</i>
anh ấy	có thể	lập trình	nhiều	tinh vi	chương trình
with	new	and	interesting	features	.
với	mới	và	hay	những đặc tính	.
A very old man has a very old computer system.					
A	very	<i>old</i>	man	has	a very <i>old</i> computer system .
một	rất	già	người đàn ông	có	một rất cũ máy tính hệ thống .
She will make up her mind about this matter soon.					
She	will	make up her mind	about	this	matter soon .
cô ấy	sẽ	quyết định	về	này	vấn đề sớm .
That book is more expensive than this book.					
That	book	is	more expensive	than	this book .
đó	cuốn sách	(thì)	mắc hơn	so với	này cuốn sách .
I amn't as intelligent as him.					
I	am	n't	as intelligent as	him	.
tôi	thì/là	không	thông minh bằng	anh ấy	.

She is not so ugly as her sister.							
She	is	not so ugly as	her	sister	.		
cô ấy	(thì)	không xấu như	của cô ấy	chị/em gái	.		
This is the most important project.							
This	is	the most important	project	.			
đây	là	quan trọng nhất	dự án	.			
computer's action is difficult.							
computer	's	action	is	difficult	.		
máy tính	của	hoạt động	(thì)	khó	.		
Normally, I always normalize normal problems.							
Normally	,	I	always	normalize	normal	problems	.
bình thường	,	tôi	luôn luôn	bình thường hóa	bình thường	những vấn đề	.
She isn't singer and mother is not either.							
She	is	n't	singer	and mother is not either	.		
cô ấy	là	không	ca sĩ	và mẹ cũng không	.		
Does a tall girl eat a ripe apple?							
Does	a	tall	girl	eat	a	ripe	apple ?
	một	cao	cô gái	ăn	một	chín	quả táo ?
Had they visited Paris in 1987?							
Had	they	visited	Paris	<i>in</i>	<i>1987</i>	?	
	chúng	thăm	Paris	vào	năm 1987	?	

CÀI ĐẶT THỬ NGHIỆM

Why do you say that matter ?											
Why	do	you	say	<i>that</i>	matter			?			
tại sao		bạn	nói	đó	vấn đề			?			
The man cutting woods is my father.											
The	man	cutting	woods	is	my	father					
	người đàn ông	đang cắt	những gỗ	là	của tôi	bố					
The man scolded by his wife went to my house.											
The	man	<i>scolded</i>	by	his	wife	went	to	my	house	.	
	người đàn ông	bị mắng	bởi	của anh ấy	người vợ	đi	tới	của tôi	ngôi nhà	.	
I send her the vase that you put in the box.											
I	send	her	the	vase	<i>that</i>	you	put	in	the	box	.
tôi	gởi	cô ấy		lọ hoa	mà	bạn	đặt	trong		cái hộp	.
He ought to help you.											
He	ought to			help	you						
anh ấy	phải			giúp	bạn						
We have to do our exercises tonight.											
We	have to	do	our	exercises	tonight						
chúng tôi	phải	làm	của chúng tôi	những bài tập	đêm nay						
She reminds him that the meeting is on saturday.											
She	reminds	him	<i>that</i>	the	meeting	is	<i>on</i>	saturday			
cô ấy	nhắc nhở	anh ấy	rằng		cuộc họp	(thì)	vào	thứ bảy			

Chương 5

**KẾT LUẬN – HƯỚNG PHÁT TRIỂN**

Khoa CNTT - ĐH KHTN TP.HCM

*Chương cuối cùng được chúng tôi dùng để tổng kết lại toàn bộ luận văn, và đề nghị hướng phát triển cho luận văn.*

## **5.1. HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN**

Dịch máy là bài toán đã được đặt ra từ hơn 50 năm nay và cũng là bài toán cực kỳ khó do tính nhập nhằng vốn có của ngôn ngữ tự nhiên. Từ đó đến nay, đã có nhiều chiến lược, cách tiếp cận mô hình để giải quyết như chưa thể trọn vẹn được. Do ý nghĩa thực tiễn của nó quá lớn nên con người vẫn không ngừng cải tiến để nâng cao chất lượng dịch của nó.

Qua luận văn này, mô hình học luật chuyên đổi từ ngữ liệu song ngữ cho hệ dịch tự động Anh-Việt đã được kiểm chứng ở phần xử lý ngữ nghĩa. Mô hình tỏ ra rất đúng đắn thông qua việc sử dụng ngữ liệu song ngữ đã được liên kết từ. Các mối liên kết từ có được không chỉ giúp thu nhận gấp đôi tri thức từ hai ngôn ngữ mà thu được hơn thế nhiều lần. Trong bài toán xử lý ngữ nghĩa, nhờ vào ngữ liệu song ngữ được liên kết từ mà việc xây dựng ngữ liệu huấn luyện ít vất vả hơn, thời gian xây dựng nhanh hơn, và đảm bảo chính xác hơn so với việc chỉ sử dụng người xây dựng.

Với những kết quả đã được thử nghiệm, phần xử lý ngữ nghĩa cho hệ dịch tự động Anh-Việt bước đầu đã hoạt động hiệu quả, góp phần phục vụ nhu cầu *thấy là hiểu* của nhiều người. Mặc dù những kết quả thử nghiệm là tốt nhưng chúng tôi hiểu rằng vấn đề mà luận văn này giải quyết là một vấn đề vô cùng khó khăn trong lĩnh vực ngôn ngữ học cũng như trong lĩnh vực áp dụng trí tuệ nhân tạo. Dưới đây là những hạn chế và cũng là những hướng mở mà luận văn phải phát triển tiếp tục trong thời gian tới để hoàn thiện hơn nữa khối xử lý ngữ nghĩa nói riêng và cho hệ dịch máy Anh-Việt nói chung.

Thứ nhất, chúng tôi muốn đề cập đến việc mở rộng lĩnh vực giải quyết cho hệ dịch tự động. Trong thời đại giao tiếp quốc tế rộng mở như hiện nay, rõ ràng là rất cần nhiều hệ dịch tự động cho các lĩnh vực khác nhau. Luận văn này chỉ mới giải quyết các nhập nhằng ngữ nghĩa cho các tài liệu tin học. Song như đã đề cập, việc



mở rộng lĩnh vực xem xét sẽ không thay đổi, ảnh hưởng nhiều đến mô hình. Khó khăn chủ yếu là nguồn ngữ liệu phục vụ tiếp cho các lĩnh vực cần mở rộng. Nhưng chúng tôi tin rằng nhờ các công cụ sử dụng trong luận văn này, việc mở rộng lĩnh vực dịch chỉ còn là vấn đề thời gian.

Thứ hai, một số vấn đề chưa được giải quyết như *thế đại từ* (anaphora), *hiện tượng tỉnh lược* (ellipsis). Những vấn đề này đã làm cho chất lượng đầu ra của câu tiếng Việt chưa được hay lắm.

Thứ ba, bên cạnh những vấn đề của riêng khối xử lý ngữ nghĩa, còn có những vấn đề khác liên quan gián tiếp làm ảnh hưởng đến chất lượng của khối xử lý ngữ nghĩa mà trong tương lai chúng tôi phải tiếp tục khắc phục như phân tích hình thái học, phân tích cú pháp...

## 5.2. KẾT LUẬN

Hệ dịch tự động Anh-Việt là một công trình rất có ý nghĩa. Đây là một trong những sản phẩm phần mềm có thể ghi dấu ấn Việt Nam trong kho tàng công nghệ thông tin thế giới, là sản phẩm thể hiện trí tuệ Việt Nam, và là sản phẩm của người Việt Nam phục vụ riêng cho người Việt Nam. Nhận thức rõ ý nghĩa to lớn này, chúng tôi nguyện cố gắng nhiều hơn nữa, tìm hiểu cải tiến chất lượng cho khối xử lý ngữ nghĩa nói riêng và hệ dịch tự động Anh-Việt nói chung. Chúng tôi mong muốn rằng, ở một thời điểm trong tương lai gần, chương trình sẽ để lại những dấu ấn rõ nét không những trong lòng người Việt Nam mà còn trong lòng bè bạn khắp năm châu.

## **Danh Mục Tài Liệu Tham Khảo**

- [1]. Dekang Lin (1993). *Principle-based parsing without overgeneration*. Proceedings of the 31<sup>st</sup> ACL Conference, Columbus, Ohio, pp 112-120.
- [2]. Dien Dinh (2002). *Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation*. Proceedings of Workshop on Machine Translation in Asia, COLING-02, Taiwan, 9/2002, pp26-32.
- [3]. Dien Dinh, Kiem Hoang (2002). *Bilingual corpus and word sense disambiguation in the English-to-Vietnamese Machine Translation*, Proceedings of APIS-02, Bangkok, Thailand, pp 8-15.
- [4]. Dien Dinh, Nguyen Luu Thuy Ngan, Do Xuan Quang, Van Chi Nam (2003). *Hybrid Approach to Word Order Transfer in the English-to-Vietnamese Machine Translation*. Paper at Machine Translation Summit IX, Louisiana, USA.
- [5]. Đinh Điền (1996). *Dịch tự động Anh-Việt*. Luận văn thạc sĩ tin học, ĐH Khoa Học Tự Nhiên, ĐH Quốc Gia TP.HCM.
- [6]. Đinh Điền (2001). *Bước đầu xây dựng kho ngữ liệu song ngữ Anh-Việt điện tử*. Luận văn thạc sĩ ngôn ngữ học so sánh, ĐH Khoa học Xã hội và Nhân văn, ĐH Quốc Gia TP.HCM.
- [7]. Đinh Điền (2003). *Mô hình học luật chuyển đổi từ ngữ liệu song ngữ cho Hệ dịch tự động Anh-Việt*. Luận án Tiến sĩ Tin học, ĐH Khoa học Tự Nhiên, Đại học Quốc gia Tp.HCM.
- [8]. Đinh Điền, Trần Lê Hồng Dũ, Văn Chí Nam (2001). *Khử nhập nhằng của từ trong văn bản tiếng Anh*. Toàn văn báo cáo khoa học, Viện công nghệ thông tin, Hà Nội, 12/2001.
- [9]. Đinh Điền, Văn Chí Nam, Trần Lê Hồng Dũ (2002). *Khai thác ngữ liệu SemCor trong xử lý ngôn ngữ tự nhiên*. Báo cáo khoa học của Hội thảo quốc gia “Một số vấn đề chọn lọc của CNTT”, Nha Trang, 6/2002.

- [10]. Dini, Luca, Vittorio Di Tomaso, Frédérique Segond (1998). *Error-driven Word sense disambiguation*. Proceedings of ACL-COLING-98, Montreal.
- [11]. Eric Brill (1993). *A Corpus-based approach to Language Learning*. Ph.D dissertation, Pennsylvania University, USA.
- [12]. Mark Stevenson, Yorick Wilks (2001). *The interaction of knowledge sources in Word sense disambiguation*. Journal of Computational Linguistics, Vol.27, Number 3, pp 321-350.
- [13]. Nancy Ide, Jean Véronis (1998). *Introduction to the special issue on Word sense disambiguation : the State of the Art*. Computational Linguistics, Vol.24, Number 1, pp 1-40.
- [14]. Natalia Zinovjeva (2000). *Learning sense disambiguation rules for Machine Translation*. Master thesis in Linguistics, Uppsala University, Sweden.
- [15]. Radu Floarian, Grace Ngai (2001). *Fast Transformation-based Learning Toolkit*. Johns Hopkins University, 9/2001.
- [16]. Radu Florian, Grace Ngai (2001). *Transformation-based learning in the fast lane*. Proceedings of North American ACL-2001.
- [17]. Samuel K. (1998). *Lazy Transformation-based learning*. Proceedings of the 11<sup>th</sup> International Florida AI Research Symposium Conference, Florida, USA, pp 227-253.
- [18]. Samuel K., Carbery S., Vijay-Shanker K., (1998). *Dialogue Act Tagging with Transformation-Based Learning*. Proceedings of COLING/ACL'98, pp 1150-1156.
- [19]. Shari Landes, Claudia Leacock, Randee I.Tengi (1999). *Building semantic concordances*. WordNet : an electronic lexical database.
- [20]. W. John Hutchins, Harold L. Somers (1992). *An Introduction to Machine Translation*, Academic Press.
-

## Phụ Lục 1. Danh Sách Nhãn Ngữ Nghĩa Cơ Bản

Mã	Diễn giải	Viết tắt của
ABS	Trừu tượng	Abstraction
ACT	Hoạt động	Act
AGT	Tác nhân	Agent
ANM	Động vật	Animal
ART	Vật nhân tạo	Artifact
ATR	Thuộc tính	Attribute
CEL	Tế bào	Cell
CHM	Hợp chất, hợp chất hoá học, nguyên tố hoá học	Chemistry compound
COM	Truyền thông	Communication
ENT	Thực thể	Entity
EVT	Sự kiện	Event
FOD	Thực phẩm	Food
FRM	Hình dáng	Form
GRB	Nhóm sinh học	Biological group
GRP	Các nhóm khác	Group
GRS	Nhóm trong xã hội	Social group
HUM	Con người, cá nhân	Human
LFR	Thực thể sống	Life form
LME	Đơn vị đo chiều dài	Linear measure
LOC	Địa điểm	Location

MEA	Số lượng	Measure
MIC	Vi sinh vật	Microorganism
NAT	Vật thể tự nhiên	Natural object
PHM	Hiện tượng	Phenomenon
PHO	Đối tượng, vật thể không có sự sống	Physical object
PLT	Cây cối	Plant
POS	Sở hữu	Possession
PRO	Quá trình	Process
PRT	Bộ phận	Part
PSY	Thuộc tính tâm lý	Psychological feature
QUD	Lượng xác định	Definite quantity
QUI	Lượng không xác định	Indefinite quantity
REL	Quan hệ	Relation
SPC	Không gian	Space
STA	Trạng thái	State
SUB	Chất liệu	Substance
TME	Khoảng thời gian, đơn vị thời gian	Time
vbody	Động từ chỉ về sự chăm sóc cơ thể, ăn mặc	Body
vchng	Động từ chỉ thay đổi	Change
vcogn	Động từ chỉ suy nghĩ, phán xét	cognition
vcomm	Động từ về truyền thông (hỏi, kể, hát...)	Communication
vcptn	Động từ chỉ về thi đua (chiến đấu...)	Competition

vcons	Động từ chỉ về tiêu thụ (ăn, uống..)	Consumption
vcont	Động từ chỉ sự tiếp xúc	Contact
vcreat	Động từ chỉ sự sáng tạo (vẽ, biểu diễn..)	creation
vemotn	Động từ chỉ cảm xúc	Emotion
vmotn	Động từ chỉ sự di chuyển (đi, chạy, nhảy..)	Motion
vpercept	Động từ chỉ nhận thức (thấy, nghe)	Perception
vpossess	Động từ chỉ sở hữu (sở hữu, mua, bán)	Possession
vsoc	Động từ chỉ các hoạt động xã hội	Social
vstat	Động từ chỉ trạng thái	State
vweath	Động từ chỉ thời tiết (mưa, bão...)	Weather

Khoa CNTT - ĐH KHTN TP.HCM

## Phụ Lục 2. Danh Sách Các Nhãn Từ Loại

Nhãn từ loại	Viết tắt của	Giải thích
CC	Coordinating conjunction	Liên từ
CD	Cardinal number	Số đếm
DT	Determiner	Định từ
EX	Existential “there”	“Có”
FW	Foreign world	Từ nước ngoài
IN	Preposition or subordinating conjunction	Giới từ hay liên từ phụ
JJ	Adjective	Tính từ
JJR	Adjective, comparative	Tính từ so sánh hơn
JJS	Adjective, superlative	Tính từ so sánh nhất
LS	List item marker	Dấu liệt kê
MD	Modal	Từ hình thái
NN	Noun, singular or mass	Danh từ số ít hoặc không đếm được
NNP	Proper noun, singular	Danh từ riêng (số ít)
NNPS	Proper noun, plural	Danh từ riêng (số nhiều)
NNS	Noun, plural	Danh từ số nhiều
PDT	Predetermine	Tiền chỉ định từ
POS	Possessive ending	Dấu cuối của sở hữu cách
PP	Personal pronoun	Đại từ nhân xưng

PR	Pronoun	Đại từ
PRP	Pronoun	Đại từ
PRP\$	Pronoun, plural	Đại từ số nhiều
RB	Adverb	Trạng từ
RBR	Adverb, comparative	Trạng từ so sánh hơn
RBS	Adverb, superlative	Trạng từ so sánh nhất
RP	Particle	Tiểu từ
SYM	Symbol	Ký hiệu
TO	“to”	Nhãn cho từ “to”
UH	Interjection	Thán từ
VB	Verb, base form	Động từ nguyên mẫu
VBD	Verb, past	Động từ ở quá khứ
VBG	Verb, gerund or present participle	Động từ (thêm -ing)
VCN	Verb, past participle	Quá khứ phân từ
VBP	Verb, non 3 <sup>rd</sup> person singular present	Động từ cho chủ từ không phải ở ngôi thứ 3 số ít
VBZ	Verb, 3 <sup>rd</sup> person singular present	Động từ cho chủ từ ở ngôi thứ 3 số ít
WDT	Wh-determiner	Định từ bắt đầu bằng WH-
WP	Wh-pronoun	Đại từ bắt đầu bằng WH-
WP\$	Possessive wh-pronoun	Đại từ sở hữu bắt đầu bằng WH-
WRB	Wh-adverb	Trạng từ bắt đầu bằng WH-



### Phụ Lục 3. Trích Một Số Luật

#### LUẬT CHUYỂN ĐỔI DỪNG TRONG XỬ LÝ NGỮ NGHĨA

1. GOOD:267 BAD:0 SCORE:267 RULE: \$SEM[-3,-1]=EMPTY  
Sense=DONTKNOW => Sense=PHM
2. GOOD:90 BAD:0 SCORE:90 RULE: \$POS[1,3]=NN POS0=VB \$POS[-3,-1]=NN => Sense=vsta
3. GOOD:61 BAD:0 SCORE:61 RULE: \$SEM[-3,-1]=EMPTY \$POS[1,3]=JJ  
POS0=VB \$POS[-3,-1]=NN Sense=PHM => Sense=vsta
4. GOOD:24 BAD:0 SCORE:24 RULE: \$SEM[-3,-1]=EMPTY  
\$POS[1,3]=PUNC POS0=NNP \$POS[-3,-1]=IN Sense=PHM =>  
Sense=HUM
5. GOOD:26 BAD:4 SCORE:22 RULE: \$SEM[-3,-1]=DEV Sense=PHM =>  
Sense=DEV
6. GOOD:18 BAD:0 SCORE:18 RULE: \$SEM[-3,-1]=EMPTY \$POS[1,3]=JJ  
POS0=VB \$POS[-3,-1]=\_ Sense=PHM => Sense=vsta
7. GOOD:18 BAD:0 SCORE:18 RULE: \$POS[1,3]=PUNC POS0=NNP \$POS[-3,-1]=PUNC => Sense=HUM
8. GOOD:23 BAD:5 SCORE:18 RULE: \$SEM[1,3]=DEV \$POS[-3,-1]=DT  
POS0=NN \$POS[1,3]=NN Sense=PHM => Sense=DEV
9. GOOD:16 BAD:0 SCORE:16 RULE: \$SEM[-3,-1]=EMPTY \$POS[1,3]=DT  
POS0=VB \$POS[-3,-1]=VB Sense=PHM => Sense=vcog
10. GOOD:16 BAD:0 SCORE:16 RULE: \$SEM[-3,-1]=EMPTY  
\$POS[1,3]=PUNC POS0=NN \$POS[-3,-1]=CD Sense=PHM => Sense=QUD
11. GOOD:12 BAD:0 SCORE:12 RULE: \$SEM[-3,-1]=EMPTY \$POS[1,3]=IN  
POS0=VB \$POS[-3,-1]=DT Sense=PHM => Sense=vsta
12. GOOD:13 BAD:2 SCORE:11 RULE: \$SEM[-3,-1]=vsta \$POS[1,3]=IN  
POS0=NN \$POS[-3,-1]=VB Sense=PHM => Sense=ATR

13. GOOD:15 BAD:5 SCORE:10 RULE: \$SEM[-3,-1]=EMPTY \$POS[1,3]=CC  
POS0=NN \$POS[-3,-1]=DT Sense=PHM => Sense=DEV
14. GOOD:13 BAD:3 SCORE:10 RULE: \$POS[1,3]=CD POS0=NN \$POS[-3,-  
1]=IN => Sense=COM
15. GOOD:9 BAD:0 SCORE:9 RULE: \$SEM[1,3]=COM \$POS[-3,-1]=NN  
POS0=VB \$POS[1,3]=IN Sense=vsta => Sense=vcre
16. GOOD:12 BAD:3 SCORE:9 RULE: \$POS[1,3]=NN POS0=VB \$POS[-3,-  
1]=VBD => Sense=vcog
17. GOOD:9 BAD:0 SCORE:9 RULE: \$POS[1,3]=IN POS0=NN \$POS[-3,-  
1]=CD => Sense=QUD
18. GOOD:8 BAD:0 SCORE:8 RULE: \$POS[1,3]=PUNC POS0=VB \$POS[-3,-  
1]=NN Sense=PHM => Sense=vsta
19. GOOD:8 BAD:0 SCORE:8 RULE: \$POS[1,3]=NN POS0=VB \$POS[-3,-  
1]=PUNC => Sense=vsta
20. GOOD:8 BAD:0 SCORE:8 RULE: \$SEM[1,3]=PHO \$POS[1,3]=CC \$POS[-  
3,-1]=IN => Sense=PHO

MỘT SỐ LUẬT DÙNG XÁC ĐỊNH NHỮNG TRƯỜNG HỢP KHÔNG  
THÊM NHỮNG CHO DANH TỪ SỐ NHIỀU

Chuyển	Thành	Điều kiện	Ghi chú
NHUNG	KHONG	TU_ $[-5,-1]$ : of	1 trong 5 từ đứng trước là <i>of</i>
NHUNG	KHONG	TU_ $[-5,-1]$ : many	1 trong 5 từ đứng trước là <i>many</i>
NHUNG	KHONG	TU_ $[-5,-1]$ : some	1 trong 5 từ đứng trước là <i>some</i>
NHUNG	KHONG	TU_ $[-5,-1]$ : multiple	1 trong 5 từ đứng trước là <i>multiple</i>
NHUNG	KHONG	TU_ $[-5,-1]$ : several	1 trong 5 từ đứng trước là <i>several</i>
NHUNG	KHONG	POS_1 : NN	Từ loại của từ đứng sau là NN
NHUNG	KHONG	POS_1 : NNS	Từ loại của từ đứng sau là NNS
NHUNG	KHONG	POS_ $[-5,-1]$ : CD	Từ loại của 1 trong 5 từ đi trước là CD

## Phụ Lục 4. Các Kết Quả Dịch Đạt Được

Những kết quả dịch dưới đây được lấy nguyên gốc từ kết quả của chương trình, không có hiệu đính. Chúng tôi chỉ chỉnh sửa một số định dạng của đầu vào cho thích hợp với chương trình. (Ví dụ như, bỏ những dấu hiệu siêu liên kết ; nối các câu bị ngắt sai lại để tạo thành câu hoàn chỉnh trong các tập tin văn bản). Những câu dịch dưới đây minh họa cho những trường hợp nhập những ngữ nghĩa mà chương trình chúng tôi đã giải quyết được.

Câu tiếng Anh	Câu dịch tiếng Việt
Ambiguity of words.	Sự nhập nhằng của từ.
He <b>is</b> a teacher and he <b>has</b> many pupils.	Anh ấy <b>là</b> một giáo viên và anh ấy <b>có</b> nhiều học sinh.
He <b>is not only</b> handsome <b>but also</b> intelligent.	Anh ấy (thì) không những đẹp trai mà còn thông minh.
He <b>is</b> at his home alone and he <b>is</b> reading a book.	Anh ấy (thì) tại nhà của anh ấy một mình và anh ấy đang đọc một cuốn sách.
He <b>was</b> scolded by his wife but he will <b>be</b> appreciated by his lover.	Anh ấy <b>bị</b> mắng bởi người vợ của anh ấy nhưng anh ấy sẽ <b>được</b> đánh giá cao bởi người yêu của anh ấy.
He <b>can program</b> many subtle <b>programs</b> with new and interesting features.	Anh ấy có thể <b>lập trình</b> nhiều <b>chương trình</b> tinh vi với những đặc tính mới và hay.
He <b>loves</b> his wife with a very faithful <b>love</b> .	Anh ấy <b>yêu</b> người vợ của anh ấy với một tình yêu rất <b>trung thành</b> .
I <b>asked</b> you to come here <b>in order that</b> I	Tôi <b>yêu cầu</b> bạn đến đây <b>ngõ hầu</b> tôi

ask you several questions.	<i>hỏi</i> bạn vài câu hỏi.
A very <b>old</b> man has a very <b>old</b> computer system.	Một người đàn ông rất <b>già</b> có một hệ thống máy tính rất <b>cũ</b> .
Meaning of idioms.	Nghĩa của thành ngữ.
Immediately, they <i>clear body out</i> in the hospital.	Ngay lập tức, <i>chúng dọn sạch</i> trong bệnh viện.
She will <i>make up her mind</i> about this matter soon.	Cô ấy sẽ <i>quyết định</i> về vấn đề này sớm.
The old man in the room sits under the <i>old glory</i> .	Người đàn ông già trong phòng ngồi dưới <i>lá cờ Mỹ</i> .
Comparison of adjectives.	Việc so sánh của tính từ.
That book is more expensive than this book.	Cuốn sách đó (thì) mắc hơn so với cuốn sách này.
I amn't as intelligent as him.	Tôi thì/là không thông minh bằng anh ấy.
She is not so ugly as her sister.	Cô ấy (thì) không xấu như chị/em gái của cô ấy.
This is the most important project.	Đây là dự án quan trọng nhất.
Possessive case.	Trường hợp sở hữu.
Computer's action is difficult.	Hoạt động <b>của</b> máy tính (thì) khó.
computers' action are difficult.	Hoạt động <b>của</b> những máy tính (thì) khó.
You <b>are</b> the sixth person.	Bạn <b>là</b> người thứ sáu.
Formation of new words.	Sự hình thành của từ mới.

<i>This</i> is an very significant modernization.	<b>Đây</b> là một sự hiện đại hóa rất có ý nghĩa.
Normally, I always normalize normal problems.	Bình thường, tôi luôn luôn bình thường hóa những vấn đề bình thường.
This is a stereo-image in a modern computer.	Đây thì/là một hình ảnh nổi trong một máy tính hiện đại.
Vice-president is a anti-war person.	Phó tổng thống là một người chống chiến tranh.
She plays tennis and he does too.	Cô ấy chơi quần vợt và anh ấy cũng vậy.
She isn't singer and mother is not either.	Cô ấy không là ca sĩ và mẹ cũng không.
A tall and beautiful girl eats a small but delicious apple.	Một cô gái cao và đẹp ăn một quả táo nhỏ nhưng ngon.
A colorless green idea sleeps furiously.	Một ý xanh không có màu ngủ giận dữ.
The Question.	Câu hỏi.
Is she a famous singer ?	Cô ấy thì/là một ca sĩ nổi tiếng phải không?
Are you free tonight ?	Bạn có tự do đêm nay không?
Does a tall girl eat a ripe apple?	Một cô gái cao có ăn một quả táo chín không?
Have I eaten many apples?	Tôi có ăn nhiều quả táo không?
Did my brother study in the library?	Anh/em của tôi đã học trong thư viện phải không?
Will she sing in the church choir next week?	Cô ấy sẽ hát trong đội đồng ca nhà thờ tuần kế phải không?

Won't she sing in the church choir next week?	Cô ấy sẽ không hát trong đội đồng ca nhà thờ tuần kế phải không?
Had they visited Paris <i>in 1987</i> ?	Chúng có thăm Paris <i>vào năm 1987</i> không?
What do you want ?	Bạn muốn cái gì?
Why do you say that matter ?	Tại sao bạn nói vấn đề đó?
Why is he angry ?	Tại sao là anh ấy giận dữ?
How is your mother ?	Mẹ của bạn có mạnh khỏe không?
What is <i>this</i> ?	<i>Đây</i> là cái gì?
Who are you ?	Bạn là ai?
What are you learning ?	Bạn đang học cái gì?
Where are you going now ?	Bạn đang đi đâu bây giờ?
I know <i>her</i> .	Tôi biết <i>cô ấy</i> .
She <i>needs</i> a new <i>dress</i> .	Cô ấy cần một cái áo mới.
She <i>did</i> her homework.	Cô ấy <i>làm</i> bài tập về nhà của cô ấy.
I opened the window.	Tôi mở cửa sổ.
He doesn't like cold weather.	Anh ấy không thích thời tiết lạnh.
She smiles <i>her thanks</i> .	Cô ấy mỉm cười <i>những lời cảm ơn của cô ấy</i> .
The man <i>cutting</i> woods is my father.	Người đàn ông <i>đang cắt</i> những gỗ là bố của tôi.
The man scolded by his wife went to my house.	Người đàn ông bị mắng bởi người vợ của anh ấy đi tới ngôi nhà của tôi.

He <i>laughed</i> with a merry <i>laugh</i> .	Anh ấy <i>cười</i> với một <i>nụ cười</i> vui vẻ.
I send her the vase <i>that</i> you put in the box.	Tôi gửi cô ấy lọ hoa <i>mà</i> bạn đặt trong cái hộp.
The man <i>that</i> beat you sleeps in the house.	Người đàn ông <i>mà</i> đánh bạn ngủ trong ngôi nhà.
The man that stealed a car sleeps in the house that is in a garden.	Người đàn ông mà ăn trộm một xe hơi ngủ trong ngôi nhà mà (thì) trong một vườn.
He ought to help you.	Anh ấy phải giúp bạn.
We have to do our exercises tonight.	Chúng tôi phải làm những bài tập của chúng tôi đêm nay.
She pretend not to see me.	Cô ấy giả vờ không để thấy/xem tôi.
I don't like to <i>ask for a favor</i> .	Tôi không thích để <i>xin một ân huệ</i> .
I forget to do what you told me.	Tôi quên để làm cái gì bạn bảo tôi.
They begin to build that bridge.	Chúng bắt đầu để xây dựng cầu đó.
He learn to ride a bicycle.	Anh ấy học để cỡi một xe đạp.
You intend to see him.	Bạn tính để thấy/xem anh ấy.
I'll <i>ask</i> him to help us.	Tôi sẽ <i>yêu cầu</i> anh ấy giúp chúng tôi.
They persuade me to believe that there is no danger.	Chúng thuyết phục tôi để tin tưởng rằng có không sự nguy hiểm.
He teaches his son to play the guitar.	Anh ấy dạy người con trai của anh ấy để chơi đàn ghi-ta.
They proved him wrong.	Chúng chứng minh anh ấy sai.
You think her a good teacher.	Bạn suy nghĩ cô ấy một giáo viên tốt.



I want you happy.	Tôi muốn bạn hạnh phúc.
I like you punctual.	Tôi thích bạn đúng giờ.
I prefer my coffee hot.	Tôi thích cà phê của tôi nóng.
I believe him right.	Tôi tin tưởng anh ấy đúng.
We consider what he said unimportant.	Chúng tôi cân nhắc cái gì anh ấy nói không quan trọng.
We also suppose him a spy.	Chúng tôi cũng giả sử anh ấy một gián điệp.
Tom's teacher thinks him the cleverest boy in the class.	Giáo viên của Tom suy nghĩ anh ấy cậu con trai khéo léo nhất trong lớp.
He was believed innocent.	Anh ấy được tin tưởng vô tội.
I smell something burning.	Tôi ngửi đang đốt cháy cái gì đó.
I saw him opening his mother's purse.	Tôi thấy/xem anh ấy mở cái ví của mẹ của anh ấy.
She heard her husband shouting.	Cô ấy nghe chồng của cô ấy hét.
We like to listen to <i>that</i> band playing in the city park every sunday.	Chúng tôi thích để lắng nghe băng <i>đó</i> chơi trong công viên thành phố mọi chủ nhật.
She helps me carry this heavy <i>bag</i> .	Cô ấy giúp tôi mang <i>túi</i> nặng này.
You've ever known her <i>lose her temper</i> .	Bạn đã từng biết cô ấy <i>mất bình tĩnh</i> .
I have known a educated person make this mistake.	Tôi biết một người được giáo dục làm lỗi này.
He painted the door green.	Anh ấy sơn cửa ra vào xanh.
The cat licked the disk clean.	Con mèo liếm đĩa sạch.

We hammered it flat.	Chúng tôi đập nó bằng phẳng.
They beat the boy <i>black and blue</i> .	Chúng đánh cậu con trai <i>bầm tím</i> .
You open your mouth wide.	Bạn mở miệng của bạn rộng.
You raise your head higher.	Bạn nâng cái đầu của bạn cao hơn.
The sun keeps us warm.	Mặt trời giữ chúng tôi ấm.
You've made your shirt dirty.	Bạn làm áo sơ-mi của bạn bẩn.
He held the door open.	Anh ấy giữ cửa ra vào mở.
The blister on my heel made me painful.	Vết bong giộp trên gót chân của tôi làm tôi đau.
He called his dog Lulu.	Anh ấy gọi con chó của anh ấy là Lulu.
They named the ship Mary.	Chúng đặt tên con tàu Mary.
The bishop elected him king of england.	Giám mục bầu anh ấy vua của nước Anh.
The pastor christened the child Jennifer.	Mục sư đặt tên thánh đứa trẻ Jennifer.
You don't throw it out of the window.	Bạn không ném nó ngoài cửa sổ.
The servant showed me to the door.	Người giúp việc chỉ tôi tới cửa ra vào.
I found the pen under the desk.	Tôi tìm thấy viết mực dưới bàn giấy.
He regards me as a good friend.	Anh ấy xem tôi như một bạn tốt.
They employ him as a clerk.	Chúng dùng anh ấy như một thư ký.
He brought his sister to see his uncle.	Anh ấy mang chị/em gái của anh ấy để thấy/xem chú của anh ấy.
I shall need an hour to finish the job.	Tôi sẽ cần một giờ để hoàn tất công việc.

They left me to do all the work.	Chúng để cho tôi để làm tất cả công việc.
They appointed an official to superintend the work.	Chúng bổ nhiệm một viên chức để coi sóc công việc.
We found the books where we had left them.	Chúng tôi tìm thấy những cuốn sách đâu chúng tôi để cho chúng.
He said that he would come and see us.	Anh ấy nói rằng anh ấy sẽ đến thăm chúng tôi.
I wish that they would tell us the truth.	Tôi ước muốn rằng chúng sẽ bảo chúng tôi sự thật.
I hear that you are going to go american next week.	Tôi nghe rằng bạn đang sắp đi tuần kế Mỹ.
We saw that the plan would fail.	Chúng tôi thấy/xem rằng kế hoạch sẽ thất bại.
I feel <i>that</i> he told the truth.	Tôi cảm <i>thấy</i> rằng anh ấy nói thực.
I hope <i>that</i> you'll be able to come.	Tôi hy vọng <i>rằng</i> bạn sẽ có khả năng đến.
I expect that you are surprised at the news.	Tôi mong đợi rằng bạn được ngạc nhiên về tin.
She suggested that we should start early.	Cô ấy gợi ý rằng chúng tôi nên bắt đầu sớm.
You think that she is coming.	Bạn suy nghĩ rằng cô ấy đang đến.
I dare say that he will come later.	Tôi dám nói rằng anh ấy sẽ đến sau đó.
She confessed to her parent that she had loved him.	Cô ấy thú nhận tới cha mẹ của cô ấy rằng cô ấy yêu anh ấy.

He admitted to his wife that he had betrayed her.	Anh ấy thu nhận tới người vợ của anh ấy mà anh ấy phản bội cô ấy.
I suggested to them that it might be better.	Tôi gợi ý tới chúng rằng nó có thể (thì) tốt hơn.
We explained to everyone that the delay was inevitable.	Chúng tôi giải thích cho mọi người mà sự trì hoãn (thì) không thể tránh khỏi.
She said to them that she would marry soon.	Cô ấy nói tới chúng rằng cô ấy sẽ cưới sớm.
I told him that my father was ill.	Tôi bảo anh ấy rằng bố của tôi bệnh.
She reminds him that the meeting is <i>on saturday</i> .	Cô ấy nhắc nhở anh ấy rằng cuộc họp (thì) <i>vào thứ bảy</i> .
The scandal taught her that silence is gold.	Chuyện tai tiếng dạy cô ấy rằng sự yên lặng là vàng.
They informed me that the prisoner had escaped.	Chúng báo tin tôi rằng tù nhân thoát.
He satisfied me that he could do the work well.	Anh ấy làm hài lòng tôi rằng anh ấy có thể làm công việc tốt.
You should ask how to get from the station to our hotel.	Bạn nên yêu cầu làm thế nào đạt từ trạm tới khách sạn của chúng tôi.
You know how to answer that question.	Bạn biết làm thế nào để trả lời câu hỏi đó.
We must find out what to do later.	Chúng tôi phải tìm ra cái gì để làm sau đó.
I forgot where to stop.	Tôi quên đâu để ngưng.
He showed me how to <i>do</i> it.	Anh ấy chỉ tôi làm thế nào để <i>làm</i> nó.

We asked the teacher how to pronounce that word.	Chúng tôi yêu cầu giáo viên làm thế nào phát âm từ đó.
You must teach your son how to behave properly.	Bạn phải dạy người con trai của bạn làm thế nào để cư xử đúng cách.
<b>old driver</b>	<b>Tài xế già</b>
driver is old	Tài xế (thì) già
<b>green field</b>	<b>Cánh đồng xanh</b>
field is green	Cánh đồng (thì) xanh
field is interesting	Trường (thì) hay
This <b>old man</b> and woman are <b>printer</b> .	Người đàn ông già đây và đàn bà là <b>thợ in</b> .
My father, a very famous teacher, sleeps, plays and sings.	Bố của tôi, một giáo viên rất nổi tiếng, ngủ, chơi và hát.
A computer, according to Webster, is an <b>programmable</b> electronic device that can process, store and retrieve data.	Theo Webster, một máy tính, là một thiết bị điện tử <b>có thể lập trình được</b> mà có thể xử lý, chứa và truy tìm dữ liệu.
I prefer "ao dai".	Tôi thích `` ao dai ".
Types of computer system.	Những loại của hệ thống máy tính.
He has 5 units of a new computer system.	Anh ấy có 5 đơn vị của một hệ thống máy tính mới.
This is a very interesting programming language.	Đây là một ngôn ngữ lập trình rất hay.
This is a new used programming language and that is the most frequently used programming language.	Đây là một ngôn ngữ lập trình sử dụng mới và đó là ngôn ngữ lập trình được sử dụng thường xuyên nhất.

these two books	Hai cuốn sách này
these 2 new interesting books	2 cuốn sách hay mới này
these 2 new interesting programming language	2 ngôn ngữ lập trình hay mới này
Programming assembly language is easier than machine language, binary code, but they are still more difficult than high level language which are more economical in terms of space because it requires the programmer to have a good knowledge of machine.	Hợp ngữ lập trình (thì) dễ hơn so với ngôn ngữ máy, mã nhị phân, nhưng chúng (thì) vẫn còn khó hơn so với ngôn ngữ mức cao mà (thì) tiết kiệm hơn về phương diện không gian bởi vì nó đòi hỏi người lập trình để có một kiến thức tốt của máy.
I <i>can can</i> a <i>can</i> .	Tôi <i>có thể đóng hộp</i> một <i>cái hộp</i> .
An <i>old</i> man is reading an <i>old</i> book.	Một người đàn ông <i>già</i> đang đọc một cuốn sách <i>cũ</i> .
An <i>old</i> police is sitting on an <i>old</i> chair.	Một cảnh sát <i>già</i> đang ngồi trên một cái ghế <i>cũ</i> .
I want to book two books.	Tôi muốn để đặt trước hai cuốn sách.
I <i>am</i> a student.	Tôi là một sinh viên.
He <i>is fat</i> .	Anh ấy ( <i>thì</i> ) <i>mập</i> .
The <i>old driver</i> is driving an <i>old car</i> .	<i>Tài xế già</i> đang lái xe một <i>xe hơi cũ</i> .
She installed <i>old driver</i> for my <i>printer</i> .	Cô ấy cài đặt <i>trình điều khiển cũ</i> cho <i>máy in</i> của tôi.
These men are <i>printers</i> .	Những người đàn ông này là <i>những thợ in</i> .
several new pens.	Vài viết mực mới.

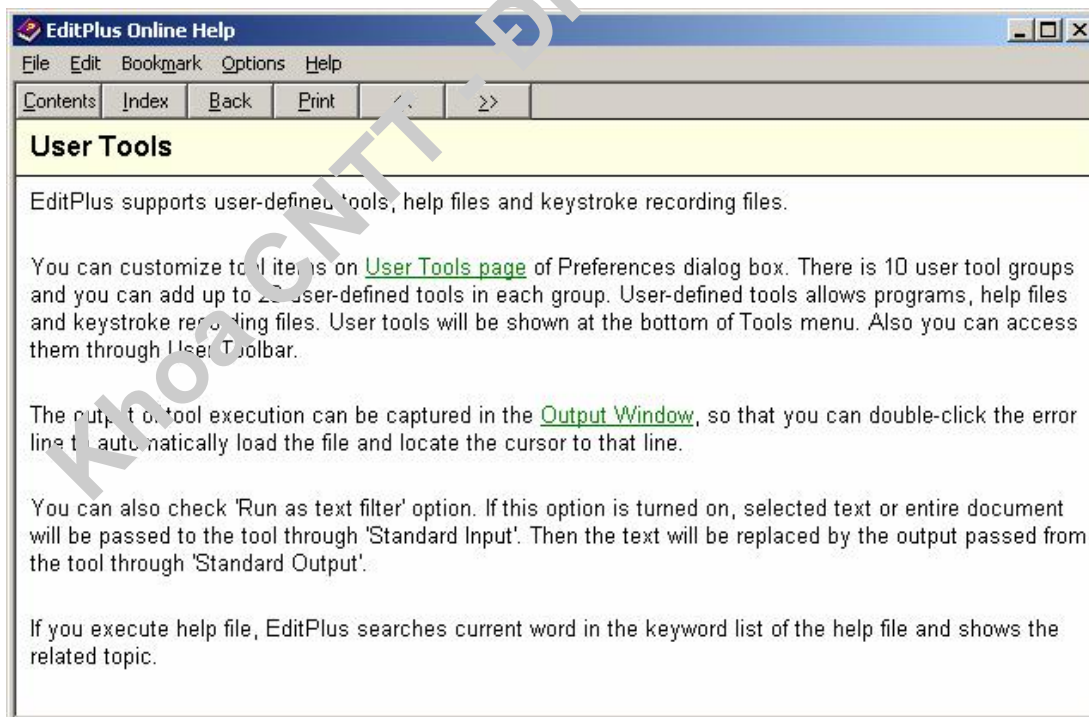
various traditional design skills.	Những kỹ năng thiết kế thông thường khác nhau.
small business and individual users.	Kinh doanh nhỏ và những người dùng riêng rẽ.
small businesses and individual users.	Những kinh doanh nhỏ và những người dùng riêng rẽ.
Vietnam will organize SEA Games <b>in 2003</b> .	Vietnam sẽ tổ chức SEA Games <b>vào năm 2003</b> .
I have waited for her <b>for</b> many years.	Tôi đợi cô ấy <b>trong</b> nhiều năm.
I am taller than him.	Tôi (thì) cao hơn so với anh ấy.
They <b>are reading</b> books.	Chúng <b>đang đọc</b> những cuốn sách.
He <b>is hated</b> by his friends.	Anh ấy <b>bị ghét</b> bởi những bạn của anh ấy.
There are a lot of beautiful places <b>in Vietnam</b> .	Có nhiều chỗ đẹp <b>ở Vietnam</b> .
These books are written by me <b>in english</b> .	Những cuốn sách này được viết bởi tôi <b>bằng tiếng Anh</b> .
I eat <b>rice</b> .	Tôi ăn <b>cơm</b> .
They plant <b>rice</b> , and he imports <b>rice</b> .	Chúng trồng <b>lúa</b> , và anh ấy nhập <b>gạo</b> .

## Phụ Lục 5. Một Số Kết Quả Dịch Thử Nghiệm

Những kết quả dịch dưới đây được lấy nguyên gốc từ kết quả của chương trình, không có hiệu đính. Chúng tôi chỉ chỉnh sửa một số định dạng của đầu vào cho thích hợp với chương trình. (Ví dụ như, bỏ những dấu hiệu siêu liên kết ; nối các câu bị ngắt sai lại để tạo thành câu hoàn chỉnh trong các tập tin văn bản). Chúng tôi cũng kèm theo hình ảnh chứa văn bản gốc của các nguồn được dùng để dịch.

Trong phụ lục này, chúng tôi có thực hiện một so sánh nhỏ giữa chương trình của chúng tôi (VCLEVT) và chương trình dịch tự động Anh-Việt thương mại đang rất phổ biến hiện nay (EVTran Technic version 2.0).

### ❑ Hướng Dẫn Sử Dụng của Trình Soạn Thảo EditPlus Version 2.01a



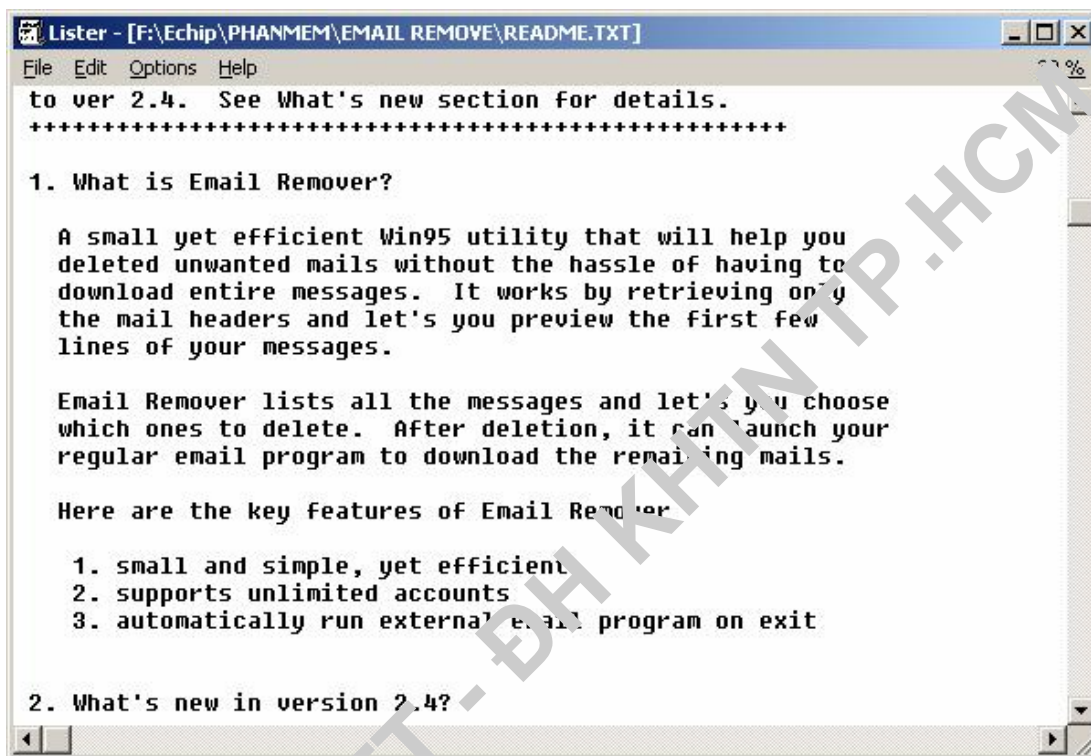


EVTran ver 2.0	VCL EVT
EditPlus supports <b>user-defined</b> tools, help files and <b>keystroke</b> recording files.	
EditPlus hỗ trợ những công cụ <b>do người dùng định ra</b> , những hồ sơ ghi file giúp đỡ và <b>nhấn phím</b> .	EditPlus hỗ trợ những công cụ <b>user-defined</b> , những tập tin sự giúp đỡ và những tập tin bản ghi <b>phím nhấn</b> .
You can customize tool <b>items</b> on User Tools page of Preferences <b>dialog box</b> .	
Bạn có thể tùy biến những <b>tiết mục</b> công cụ trên (về) trang những công cụ Người dùng (của) <b>hộp thoại</b> những ưu tiên.	Bạn có thể điều chỉnh những <b>mục</b> công cụ trên trang công cụ người dùng của <b>cái hộp đối thoại</b> sở thích.
There is 10 user tool groups and you can <b>add up to</b> 20 user-defined tools in each group.	
Có 10 nhóm công cụ người dùng và bạn có thể <b>lấy tổng tới</b> 20 công cụ do người dùng định ra trong mỗi nhóm.	Có 10 nhóm công cụ người dùng và bạn có thể <b>thêm lên tới</b> 20 công cụ user-defined trong mỗi nhóm.
User-defined tools allows programs, <b>help files</b> and keystroke recording files.	
Những công cụ do người dùng định ra cho phép những chương trình, <b>những hồ sơ ghi file giúp đỡ</b> và nhấn phím.	Những công cụ User-defined cho phép những chương trình, <b>những tập tin sự giúp đỡ</b> và những tập tin bản ghi phím nhấn.
User tools will <b>be shown</b> at the bottom of Tools menu.	
Những công cụ Người dùng sẽ được <b>cho thấy</b> ở (tại) đáy (của) thực đơn những công cụ.	Những công cụ người dùng sẽ <b>được chỉ</b> tại đáy của thực đơn công cụ.
Also you can access them through User Toolbar.	
Cũng bạn có thể truy nhập chúng xuyên qua Thanh công cụ Người dùng.	Cũng bạn có thể truy cập chúng xuyên qua Toolbar người dùng.

<p>The output of tool execution can <b>be captured</b> in the <b>Output Window</b>, so that you can double-click the error <b>line</b> to automatically load the file and locate the cursor to that <b>line</b>.</p>	
<p>Đầu ra (của) sự thực hiện công cụ có thể <b>được bắt</b> trong <b>lỗi ra</b>, để bạn có thể nhấn đúp <b>hàng</b> lỗi để tự động tải hồ sơ và định vị con trỏ tới <b>hàng</b> đó.</p>	<p>Đầu ra của sự thi hành công cụ có thể <b>bị bắt giữ</b> trong <b>cửa sổ đầu ra</b>, để bạn có thể nhấp kép <b>đường</b> lỗi sai để một cách tự động nạp tập tin và định vị con trỏ tới <b>đường</b> đó.</p>
<p>You can also check Run as text filter option.</p>	
<p>Bạn có thể cũng kiểm tra tùy chọn được chạy như văn bản lọc.</p>	<p>Bạn cũng có thể kiểm tra chạy khi sự lựa chọn bộ lọc văn bản.</p>
<p>If this option is turned on, selected text or entire document will be passed to the tool through <b>Standard Input</b>.</p>	
<p>Nếu tùy chọn này được bật, văn bản được lựa chọn hoặc toàn bộ tài liệu sẽ được đi qua cho công cụ xuyên qua <b>Chuẩn</b> được nhập vào.</p>	<p>Nếu sự lựa chọn này (thì) bật thì văn bản được lựa chọn hay là toàn thể tài liệu sẽ được băng qua tới công cụ xuyên qua <b>Standard đầu vào</b>.</p>
<p>Then the text will be replaced by the output passed from the tool through Standard Output.</p>	
<p>Rồi văn bản sẽ được thay thế bởi đầu ra đi qua từ công cụ xuyên qua Đầu ra Chuẩn.</p>	<p>Sau đó văn bản sẽ được thay thế bởi đầu ra được băng qua từ công cụ xuyên qua đầu ra chuẩn.</p>
<p>If you execute help file, EditPlus searches current word in the keyword list of the help file and <b>shows</b> the <b>related</b> topic.</p>	
<p>Nếu bạn thực hiện file giúp đỡ, EditPlus tìm kiếm từ hiện thời trong danh sách từ</p>	<p>Nếu bạn thực hiện giúp hồ sơ thì EditPlus tìm kiếm từ hiện hành trong danh sách từ</p>

khóa của file giúp đỡ và <i>những sự trung bày</i> của chủ đề <i>liên quan</i> .	khóa của tập tin sự giúp đỡ và <i>chỉ</i> chủ đề <i>được có quan hệ</i> .
--	---

❑ **Nguồn : Tập tin Readme.txt trong phần mềm Email Remover version 2.4**



EVTran ver 2.0	VCL EVT
1. What is Email Remover?	
1. Cái gì là <i>Người dọn đồ Email</i> ?	1. cái gì là <i>Email Remover</i> ?
A small yet efficient Win95 utility that will help you deleted unwanted mails without the hassle of having to download entire messages.	
Một tiện ích Win95 nhỏ <i>tuy thế</i> hiệu quả mà sẽ giúp đỡ bạn xóa những thư từ không cần đến mà không có chuyện phiền (của) việc phải tải xuống những	Một tiện ích Win95 hiệu quả <i>nhưng</i> nhỏ mà sẽ giúp bạn xóa những thư vô ích mà không có hassle của phải nạp xuống toàn thể thông báo.

toàn bộ thông báo.	
It works by <b>retrieving only</b> the mail headers and let you preview the first few <b>lines</b> of your messages.	
Nó làm việc bởi <b>khôi phục chỉ</b> những đầu mục thư từ và để cho bạn xem trước ít <b>dòng</b> những thông báo (của) bạn đầu tiên.	Nó làm việc bởi <b>truy tìm chỉ</b> những đầu thư và để bạn một ít <b>đường</b> đầu tiên của thông báo của bạn xem trước.
<b>Email Remover</b> lists all the messages and let you choose <b>which ones to delete</b> .	
Người dọn đồ Email liệt kê tất cả các thông báo và để cho bạn chọn <b>những xóa một nào</b> .	<b>Email Remover</b> liệt kê tất cả những thông báo và để bạn chọn <b>những cái mà để xóa</b> .
After deletion, it can <b>launch</b> your regular email program to download the <b>remaining</b> mails.	
Sau sự xóa, nó có thể <b>giới thiệu</b> chương trình email bình thường (của) bạn để tải xuống <b>Còn lại</b> Những thư từ.	Sau sự xóa, nó có thể <b>khởi chạy</b> chương trình thư tín điện tử thông thường của bạn để nạp xuống những thư <b>đang còn lại</b> .
<b>Here</b> are the <b>key</b> features of Email Remover	
ở đây <b>chìa khóa</b> là những đặc tính (của) Người dọn đồ Email	<b>Đây</b> là những đặc tính <b>chủ chốt</b> của Email Remover
1. small and simple, yet efficient	
1. nhỏ và đơn giản, tuy thế hiệu quả	1. nhỏ và đơn giản, nhưng hiệu quả
2. supports unlimited accounts	
2. hỗ trợ những tài khoản vô tận	2. hỗ trợ những tài khoản không giới hạn

3. automatically run external email program <i>on</i> exit	
3. tự động chạy chương trình email ngoài <i>trên (về)</i> lỗi ra	3. chạy chương trình thư tín điện tử bên ngoài <i>trên</i> lỗi ra một cách tự động
2. What's new in version 2.4?	
2. Cái gì <i>Có</i> mới trong phiên bản 2.4?	2. cái gì ( <i>thì</i> ) mới trong phiên bản 2.4?
This release's main purpose <i>is to fix</i> a <i>bug</i> which may cause possible deletion of wrong mail when the sorting feature <i>is used</i> .	
Mục đích chính (của) phiên bản này <i>Sẽ cố định</i> một <i>con rệp</i> mà có thể gây ra sự xóa có thể (của) thư từ sai khi đặc tính sắp xếp <i>được sử dụng</i> .	Mục đích chính của ấn bản này <i>là để khắc phục</i> một <i>lỗi</i> mà có thể gây ra sự xóa có thể của thư sai khi mà đặc tính sắp xếp ( <i>thì</i> ) <i>được sử dụng</i> .
The problem <i>was reported</i> to me <i>by</i> Ivan Hamilton and Brett last week ( <i>End</i> of June).	
Vấn đề <i>được báo cáo</i> tới tôi <i>trước</i> Hamilton và Brett Ivan tuần trước ( <i>Kết thúc</i> (của) <i>Tháng sáu</i> ).	Vấn đề <i>được báo cáo</i> tới tôi <i>bởi</i> Ivan Hamilton và Brett tuần vừa qua ( <i>End</i> của <i>June</i> ).
I've tried <i>fixed</i> it but have <i>posted</i> the wrong version to the web as ver 2.3. <i>Thanks to</i> Dave Rigg <i>for</i> notifying me again.	
Tôi thử <i>cố định</i> nó nhưng đã <i>bố trí</i> phiên bản sai tới mạng Nh V 2.3. <i>Nhờ</i> vào Dave Rigg <i>để</i> thông báo tôi lần nữa.	Tôi thử được <i>khắc phục</i> nó nhưng <i>gởi</i> phiên bản sai tới những web như những <i>cám ơn</i> 2.3. ver <i>tới</i> Dave Rigg <i>cho</i> thông báo tôi một lần nữa.
I hope ver 2.4 finally correct the problem.	
Tôi hy vọng Ver 2.4 cuối cùng sửa chữa vấn đề.	Tôi hy vọng ver 2.4 cuối cùng sửa sai vấn đề.

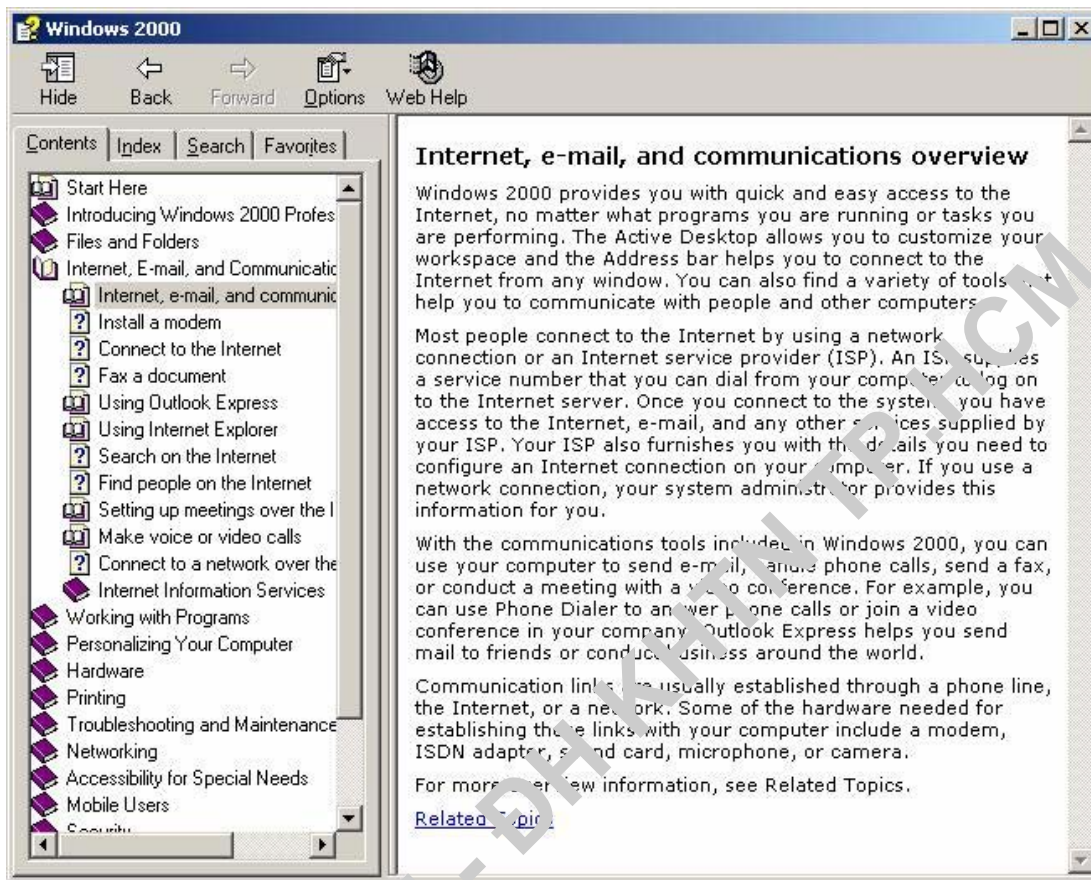
3. <b>System</b> Requirements	
3. Những yêu cầu <b>Hệ thống</b>	Những yêu cầu 3. <i>System</i>
Email Remover is small (about 460K only), thus use only little resources of your system.	
Người dọn đồ Email (thì) nhỏ ( về 460 K chỉ), như vậy sử dụng những tài nguyên nhỏ bé chỉ (của) hệ thống (của) bạn.	Email Remover (thì) nhỏ (về 460K chỉ), như vậy sử dụng chỉ nhỏ những tài nguyên của hệ thống của bạn.
It has been tested <b>to</b> run on below requirements.	
Nó đã được kiểm tra chạy trên (về) ở bên dưới những yêu cầu.	Nó được kiểm tra <b>để</b> chạy trên phía dưới những yêu cầu.
Windows 95 / NT	
Windows 95 / NT	Windows 95 / NT
TCP / IP installed	
TCP / IP được thiết đặt	TCP / IP cài đặt
Less than 400K of Hard disk space	
ít hơn 400 K (của) không gian đĩa cứng	Kém hơn hơn 400K của không gian đĩa cứng
16MB RAM	
16 MB <i>Nhồi nhét</i>	<b>RAM</b> 16MB
<b>To</b> install, simply extract above files, <b>preferably</b> store under single subdirectory name "Email Remover".	
Thiết đặt, đơn giản rút ở trên những hồ sơ, <b>preferably</b> cất giữ tên danh mục con đơn ở dưới " Người dọn đồ Email ".	<b>Để</b> cài đặt, đơn giản rút ra trên những tập tin, <b>có thể thích</b> chứa dưới tên thư mục con đơn " Email Remover ".

Create an icon by <b>dragging</b> eremove.exe from <b>Explorer</b> and <b>drop</b> onto the Desktop.	
Tạo ra một biểu tượng bởi <i>cán</i> (sự kéo) eremove.exe từ <i>Người thăm dò</i> và <i>giọt</i> lên trên Desktop.	Tạo ra một biểu tượng bởi <i>kéo</i> eremove.exe từ <b>Explorer</b> và <i>làm rớt</i> vào Desktop.
<b>To</b> run, double click on the icon created.	
<u>Để</u> chạy , nhấn đúp trên (về) biểu tượng tạo ra.	<u>Để</u> chạy, nhấp kép trên biểu tượng tạo ra.
<b>To</b> uninstall, simply delete all the files you have extracted.	
<i>Tôi uninstall</i> , Đơn giản xóa tất cả các hồ sơ (mà) bạn rút.	<b>Không cài đặt</b> , đơn giản xóa tất cả những tập tin bạn rút ra.
5. Legal Matters	
5. Vấn đề pháp lý	5. vấn đề hợp pháp
This program is free.	
Chương trình này (thì) tự do.	Chương trình này (thì) tự do.
It <b>is written</b> with my best ability <b>at</b> my own leisure time.	
Nó <u>được viết</u> với khả năng tốt nhất (của) tôi ở ( <i>tại</i> ) thì giờ nhàn rỗi (của) chính mình thời gian.	Nó <u>được viết</u> với khả năng tốt nhất của tôi <b>lúc</b> thời gian thời gian rảnh rỗi của tôi.
I am proud to share it with the Internet community; however, I SHOULD NOT <b>BE HELD LIABLE</b> FOR ANY DAMAGE DIRECTLY OR INDIRECTLY CAUSED BY USE OR MISUSE OF THIS SOFTWARE.	
Tôi (thì) tự hào để chia sẻ nó với cộng đồng Internet; tuy nhiên, Tôi không cần <i>bị giữ Có trách nhiệm</i> Cho bất kỳ Thiệt	Tôi (thì) hãnh diện để chia sẻ nó với cộng đồng Internet; tuy nhiên, tôi không nên ( <b>thì</b> ) <b>giữ chịu trách nhiệm</b> cho bất

hại nào Trực tiếp Hoặc Gián tiếp gây ra Bởi sự Sử dụng Hoặc Dùng sai (của) Phần mềm này.	kỳ sự hư hại nào trực tiếp hay là gián tiếp gây ra bởi việc sử dụng hay là việc sử dụng lầm của phần mềm này.
I must be informed if anybody <b>would like to</b> distribution this <b>program for</b> commercial purpose.	
Tôi phải được thông tin rằng nếu bất cứ ai <i>thích tôi</i> phân phối (cái) này <i>lập trình cho</i> mục đích thương mại.	Tôi phải được báo tin nếu bất kỳ ai <b>xin</b> sự phân phối <b>chương trình</b> này <b>vì</b> mục đích thương mại.
6. Registration	
6. Sự Đăng ký	6. sự đăng ký
I don't earn any tangible benefit <b>out of</b> this program, but as a human being, I still crave for some recognition.	
Tôi không kiếm được bất kỳ lợi ích hữu hình nào <i>ra khỏi</i> chương trình này, nhưng như một con người, Tôi vẫn còn khao khát cho sự đoán nhận nào đó.	Tôi không kiếm được bất kỳ lợi ích có thể sờ mó được nào <b>ngoài</b> chương trình này, nhưng như một con người, tôi vẫn còn nài xin cho vài sự nhận ra.
The <b>thought that</b> my program is being used by more people from all over the world motivates me.	
<i>Tư duy mà</i> chương trình (của) tôi đang được sử dụng bởi nhiều người hơn từ khắp (nơi) thế giới thúc đẩy tôi.	<b>Ý nghĩ mà</b> chương trình của tôi đang được sử dụng bởi người nhiều từ khắp thế giới làm cho động cơ thúc đẩy tôi.
Registration also allow me to send you updates <b>about</b> this program.	
Sự Đăng ký cũng cho phép tôi gửi cho bạn cập nhật <i>khoảng</i> chương trình này.	Sự đăng ký cũng cho phép tôi gửi bạn những cập nhật <b>về</b> chương trình này.



❑ Nguồn : Phần Hướng Dẫn Sử Dụng Trong Windows 2000

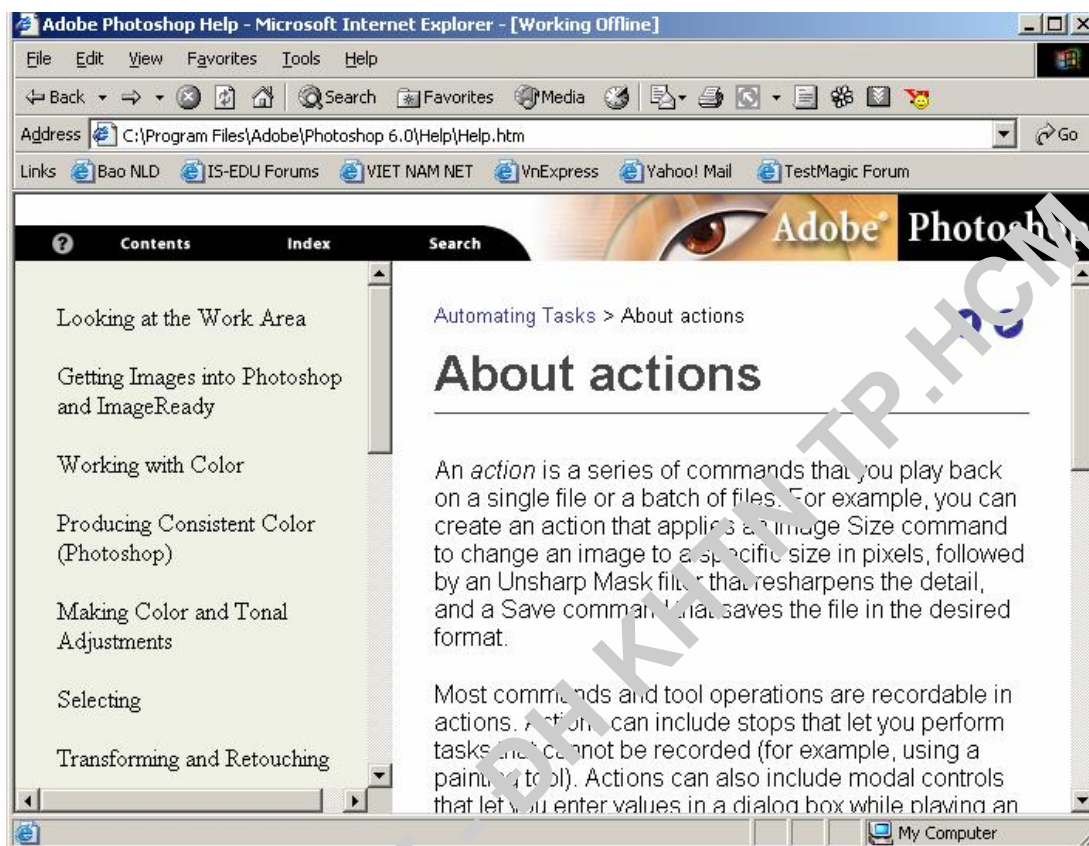


Câu tiếng Anh	Câu dịch của chương trình VCLEVT
Internet, e-mail, and communications overview	Mạng INTERNET, thư tín điện tử, và tổng quan truyền thông
Windows 2000 provides you with quick and easy access to the Internet, no matter what programs you are running or tasks you are performing.	Windows 2000 cung cấp bạn với (sự) truy cập dễ và nhanh tới Internet, dù những chương trình cái gì bạn đang chạy hay là những nhiệm vụ bạn đang thực hiện.
The Active Desktop allows you to customize your workspace and the Address bar helps you to connect to the	Loại để bàn đang hoạt động cho phép bạn điều chỉnh không gian làm việc của bạn và thanh địa chỉ giúp bạn dễ nối tới

Internet from any window.	Internet từ bất kỳ cửa sổ nào.
You can also find a variety of tools that help you to communicate with people and other computers.	Bạn cũng có thể tìm thấy nhiều dạng công cụ mà giúp bạn để truyền tin với người và những máy tính khác.
Most people connect to the Internet by using a network connection or an Internet service provider (ISP).	Hầu hết người nối tới Internet bởi sử dụng một kết nối mạng hay là một nhà cung cấp dịch vụ Internet ( ISP ).
An ISP supplies a service number that you can dial from your computer to log on to the Internet server.	Một ISP cung cấp một số dịch vụ mà bạn có thể quay số từ máy tính của bạn tới truy cập vào tới máy chủ Internet.
Once you connect to the system, you have access to the Internet, e-mail, and any other services supplied by your ISP.	Một khi bạn nối tới hệ thống, bạn có (sự) truy cập tới Internet, thư tín điện tử, và bất cứ những dịch vụ khác được cung cấp bởi ISP của bạn.
Your ISP also furnishes you with the details you need to configure an Internet connection on your computer.	ISP của bạn cũng trang bị bạn với những chi tiết bạn cần để cấu hình một kết nối Internet vào máy tính của bạn.
If you use a network connection, your system administrator provides this information for you.	Nếu bạn sử dụng một kết nối mạng thì người quản trị hệ thống của bạn cung cấp thông tin này trong bạn.
With the communications tools included in Windows 2000, you can use your computer to send e-mail, handle phone calls, send a fax, or conduct a meeting with a video conference.	Với những công cụ truyền thông bao gồm trong Windows 2000, bạn có thể sử dụng máy tính của bạn để gửi thư tín điện tử, xử lý những cuộc gọi điện thoại, gửi một fax, hay là dẫn (điện) một cuộc họp với một hội nghị hình.
For example, you can use Phone Dialer	Ví dụ như, bạn có thể sử dụng quay số

to answer phone calls or join a video conference in your company.	điện thoại để trả lời những cuộc gọi điện thoại hay là nối một hội nghị hình trong công ty của bạn.
Outlook Express helps you send mail to friends or conduct business around the world.	Express toàn cảnh giúp bạn gửi thư tới những bạn hay là dẫn (điện) kinh doanh quanh thế giới.
Communication links are usually established through a phone line, the Internet, or a network.	Những mối liên kết Communication thường thường được thiết lập xuyên qua một đường điện thoại, Internet, hay là một mạng.
Some of the hardware needed for establishing these links with your computer include a modem, ISDN adapter, sound card, microphone, or camera.	Một vài phần cứng được cần cho thiết lập những mối liên kết này với máy tính của bạn bao gồm một bộ điều giải, ISDN bộ thích ứng, âm thanh tám mạch, micrô, hay là máy ảnh.

❑ **Nguồn : Trang Web hướng dẫn sử dụng của Phần mềm Photoshop 6.0**



EVTran Version 2.0	VCL EVT
<b>About actions</b>	
<i>Khoảng</i> những hoạt động	<b>Về</b> những hoạt động
An action is <i>a series of commands</i> that you play back on a single file or a batch of files.	
Một hoạt động là <i>một đợt (của) những lệnh</i> mà bạn chơi sau trên (về) một hàng một hoặc một lô (của) những hồ sơ.	Một hoạt động là <i>một loạt của lệnh</i> mà bạn chơi vào một tập tin đơn hay là một lô của tập tin lại.
For example, you can create an action that applies an Image Size command to change an image to a <i>specific</i> size in <i>pixels</i> , <i>followed</i> by an Unsharp Mask <i>filter</i>	

that resharpen the detail, and a Save command that saves the file in the desired format.	
<p>Chẳng hạn, bạn có thể tạo ra một hoạt động mà áp dụng một lệnh Kích thước Hình ảnh để thay đổi một Hình ảnh tới một Kích thước <i>đặc biệt</i> trong những <i>điểm, đi theo</i> bởi một Mặt nạ Unsharp <i>lọc</i> mà mài sắc lại (mà) chi tiết, và một lệnh Lưu trữ mà <i>Cất giữ hồ sơ</i> trong mong muốn định dạng.</p>	<p>Ví dụ như, bạn có thể tạo ra một hoạt động mà áp dụng một lệnh kích thước hình ảnh để thay đổi một hình ảnh tới một kích thước <i>cụ thể</i> trong những <i>điểm sáng, được theo</i> bởi một <i>bộ lọc</i> cái mặt nạ Unsharp mà mài sắc lại chi tiết, và một lệnh Save mà lưu tập tin trong dạng được mong muốn.</p>
Most commands and tool <i>operations</i> are <i>recordable</i> in actions.	
<p>Đa số các <i>thao tác</i> lệnh và công cụ (thì) <i>có thể bản ghi</i> trong những hoạt động.</p>	<p>Hầu hết những lệnh và những <i>hoạt động</i> công cụ (thì) <i>có thể ghi được</i> trong những hoạt động.</p>
Actions can include <i>stops</i> that let you perform tasks that <i>cannot be recorded</i> (for example, using a painting tool).	
<p>Những hoạt động có thể bao gồm <i>dừng</i> mà để cho bạn thực hiện những nhiệm vụ mà <i>không thể là lại cột bằng dây</i> (chẳng hạn, sử dụng một công cụ bức tranh).</p>	<p>Những hoạt động có thể bao gồm những <i>điểm dừng</i> mà để bạn thực hiện những nhiệm vụ mà <i>không thể được ghi</i> ( ví dụ như, sử dụng một công cụ bức họa ).</p>
Actions can also include modal controls that let you <i>enter</i> values in a dialog box while playing an action.	
<p>Những hoạt động có thể cũng bao gồm những điều khiển phương thức mà để cho bạn <i>vào</i> những giá trị trong một hộp thoại trong khi chơi một hoạt động.</p>	<p>Những hoạt động cũng có thể bao gồm những sự điều khiển kiểu mà để bạn <i>nhập</i> những giá trị trong một cái hộp đối thoại trong khi chơi một hoạt động.</p>

<p>Actions form the basis for droplets, small applications that automatically process all files that are dragged onto <i>their</i> icon.</p>	
<p>Những hoạt động hình thành cơ sở cho những giọt nhỏ, những ứng dụng nhỏ mà tự động xử lý tất cả các hồ sơ mà được kéo lên trên biểu tượng của <i>họ</i>.</p>	<p>Những hoạt động hình thành cơ sở cho những giọt nhỏ, những ứng dụng nhỏ mà một cách tự động xử lý tất cả những tập tin mà được kéo vào biểu tượng <i>của chúng</i>.</p>
<p><b>Both</b> Photoshop <b>and</b> ImageReady <b>ship</b> with a number of predefined actions, although Photoshop has significantly more actions than ImageReady.</p>	
<p><b>Cả</b> Photoshop <b>lẫn</b> con tàu ImageReady với một số hoạt động đặt sẵn, mặc dù Photoshop Có một cách đáng kể nhiều hoạt động hơn hơn ImageReady.</p>	<p><b>Cả hai</b> Photoshop <b>và</b> ImageReady <b>giao</b> hàng với một số của hoạt động tiền định nghĩa, mặc dù Photoshop có những hoạt động nhiều có ý nghĩa so với ImageReady.</p>
<p>You can use these actions as is, customize them to meet your <b>needs</b>, or create new actions.</p>	
<p>Bạn có thể sử dụng những hoạt động này như nó có, tùy biến chúng để đáp ứng <u>yêu cầu</u> (của) Bạn, hoặc tạo ra những hoạt động mới.</p>	<p>Bạn có thể sử dụng những hoạt động này như thì/là, điều chỉnh chúng để đáp ứng những <u>nhu cầu</u> của bạn, hay là tạo ra những hoạt động mới.</p>

## Phụ Lục 6. Một Số Ví Dụ So Sánh

EVTran Technic version 2.0	VCLEVT
He is a teacher and he has many pupils.	
Anh ta là một giáo viên và Anh ta có nhiều học sinh.	Anh ấy là một giáo viên và anh ấy có nhiều học sinh.
He is not only handsome but also intelligent.	
Anh ta (thì) không chỉ dễ coi mà còn thông minh.	Anh ấy (thì) không những đẹp trai mà còn thông minh.
He is <b>at</b> his home alone and he is reading a book.	
Anh ta ở (tại) một mình <i>về(ở)</i> nhà (của) anh ấy và Anh ta đang đọc một (quyển) sách.	Anh ấy (thì) <b>tại</b> nhà của anh ấy một mình và anh ấy đang đọc một cuốn sách.
He <b>was scolded</b> by his wife but he will be appreciated by his lover.	
Anh ta <i>được trách mắng</i> bởi vợ (của) anh ấy nhưng Anh ta <i>sẽ được đánh giá</i> bởi người yêu (của) anh ấy.	Anh ấy <b>bị mắng</b> bởi người vợ của anh ấy nhưng anh ấy <i>sẽ được đánh giá cao</i> bởi người yêu của anh ấy.
I <b>asked</b> you to come here <b>in order that</b> I <b>ask</b> you several questions.	
Tôi <i>hỏi</i> bạn để đến ở đây <i>trong thứ tự</i> mà Tôi <i>hỏi</i> bạn vài câu hỏi.	Tôi <b>yêu cầu</b> bạn đến đây <b>ngõ hầu</b> tôi <i>hỏi</i> bạn vài câu hỏi.
Immediately, they <b>clear body out</b> in the hospital.	
Ngay lập tức, chúng <i>làm sạch thân thể ở</i> ngoài trong bệnh viện.	Ngay lập tức, chúng <b>dọn sạch</b> trong bệnh viện.
This is the most important project.	

Đây là dự án quan trọng nhất.	Đây là dự án quan trọng nhất.
She plays tennis and he <b>does too</b> .	
Cô ấy cũng chơi quần vợt và anh ta <i>làm</i> .	Cô ấy chơi quần vợt và anh ấy <b>cũng vậy</b> .
She needs a new <b>dress</b> .	
Cô ấy cần một mới <i>mặc quần áo</i> .	Cô ấy cần một <b>cái áo</b> mới.
We like to listen to <b>that band</b> playing in the city <b>park</b> every Sunday.	
Chúng ta thích nghe <i>mà rằng buộc</i> chơi trong thành phố <i>bãi</i> mỗi Chủ nhật.	Chúng tôi thích để lắng nghe <b>băng đờ</b> chơi trong <b>công viên</b> thành phố mọi chủ nhật.
You've ever known her <b>lose her temper</b> .	
Bạn đã từng biết cô ấy <i>mất tâm tính (của) cô ấy</i> .	Bạn đã từng biết cô ấy <b>mất bình tĩnh</b> .
The blister on my heel made me painful	
Vết bong rộp trên (về) gót (của) tôi làm cho tôi là đau đớn.	Vết bong giộp trên gót chân của tôi làm tôi đau.
She confessed to me <b>that</b> she had loved him.	
Cô ấy thú tội tới tôi <i>mà</i> Cô ấy đã yêu anh ấy.	Cô ấy thú nhận tới tôi <b>rằng</b> cô ấy yêu anh ấy.
She reminds him <b>that</b> the meeting <b>is on</b> saturday.	
Cô ấy nhắc nhở anh ấy <b>rằng</b> cuộc gặp (đã) <i>bật</i> thứ bảy.	Cô ấy nhắc nhở anh ấy <b>rằng</b> cuộc họp (thì) <b>vào</b> thứ bảy.
The scandal taught <b>her that</b> silence is gold.	
Vụ bê bối dạy sự yên lặng <i>đó (của) cô ấy</i> là vàng.	Chuyện tai tiếng dạy <b>cô ấy rằng</b> sự yên lặng là vàng.



old driver	
bộ(người) điều khiển cũ (già)	Tài xế (thì) già
green field	
đồng xanh	Cánh đồng xanh
<b>field</b> is green	
<i>lĩnh vực</i> (thì) xanh lục	<b>Cánh đồng</b> (thì) xanh
<b>This</b> old man and woman are <b>printer</b>	
Ông già và phụ nữ <b>này</b> là <b>máy in</b> .	Người đàn ông già <b>đấy</b> và đàn bà là <b>thợ in</b>
I <b>can can</b> a <b>can</b> .	
Tôi <b>Có thể</b> <b>Có thể</b> Một <b>Có thể</b> .	Tôi <b>có thể</b> <b>đóng hộp</b> một cái <b>hộp</b> .
An <b>old</b> police <b>is sitting</b> on an <b>old chair</b> .	
Một cảnh sát <b>cũ</b> (già) <u>đang ngồi</u> trên (về) một <b>cũ</b> (già) <b>chủ trì</b> .	Một cảnh sát <b>già</b> <u>đang ngồi</u> trên một <b>cái ghế cũ</b> .
The <b>old driver</b> <b>is driving</b> an <b>old</b> car	
Bộ(người) điều khiển cũ (già) <u>đang điều khiển</u> một ô tô cũ (già).	<b>Tài xế già</b> <u>đang lái xe</u> một xe hơi <b>cũ</b> .
She installed <b>old driver</b> for my <b>printer</b> .	
Cô ấy thiết đặt bộ(người) điều khiển cũ (già) cho <b>máy in</b> (của) tôi.	Cô ấy cài đặt <b>trình điều khiển cũ</b> cho <b>máy in</b> của tôi.
These men are <b>printers</b> .	
Những người đàn ông này là <b>những máy in</b> .	Những người đàn ông này là <b>những thợ in</b> .

Vietnam will organize <b>SEA Games</b> in 2003.	
Việt nam sẽ tổ chức <i>những trò chơi Biển</i> vào 2003.	Vietnam sẽ tổ chức <b>SEA Games</b> vào <b>năm 2003</b> .
I have waited for her for many years.	
Tôi đã đợi cô ấy nhiều năm.	Tôi đợi cô ấy trong nhiều năm.
They <b>are reading</b> books	
Chúng là những (quyển) sách đọc.	Chúng <b>đang đọc</b> những cuốn sách.
He <b>is hated</b> by his friends.	
Anh ta <i>được căm thù</i> bởi những bạn (của) anh ấy.	Anh ấy <b>bị ghét</b> bởi những bạn của anh ấy.
There are a lot of beautiful places <b>in Vietnam</b> .	
Có nhiều chỗ đẹp <i>trong Việt nam</i> .	Có nhiều chỗ đẹp <b>ở Vietnam</b> .
These books <b>are written</b> by me <b>in english</b> .	
Những (quyển) sách này <i>được viết</i> bởi tôi <i>trong tiếng Anh</i> .	Những cuốn sách này <i>được viết</i> bởi tôi <b>bằng tiếng Anh</b> .
I eat <b>rice</b> .	
Tôi ăn <i>gạo</i> .	Tôi ăn <b>com</b> .
They plant <b>rice</b> , and he imports <b>rice</b> .	
Chúng gieo trồng <i>gạo</i> , và anh ta nhập khẩu <i>gạo</i> .	Chúng trồng <b>lúa</b> , và anh ấy nhập <b>gạo</b> .