MỤC LỤC

DANH SÁCH HÌNH VẼ

LỜI CẢM ƠN

Trước hết em xin gửi lời cảm ơn đến thầy Ths. Nguyễn Trịnh Đông, người đã hướng dẫn em rất nhiều trong suốt quá trình tìm hiểu nghiên cứu và hoàn thành khóa luận này từ lý thuyết đến ứng dụng. Sự hướng dẫn của thầy đã giúp em có thêm được những hiểu biết về WordNet và ứng dụng của từ điển trong thực tiễn.

Đồng thời em cũng xin cảm ơn các thầy cô trong khoa Công nghệ thông tin - Trường ĐHDL Hải Phòng, những người đã nhiệt tình giảng dạy và truyền đạt những kiến thức cần thiết trong suốt thời gian em học tập tại trường để em có thể hoàn thành tốt khóa luận này.

Sau cùng, em xin gửi lời cảm ơn đến gia đình, bạn bè đã tạo mọi điều kiện để em xây dựng thành công khóa luận này.

Em xin chân thành cảm ơn !

Hải Phòng, ngày 09 tháng 07 năm 2010.

Người viết báo cáo

Trang

Nguyễn Thị Thu Trang

MỞ ĐẦU

Trong những năm gần đây, trong lĩnh vực xử lý ngôn ngữ tự nhiên thì xử lý ngữ nghĩa chiếm vai trò quyết định tính chính xác của các mô hình xử lý ngôn ngữ tự nhiên. Để xử lý ngữ nghĩa chúng ta phải có cơ sở tri thức ngữ nghĩa từ vựng của ngôn ngữ cần xử lý, hiện nay tiếng Anh đã có cơ sở tri thức ngữ nghĩa hoàn chỉnh. Trong đó hệ thống cơ sở tri thức ngữ nghĩa từ vựng WordNet là phổ biến nhất hiện nay. Đây là mạng ngữ nghĩa đồ sộ hơn 110.000 synset tiếng Anh. Các nhà ngôn ngữ học, tâm lý học và tin học đã bỏ ra gần 20 năm để xây dựng hệ thống này và ngày nay chúng vẫn còn được nâng cấp về số lượng và chất lượng. Tuy nhiên với các ngôn ngữ khác, hệ thống như vậy vẫn chưa có nhiều. Điển hình là tiếng Việt, hiện nay chúng ta vẫn chưa có hệ thống cơ sở tri thức ngữ nghĩa từ vựng như vậy. Do đó vấn đề cấp bách hiện nay là phải xây dựng một hệ thống ngữ nghĩa của tiếng Việt cho máy tính nếu chúng ta muốn phát triển các ứng dụng về xử lý ngôn ngữ tự nhiên.

WordNet phân biệt giữa danh từ, động từ, tính từ và trạng từ, vì họ làm theo quy tắc ngữ pháp khác nhau. Danh từ là một loại từ phổ biến và phổ dụng trong mọi ngôn ngữ. Đến nay, đã có nhiều cách phân lớp danh từ tiếng Việt theo các tiêu chí khác nhau, nhưng ít nhiều các cách này đều mang tính chủ quan và chỉ được thực hiện trên một số ít các ví dụ cụ thể. Tuy nhiên, trong thực tế, khi phân giải ngữ nghĩa của một danh từ tiếng Việt trong một ngữ cảnh bất kì, chúng ta lại cần đến một hệ thống phân lớp hoàn chỉnh cho tất cả các danh từ tiếng Việt theo những ý niệm chung nhất trong tư duy của con người. Việc xây dựng một hệ thống phân lớp như thế đã được thực hiện thành công lần đầu tiên đối với tiếng Anh qua mạng WordNet, và cũng chính từ đây, các mạng tương tự cho tiếng Pháp, Tây Ban Nha, Đức, Hoa, Nhật, . đã được hình thành trên cơ sở mạng này. Việc xây dựng một mạng từ vựng tương tự WordNet có nhiều ý nghĩa. Nó cho việc phát triển các ứng dụng xử lý ngôn ngữ tiếng Việt, cho các nghiên cứu về ngôn ngữ học tiếng Việt. Do vậy, trong bài báo cáo này, em trình bày về phương pháp xây dựng từ điển danh từ Tiếng Việt dựa theo từ điển WordNet.

Đồ án được chia thành các phần như sau:

Chương 1: *Tìm hiểu đề tài và phương pháp tiếp cận.*

Chương 2: *Tìm hiểu về tiếng Việt và WordNet áp dụng trong việc xây dựng từ điển danh từ tiếng Việt*

Chương 3: *Xây dựng mô hình tổ chức dữ liệu cho WordNet tiếng Việt và thực nghiệm.*

CHƯƠNG 1: GIỚI THIỆU

1.1 Đặt vấn đề

Vấn đề xử lý ngôn ngữ tự nhiên, xử lý ngữ nghĩa chiếm vai trò rất quan trọng trong ứng dụng xử lý ngôn ngữ tự nhiên. Để xây dựng được một từ điển Tiếng Việt cho máy tính đáp ứng được nhu cầu cấp thiết được rất nhiều nhà nghiên cứu quan tâm. Hiện nay trên thế giới, WordNet là một hệ thống cơ sở tri thức khổng lồ về ngôn ngữ học của từ vựng tiếng Anh, được coi là nguồn tài nguyên quan trọng nhất có sẵn cho các nhà nghiên cứu ngôn ngữ học, tính toán, phân tích văn bản, và nhiều lĩnh vực liên quan. Cũng chính từ đây, các cơ sở dữ liệu tri thức ứng dụng trong việc xây dựng các từ điển tiếng Pháp, Tây Ban Nha, Đức, Hoa, Nhật…, theo cấu trúc lưu trữ từ vựng của WordNet. Để áp dụng WordNet xây dựng từ điển danh từ Tiếng Việt chúng ta cần giải quyết vấn đề sau:

- Nắm được cấu trúc cơ sở tri thức từ vựng trong WordNet.
- Xây dựng mô hình tổ chức dữ liệu cho từ điển tiếng Việt dựa trên WordNet và sau đây được gọi là WordNet tiếng Việt.

1.1.1 Tổ chức cơ sở ngữ nghĩa từ vựng

Để xử lý ngôn ngữ tự nhiên trên máy tính, chúng ta cần có những cơ sở ngữ nghĩa từ vựng của ngôn ngữ đó .Thông thường các cơ sở ngữ nghĩa từ vựng này là một từ điển phân loại của các từ hay nhóm từ, tức là mỗi từ sẽ được gắn một hay nhiều nghĩa. Đặc biệt có nhiều cơ sở tri thức còn đưa ra mối quan hệ về ngữ nghĩa giữa các nhãn ngữ nghĩa đó. Các mối quan hệ này có thể là quan hệ toàn thể, bộ phận, thừa kế .... Có một số mô hình cơ sở tri thức ngữ nghĩa từ vựng lại chú trọng vào một số lĩnh vực hẹp hay phạm vi nhỏ. Nhưng các cơ sở tri thức ngữ nghĩa từ vựng là thành phần không thể thiếu được với một hệ thống xử lý ngôn ngữ tự nhiên và sự ra đời của WordNet.

Hệ thống mạng ngữ nghĩa WordNet: Hệ cơ sở tri thức ngữ nghĩa từ vựng này được bắt đầu phát triển vào năm 1993. Bao gồm 152059 cụm từ được phân bố vào 115.424 synsets và 44 chủ đề. Quan trọng hơn nữa hệ thống này còn xây dựng một mạng lưới các mối quan hệ giữa các ý niệm với nhau. Đây có thể xem là một mạng ngữ nghĩa đầy đủ và hoàn thiện nhất.

Hiện nay mỗi khi sử dụng các cơ sở tri thức ngữ nghĩa từ vựng về thế giới thực, người ta thường sử dụng WordNet. Trong đề tài này em tập trung vào tìm hiểu cấu trúc cơ sở lưu trữ tri thức của WordNet từ đó ứng dụng vào việc xây dựng từ điển danh từ tiếng Việt.

1.1.2 Mô hình gán nhãn ngữ nghĩa

Sau khi đã chọn được quy tắc phân chia của mạng ngữ nghĩa, chúng ta phải tìm mô hình để gán nhãn của các (cụm) từ tiếng Việt vào mạng ngữ nghĩa WordNet .

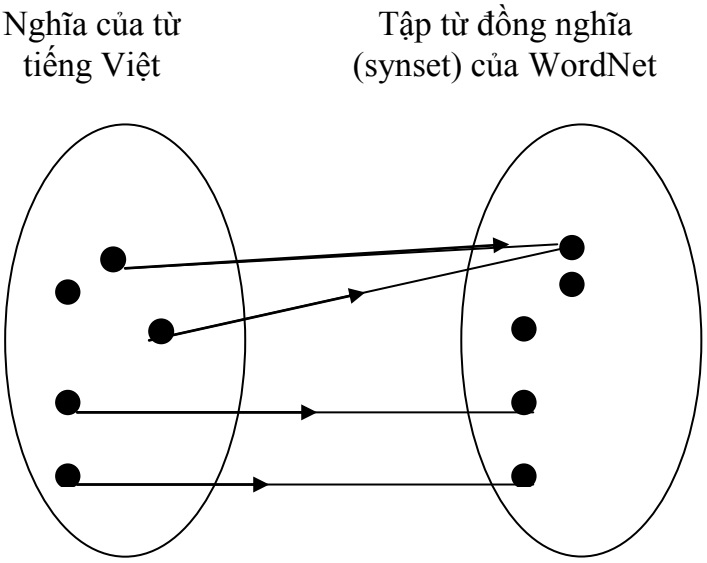
Chúng ta có thể đặc tả bài toán như sau :

V : tập hợp các từ tiếng Việt

Ω: là tập hợp các synnet trong WordNet. Synnet trong WordNet có thể được xem là hình vị hóa của ý niệm. Hay nói rõ hơn synnet là một nhóm các từ có chung một ý niệm trong WordNet.

δ : là ánh sạ từ $V \rightarrow \Omega$

Với δ [illegible] $\in \Omega, v \in V$



Hình 1: Ánh xạ n-1 từ nghĩa của từ tiếng Việt và synset trong tiếng Anh

Giả thiết, chúng ta có tiên đề sau:

Ánh xạ từ $V \rightarrow \Omega$ là ánh xạ 1-n. Một số nghĩa của từ tiếng Việt có thể cùng chung một synset trong WordNet. Tuy nhiên, một synset trong WordNet chỉ có thể ánh xạ thành một nghĩa trong tiếng Việt. Do đó bài toán được quy về là tìm ánh xạ δ.

1.2 Các hướng tiếp cận truyền thống

Trên thế giới đã có nhiều cách tiếp cận để giải quyết cho từng ngôn ngữ cụ thể. Mỗi phương án được đề xuất đều xuất phát từ nguồn tài nguyên hiện có của ngôn ngữ đó. Với các ngôn ngữ phổ biến, đã có nhiều hệ thống phân loại từ vựng, hệ thống WordNet của ngôn ngữ ấy được xây dựng theo cách tiếp cận sử dụng các từ điển phân loại hiện có và xây dựng bản ánh xạ tương ứng. Tuy nhiên

với các ngôn ngữ ít phổ biến, chưa có các từ điển phân loại, thì mô hình khả thi được đề xuất là xây dựng từ điển phân loại dựa trên từ điển đơn ngữ,… dĩ nhiên, độ chính xác cũng kém hơn.

1.2.1 Xây dựng từ điển phân loại dựa trên từ điển MRD

Phương pháp này sử dụng một từ điển đơn ngữ để rút trích các liên kết giữa các từ và các nghĩa. Các mô hình dạng này sẽ phân tích phần giải thích của một từ đơn trong từ điển đơn nghĩa để tìm ra các thuật ngữ chính. Dựa vào phân loại của các thuật ngữ này chúng ta có thể xác định được phân loại của các từ.

Hướng tiếp cận này có thể áp dụng cho mọi ngôn ngữ, do hầu như ngôn ngữ nào cũng có từ điển đơn ngữ của ngôn ngữ mình.Tuy nhiên các phương pháp này không cho kết quả chính xác do chúng ta cần phải giải quyết các vấn đề của từ điển đơn ngữ như phân loại thiếu phân loại không hợp lý và có rất ít kĩ thuật khử nhập nhằng của các phân loại.

1.2.2 Sử dụng các liên kết trong các từ điển đã có

Các phương pháp này thuộc các tiếp cận dạng này sử dụng cho các ngôn ngữ đã có một từ điển đã được phân loại. Khi đó chúng ta có thể sử dụng từ điển dạng này kết hợp với các phân loại khác nhau để tạo lên một cấu trúc hoàn chỉnh đa ngôn ngữ .

Tuy nhiên, khi áp dụng phương pháp này để tìm ánh xạ giữa hai ngôn ngữ khác nhau kết quả thu được độ chính xác không cao, không khả quan nhiều.

1.2.3 Sử dụng ánh xạ từ điển MRD song ngữ

Phương pháp này sẽ tìm cách liên kết từ tiếng Anh tương ứng trong từ điển song ngữ với synset tương ứng trong WordNet. Hướng tiếp cận này thu được kết quả rất tốt nếu chúng ta sử dụng các quan hệ giữa các Synset như đồng nghĩa, phản nghĩa bao hàm ...

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Các vấn đề về Ngôn Ngữ học

So với các ngôn ngữ khác, hiện nay, tiếng Việt chúng ta còn nhiều quan điểm khác nhau về các vấn đề ngôn ngữ học. Có nhiều trường phái thiên về vấn đề làm sao cho máy tính dễ xử lý và có nhiều trường phái lại rất khó áp dụng máy tính để xử lý .

2.1.1 Từ trong Tiếng Việt

2.1.1.1 Hình vị

Trong tiếng Việt đơn vị này còn được gọi là tiếng.Về các mặt ngữ âm ngữ nghĩa, ngữ pháp nó đều có giá trị quan trọng .

Hình vị là đơn vị ngôn ngữ nhỏ nhất có nghĩa và/hoặc có giá trị (chức năng) về mặt ngữ pháp.

Về giá trị ngữ âm

Đứng về mặt ngữ âm thì hình vị thường trùng với âm tiết. Xét về mặt ngữ âm, âm tiết là đơn vị ngữ âm rất dễ nhận diện, vì nó là đơn vị phát âm tự nhiên ứng với sự căng lên và trùng xuống của dây thanh, và được phân cách bởi một khoảng ngắt hơi.

Về bình diện về chữ viết

Trong chữ quốc ngữ tức chữ Việt hiện nay, mỗi âm tiết được ghi thành một chữ, nên ở mặt chữ viết, âm tiết cũng dễ được nhận ra. Mỗi âm tiết trong tiếng Việt đều có một thanh.

Về giá trị ngữ nghĩa

Đứng về mặt ngữ nghĩa thì hình vị cũng là đơn vị nhỏ nhất có thể có nghĩa. Đơn vị ngữ âm ở bậc thấp hơn, là âm vị, thì không thể có nghĩa, mà chỉ có giá trị khu biệt nghĩa. Chẳng hạn, âm vị */-a-/*và âm vị */-t-/* riêng lẻ tự nó không có nghĩa gì, nó chỉ có giá trị khu biệt nghĩa: *ta-ma-xa-na ...; ta-tu-ti-to...* thanh điệu cũng có giá trị như một âm vị tự nó không có nghĩa . nhưng nếu được kết hợp lại thành tiếng hoàn chỉnh , thành âm tiết như ta hay tạ, má hay ma... thì có thể thành những đơn vị nhỏ nhất có nghĩa. Trong tiếng Việt, có những loại hình vị khác nhau như sau:

- Loại hình vị độc lập, như: *đất, nước, nhà, xe, máy; làm, ăn, ngủ, nhìn, học; xấu, tốt, mới , cũ...* đó là loại hình vị tự nó có nghĩa có thể dùng để gọi tên sự vật, hiện tượng, tính chất và có thể được dùng để tạo từ , từ một tiếng, đơn vị ở bậc trực tiếp cao hơn

- Loại hình vị không độc lập, như *thủy, thổ, hỏa, sơn; thực, khán, thính, tọa ; mỹ, lạc hí, nộ*... Đây là loại hình vị, tuy tự nó có nghĩa nhưng không dùng để gọi tên sự vật, hiện tượng, không có khả năng vận dụng tự do để tạo thành câu dược. Chúng ta không chỉ vào nước mà nói rằng: đó là *thủy,* mà nói đó là *: nước*; chúng ta cũng không thể nói là *uống thủy* mà nói: *uống nước*. Nhưng loại tiếng này có thể được dùng để cấu tạo những đơn vị ở bặc trực tiếp cao hơn, tức là từ, như *thực phẩm, mỹ nghệ; tàu thủy, lính thủy*. Và đó là từ hai tiếng.
- Loại hình vị không có nghĩa tự thân, như *long, lanh (long lanh), bâng, khuâng (bâng khuâng), lẽ (lặng lẽ), dàng (dễ dàng)… …*tuy không tự nó có nghĩa, nhưng có tác dụng tạo nghĩa khu biệt hoặc tạo nghĩa cho đơn vị ở bậc trực tiếp cao hơn, tức là từ, như *long lanh, bâng khuâng, lặng lẽ, dễ dàng.* Đây cũng là từ hai tiếng.

Về giá trị ngữ pháp

Ngữ pháp bao gồm những quy tắc cấu tạo từ, cấu tạo câu. Hình vị là đơn vị ngữ pháp được dùng để cấu tạo từ. Có một số trường hợp cấu tạo từ sau đây:

Cấu tạo từ một tiếng. Đây là một trường hợp một hình vị độc lập được dùng làm một từ. Chẳng hạn: *nước* là một hình vị được dùng làm từ. Có thể dùng từ một tiếng này để cấu tạo câu. Ví dụ: có thể nói câu *tôi uống nước* hay nói *nước rất trong.*

Cấu tạo từ hai tiếng hay nhiều tiếng. Đó là trường hợp có sự kết hợp giữa hai thành tố, mà hai thành tố này có thể là hai hình vị độc lập, hoặc không độc lập, hay không có nghĩa tự thân kết hợp với nhau, và có sự gán bó tương dối chặt chẽ về mặt nội dung và hình thức. Chẳng hạn: *Nhà nước, xóm làng, quàn áo; thợ sơn, hoa hồng, cá thu; quốc gia, giang sơn, huynh đệ; tàu thủy,bình thủy, lính thủy; dễ dàng, gọn nhẹ, lẹ làng, long lanh, lai rai, lơ thơ; bồ hóng, bù nhìn, cà phê; chợ búa, tre pheo, khách khứa...*

Cũng có những trường hợp hơn hai tiến kết hợp với nhau thành từ. Ví du: *hợp tác xã, câu lạc bộ, cộng sản chủ nghĩa, chủ nghĩa xã hội...*

2.1.1.2 Từ

Từ là đơn vị sẵn có trong ngôn ngữ. Từ là đơn vị nhỏ nhất, cấu tạo ổn định, mang nghĩa hoàn chỉnh, được dùng để cấu thành nên câu. Từ có thể làm tên gọi của sự vật (danh từ), chỉ các hoạt động (động từ), trạng thái, tính chất (tính từ)... Từ là công cụ biểu thị khái niệm của con người đối với hiện thực.

Trong ngôn ngữ học, từ là đối tượng nghiên cứu của nhiều cấp độ khác nhau, như cấu tạo từ, hình thái hoc, ngữ âm học, phong cách học, cú pháp học...

"Từ tiếng Việt được cấu tạo bởi những hình vị tiếng Việt''. Từ tiếng Việt ở đây cũng bao gồm; từ đơn, từ ghép, từ láy và từ ngẫu hợp. Ngoài quan niệm chính về từ tiếng Việt như trên, họ còn gán tư cách từ cho một số ít đơn vị tiếng Việt còn đang tranh cãi về tư cách từ của nó dựa theo sự từ vựng hóa trong tiếng Anh, Chẳng hạn: *nhà_tranh (line), xe_đạp (bicycle), máy tính (computer), đường_thẳng (line)* ...là từ *;còn nhà gạch (brick house),* .. không là từ.

Giống như cách trình bày của WordNet, trong luận văn, em sẽ dùng thêm kí hiệu dấu gạch liền ở dưới (underline "_") để nối các hình vị của từ tiếng Việt đó. Ví dụ: *học_sinh, máy_tính, màn hiển_thị , đo_lường từ xa,...*

2.1.2 Từ trong Tiếng Anh

Tiếng Anh thuộc loại ngôn ngữ biến hình (inflextion), do đó từ trong tiếng Anh có thể dễ dàng xác định thông qua dấu khoảng cách. Từ trong tiếng Anh có thể có nhiều cách biến đổi như sau:

Biến cách:

Có 8 loại biến cách như sau:

- Số nhiều (danh từ) (thêm-s)
- Ngôi thứ ba số ít (động từ) (thêm-s)
- Sở hữu cách (tính từ) (thêm – ' s)
- Hiện tại phân từ (thêm-ing)
- Quá khứ (thêm-ed)
- Quá khứ phân từ (thêm-ed)
- So sánh hơn (thêm-er)
- So sánh nhất (thêm-est)

Đặc điểm của cách biến đổi này là sự biến đổi này không được nối tầng và có thể áp dụng cho tất cả các từ. Quan trọng hơn, cách biến đổi này không làm thay đổi từ loại của từ gốc.

Dẫn xuất :

Có 2 dang của cách biến đổi này là dạng biến đổi tiền tố và hậu tố:

- Tiền tố :không làm thay đổi từ loại của từ
- Hậu tố : thường làm thay đổi từ loại của từ

2.1.3 Nghĩa của từ:

Theo[5] thì ngôn ngữ có hai mặt: mặt biểu hiện (âm thanh) và mặt được biểu hiện (nội dung). Nghĩa của từ thuộc về mặt thứ hai.

Ví dụ, từ CÂY trong tiếng Việt cóVỏ ngữ âm như ta đọc lên ([kej 1]), và từ này có nội dung, có ý nghĩa của nó.

2.1.3.1 Nghĩa của từ là gì?

Khái niệm nghĩa (sense) của từ đã được nêu ra từ lâu và cũng đã có nhiều cách hiểu, nhiều định nghĩa khác nhau. Để trả lời câu hỏi chính: " nghĩa của từ là gì" trước hết ta phải trở lại bản chất tín hiệu của từ. Từ là tín hiệu; nó phải "nói lên", phải đại diện cho, phải được người sử dụng quy chiếu về một cái gì đó.

Khi một người nghe hoặc nói một từ nào đó, họ gán nó vào đúng sự vật có tên gọi là từ đó như cả cộng đồng vẫn gọi; đồng thời ít nhiều họ cũng biết được những đặc trưng bản chất của sự vật đó; và họ sử dụng từ đó trong giao tiếp đúng với các mẹo luật mà ngôn ngữ có từ cho phép; ta nói rằng họ hiểu được nghĩa của từ đó.

Ví dụ: một người Việt hoặc không phải là người Việt, nói hoặc nghe một từ như CÂY chẳng hạn; mà họ có thể :

- Quy chiếu, gắn được từ cây vào mọi cái bất kì trong thực tại đời sống.
- Ít nhiều cũng biết được đại khái như: cây là loài thực vật mà phần thân, lá đã phân biệt rõ; ví dụ như: cây mía, cây tre,...
- Dùng từ CÂY trong giao tiếp, phát ngôn...đúng với các quy tắc tiếng Việt.

Ta nói rằng : họ hiểu được nghĩa của từ CÂY trong tiếng Việt.

Cho tới nay, đa số nhà nghiên cứu đều quan niệm nghĩa của từ là những liên hệ. Tuy nhiên, đó không phải là những liên hệ logic tất yếu; mà là những liên hệ phản ánh, mang tính quy ước, được xây dựng bởi những cộng đồng người bản ngữ.

Mỗi khi học nghĩa của một từ, chúng ta đều học bằng cách liên hội từ với những cái mà từ đó chỉ ra (trước hết là sự vật, hiện tượng, hành động hoặc thuộc tính… mà từ đó làm tên gọi cho nó). Mặt khác, nghĩa của từ cũng được học thông qua hoặc liên quan với vô vàn tình huống giao tiếp ngôn ngữ mà từ đó được sử dụng.

Thuở nhỏ, ta thấy một cái cây bất kì chẳng hạn. Ta hỏi đó là cái gì và được trả lời là cái cây. Dần dần, nay với cây này mai với cây khác, ta liên hội được từ CÂY của tiếng Việt với chúng. Thế rồi bước tiếp theo nữa, ta dùng được từ "cây" trong các phát ngôn như *trồng cây, chặt cây, tưới cây, cây đổ, cây rau, cây hoa… v*à tiến tới hiểu cây là loài thực vật, có thân, rễ, lá hoặc hoa, quả… vậy là ta hiểu được nghĩa của từ CÂY.

Đến đây có thể phát biểu vắn tắt lại như sau: nói chung, *nghĩa của từ là những liên hệ được xác lập trong nhận thức của chúng ta giữa từ và những cái mà nó (từ) chỉ ra (những cái mà nó làm tín hiệu cho).*

2.1.3.2 Nghĩa của từ tồn tại ở đâu?

Ta đã thừa nhận và chứng minh bản chất tín hiệu của từ, rằng nó có hai mặt; mặt hình thức vật chất âm thanh và mặt nội dung ý nghĩa; hai mặt này gắn bó với nhau như hai mặt của một tờ giấy, nếu không có mặt này thì cũng không có mặt kia. Vậy nghĩa của từ tồn tại trong từ; nói rộng ra là trong hệ thống ngôn ngữ. Nó là cái phần nửa làm cho ngôn ngữ nói chung, và từ nói riêng, trở thành những thực thể vật chất - tinh thần.

Nghĩa của từ tồn tại trong ý thức, trong bộ óc của con người. Trong ý thức, trong tư duy của con người chỉ có những hoạt động nhận thức, hoạt động tư duy mà thôi. Điều này ngụ ý rằng: trong ý thức, bộ óc trí tuệ của con người chỉ tồn tại sự hiểu biết về nghĩa của từ chứ không phải là nghĩa của từ.

Từ những điều trên đây, suy tiếp ra rằng những lời trình bày. Giải thích trong từ điển, cái mà ta vẫn quen gọi là nghĩa của từ trong từ điển, thực chất là những lời trình bày tương đối đồng hình với sự hiểu biết của ta về nghĩa của từ mà thôi.

2.1.3.3 Các thành phần nghĩa của từ

Từ có liên hệ với nhiều nhân tố, nhiều hiện tượng. Bởi thế, nghĩa của từ cũng không phải chỉ có một thành phần, một kiểu loại. Khi nói về nghĩa của từ, người ta thường không phân biệt các thành phần nghĩa sau đây:

- Nghĩa biểu vật (denotative meaning): là liên hệ giữa từ với sự vật (hoặc hiện tượng. thuộc tính, hành động…) mà nó chỉ ra. Bản thân sự vật, hiện tượng, thuộc tính, hành động…đó, người ta gọi là biểu vật hay cái biểu vật (detonat). Biểu vật có thể hiện thực hoặc phi hiện thực;hữu hình hay vô hình;có bản chất vật chất hoặc phi vật chất. ví dụ:đất, trời, mưa, nắng, nóng, lạnh, ma, quỷ, thánh ,thần, thiên đường, địa ngục…
- Nghĩa biểu niệm (fignificative meaning): là liên hệ giữa từ với ý (hoặc ý nghĩa, ý niệm-sigification- nếu chúng ta không cần phân biệt nghiêm ngặt mấy tên gọi này). Cái ý đó người ta gọi là cái biểu niệm hoặc biểu niệm (sự phản ánh các thuộc tính của biểu vật vào trong ý thức của con người).

Ngoài hai thành phần trên đây, khi xác định nghĩa của từ,người ta còn phân biệt hai thành phần nghĩa nữa. Đó là nghĩa ngữ dụng và nghãi cấu trúc.

Nghĩa ngữ dụng (pragmatical meaning), còn được gọi là nghĩa 90 biểu thái, nghĩa hàm chỉ (connotative meaning) 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 trừ 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 (paradigmatial axis), trục ngữ đoạn (syntagmatical axis). Quan hệ trên trục đối vị cho ta xác định được giá trị của từ, khu biệt này với từ khác; còn quan hệ trên trục ngữ đoạn cho ta xác định được ngữ trị (valence)- khả năng kết hợp- của từ.

Trọng tâm chú ý phân tích, miêu tả của từ vựng - ngữ nghĩa học là biểu niệm chứ không phải là các thành phần khác (chúng chỉ được lưu ý trong những trường hợp cần thiết mà thôi). Vì vậy, ở đây khi không thật bắt buộc xác định rành mạch về mặt thuật ngữ, thì chúng ta sẽ nói đến nghĩa với nội dung được hiểu là nghĩa biểu niệm cho giản tiện.

2.1.3.4 Phân biệt nghĩa của từ với khái niệm

Cần phân biệt nghĩa của từ với khái niệm. Nghĩa và khái niệm gắn bó với nhau rất mật thiết, nhưng nói chung là chúng không trùng nhau.

Khái niệm là kết quả của quá trình nhận thức, phản ánh những đặc trưng chung nhất, khái quát nhất và bản chất nhất của sự vật, hiện tượng. Người ta có được khái niệm chủ yếu là nhờ những khám phá, tìm tòi khoa học. Nội dung của một khái niệm có thể rất rộng, rất sâu, tiệm cận đến chân lý khoa học; và có thể được diễn đạt bằng hàng loạt các ý kiến, nhận xét. Mặt khác, rõ ràng là không phải khái niệm nào cũng được phản ánh bằng từ; nó có thể được biểu hiện bằng hơn một từ. Ví dụ: *nước cứng; tổ hợp quỹ đạo; máy gặt đập liên hoàn; công nghệ sinh học…*

Nghĩa của từ cũng phản ánh những đặc trưng chung, khái quát của sự vật, hiện tượng do con người nhận thức được trong đời sống thực tiễn tự nhiên và xã hội. Tuy nhiên, nó có thể chưa phải là kết quả của nhận thức đã tiệm cận tới chân lý khoa học. Vì thế, sự vật, hiện tượng nào mà càng ít được nghiên cứu khám phá thì nhận thức về nó được phản ánh trong nghĩa của từ gọi tên nó, càng xa với khái niệm khoa học.

Bên cạnh đó, ta thấy rằng không phải từ nào cũng phản ánh khái niệm (các thán từ và các từ công cụ ngữ pháp chẳng hạn) và trong nghĩa của từ còn có thể hàm chứa cả sự đánh giá về mặt này hay mặt khác, có thể chứa cả cảm xúc và thái độ của con người…

Để tiện so sánh, chúng ta phân tích từ nước của tiếng Việt. Khái niệm khoa học [hóa học] về *nước* là: hợp chất của oxy và hidro mà trong thành phần của mỗi phân tử nước, có hai nguyên tử hidro và một nguyên tử oxy.

Nghĩa "nôm" của từ *nước* có thể được miêu tả dưới dạng từ điển ngắn gọn là: chất lỏng không màu, không mùi và hầu như không vị, sẵn có trong ao hồ, sông suối…

Miêu tả như thế thạt chưa đủ. Rất nhiều thứ, loại (biểu vật) được người Việt quy về loại nước mà chỉ cần chúng bảo đảm thuộc tính lỏng; còn có nước nhiều hay ít; mùi vị thế nào; thậm chí có nước hay không..đều không quan trọng. Chẳng hạn: *nước biển, nước mắn, nước sốt, nước dứa, nước ép hoa quả.*

Phở nước (đối lập với phở xào)

Mỡ nước (đối lập với mỡ khô)

Phân tích như trên đây chứng tỏ rằng nghĩa và khái niệm không đồng nhất.

2.1.3.5 Phân tích nghĩa của từ

Phân tích, miêu tả cho được cấu trúc nghĩa của từ là một trong những nhiệm vụ và mục đích hàng đầu của việc nghiên cứu từ vựng ngữ nghĩa.Trong lĩnh vực này, người ta đã đề xướng nhiều phương pháp phân tích nghĩa của từ, nhưng thường gặp và dễ dùng nhất là phương pháp sử dụng ngữ cảnh.

- **Ngữ cảnh là gì?**

Khi dùng ngôn ngữ để giao tiếp, người ta thường nói ra những câu, những phát ngôn, chứ không phải là những từ rời rạc. Tại đó, các từ kết hợp với nhau theo những quy tắc và chuẩn mực của ngôn ngữ. Cũng trong câu hoặc phát ngôn cụ thể, người ta mới biết được rằng: Tại trường hợp, hoàn cảnh cụ thể này, từ có nghĩa gì (tức là nó bộc lộ nghĩa nào trong số các nghĩa của nó).

Ví dụ: khi ta nghe thấy chỉ một từ "chắc" trong tiếng Việt thôi, thì không thể biết được người nói muốn nói gì tới nghĩa nào đó của từ này. Thế nhưng, từng nghĩa một của từ "chắc" sẽ xuất hiện rất rõ ràng, nếu ta nghe thấy nó trong các phát ngôn, những chuỗi từ đại loại như sau:

Lúa đã chắc hạt; Nhà xây rất chắc; Lời nói chắc như đinh đóng cột; Ông này chắc đã có con lớn; Anh làm thế, dễ người ta không biết đấy chắc;...

Định nghĩa về ngữ cảnh được phát biểu như sau:

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

(Định nghĩa này cho thấy rằng ngữ cảnh có thể tối thiểu là một từ, tối đa là một chuỗi lớn hơn, có khả năng ứng với một câu, một phát ngôn,...)

Sở dĩ từ bộc lộ một nghĩa xác định nào đó trong ngữ cảnh chứa nó là vì trong mỗi ngữ cảnh, từ thể hiện khả năng kết hợp từ vựng và khả năng kết hợp ngữ pháp của mình. Khả năng kết hợp ngữ pháp của từ là khả năng nó có thể đứng vào một vị trí nhất định trong những cấu trúc nhất định nào đó. Nói khác đi, đó là khả năng từ có thể tham gia vào những cấu trúc ngữ pháp nào đó.

Ví dụ: trong tiếng Việt, động từ có khả năng kết hợp với các từ: *đã, đang, sẽ, sắp, lại, vừa*,... ở đằng trước; và: *xong, rồi, mãi*,... ở đằng sau (ví dụ: *đang đi, làm mãi*,...).

Nếu từ thuộc lớp ngữ pháp nào, thuộc từ loại nào,... thì sẽ được quy định cho những khả năng tương ứng, những vị trí tương ứng trong các mô hình cấu trúc ngữ pháp.

Ngược lại, khả năng kết hợp từ vựng của từ lại là khả năng kết hợp giữa một nghĩa của từ này với một nghĩa của từ khác, sao cho tổ hợp được tạo thành phải đúng với thực tại, phù hợp với logic và thói quen sử dụng ngôn ngữ của người bản ngữ.

Ví dụ, người Việt vẫn nói: *ăn cơm, học bài, nhắm mắt*,... và cũng nói: *bây giờ đang mùa thu, trông vẫn còn con gái như ai, nhà này cũng năm tầng*;... mà không thể nói: *ăn bài, học cơm, nhắm miệng, bây giờ đang nhà*,...

Có những từ có khả năng kết hợp từ vựng rất rộng, nhưng có những từ thì khả năng đó lại hẹp hoặc vô cùng hẹp. Chẳng hạn, các động từ: *nhắm, nháy, nghển, kiễng, phưỡn, mấp máy*,... có khả năng kết hợp với từ vựng rất hẹp. Mỗi động từ đó chỉ kết hợp được với một hoặc vài danh từ khác mà thôi.

Có thể diễn giải khả năng kết hợp từ vựng của từ như sau:

– Khi có hai từ A và B kết hợp với nhau chẳng hạn, thì không phải lúc đó tất cả các nghĩa của A đều hiện lên và kết hợp với tất cả các nghĩa của B.

– Nếu ta hình dung mỗi từ có một "phổ" nghĩa:

A = a, b, c,... B = x, y, z,... thì sự kết hợp AB về mặt từ vựng có thể là kết hợp nghĩa *ax, by, bz, bx, ay, az*,... tuỳ trường hợp cụ thể mà AB phản ánh.

Ví dụ: Xét kết hợp "*che đầu*" trong câu *Trời mưa một mảnh áo bông che đầu*, ta thấy:

Từ "*che*" có hai nghĩa:

1.(...)

2.Ngăn cho khỏi bị một tác động nào đó từ bên ngoài.

Từ "*đầu*" có 4 nghĩa:

1.Bộ phận thân thể người, động vật nằm ở vị trí trên cùng hoặc trước nhất.

2.(...)

3.(...)

4.(...)

Khả năng kết hợp từ vựng của "*che*" với "*đầu*" trong trường hợp này là kết hợp của nghĩa *che (2)* với nghĩa *đầu (1)*.

Những phân tích vừa nêu trên chứng tỏ rằng: Khả năng kết hợp từ vựng của các từ quy định và cho phép chúng có kết hợp với nhau được hay không. Ngược lại, thông qua các kết hợp cụ thể từ này với các từ khác, ta có thể phát hiện dần từng nghĩa riêng của từ, tiến tới xác định được cả một "phổ", cả một cơ cấu của nghĩa từ. Điều này cũng tương tự như hình thái học phát hiện tất cả các từ hình của từ trong hoạt động lời nói để rồi quy chúng về cái gọi là từ vị vậy.

- **Cách dùng phương pháp phân tích theo ngữ cảnh**

Phân tích ngữ cảnh

Đây là bước đầu tiên, bắt buộc phải thực hiện vì đó là tư liệu làm việc. Trước hết phải xác định được các ngữ cảnh (có chứa từ mà ta cần phân tích) trong các loại văn bản thành văn thuộc các loại hình phong cách chức năng khác nhau. Sau đó trích các ngữ cảnh đó ra và tập hợp lại.

Phân loại ngữ cảnh

Khi đã thu được số lượng ngữ cảnh đủ nhiều, đáng tin cậy, phản ánh đủ hết các nghĩa của từ, chúng ta sẽ phân loại. Những ngữ cảnh nào cùng làm hiện thực hoá một nghĩa của từ (tức là trong những ngữ cảnh đó, từ xuất hiện với cùng một nghĩa), thì được xếp vào một nhóm gọi là nhóm ngữ cảnh cùng loại.

Nếu việc phân loại ngữ cảnh làm càng chuẩn xác thì sẽ càng tạo điều kiện thuận lợi cho việc tách nghĩa của từ đa nghĩa, bởi vì, từ càng đa nghĩa thì càng phức tạp, càng khó xử lí.

Xét các ngữ cảnh chứa từ "*say*" như sau đây làm ví dụ:

1. Má hồng không thuốc mà *say*.
2. Đất *say* đất cũng lăn quay
   Trời *say* trời cũng đỏ gay ai cười.
3. *Say* thuốc lào.
4. *Say* xe.
5. *Say* sóng.
6. Da anh đen cho má em hồng
   Cho duyên em thắm, cho lòng anh *say*.

Các ngữ cảnh trên đây được phân tích thành hai nhóm:

- Nhóm 1 gồm ngữ cảnh 1, 6,

- Nhóm 2 gồm ngữ cảnh 2, 3, 4, 5

Phân tích nghĩa

Đối với từ đơn nghĩa, nhiệm vụ ở bước này là so sánh với các từ khác cùng nhóm (tương đồng, tương cận hoặc tương phản với nó) để phát hiện các nghĩa tố cần yếu trong cấu trúc nghĩa của từ.

Riêng từ đa nghĩa, vấn đề phức tạp hơn. Cùng với việc so sánh, phát hiện các nghĩa tố cần yế của từng nghĩa, thì việc tách ra bao nhiêu nghĩa trong toàn bộ cơ cấu nghĩa từ phải được tiến hành trước một bước. Ta cần phải làm những bước sau đây:

- Xác định nghĩa gốc của từ (trong thế tương quan lưỡng phân nghĩa gốc – nghĩa phái sinh). Nghĩa gốc của từ có thể là một nghĩa từ nguyên, nhưng cũng có thể chỉ là một nghĩa phái sinh rồi phái sinh tiếp tục ra nghĩa khác. Ví dụ tính từ "*bạc*" có 3 nghĩa:

1. Mỏng manh, ít ỏi, không trọn vẹn: *Mệnh bạc,...*
2. Ít ỏi, sơ sài (trái với *hậu*): *Lễ bạc lòng thành,...*
3. Không nhớ ơn nghĩa, không giữ được tình nghĩa trọn vẹn trước sau như một: *Ăn ở bạc với bố mẹ,...*

Nghĩa (1) của tính từ "*bạc*" là nghĩa từ nguyên, vốn từ gốc Hán.

Nghĩa (2) và (3) của nó đều được phái sinh từ nghĩa (1).

Thế nhưng trong tiếng Việt hiện đại, nghĩa (3) mới là nghĩa phổ biến nhất. Dựa vào nghĩa gốc, ta phát hiện các nghĩa phái sinh và các quy tắc chuyển nghĩa của chúng.

- Xác định nghĩa không thường trực (nghĩa ngữ cảnh) nếu có, để loại trừ khỏi phạm vi mà chúng ta đang quan tâm. Như vậy, chỉ những nghĩa thường trực mới được đưa vào phân tích xử lí. Ngay trong khi phân loại ngữ cảnh, thực chất là đã bao hàm việc tách nghĩa từ trong đó rồi. Vì vậy, nếu phân loại ngữ cảnh mà chuẩn xác thì số nhóm ngữ cảnh cùng loại nói chung là ứng với số nghĩa khác nhau của từ.

2.1.4 Quan hệ đồng âm, đồng nghĩa

2.1.4.1 Từ đồng âm (hynonymic words)

Là những từ trùng nhau về hình thức ngữ âm nhưng khác nhau vè nghĩa. Ví dụ: nhóm đồng âm: *to, too, two* trong tiếng Anh; *đường (sắt), đường (ăn)* trong tiếng Việt. Hiện tượng đồng âm xảy ra chủ yếu giữa các từ với nhau, ít khi nào quan sát được hiện tượng này ở các cụm từ với nhau. Từ đồng âm có mặt trong ngôn ngữ là một tất yếu vì số lượng âm thanh mà con người phát ra được

và dùng làm vỏ ngữ âm cho các từ, dù có nhiều đến mấy cũng chỉ có giới hạn của nó.

Trong tiếng Việt, do đặc trưng loại hình đơn lập của tiếng Việt quy định nên nó có những đặc điểm sau:

- Những từ là đồng âm với nhau thì luôn đồng âm trong tất cả các bối cảnh được sử dụng.
- Đồng âm giữa từ với từ là kết quả của đồng âm tiếng với tiếng, điều này được khai thác tong nghệ thuật chơi chữ như tên riêng "Hitle" được tách thành hai tiếng và liên hội với hai động từ "hit" và "le".

Tùy theo từng ngôn ngừ mà các từ đồng âm được phân loại khác nhau:

- Tiếng Anh:
  - Đồng âm, đồng tự, như: *can (có thể) – can (đóng hộp)*
  - Đồng âm, không đồng tự, như: *son – sun*
  - Đồng tự, không đồng âm, như: *tear (xé) – tear (nước mắt).*
- Tiếng Việt:
  - Đồng âm từ giới từ, như: *đường (con đường)- đường (ăn)* (về mặt từ vựng) và *chỉ (khâu)- chỉ (trỏ)* (về mặt từ vựng - ngữ pháp)
  - Đồng âm tiếng với tiếng:như *"than"* trong câu: *"nhà của đề lầm than con thơ trẻ lấy ai rèn cặp"*.

Các nhóm từ đồng âm có thể không tìm được lý do (tù bản ngữ) hay tìm được lý do (từ vay mượn, tách nghĩa cảu từ đa nghĩa, biến đổi ngữ âm). Ta cần phân biệt từ đa nghĩa với từ đồng âm:

- Đồng âm: khác nguồn gốc nhưng trùng ngữ âm.
- Đa nghĩa: cùng nguồn gốc và trùng ngữ âm.

Việc nghiên cứu và khảo sát kỹ các từ đồng âm cả về lý thuyết và thực tiễn đều rất cần thiết; đặc biệt trong lĩnh vực từ điển và dịch máy.

2.1.4.2 Từ đồng nghĩa (synonymic words)

Là những từ tương đồng nhau về nghĩa; khác nhau về âm thanh và có 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 các nhóm đồng nghĩa, trong:

- Tiếng Anh*: start, begin, commence (bắt đầu)*
- 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: nhóm: *"yếu, yếu ớt, yếu đuối,.."* có từ *"yếu"* là từ trung tâm.

2.1.5 So sánh từ tiếng Việt và từ tiếng Anh về hình thái

Theo [4], do sự khác nhau về loại hình (biến cách và đơn lập) nên từ tiếng Việt và từ tiếng Anh khác nhau cả về mặt từ vựng hóa (lexicalization) và hình thái học (morphology). Do đó, không thể lúc nào cũng có sự tương ứng (1-1) giữa từ tiếng Anh với từ tiếng Việt. Trái lại, ánh xạ này phải là m-n, nghĩa là 1 hay nhiều từ tiếng Anh có thể tương ứng với một hay nhiều từ tiếng Việt.

2.1.5.1 Sự khác biệt về từ vựng hóa

Một từ tiếng Anh có thể được dịch thành một cụm gồm nhiều từ tiếng Việt và ngược lại. Đây là ánh xạ m-n. Ví dụ:
Ánh xạ *1-1*: từ *"display"* và nghĩa tiếng Việt tương ứng của nó là *"hiển_thị"*
Ánh xạ *1-n*: từ *"display"* và nghĩa tiếng Việt tương ứng của nó là *"màn hiển_thị"*
Ánh xạ *m-1*: từ *"display*" và nghĩa tiếng Việt tương ứng của nó là *"thực_hiện"*
Ánh xạ *m-n*: từ *"display*" và nghĩa tiếng Việt tương ứng của nó là *"gọi điện_thoại"*

2.1.5.2 Sự khác biệt về hình thái học

Bên cạnh về từ vựng, sự khác nhau về loại hình ngôn ngữ cũng tạo nên sự khác nhau về hình thái của từ tiếng Anh và tiếng Việt. Chính điều này hình thành nên ánh xạ m_n khi dịch các từ mở rộng này sang tiếng Việt.

Xét về mặt biến cách của từ tiếng Anh

Trong khi từ tiếng Anh được mở rộng theo kiểu biến cách bằng các hình vị phụ tố thì các từ tiếng Việt mở rộng bằng các tư hư. Vì vậy, ứng với một từ trong tiếng Anh, khi chưa biến cách, ánh xạ của tiếng Việt tương ứng là 1-1 (nếu không tính yếu tố khác biệt về từ vựng hóa), nhưng sau khi biến cách nó lại là 1-n.

Bảng 2-1: *Sự khác biệt về mặt biến cách giữa từ tiếng Anh và từ tiếng Việt.*

	Ý nghĩa ngữ pháp	**Tiếng Anh**		**Tiếng Việt**	
		Hậu tố	Ví dụ	Từ Hư	Ví dụ
1	Danh từ số nhiều	N + - s	Books; Two students	Những/ các + N;Φ	Những/ các cuốn_sách; hai_sinh viên
2	Động từ ngôi 3 số it	V + - s	He sleeps	Φ	Anh ấy ngủ

3	Sở hữu cách	X's Y	John's book; teachers' book	Y của X	Cuốn sách của John; các cuốn_sách của những giáo viên
4	Hiện phân từ	V-ing	Sleeping	Đang V	Đang ngủ
5	Quá khứ	V- ed	Worked	Đã V	(Đã) làm việc
6	Quá khứ phân từ	V- en	Spoken	Đã V	(Đã) nói
7	So sánh hơn	Adj-er Adv-er	Shorter Slower	Adj- hơn	Ngắn hơn Chậm hơn
8	So sánh hơn nhất	Adj-est Adv-est	Shortest Slowest	Adj- nhất	Ngắn nhất Chậm nhất

Xét về mặt dẫn xuất của từ tiếng Anh

Bên cạnh sự khác biệt về mặt biến cách như trên, các từ dẫn xuất trong tiếng Anh được hình thành bàng cách sử dụng các hình vị phụ tố dẫn xuất (derivational affixes), còn tiếng Việt dùng từ độc lập hoặc trật tự từ để thể hiện các ý nghĩa từ vựng mới. Điều này khiến từ ánh xạ tiếng Anh và từ tiếng Việt trong trường hợp này trở thành 1-n nếu phần nghĩa tiếng Việt tương ứng của phụ tố dẫn xuất này là từ thuần Việt. Nếu phần nghĩa tiếng Việt tương ứng của phụ tố này là những từ Hán-Việt, thì ánh xạ liên kết từ Anh- Việt trong trường hợp này vẫn là 1-1.

Ví dụ: Ánh xạ 1-1: *Reader: độc_giả, illegal: bất_hợp_pháp, normalize: bình_thường_hóa, non-government: Phi_chính_phủ,…*

Ánh xạ 1-n*: caller: người gọi, illegal: không hợp_pháp, normalize: làm cho bình_thường, readable: có_thể đọc được,..*

Những khác biệt do dặc thù của tiếng Việt

Cuối cùng, do đặc thù của ngôn ngữ tiếng Việt, nên các danh từ đơn thể trong tiếng Việt thường đi kèm với loại từ (classifier) tương ứng của nó, như: *cuốn/ quyển + sách, bức/ lá + thư,…* (tiếng Hoa cũng có đặc điểm này). Các loại từ này *(cuốn, quyển, bức, lá, cái, con, ..)* là các phó danh từ và gắn nó với từ tiếng Việt tương ứng để hình thành nên một cụm từ.

2.1.6 So sánh từ tiếng Việt và từ tiếng Anh về mặt ngữ pháp

Thường trong một ngôn ngữ, người ta có thể phân ra hai lớp từ cơ bản mà người ta gọi là thực từ và hư từ. Mỗi lớp thực từ và hư từ bao gồm một số loại như: danh từ (noun, nom); động từ (verb, verbe); tính từ (adjective, adjectif); đại từ (pronoun, pronom…).

2.1.6.1 Hệ thống nhãn từ loại trong tiếng Anh

Đã ổn định và gồm 8 từ loại: danh từ (noun); động từ (verb); tính từ (adjective), đại từ (pronoun), trạng từ (adverb), giới từ (preposition), liên từ (conjunction) và thán từ (interjection).

2.1.6.2 Hệ thống nhãn từ loại trong tiếng Việt

Hiện nay, có nhiều xu hướng về cách phân chia từ loại trong tiếng Việt. Tuy nhiên, các cách phân chia phổ biến nhất vẫn được các nhà ngôn ngữ học chấp nhận đó là chia từ loại tiếng Việt thành hai loại: thực từ và hư từ.

Thực từ (từ có nghĩa thực sự) gồm danh từ, động từ, tính từ.

Hư từ (từ chỉ có nghĩa ngữ pháp) gồm một số nhỏ các từ bao gồm phụ từ (phó từ), kết từ (liên từ và giới từ), ngoài ra còn có đại từ, trợ từ, số từ, loại từ, cảm từ và từ chỉ hướng.

2.1.6.3 Đối chiếu nhãn từ loại tiếng Anh và tiếng Việt

Do tiếng Anh và tiếng Việt khác nhau về loại hình nên khi xét về từ loại, hai ngôn ngữ này cũng có sự khác nhau.

Về từ loại

Tiếng Việt có 12 đơn vị từ loại trong khi tiếng Anh có 8 đơn vị. Trong đó, sự khác biệt lớn nhất giữa hai ngôn ngữ này là ở các hư, với các thực từ thì sự khac biệt này không lớn lắm. May mắn, WordNet tiếng Anh chỉ gồm 4 từ loại (danh từ, động từ, tính từ và phó từ) và luận văn chỉ đề cập đến phần danh từ nên chúng tôi không đi sâu vào sự khác nhau của các hư từ.

Bảng đối chiếu nhãn từ loại

Ánh xạ giữ từ loại tiếng Anh và từ loại tiếng Việt không là ánh xạ 1-1, nghĩa là từ X trong tiếng Anh có nghĩa là Y thì không chắc từ loại X là từ loại của Y. Bản đối chiếu từ loại giữa hai ngôn ngữ như sau [4]:

Với từ gốc

Bảng 2-2: Bảng đối chiếu nhãn từ loại của từ gốc tiếng Anh và tiếng Việt

Từ pháp tiếng Anh	Từ pháp tiếng Việt
Danh từ (NN):*table, person,,..* Danh từ riêng (NP): *John, Hanoi,..* Danh từ (NN):*attention, help,…*	Danh từ (N): *bàn, người,..* Danh từ riêng (Nn): *Tuấn, Hà_nội..* Động từ (V): *chú_ý, giúp_đỡ,..*
Trạng từ (RB): *above, below, here,…*	Danh từ vị trí(Np): *trên, dưới, đây,..*
Động từ (VB):*eat, learn, …*	Động từ (V):*ăn, học,….*

Tính từ (JJ):*big, good,…* Tính từ (JJ):*every, each,..* Tính từ (JJ):*electrich, national,…*	Tính từ (J):*lớn, tốt,…* Phó từ (R):*mọi, từng,..* Danh từ (N):*điện, quốc_gia,..*
Đại từ (PP):*I,you,he,…*	Đại từ (P):*tôi, anh, anh ấy,…*
Trạng từ (RB):*strongly, slowly,…* Trạng từ (RB):*still, just, …; already,..* Trạng từ (RB):*perhaps, of course,…* Trạng từ (RB):*even,…* Trợ động từ (MD): *can, may, will,..*	Tính từ (J):*mạnh_mẽ, chậm chạp,…* Phó từ (R):*vẫn, vừa,…;đã, đang, sẽ,…* Phó từ (R):*có_lẽ, tất_nhiên* Trợ từ (M):*cả, chính,…* Phó từ / tính từ:*có _thể, sẽ,…*
Giới từ (IN): *in, on, by, of,…*	Giới từ (I):*trong, tại, bởi, của*
Liên từ (CC): *and, or, although,…*	Liên từ (C): *và, hay, dù,…*
Thán từ (UH): *oh !..*	Cảm từ (U): *ôi!*
Cardinal (CD):*one. Two,…* Tính từ (JJ): *few, several, some,…*	Số từ (Q): *một, hai,…* Số từ (Q): *các, những, vài*
Định từ (DT*): a, an, the,…*	Loại từ (L):*cái, con, cuốn,…*
Tiền chỉ định từ (PDT):*this, that,…*	Đại từ (P);*đây, đó, này, nọ,…*
Tiểu từ (RP):*up, on, off, to,…*	Từ chỉ hướng (D): *lên, xuống,…*

Với từ biến cách

Bảng 2-3 bảng đối chiếu từ loại của từ biến cách của từ tiếng Anh và từ tiếng Việt

	Ý nghĩa ngữ pháp	**Từ pháp tiếng Anh**	**Từ pháp tiếng Việt**
1	Danh từ số nhiều	Books/NNS; Two/CD students/NNS	*Những/*Qcuốn/Lsách/N;hai/Q sinh_viên/N
2	Động từ ngôi 3 số ít	He/PP sleeps/VBZ	
3	Sở hữu cách	John/NP’s/POSbook/NN; eachers/NNS’/POS Books/NNS	Cuốn/L-sách/N *của*/I John/Nn; Các/Q cuốn/L-sách/N *của*/I những/Q giáo_viên/N
4	Hiện phân từ	Sleeping/ VBD	Đang/R ngủ/V
5	Quá khứ	Worked/VBD	(đã/R) làm_việc/V
6	Quá khứ phân từ	Spoken/ VBN	(đã/R) nói/V
7	So sánh hơn	Shorter/ JJR Slower/ RBR	Ngắn/J hơn/J Chậm/J hơn/J
8	So sánh hơn nhất	Shortest/ JJS Slowest/ RBS	Ngắn/J nhất/J Chậm/ J nhất/J

Với tù dẫn xuất

Như đề cập ở phần trên, với các trường hợp dẫn xuất sử dụng tiền tố, sẽ không xảy ra sự biến đổi từ loại của từ. Trong khi, với các trường hợp dẫn xuất hậu tố, sự chuyển đổi từ loại của từ sẽ thay đổi.

2.1.7 So sánh từ tiếng Việt và tiếng Anh về mặt nhãn ngữ nghĩa

Như đã trình bày ở phần trên, mỗi từ có thể mang nhiều nghĩa khác nhau, và tùy thuộc vào ngữ cảnh cụ thể mà từ này sẽ mang một nghĩa nhất định nào đó. Chẳng hạn, từ *"bank"* trong tiếng Anh có thể là *"ngân hàng",* hoặc "*bờ sông"* hoặc *"dãy";* từ "*đường"* trong tiếng Việt có thể có nghĩa là "*đường ăn*", hay *"đường đi",*…thậm chí, với các nền văn hóa khác nhau, sẽ xảy ra tình trạng phân chia nhỏ ý niệm. Ví dụ: tiếng Anh chỉ có một từ *"rice"* nhưng ý niệm này trong tiếng Việt lại được chia thành *"thóc", "cơm", "gạo",*...để dễ phân biệt các nghĩa từ vựng khác nhau, các nhà ngôn ngữ học, tâm lý học và tin học đã phân chia toàn bộ các ngữ nghĩa từ vựng thành hệ thống các ý niệm (cây ý niệm) và mỗi ý niệm như vậy được coi là một nhãn ngữ nghĩa của từ.

Chẳng hạn, với danh từ *"bank"* đã đề cập ở trên sẽ có các nhãn ngữ nghĩa là: HOU (*công trình xây dựng nhân tạo*) nếu nó mang ý nghĩa *"ngân hàng";* NAT (*công trình thiên tạo*) nếu nó mang ý nghĩa *"sông*"; GRP (*sự sắp xếp tổ chức*) nếu nó mang ý nghĩa là "*dãy"*. Tương tự từ *"đường*" trong tiếng Việt sẽ có các nhãn ngữ nghĩa như CHM (*hóa chất*) nếu nó mang ý nghĩa là *"đường ăn"*; LIN (*đường nét, dấu vết*) nếu nó mang ý nghĩa là *"đường đi"*...

Đây cũng chính là nền tảng lý luận 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.

Hệ thống các ý niệm (concept) này sẽ là chung nhất cho mọi ngôn ngữ, vì: hệ thống các ý niệm này được xây dựng dựa trên sự phân chia của thế giới khách quan. Trong khi đó, ngôn ngữ là công cụ tư duy, mà tư duy là phản ánh hình ảnh của thế giới khách quan. Chẳng hạn: khái niệm "người chồng" trong các ngôn ngữ khác nhau chắc chắn sẽ được xây dựng từ các ý niệm là "người nam", "người đã trưởng thành", "có gia đình", "có vai trò là chồng trong quan hệ với vợ". Nghĩa là cái biểu đạt trong các ngôn ngữ khác nhau là khác nhau (như : tiếng Việt là CHỒNG, tiếng Anh là "HUSBAND", tiếng hoa là /fu/), nhưng cái được biểu đạt là như nhau. *Vì ý niệm và từ không trùng nhau nên hệ thống ý niệm này đảm bảo được sử dụng cho mọi ngôn ngữ.*

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, phụ thuộc vào mối quan hệ giữa

ngôn ngữ và từ duy của con người. Một số phổ quát ngôn ngữ khác lại là những hiện tượng về dân tộc- ngôn ngữ học, phụ thuộc vào mối quan hệ giữa ngôn ngữ và văn hóa. Các nhà ngôn ngữ chia phổ quát ngôn ngữ thành hai dạng sau:

- Các phổ quát về thực thể: là những nét chung về sự tổ chức các thực thể ngôn ngữ. Chẳng hạn, môi 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 kí hiệu hoặc tới cái biểu đạt hoặc tới cái được biểu đạt, người ta còn chú ý tới các phổ quát ngôn ngữ về kí hiệu, chúng đề cập tới cái quan hệ giữa cái biểu đạt và cái được biểu đạt. Ngoài ra trường nghĩa biểu vật là tập hợp những từ đồng về ý nghĩa biểu vật và trường biểu niệm là một tâp hợp các từ có chung cấu trúc biểu niệm.

Phương pháp đối chiếu nhãn ngữ nghĩa của tiếng Anh và tiếng Việt như sau:

2.1.7.1 Với liên kết 1-1

Với trường hợp này, chỉ việc ánh xạ nhãn ngữ nghĩa giữa hai từ tiếng Anh và tiếng Việt. Tuy nhiên, do có sự chuyển loại từ giữa hai ngôn ngữ Anh- Việt nên có hai trường hợp chúng ta phải quan tâm: nếu từ tiếng Anh là danh từ và từ tiếng Việt là động từ (ví dụ *"assistance, NN*" và *"trợ giúp, V"; "hepl,NN"* và *"giúp đỡ, V")*: Khi đó chuyển từu Tiếng Anh và Việt về dạng gốc (động từ). Sau khi lấy được nhãn ngữ nghĩa của động từ gốc tiếng Anh, ta phải chuyển nhãn ngữ nghĩa này về dạng danh từ tương ứng.

Nếu từ tiếng Anh là tính từ và từ tiếng Việt là danh từ *("electronic,JJ"* và *"điện tử, N"*): khi đó chuyển từ tiếng Anh và Việt về dạng gốc (danh từ). Sau khi lấy được nhãn ngữ nghĩa của danh từ gốc tiếng Anh, ta phải chuyển nhãn ngữ nghĩa này về dạng danh từ tương ứng.

2.1.7.2 Với liên kết 1-n

Với trường hợp này, một từ tiếng Anh được dịch ra bởi nhiều từ tiếng Viêt. Khi đó, vấn đề làm thế nào để chọn đúng nhãn ngữ nghĩa cuả chúng các từ này. Trong trường hợp này, Theo [4] đưa ra phương pháp xem ánh xạ 1-n là n ánh xạ 1-1 và xem xét các ánh xạ nào là ánh xạ hợp lệ (ánh xạ chính).

Nếu chỉ có một ánh xạ hợp lệ, chúng ta sẽ đưa về trường hợp liên kết 1-1. (ví dụ *: ánh xạ "planes/NNS" → "các/ Q máy_bay/N"*thì ánh xạ *planes → máy_bay* là ánh xạ chính).

Nếu có nhiều ánh xạ hợp lệ chúng ta sẽ căn cứ váo nghĩa chính của từ tiếng Việt để xác định ánh xạ hợp lệ, sau đó, chúng ta xem trường hợp này như trường hợp liên kết 1-1. (Ví dụ:*ánh xạ "computerization/NN"→ "sự/N điện_hóa/V"* có ánh xạ hợp lệ là "*computerization/NN"→ "sự/N điện_hóa/V"*)

2.1.7.3 Với liên kết m-1

Với trường hợp này, cụm từ gồm nhiều từ tiếng Anh được dịch ra một từ tiếng Việt. Khi đó, vấn đề là làm thế nào để chọn đúng nhãn ngữ nghĩa của chúng các từ này. Trong trường hợp này, đưa ra hai trường hợp xem ánh xạ m-n là m ánh xạ 1-1 giữa các m từ tiếng Anh và 1 từ tiếng Việt và xem xét các ánh xạ nào là ánh xạ chính [4].

Nếu trong m ánh xạ trên, chỉ có 1 ánh xạ hợp lệ: khi đó ta sẽ chọn ánh xạ này làm ánh xạ chính và đưa trường hợp này trở về trường hợp của ánh xạ 1-1. (*ánh xạ " carry/VB out/RP" → "thực_hiện/V"* có ánh xạ hợp lệ là *" carry/VB out/RP" → "thực_hiện/V").*

Nếu có nhiều ánh xạ hợp lệ, chúng ta sẽ căn cứ vào độ tương đồng hình vị của các nghĩa tiếng Việt của từ tiếng Anh và từ tiếng Việt để xác định ánh xạ hợp lệ , sau đó, chúng ta xem trường hợp này như trường hợp liên kết 1-1. (Ví dụ : ánh xạ *"elder/JJ brother/NN" → "anh/N"* có ánh xạ hợp lệ là *"elder/JJ brother/NN" → "anh/N")*

2.1.7.4 Với liên kết m-n

Với trường hợp này, cụm từ gồm nhiều từ tiếng Anh được dịch thành một cụm từ gồm nhiều từ tiếng Việt. khi đó, vấn đề là làm thế nào để chọn đúng nhãn ngữ nghĩa của chúng các từ này. Trường hợp này bao gồm m ánh xạ 1-n giữa các m từ tiếng Anh và n từ tiếng Việt và xem xét các ánh xạ nào là ánh xạ chính và đưa về một trong ba trường hợp trên.

2.2 WordNet

Năm 1980, Miller và cộng sự tại trường Đại học Princeton (Mỹ) đã xây dựng lên một hệ cơ sở tri thức ngữ nghĩa từ vựng mang tên WordNet. WordNet là một cơ sở dữ liệu tri thức ngữ nghĩa từ vựng bằng tiếng Anh. Người ta xây dựng WordNet dựa trên những lý thuyết về ngôn ngữ_tâm lý theo cách liên tưởng từ ngữ của con người. Từ trong WordNet được phân loại thành danh từ,

động ừ, tính từ, và trạng từ. Chúng được tổ chức thành những tập đồng nghĩa (synset), mỗi tập đồng nghĩa miêu tả, tượng trưng cho một ý niệm cơ bản. Mỗi synset được nối với nhau bởi nhiều loại quan hệ (relation) khác nhau. Hiên nay WordNet đã phát triển lên đến version 2.0 bao gồm hơn 110.000 synsets với hơn 150.000 từ và hệ cơ sở tri thức này miễn phí (cung cấp cả chức năng online và offline) cho các công tác học tập và nghiên cứu. WordNet là một kho tàng tri thức ngữ nghĩa từ cựng khổng lồ và đã được rất nhiều các nhà ngôn ngữ học và ngôn ngữ học_ máy tính khác, ứng dụng thành công trong nhiều bài toán xử lý ngữ nghĩa. Hiện nay, WordNet đang được các nhà khoa học về ngôn ngữ, tâm lý, máy tính trên toàn thế giới tiếp tục khai thác, đóng góp để cải tiến ngày càng hoàn thiện hơn. WordNet có nhiều ưu điểm như: tính khoa học,tính hệ thống, tính mở (open), tính dễ sử dụng, tính phổ thông, tính phát triển… Chính vì vậy, đến nay, đã có một số công trình bản địa hóa WordNet theo ngôn ngữ của một số nước (Pháp, Nhật, Tây ban Nha, Hoa…)

2.2.1 Mô hình WordNet

WordNet là một loại từ điển tương tự từ điển đồng nghĩa. WordNet phân chia từ vựng thành 5 loại: noun, verb, adjective, adverb và funtion words, nhưng thực tế nó chỉ chứa noun, verb, adjective, adverb.

- Danh từ được tổ chức thành các hệ thống phân cấp.
- Động từ được tổ chức theo các mối quan hệ thừa kế có thứ tự.
- Tính từ và trạng từ được tổ chức siêu không gian n chiều (N-dimensional hyperspace)

WordNet phân biệt 2 mối quan hệ: Quan hệ ngữ nghĩa và quan hệ từ vựng.

- Mối quan hệ ngữ nghĩa là mối quan hệ theo nghĩa với nhau, các nghĩa này biểu hiện bằng các synset.
- Mối quan hệ từ vựng là quan hệ giữa các hình thức từ với nhau.

2.2.1.1 Ma trận từ vựng

WordNet không chỉ đơn thuần là nhóm các từ đồng nghĩa hay các từ có quan hệ ngữ nghĩa với nhau thành từng lớp như một số như từ điển LDOCE, LLOCE…mà WordNet còn là một hệ thống các ý niệm có quan hệ nhiều mặt với nhau, tạo thành một mạng lưới phức tạp. Mục tiêu cơ bản của WordNet là chứa các thông tin về ngữ nghĩa của từ, mà hễ nói đến khái niệm hay định nghĩa về "từ" thì chắc chắn lại dẫn đến nhiều ý kiến khác nhau. Chính vì vậy, ngay từ đầu, ta phải xác định các hiểu về đơn vị từ trong WordNet là như thế nào, sau đó ta

tìm hiểu về tập đồng nghĩa (synset) – một thành phần cơ bản của WordNet để áp dụng vào Tiếng Việt.

"Từ" trong WordNet

Trên phương diện ngữ nghĩa học từ vựng, WordNet xem "từ" là một sự kết hợp giữa một ý niệm được từ vựng hóa và một phát ngôn có một vai trò cú pháp trong định nghĩa về "từ" như vậy, chúng ta cần làm rõ thêm: thứ nhất, loại phát ngôn nào có thể tham gia vào trong kết hợp này; thứ hai: bản chất và tổ chức của ý niệm dược từ vựng hóa mà từ thể hiện và thứ ba: những vai trò cú pháp của các từ khác nhau. Chúng ta cần làm ra ba vấn đề trên, nhưng vì mục tiêu của WordNet là tổ chức ngữ nghĩa của từ vựng, chính vì vậy trong khuôn khổ của luân văn này sẽ đề cập đến vấn đề thứ hai, đó là cấu trúc ngữ nghĩa của từ vựng tiếng Anh.

Vì từ "từ" lại được dùng chung cho tất cả phát ngôn (mặt thể hiện, mặt hình thức) và cho cả ý niệm được kết hợp trong nó (mặt ý nghĩa, mặt nội dung), chính vì vậy để tránh hiểu nhầm, trong WordNet sẽ dùng thuật ngữ "dạng từ", hay là "hình thức từ" (word from) để chỉ đến mặt hình thức, thể hiện vật chất của "từ", còn thuật ngữ "nghĩa từ" (word meaning) để chỉ đến mặt nội dung, ý niệm được từ vựng hóa của "từ". Xuất phát từ 2 khái niệm trên, ta có thể nói rằng: "ngữ nghĩa học từ vựng là sự ánh xạ giữa hình thức và nghĩa" và "mỗi từ loại cú pháp khác nhau, sẽ có các kiểu ánh xạ khác nhau".

Bảng 2-4: Ma trân từ vựng trong WordNet

Từ / Nghĩa	Dạng thức từ				
	F_1	F_2	F_3	….	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
...				…	
M_m					$E_{m,n}$

Ta thử xem xét một ma trân từ vựng (lexical matrix) như trong Bảng 2-4 trên đây. Mỗi hàng M_1, M_2,….M_m là các nghĩa khác nhau của một dạng từ (word form) F nào đó. Các cột F_1,F_2,..F_n là các dạng thể hiện khác nhau của cùng một nghĩa từ (word meaning) M nào đó. Giao giữa hàng M và cột F cho một mục E có nghĩa dạng từ F đó dùng thể hiện nghĩa M đó. Ví dụ : $E_{1,2}$ là dạng từ F_2 dùng để thể hiện nghĩa M_1. Nếu cột F nào có nhiều hơn hai mục E thì ta nói dạng từ đó là đa nghĩa (polysemous). Nếu hai mục E cùng nằm trên một hàng M thì ta nói

hai dạng từ đó đồng nghĩa (synonym) với nhau. Ví dụ : Bảng 2-4 trên, thì F_2 là đa nghĩa, F_1 và F_2 là đồng nghĩa.

Phép ánh xạ giữa dạng thức từ và nghĩa của từ là ánh xạ n-m (nhiều-nhiều) nghĩa là: có dạng (thức) từ mà có nhiều nghĩa và cũng có nghĩa từ được thể hiện thành nhiều dạng. Ở phần cơ sở lý thuyết về ngôn ngữ học, chúng tôi đã trình bày về vấn đề đa nghia và đồng nghĩa của từ. Như thế, nghĩa của từ được trình bày như thế nào trong WordNet? Muốn bắt trước một ma trận từ vựng cần thiết phải có một cách để trình bày cả hình thái và nghĩa trong cơ sở dữ liệu. Những câu giải thích có thể cung cấp một giải pháp thỏa mãn một cách hợp lý cho những hình thức, nhưng nghĩa nên được miêu tả kiểu như thế nào là một câu hỏi đặt ra cho một giả thuyết nào đó của ngữ nghĩa từ vựng. Không có một giả thuyết tâm lý thích hợp, những phương pháp phát triển bởi những nhà từ điển học có thể được cung cấp một giải pháp tạp thời: Những định nghĩa có thể đống cùng một vai trò trong một sự bắt trước mà những nghĩa biểu diễn trong tâm trí của một người sử dụng ngôn ngữ.

Những khái niệm từ vựng là thế nào để được diễn tả bởi những định nghĩa trong một giả thuyết cảu ngữ nghĩa từ vựng phụ thuộc vào có hay không có giả thuyết được định tính cách xây dựng hay chỉ khác nhau. Trong một giả thuyết có tính xây dựng, sự miêu tả nên chứa thông tin đầy đủ để hỗ trợ một cấu trúc chính xác của khái niệm (bởi hoặc một người hoặc một máy). Những yêu cầu của một giả thuyết có tính cách xây dựng không dễ được gặp, và có một số lý do để tin rằng những định nghĩa đã tìm trong hầu hết những từ điển chuẩn mà không thấy chúng [10]. Mặt khác, trong một giả thuyết khác nhau, những nghĩa có thể được miêu tả bởi một số biểu tượng mà cho phép một nhà luận lí phân biệt giữa chúng. Nhứng yêu cầu cho một giả thuyết khác nhau là mẫu mực hơn, là yêu cầu cách đặt câu theo những phép ánh xạ. Người đọc luôn nắm định nghĩa khái niệm về những yêu cầu để xác định số lượng từ đồng nghĩa (hoặc từ gần nghĩa). Mặt khác, nghĩa từ M_1 trong bảng 1 có thể được miêu tả bởi sự liệt kê đơn giản những hình thái từ mà có thể được sử dụng diễn tả nó: (F_1, F_2,…F_n).

Ví dụ: một người nào đó mà biết *Board* có thể biểu thị bằng một *lumber* hoặc *plank* hoặc *committee*. Những bộ từ đồng nghĩa,*{board, plank}* hoặc *{board, committe}* có thể phục vụ như chỉ định rõ ràng về hai nghĩa của board. Những tập từ đồng nghĩa (synsets) không giải thích những khái niệm là gì mà chúng chỉ biểu thị sự tồn tại. Những người biết tiếng Anh phải nắm rõ những khái niệm và dễ dạng nhận ra chúng từ những từ đã liệt kê trong tập từ đồng nghĩa (synset).

Vì thế, một ma trận từ vựng có thể được miêu tả cho những mục đích lý thuyết bởi một phép ánh xạ giữa những từ được viết và synset. Khi tiếng Anh phong phú trong những từ đồng nghĩa, synsets đủ cho những mục đích khác nhau. Tuy nhiên thỉnh thoảng một từ đồng nghĩa thích hợp không có sẵn, trong trường hợp từ đa nghĩa có thể giải quyết bởi một lời chú thích ngắn, ví dụ., *{board,* (những bữa ăn của một người, thường cung cấp bởi tiền)} có thể phục vụ để phân biệt nghĩa của *board* này từ những từ khác; nó có thể được xem như một synset với một thành viên đơn. Nơi chú thích không được chỉ định sử dụng cho việc đặt một khái niệm từ vựng mới bởi một người không quen thuộc với nó, và nó khác với một từ đồng nghĩa trong đó nó không được sử dụng để thêm vào thông tin mà lưu trữ trong từ vựng tâm lý. Nó thực hiện mục đích của nó cho phép người sử dụng WordNet tiếng Anh để phân biệt nghĩa từ này với những nghĩa từ khác khi được đảo lộn.

Hiển nhiên, từ đồng nghĩa là một quan hệ từ vựng giữa những hình thái từ, vì nó được phân công vai trò trung tâm này trong WordNet, sự phân biệt lời chú thích được làm giữa những từ có quan hệ bởi từ đồng nghĩa mà được đính kèm trong dấu ngoặc móc '{' và '}', và những quan hệ được đính kèm trong dấu ngặc vuông '[' và ']'. Quan hệ ngữ nghĩa được chỉ định bởi những con trỏ.

WordNet được tổ chức bởi các quan hệ ngữ nghĩa, và khi những nghĩa có thể được miêu tả bởi synset như những con trỏ giữa synset. Nó có đặc điểm bởi những quan hệ ngữ nghĩa mà chúng được trao đổi lẫn nhau: nếu có một quan hệ ngữ nghĩa R giữa nghĩa {x,x',…} và nghĩa {y, y',…}, sau đó cũng có một quan hệ R' giữa {y, y',…}và {x,x',…}. Mục đích thảo luận là quan hệ ngữ nghĩa sẽ phục vụ một vai trò đôi: nếu quan hệ giữa nghĩa {x,x',…} và {y, y',…} được gọi là R, thì sau R cũng sẽ được sử dụng để đặt tên quan hệ giữa những hình thái từ riêng phụ thuộc vào những synset đó. Nó có trật tự như một cách luận lý để hướng dẫn những thuật ngữ khác nhau cho mối quan hệ những nghĩa, và quan hệ giữa những từ, sự đảo lộn lớn có thể rút kết từ sự hướng dẫn của nhiều thuật ngữ kỹ thuật mới.

2.2.1.2 Tập đồng nghĩa (synsets) trong WordNet

Trọng tâm của WordNet là những ý niệm đã được từ vựng hóa (ngữ nghĩa của từ, tạm gọi gọi là: ý niệm từ vựng), chính vì vậy WordNet quan tâm đến cách biểu diễn những nghĩa (hay ý niệm) này. Bảng 3-4 ở trên dùng ma trận từ vựng để thể hiện các dạng và các nghĩa của từ. Tuy nhiên, phương pháp dùng kí hiệu

chữ viết chỉ có thể dùng để biểu diễn dạng thức của từ (word form) mà thôi, chứ không thể dùng để biểu diễn nghĩa.

Việc biểu diễn ý niệm từ vựng này phụ thuộc vào mục tiêu phục vụ của WordNet: nếu dự tính dùng để xây dựng lên ý niệm từ vựng thì WordNet phải đảm bảo chứa tất cả các thông tin ngữ nghĩa có liên quan của từ sao cho chính từ WordNet, người ta có thể xây dựng chính xác ý niệm đó (theo quan điểm lý thuyết xây dựng nghĩa). Tuy nhiên, ý định này khó mà đáp ứng được, vì ngay cả các nghĩa chứa trong các từ điển hiện nay cũng chưa đáp ứng được yêu cầu tái hiện nghĩa nói trên. Còn nếu dự tính dùng WordNet chỉ để phân biệt nghĩa này với nghĩa khác, ý niệm từ vựng này với ý niệm từ vựng khác thì trong WordNet chỉ cần chứa các thông tin dưới dạng kí hiệu chữ sao cho người sử dụng có thể dựa vào đó để phân biệt được nghĩa này với nghĩa khác của cùng một từ đa nghĩa. Ví dụ: từ *"letter"*có hai nghĩa là *"là thư"* và *"chữ cái"*. Nếu ta lưu thành hai tập như sau*: {letter, message,…}* và *{letter, alphabet,..}* thì người sử dụng lập tức biết ngay dạng từ *"letter"* nào có nghĩa gì. Vậy hai tập đồng nghĩa (synset) nói trên chính là cách biểu diễn hai nghĩa của dạng từ *"letter"*.

Những tập đồng nghĩa (SYN onym SET = synset) tự thân chúng không giải thích về nghĩa (hay ý niệm) mà chúng mang là gì, chúng chỉ cho biết là chúng có mang một nghĩa (ý niệm) duy nhất nào đó mà tất cả các từ có dạng từ được chứa trong tập đó cùng mang. Ví dụ: lớp SS_i= {WF_{i1}, WF_{i2},….,WF_{in}} sẽ mang 01 nghĩa duy nhất mà các từ W_1, W_2,…W_n cùng mang. (Lưu ý: tập đồng nghĩa trong WordNet được đặt giữa hai dấu ngoặc móc: {…}). Vì tiếng Anh là ngôn ngữ giàu từ đồng nghĩa, nên trong mỗi synset có nhiều (dạng) từ. Nếu trong synset nào chỉ có một (dạng) từ, thì trong WordNet nhất thiết phải có mở ngoặc giải thích thêm về nghĩa của dạng từ đó (hiện nay, đa số synset đều có giải thích). Các synset được tổ chức thành dạng file insdex mà chúng ta hay gặp. Khi đó, mỗi synset trong WordNet được gán cho một mã số duy nhất (synset id) (có thể xem đây là khóa) để dễ truy xuất khi xử lý tự động bằng máy tính và đây cũng chính là nhãn ngữ nghĩa. Mã số này được tính bằng cách sử dụng chính là độ dời (offset) tính từ đầu của tệp tin *.dat của synset đó , vì vậy chúng ta có thể định vị synset đó một cách nhanh chóng (bằng cách sử dụng hàm fseek trong ngôn ngữ C chẳng hạn) .

2.2.1.3 Các quan hệ trong WordNet

Vì trọng tâm của WordNet là ngữ nghĩa, nên các quan hệ trong WordNet cũng chủ yếu là các quan hệ liên quan đến nghĩa, nhưng vì nghĩa của từ trong

WordNet thì được biểu diễn bởi các synset (thành phần cơ bản trong WordNet), chính vì vậy quan hệ chủ yếu trong WordNet cũng chính là các quan hệ giữa các synset. Nếu giữa các synset SS_i = { WF_{i1}, WF_{i2},….,WF_{in} } và synset SS_j = { WF_{j1}, WF_{j2},….,WF_{jn} } có quan hệ R_{ij} với nhau, thì synset SS_j = { WF_{j1}, WF_{j2},….,WF_{jn} } cũng sẽ có quan hệ R_{ji} với synset SS_i = { WF_{i1}, WF_{i2},….,WF_{in} }. Tính chất này của quan hệ được gọi là tính hỗ tương (reciprocate). Ngoài ra, nếu giữa hai synset SS_i = { WF_{i1}, WF_{i2},….,WF_{in} } và synset SS_j = { WF_{j1}, WF_{j2},….,WF_{jn} } có quan hệ R với nhau, thì WordNet cũng dùng quan hệ R đó để nói nên quan hệ giữa các dạng từ (Word form) $WF_i \in SS_i$ và $WF_j \in SS_j$ với nhau. Các quan hệ trong WordNet được diễn tả trực quan bằng các con trỏ (pointer) liên kết giữa synset này với synset kia. Dưới đây là các quan hệ được sử dụng trong WordNet:

- Quan hệ đồng nghĩa (synonymy)
- Quan hệ trái nghĩa (antonymy)
- Quan hệ hạ danh (thuộc cấp hyponym) và quan hệ thượng danh (bao hàm, hypernym)[5]
- Quan hệ bộ phận (meronymy/ holonymy)
- Quan hệ kéo theo (entailment)
- Quan hệ cách thức đặc biệt (troponymy)

Tất nhiên, với mỗi từ laoij sẽ có một số các quan hệ mà từ loại khác không thể có được. Ví dụ danh từ trong WordNet có hai mối quan hệ : Quan hệ ngữ nghĩa và quan hệ từ vựng. Mối quan hệ ngữ nghĩa là mối quan hệ theo nghĩa với nhau, các nghĩa này biểu hiện bằng các synset: quan hệ hạ danh, quan hệ thượng danh, quan hệ bộ phận. Ngoài mối quan hệ ngữ nghĩa, danh từ trong WordNet còn có quan hệ từ vựng (là quan hệ giữa các hình thức từ) với nhau: quan hệ trái nghĩa (antonyms). Trong khi đó, động từ trong WordNet lại phổ biến các mối quan hệ kéo theo, cách thức đặc biệt…

Synonymy

Quan hệ quan trọng nhất trong WordNet có thể được coi là quan hệ đồng nghĩa, biểu diễn mối quan hệ giống nhau về nghĩa.Từ đó, có khả năng phán xét mối quan hệ giữa hình thức từ là điều kiện tiên quyết để biểu diễn nghĩa của từ trong ma trận từ vựng. Theo một định nghĩa (thường quy cho Leibniz) là hai sự diễn đạt về đồng nghĩa nếu được thay thế vào một câu bằng một sự đánh giá chính xác sẽ chọn ra nghĩa đúng nhất. Theo định nghĩa đó, tìm ra từ đồng nghĩa chính xác là rất tốt nếu chúng tồn tại. Tính đồng nghĩa liên quan đến ngữ cảnh :

hai sự diễn đạt đồng nghĩa trong ngữ cảnh ngôn ngữ C nếu thay thế cho một ngôn ngữ khác trong C không làm thay đổi sự đánh giá chính xác. Ví dụ: "plant" thay thế cho "board" cũng ít có thể thay đổi trong ngữ cảnh "carpentry" mặc dù có các ngữ cảnh khác của "board" sẽ được thay thế hoàn toàn không thích hợp.

Lưu ý rằng, định nghĩa của tính đồng nghĩa là điều kiện cần thiết làm thay đổi phân WordNet thành danh từ, động từ, tính từ và phó từ. Điều đó nói nên, các ý niệm được biểu diễn bởi tập đồng nghĩa (synsets), và các từ đồng nghĩa được thay thế cho nhau, lúc đó các từ khác nhau về cú pháp từ loại không thể là từ đồng nghĩa (không thể hình thành nên tập đồng nghĩa (synsets)) chúng không thể hoán đổi cho nhau. Vì vậy danh từ hình thành ý niệm (concepts) của danh từ, tính từ hình thành ý niệm của tính từ, động từ hình thành ý niệm của động từ..và cung cấp cách thức hội đủ điều kiện những ý niệm. Nói cách khác, việc sử dụng các synsets để đại diện cho nghĩa của từ phù hợp với thuộc ngôn ngữ tâm lý bằng chứng là danh từ, động từ và bổ ngữ được tổ chức độc lập trong bộ nhớ ngữ nghĩa.

Antonymy

Mối quan hệ quen thuộc trong ngữ nghĩa nữa là antonymy, hóa ra khó xác định.Từ trái nghĩa với từ x là not-x nhưng không phải luôn luôn là như vậy. Ví dụ từ "rick" và từ "poor" là hai từ trái nghĩa. Nhưng để nói một người nào đó không "rick" không có nghĩa là họ "poor". Nhiều người tự coi mình là người không "poor" và không "rick". Antonymy tưởng chừng như là một quan hệ đối xứng đơn giản nhưng thực tế khá phức tạp nhưng người nói tiếng Anh công nhận từ trái nghĩa ít khó khăn khi họ nhìn thấy chúng.

Antonymy là một mối quan hệ hình thức từ vựng không phải là mối quan hệ ngữ nghĩa giữa từ với ý nghĩa với nhau. Ví dụ: ý nghĩa {rise,ascend} và {fall, descend} là có thể trái ngược nhau nhưng chúng không phải là quan hệ antonymy, [rise/fall] là quan hệ trái nghĩa và [ascend/descend] cũng vậy. Nhưng hầu hết mọi người không nhất quyết và ngần ngại khi được hỏi [rise] /[ascend] và [fall]/[descend] là từ trái nghĩa. Như vậy việc cần làm rõ ràng thật sự cần thiết phải phân biệt mối quan hệ ngữ nghĩa giữa hình thức từ và quan hệ ngữ nghĩa giữa nghĩa của từ.

Hyponymy

Không giống với quan hệ synonymy và antonymy, hyponymy/hypernymy là mối quan hệ ngữ nghĩa giữa nghĩa của từ. Ví dụ: {maple} là hyponymy của {tree} và {tree} là hyponymy của {plant}. Phần lớn sự chú ý được dành cho hyponymy/ hypernymy. Một ý niệm đại diện bởi synset {x, x',…} được gọi là

hyponymy của synset {y,y',…} nếu người nói tiếng Anh chấp nhận xây dựng từ khung *An x is a (kind of) y* .Mối quan hệ này có thể được đại diện kể cả trong {x, x',…} là một con trỏ tới superordinate của nó và kể cả trong {y,y',…} là một con trỏ tới hyponyms của nó.

Hyponymy là quan hệ bắc cầu và không đối xứng (Lyons, 1977, q.1) và từ đó thường chỉ có mooth superordinate duy nhất, nó tạo ra cấu trúc thứ bậc ngữ nghĩa trong đó một hyponymy được cho là cấp dưới của superordinate của nó. Như vậy, đại diện cho thứ bặc tiêu biểu được sử dụng rộng rãi trong xây dựng hệ thống truy vấn thông tin, và được gọi là hệ thống kế thừa (Touretsky, 1986): hyponymy thừa hưởng tất cả đặc tính chung các ý niệm và cho biết thêm một tính năng khác biệt superordinate của nó và bất kì hyponymy khác của superordinate đó. Ví dụ: "maple" thừa hưởng đặc trưng của "tree" superordinate, nhưng khác biệt từ "trees" khác bởi "hardness of its wood, the shape of its leaves, the use of its sap for syrup,.."cung cấp nguyên tắc tổ chức chính cho danh từ trong WordNet.

2.2.2 Danh từ trong WordNet

Hiện nay, WordNet đã bao gồm hơn 110.000 danh từ được phân chia vào gần 80.000 synset. Rất nhiều từ trong số đó là từ ghép và có một số danh từ riêng thông dụng. WordNet được xây dựng dựa trên các nguyên tắc về tâm lý học. Do đó nó hơi khác với từ điển thông thường. Các từ điển thông thường cung cấp cho chúng ta các thông tin về cách phát âm, định nghĩa, các dạng dẫn xuất và biến cách của từ, từ loại, định nghĩa… tuy nhiên WordNet được tổ chức theo một cách khác. Để đơn giản ta xem ví dụ sau:

Trong các từ điển thông thường, nếu chúng ta tra danh từ *"tree"* thì sẽ nhận được kết quả là *"**tree is** a plant that is large, woody, perennial and has a distinct trunk"*(cây là một loại thực vật lớn, thân gỗ, sống lâu năm và có thân rõ rệt).

Đây là cách giải thích tương đối đầy đủ và thích hợp.Từ điển thông thường sủ dụng cách giải thích này để giải thích các danh từ: sử dụng từ có tính tổng quát để định nghĩa từ riêng biệt bằng cách liệt kê thêm một số dặc điểm riêng biệt.

Tuy nhiên, định nghĩa như trên không hoàn chỉnh. Ví dụ : nó không cho chúng ta biêt *"tree*"(cây) có rễ, cây bao gồm các tế bào xen-lu-lo, thậm chí chúng ta không biết được cây là vật thể sống. Dĩ nhiên, chúng ta có thể biết được các

thông tin trên nếu chúng ta tìm nghĩa của từ tổng quát hơn: trong trường hợp này là *" plant" (thực vật).*

Thứ nhì, cách định nghĩa như trên sẽ không cho chúng ta biết được các loại thực vật khác: cùng có nghĩa tổng quát với từ cần tra. Ví dụ : từ điển thông thường không cho chúng ta biết ngoài *"tree"(cây)* da, có bao nhiêu từ nữa có cùng từ tổng quát là *" plant" (thực vật)* (Trong trường hợp này buộc người muốn tìm hiểu phải duyệt từ đầu từ điển đến cuối để xem có bao nhiêu từ co định nghĩa là *"is a plant..",* tuy nhiên cách tìm này bao giờ cũng có kết quả nếu người ta sử dụng từ tổng quát khác).

Thứ ba, với từ điển thông thường chúng ta không thể biết được có bao nhiêu loài *"tree" (cây),* hay nói cách khác *"tree" (cây*) sẽ được phân thành bao nhiêu loại con. Ví dụ : cây sẽ được phân loại thành *"cây sồi", "cây ổi"*(Trong trường hợp này buộc người muốn tìm hiểu phải duyệt từ đầu đến cuối để tìm xem có bao nhiêu từ định nghĩa là *"is a tree…"*). Trong các từ điển thông thường, phần giải thích nghĩa chỉ cung cấp cho chúng ta các thông tin về từ tổng quát hơn chứ không cung cấp cho chúng ta các từ cụ thể của từ đang tra.

Thứ tư, từ điển thông thường không cung cấp cho chúng ta các thông tin về *"tree" (cây)* mà mọi người đều biết như: cây có vỏ và nhánh con, chúng mọc lên từ hạt, cây lớn cao hơn chiều cao của con người,chúng tự sản xuất thức ăn cho chúng bằng quá trình quang hợp, chúng cung cấp bóng mát và chống gió, chúng tạo nên rừng, gỗ của chúng được dùng để xây dựng hay tạo ra năng lượng…nếu một người không biết về *"tree" (cây)* sẽ không thể xây dựng một ý niệm chính xác về *"tree" (cây)* từ các thông tin mà từ điển thông thường cung cấp. Các từ điển thông thường chỉ *"vẽ"* ra các đặc điểm khác biệt quan trọng, giúp cho người đọc nhớ lại ý niệm rõ hơn. Từ điển thông thường không cung cấp các kiến thức tổng quat như các bách khoa toàn thư.

Chúng ta lưu ý: sự thiếu các thông tin như của từ điển thông thường tập trung vào các thông tin thuộc vào cấu trúc chứ không thiếu các dữ liệu. Các nhà từ điển học thường tạo ra các sự liên tưởng giữa từ và các thông tin hình dung về chúng. Do đó các từ trong từ điển thường rời rạc, xếp theo thứ tự ABC,… vì vậy, để tránh sự lặp lại thông tin, từ điển thông thường sẽ không chứa mọi thông tin liên quan đến từ được định nghĩa.

2.2.2.1 WordNet là một hệ thống kế thừa từ vựng

Nếu chú ý kĩ, chúng ta sẽ có nhận xét là các từ trong từ điển đôi khi được định nghĩa vòng. Đó là từ w_a được sử dụng để định nghĩa từ w_b và từ w_b lại được

sử dụng để định nghĩa từ w_a. Có thể đây là sự định nghĩa từ đồng nghĩa của các nhà từ điển học, nhưng khi sử dụng từ điển này trên máy tính là không được.

Nguyên tắc thiết kế cơ bản mà các nhà từ điển cố gắng làm sao để định nghĩa của danh từ sẽ không mắc phải tình trạng như trên, cách định nghĩa này là một cây (nghĩa cây trong lý thuyết đồ thị không phải khái niệm cây trong cấu trúc dữ liệu). Một cây từ vựng được xây dựng dựa vào một dãy các thuật ngữ phân cấp.

Ví dụ:

Oak @→tree @→plant @→ organism

(cây sồi @→cây @→thực vật @→ sinh vật)

Quan hệ ngữ nghĩa này bằng kí hiệu @→, nó dùng để chỉ một từ đến một từ tổng quát hơn nó. Đay là quan hệ có tính chất bắc cầu và không đối xứng. Quan hệ này được đọc là "is –a "hay "is a kind of", nó đi từ cụ thể đến tổng quát (quá trình này gọi là tổng quát hóa). Cách thiết kế này tạo ra một hệ thống các cấp bậc, phân cấp, đi từ các thành phần chi tiết đến các thành phần tổng quát hơn ở phía trên. Đây chính là quan hệ thượng danh (hypernym) trong WordNet, quan hệ này chỉ đến thành phần tổng quát hơn.

Với bất kì danh từ w_h @→w_s sẽ luôn có quan hệ ngược lại w_s ~ → w_h , nói cách khác w_s là cấp cha (hay còn gọi là cấp tổng quát) (superordinate) của w_h, và w_h là cấp con (subordinate) hay còn gọi là cấp dưới (hyponym) của w_s. Quan hệ "~→" đi từ tổng quát đến cụ thể nên được gọi là quan hệ cụ thể hóa.

Thông thường một danh từ có một từ tổng quát, các từ điển thường thể hiện từ cấp tổng quát này trong phần định nghĩa, một danh từ có thể có nhiều từ thuộc cấp con, từ điển thông thường tiếng anh sẽ không liệt kê chúng. Trong WordNet quan hệ tổng quát hóa "@→" được liệt kê một cách rõ ràng với con trỏ nhãn giữa các khái niệm từ vựng hay ngữ nghĩa. Tuy nhiên, WordNet không lưu các quan hệ "~→" bởi vì chúng ta có thể suy ra quan hệ cụ thể hóa "~→" từ quan hệ "@→".

Thực ra, quan hệ phân cấp kiểu này không mới, Touretzky đã đưa ra giải pháp này cho lập trình viên tổ chức các cơ sở dữ liệu lớn. Khi đó, các dữ liệu chung của các từ sẽ không bị lưu lặp lại. Nói cách khác, WordNet và cơ sở dữ liệu kể trên đều sử dụng cấu trúc phân cấp để tiết kiệm không gian lưu trữ. Điều này đã tạo nên một thuật ngữ "hệ thống kế thừa" (inheritance system). Khi đó tất cả các thuộc tính cảu thành phần cha sẽ có thành phần con. Điều này sẽ giảm đi dư thừa khi ta liệt kê chúng, và chúng ta chỉ liệt kê những thuộc tính ở những nút cha và nút con trỏ đến nút cha được hiểu là theo nghĩa nào những thuộc tính được

thêm vào xem từ nút cha. Do đó, thành phần con sẽ không cần nêu đầy đủ các tính chất của mình, muốn biết đày đủ thuộc tính của thành phần con phải tìm ở thành phần cha.

Nói cách khác, hệ thống thừa kế là ẩn trong định nghĩa từ vựng ,đầu tiên nhà từ điển không lưu trữ thông tin chung của "tree" và "plant"ở cả hai mục, nếu lưa trữ hết sẽ gây ra dư thừa dữ liệu, các thuộc tính chung sẽ được lưu trữ tại "plant" sau đó, chúng ta định nghĩa của "tree" theo cách để người đọc tự tìm xem nó có thuộc tính gì? Do đó, cách cha từ này không thích hợp cho các từ điển thông thường (do người sử dụng phải cha rất nhiều), nhưng lại thích hợp khi tổ chức trên máy tính.

WordNet được xây dựng theo kiểu hệ thống kế thừa từ vựng như thế. Hệ thống này xây dựng dựa trên các liên kết giữa các thành phần con (hyponym) và các thành phần cha (superordinate) và ngược lại. Trong cơ sở WordNet, một mục từ (entree)cho từ "tree"sẽ chứa một tham chiếu (hay còn gọi là con trỏ "@→") đến một từ *"plant":* Con trỏ được gán nhãn "cấp trên" bởi kí hiệu "@". Do đó, synset *"tree"* sẽ có dạng như sau:

{tree, plant,@ connifer,~alder,~...}

Với '…' biểu thị con trỏ hyponym nữa. Trong cơ sở WordNet, con trỏ "@" trỏ từ *"tree"* tới cấp cha *"plant"* và sẽ có ánh xạ ngược là con trỏ "~" từ *"plant"* tới *"tree"* trong synset *"plant"* ; con trỏ "~" được gọi là con trỏ " hyponym" :

{plant,flora, organism, @ tree,~ ....}

Và *{tree}* không chỉ là cấp con dua nhất của *{plant, flora}.* Tất nhiên, thứ tự sắp xếp khi liệt kê các con trỏ này không quan trọng.

Tuy nhiên, có nhiều từ là hyponym của chính nó. Trong các từ điển thông thường, vấn đề này không gây lên vấn đề gì cả. Ví dụ: một từ điển thông thường, từ *"board"* (tấm ván) có thể được sử dụng để chỉ các ý niệm hẹp *"surf board" (ván lướt sóng)* và *"skate board" (ván trượt).* điều này dẫn tới khả năng: từ "board" mang nghĩa hẹp sẽ là "hyponym" của từ "board" mang nghĩa tổng quát. Vấn đề này trong WordNet giải quyết theo cách phân chia từ *"board"* thành hai phần và phân biệt như sau*: {board,surf board} @→ board*. Đây là phương án đẻ tránh vấn đề một từ là hyponym của chính nó trong WordNet . Một ví dụ tương tự cho trường hợp này là từ *"cat"*. Trong WordNet, từ *"cat" (họ mèo)* có nghĩa tổng quát của từ *"big cat" (sư tử, báo)* và *"house cat" (mèo nhà),* trong khi đó, thực tế, mỗi khi nhắc đến " cat" chúng ta hay liên tưởng ngay (thường sử dụng) đến ý niệm con mèo *(house cat, tabby, pussy cat, domesticated cat).* Nguyên nhân của vấn đề này là do WordNet không phân biệt giữa tần xuất sử dụng của từ

đó (tuy nhiên WordNet 2.0 cũng có cung cấp thêm các thông tin về tần xuất của từng nghĩa trong các tập ngữ liệu (corpus) thông dụng). Thêm vào đó, WordNet sử dụng thêm các từ có nghĩa hẹp đi kèm với từ có nghĩa tổng quát đẻ tránh vấn đề không chắc chắn khi biểu diễn một ý niệm.

2.2.2.2 Vấn đề tâm lý học trong việc tổ chức WordNet

WordNet được tổ chức dựa trên các nguyên tắc tâm lý học, tổ chức theo cách tổ chức về từ vựng trong bộ não con người.

Bộ não của con người tổ chức các danh từ theo hệ thống kế thừa phản ánh sự phán đoán về tâm lý học từ vựng. Bằng chứng là quá trình con người dễ dàng phán đoán ra các từ tổng quát được lặp lại với từ chi tiết hơn đã đề cập phía trước.

Thứ nhất: các danh từ tổng quát có thể được sử dụng để chỉ các từ ở mức độ chi tiết hơn. Ví dụ: He *owned a **rifle**, but the gun had not been fired (Anh ta sở hữu cây súng trường, nhưng cây súng không nổ)*. Chúng ta có thể hiểu được the gun là một danh từ trùng lặp của a rifle đã được nêu ở trước.

Trong khi đó, từ tổng quát từ các quan hệ hạ danh (hyponym) của chúng không thể được xem là tương đương. Ví dụ: *A rifle is safer than a gun (súng trường thì an toàn hơn súng) và A gun is safer than a rifle (súng thì an toàn hơn súng trường)* sẽ không đúng ý nghĩa.

Do đó, chúng ta có thể kết luận về mối quan hệ ngữ nghĩa là quan hệ kế thừa. Người đầu tiên đề cập vấn đề này là Quillian (1967,1968). Các phát hiện của các tác giả này được trình bày trong bài thảo luận vào năm 1969 bởi Quillian và Collins. Họ giả thiết rằng thời gian phản ứng (suy nghĩ) có thể được xử dụng để xác định mức độ gần nghĩa giữa hai ý niệm. Ví dụ: thời gian để chúng ta suy xét câu *"A canary can sing"(chin hoàng yến có thể hót*) là đúng (TRUE) sẽ ngắn hơn thời gian chúng ta sử dụng để xét câu *"A canary can fly"(chim hoàng yến có thể bay).* Và thời gian chúng ta xét câu *"A canary can fly"(chim hoàng yến có thể bay)* là đúng lại ngắn hơn thời gian chúng ta xét câu *"A canary has skin" (chim hoàn yến có da).* Các tác giả trên giả thuyết rằng thuộc tính *can sing (có thể hót)* được lưu trong đối tượng *canary (chim hoàng yến), can fly (có thể bay)* là đặc tính *của bird (loài chim*), và thuộc tính *has skin* thuộc đối tượng *animal (động vật).* Nếu cả ba đặc tính này được lưu trực tiếp là đặc tính của *canary* thì thời gian suy luận đúng sai phải bằng nhau. Tuy nhiên, thực tế cho thấy, cả ba thời gian này khác nhau. Điều này cho thấy dặc tính can fly và has skin là thuộc tính của đối tượng tổng quát hơn. Collins và Quillian đã rút ra kết luận từ các

quan sát trên rằng các thông tin chung sẽ chỉ được ở các đối tượng tổng quát. Tuy nhiên, các đối tượng hyponym có thể lấy được các thuộc tính chung này. (Thực tế, trong WordNet cách tổ chức cũng theo ý tưởng *đó canary @→finch @→ bird @→vertebrate @→ animal (chim hoàng yến @→chim họ sẻ @→ họ chim @→động vật có xương sống @→ động vật)).*

Hầu hết các nhà tâm lý học đều khẳng định rằng các danh từ tiếng anh thông thường được tổ chức thành hệ thống kế thừa trong bộ nao con người, tuy nhiên các thuộc tính chung được kế thừa hay lưu ở đối tượng tổng quát là vấn đề còn nhiều tranh cãi (Smith, 1978). Năm 1969 hai tác giả Collins và Quillian cũng đưa ra một số cần xem xét. Ta hãy xét *robin (chim cổ đỏ) và ostrich (chim cổ đỏ) cùng một loài con của bird (loài chim)*. Nhưng thời gian để chúng ta nhận xét câu *"A robin is a bird" (chim cổ đỏ thuộc loài chim) sẽ nhanh hơn câu "An ostrich is a bird" (chim cổ đỏ thuộc loài chim).* Hay ví dụ khác: các thuộc tính *can move (co*

thể di chuyển) và has ears (có tai) của animal (động vật). Nhưng thời gian nhận xét câu *"An animal can move " (động vật có thể di chuyển)* lại nhanh hơn câu *"An anlmal has ears" (động vật thì có tai)*.

WordNet sử dụng giả thiết danh từ được tổ chức theo hệ thống kế thừa nhưng không sử dụng quan điểm độ đo thời gian của Collins và Quillian. Các tác giả WordNet xem độ đo thời gian này là võ đoán hơn là khoảng cách ngữ nghĩa, sự khác nhau này có thể xem như sự khác nhau giữa cách sử dụng từ và nghĩa của từ.

2.2.2.3 Ý niệm nguyên thủy

Chúng ta giả sử hệ thống WordNet là một hệ thống kế thừa, như thế thành phần cao nhất, tổng quát nhất sẽ không mang ý nghĩa gì cả. Thật vậy, nếu chúng ta sử dụng *{entity}* làm ý niệm gốc duy nhất, sau đó các ý niệm kế thừa từ ý niệm gốc là *{object, thing} và {idea}* thì hệ thống phân cấp của chúng ta rất lớn. Hơn nữa,với cách trình bày như vậy, các ý niệm gốc sẽ mang rất ít thông tin và các ý niệm con sẽ rất nặng nề về các thuộc tính.

Do đó, WordNet được tổ chức thành 25 ý niệm gốc. Việc chia nhỏ như vậy còn tạo điều kiện cho việc biên soạn từ điển, làm giảm kích thước của các tệp tin mà các nhà từ điển biên soạn, tạo điều kiện cho nhiều nhà từ điển làm việc song song với nhau.

Tuy nhiên, vấn đề nảy sinh là làm sao chọn được những ý niệm nào là ý niệm nguyên thủy. Cuối cùng người ta đã chọn được 25 ý niệm nguyên thủy bao trùm mọi danh từ trong tiếng Anh trong bảng 2-5.

Bảng 2-5: Danh sách 25 ý niệm nguyên thủy cho các file danh từ

{act, activity}	{food}	{possesion}
{animal, fauna}	{group,grouping}	{process}
{artifact}	{location}	{quantity, amoumt}
{attribute}	{motivation,motive}	{relation}
{body}	{natul object}	{shepe}
{cognition, knowledge}	{natural phenomenon}	{state}
{communication}	{person, Human being}	{substance}
{event,happening}	{plant, flora}	{time}
{feeling,emotion}		

Tuy nhiên, trong 25 ý niệm nguyên thủy có một số ý niệm chứa trong ý niệm khác. Ví dụ: 8 ý niệm liên quan đến định nghĩa sự vật, 5 ý niệm có thể được nhóm lại để chỉ về các khái niệm trừu tượng, 3 ý niệm miêu tả về các đặc điểm của tâm lý học. Do đó, chúng ta có thể giảm xuống còn 11 ý niệm cơ bản như bảng.

Bảng 2-6: Sơ đồ của việc giảm 25 file danh từ gốc thành 11 ý niệm cơ bản (các ý niệm co bản được in nghiêng)

Entity	Organism	Animal	
		Person	
		Plant	
	Object	Artifact	
		Natural Object	Body
		Substance	Food
	Abstracttion	Attribute	
		Quantity	
		Relation	Communication
		Time	
	Psychol,feature	Cognition	
		Feeling	
		Motivation	
		Natural Phenonmenon	Process
		Activity	
		Event	
		Group	

Location	
Possession	
Shape	
State	

Khi chọn 25 ý niệm cơ bản, độ sâu của cây kế thừa thu được ở mức chấp nhận được (10-12 cấp) và các từ ở cấp thấp thông thường chỉ là các từ chuyên môn, chúng ta ít sử dụng các từ này thường ngày. Ví dụ: *sheland pony @→ pony @→ horse @→ quid @→ oddg-toed ungulate @→ placental mammal @→ mammal @→ vertebrate @→ chordate @→ animal @→ organism @→ entity (ngựa nhỏ Sheland @→ ngựa nhỏ @→ ngựa @→ họ ngựa @→ loài có móng guốc lẻ @→ động vật có vú mang thai @→ động vật có vú @→ động vật có xương sống @→ động vật có dây sống @→ động vật @→ sinh vật @→ thực thể)*: 12 cấp độ, 1 nửa trong số chúng là từ chuyên môn (kĩ thuật).

2.2.2.4 Một vài giả thuyết về tâm lý học

Mặc dầu cấu trúc tổng quát của hệ thống phân cấp danh từ được tạo ra bằng mối quan hệ hyponym/ hypernym nhưng mối quan hệ đó không chỉ rõ các kiến thức này được biểu diễn như thế nào trong kí ức từ vựng của con người. Dường như các đặc tính riêng là dấu hiệu để phân biệt các khái niệm với nhau. Ví dụ: Con *chim cổ đỏ* (robin) phải kế thừa từ *chim* (bird) những thuộc tính như *mỏ và lông*, hơn thế nữa, nó còn kế thừa các thuộc tính của *động vật có xương sống* (là cấp cha của chim (bird)) như *có máu có màu đỏ*…, tuy nhiên, chim cổ đỏ khác với chim (bird) ở các đặc điểm như nhiều màu sắc, có thể hát và có thể bay. Có ba loại đặc tính tạo nên sự phân biệt này là:

- Thuộc tính (attribute*): máu đỏ (warm-booded), có xương sống (vertebrate).*
- Bộ phận (parts): *(beak), cánh (wing)*
- Chức năng (function): *hót (sing), bay (fly).*

Với mỗi loại đặc điểm phân biệt này có vai trò khác nhau (thuộc tính là tính từ, bộ phận là danh từ, về chức năng là động từ). Do đó, cách định nghĩa một hyponym như sau: khi một đặc tính đặc trưng của synset{A} được bao bởi các đặc tính đặc trưng của synset{B}là hyponym của {A}.Nếu quan hệ hyponym được định nghĩa thông qua các đặc tính, thì các đặc tính trở nên đặc biệt quan trọng. với mọi synset, các đặc tính riêng biệt của nó phải đảm bảo tính cần và đủ

2.2.2.5 Quan hệ bộ phận (parts and meronymy)

Ngoài hai quan hệ hình thức và chức năng đề cập ở trên, Pustejovsky (1991) còn đưa ra mối quan hệ đóng vai trò "constitutive" ("cấu thành"). Quan hệ này chỉ ra mối quan hệ giữa đối tượng và các thành phần của nó. Quan hệ này sẽ liên kết giữa một danh từ biểu thị toàn thể và một danh từ biểu thị bộ phận.

Quan hệ bộ phận-toàn thể giữa các danh từ là một quan hệ ngữ nghĩa gọi là meronymy (từ này bắt nguồn từ tiếng Hy lạp cổ "meros"). Quan hệ này khác với các quan hệ synonym, antonym và hypornym. Quan hệ này có tính chất phản xạ tức là nếu w_m là meronym của w_h thì w_h là holonym của w_m. Chúng ta có thể sử dụng thuật ngữ IS_A_PART_OF và HAS_A để chỉ quan hệ meronym và holonym. Cụ thể nếu w_h là HAS_IS_PART_OF w_h thì $w_{m\ là}$ meronym của w_h. Nếu w_h HAS_A w_m thì w_h là holonym của w_m.

Trong WordNet quan hệ này phổ biến trong các tập tin noun.body, noun.artifact, noun.quantity. Với các đối tượng cụ thể như cơ thể, các vật nhân tạo, quan hệ meronym được sử dụng để định nghĩa các thành phần cơ bản.

Quan hệ meronym giống với quan hệ hyponym ở đặc điểm cả hai đều có tính chất không đối xứng, bắc cầu và cả hai đều là quan hệ có tính chất kế thừa.
Thí dụ :

Mỏ và cánh là meronym của chim, nếu chim hoàng yến là hyponym của chim, theo sự kế thừa thì mỏ và cánh là meronym của chim hoàng yến.

Tuy nhiên quan hệ meronym có nhiều loại, thí dụ như *một ngón tay là bộ phận của bàn tay, bàn tay là bộ phận của cánh tay, cánh tay là bộ phận của con người có nghĩa là ngón tay là meronym của bàn tay, bàn tay là meronym của cánh tay, cánh tay là meronym của cơ thể* . Khi đó chúng ta có thể nói ngón tay là bộ phận của cơ thể. Nếu chúng ta bắt đầu từ ý niệm tổng quát như {automobice} (xe máy) hay { human_ body} (cơ thể con người) thì sẽ có nhiều cấp của quan hệ meronym. Nhưng các meronym này sẽ lại là meronym cho ý niệm tổng quát hơn nữa. Quan hệ kế thừa theo kiểu "tangle" (rối) này hiếm khi xuất hiện trong mối quan hệ hypornym nhưng lại phổ biến trong quan hệ meronym.

Quan hệ meronym và hyponym có quan hệ mật thiết với nhau. Ví dụ : *{mỏ chim} (bech) và {cánh chim} (wing) là meronym của {chim} (bird), và nếu {chim cổ đỏ} (robin) là hyponym của {chim} (bird) thì nó sẽ được kế thừa các quan hệ meronym với ý niệm {cánh chim} (wing) và {mỏ chim} (beck).*

Tuy nhiên, cấu trúc của IS_PART_OF không phải lúc nào cũng là quan hệ meronym. Chúng ta xem ví dụ sau: *" cái tay nắm là meronym của cái cửa " và "cái cửa là meronym của căn nhà" ,* khi đó sẽ có hai khả năng sau : *"căn nhà*

có cái tay nắm cửa" hay "tay nắm cửa nhà là một phần của căn nhà" (Lyons, 1977). Winston (1987) cũng đưa ra một ví dụ tương tự khi xem xét mối quan hệ bộ phận-toàn thể. Ví dụ: *"nhánh cây là bộ phận của cây" và "cây là một phần của rừng"* nhưng chúng ta không nói *"nhánh cây là bộ phận của rừng",* bởi vì quan hệ nhánh cây/rừng không giống như quan hệ cây/ rừng. Nói rõ hơn, chúng ta có thể sử dụng quan hệ IS_PART_OF để chỉ quan hệ IS_ATTACHED_TO (thành phần), nhưng quan hệ IS_PART_OF là quan hệ có tính bắc cầu, còn quan hệ IS_ATTACHED_TO không có tính chất đó. Lấy lại ví dụ của Lyons ở trên, chúng ta nói *"căn nhà có cái tay nắm cửa"* hợp lý hơn bởi vì tay nắm arcos quan hệ IS_ATTACHED_TO với căn nhà.

Trong WordNet chỉ có 3 loại meronym:

W_m # p→ w_h : w_m là component của w_h

W_m # m→ w_h : w_m là member của w_h

W_m # s→ w_h : w_m là stuff của w_h được làm từ.

Một trong 3 meronym thì meronym # p (IS_A_COMPONENT_OF) được sử dụng nhiều nhất.

2.2.3 Định dạng file cơ sở dữ liệu trong WordNet

Định dạng file index

Mỗi file index bắt đầu với nhiều dòng có chứa một thông báo bản quyền, số phiên bản và các thỏa thuận cấp phép. Những dòng này tất cả bắt đầu với hai không gian và số dòng để họ không can thiệp với các thuật toán tìm kiếm nhị phân được sử dụng để tìm kiếm các mục trong các file index. Tất cả các dòng khác có định dạng sau đây. Trong lĩnh vực mô tả, số luôn luôn đề cập đến một số nguyên thập phân trừ trường hợp được xác định.

Lemma pos synset_cnt p_cnt [ptr_symbol ...] sense_cnt tagsense_cnt synset_offset [synset_offset ...]

Trong đó:

- ***Lemma*:** Trường hợp thấp hơn văn bản ASCII của từ hoặc sắp xếp có thứ tự. Cách sắp xếp được hình thành bằng cách các từ riêng lẻ kết hợp bằng một kí tự gạch dưới (_).
- ***Pos*:** thể loại cú pháp: *n* cho các tệp tin danh từ, *v* cho các tệp tin động từ, *a* cho các tệp tin tính từ, *r* cho các tệp tin trạng từ. Tất cả các trường còn lại là đối với các giác quan của bổ đề trong *Pos*.

- ***Synset_cnt :***số synset mà lemma nhập này là số lượng các nghĩa của các từ trong WordNet. Số giác quan là cách thức con số ý nghĩa được giao và thứ tự của synset_offset s trong file index.
- ***P-cnt:***số lượng các con trỏ khác nhau mà lemma có trong tất cả các synsets có chứa nó.
- ***Ptr_symbol:*** một khoảng trống tách ra các loại danh sách khác nhau của con trỏ *P_cnt* mà *lemma* có trong tất cả các synset chứa nó. Nếu tất cả các giác quan của *lemma* không có con trỏ, trường này bỏ đi và P_cnt là 0.
- ***Sense_cnt:*** Tương tự như *sense_cnt* .Điều này là không cần thiết, nhưng lĩnh vực này được bảo tồn vì các lý do tương thích.
- ***Tagsense_cnt***: Số lượng các nghĩa của *lemma* được xếp hạng theo tần số của chúng về sự xuất hiện trong các văn bản ngữ nghĩa.
- ***Synset_offset:*** *Byte offset* trong file *dữ liệu. Pos* của một synset chứa *lemma.* Mỗi *synset_offset* trong danh sách tương ứng với một ý nghĩa khác nhau của *bổ đề* trong WordNet. *Synset_offset* là 8 chữ số, điền số nguyên thập phân, số không, có thể được sử dụng với hàm *fseek* (trong C) để đọc một synset từ tập tin dữ liệu. Khi được thông qua để đọc các synset cùng với các thể loại cú pháp, một cấu trúc dữ liệu phân tích cú pháp có chứa các synset được trả lại.

Định dạng file dữ liệu

Mỗi file dữ liệu bắt đầu với nhiều dòng có chứa một thông báo bản quyền, số phiên bản và các thỏa thuận cấp phép. Những dòng này tất cả bắt đầu với hai không gian và số dòng. Tất cả các dòng khác có định dạng sau đây. Integer các trường là chiều dài cố định, và là số không đầy.

synset_offset lex_filenum ss_type w_cnt word lex_id [word lex_id...] p_cnt [ptr...] [frames...] | gloss

Trong đó:

- ***synset_offset*** : Hiện tại byte offset trong tập tin được đại diện là 8chữ số nguyên thập phân.
- ***lex_filenum*** : Hai chữ số nguyên tập phân tương ứng với tên file có chứa các synset người nghiên cứu từ ngữ học.
- ***ss_type*** : các loại mã synset:
  n Danh từ
  v Động từ
  a Tính từ
  s Tính từ vệ tinh
  r Trạng từ

- ***w_cnt*** :Hai chữ số nguyên thập lục phân chỉ số từ trong synset này.
- ***word*** :Hình thức của một từ như đã nhập trong synset bằng người nghiên cứu từ ngữ học, với khoảng trống thay thế bởi dấu gạch dưới (-_).
- ***lex_id:***số nguyên thập lục phân, khi được phụ thêm vào *lemma*, số *lex-_id* thường bắt đầu bằng số 0 (giá trị 0 là mặc định).
- ***p_cnt*** : Ba chữ số nguyên thập phân chỉ số lượng các con trỏ từ synset này để synsets khác. Nếu *p_cnt* là *000* các synset không có con trỏ.
- ***ptr*** : *pointer_symbol synset_offset pos source/target*
  pointer_symbol: con trỏ, trỏ từ synset này đến synset khác
  synset_offset: Hiện tại byte offset trong tập tin được đại diện là 8chữ số nguyên thập phân.
  pos: loại mã synset
  source/target: Một giá trị 0000 *pointer_symbol* có nghĩa là đại diện cho một mối quan hệ ngữ nghĩa giữa nguồn hiện tại của synset và đích của synset các chỉ báo bởi *synset_offset.*
- ***frames:*** chỉ trong data.verb
  f_cnt + f_num w_num [+ f_num w_num...]
  f_cnt: hai số nguyên thập phân, liệt kê chỉ số chung chung của frames.
  f_num là hai chữ số nguyên thập phân hình số khung.
  w_num là một số nguyên hệ thập lục phân hai chữ số chỉ ra các từ trong synset mà khung áp dụng.
- ***Gloss:*** Mỗi synset chứa một Gloss. Một Gloss được đại diện như là một thanh dọc (|)**,** tiếp theo là một chuỗi văn bản đó tiếp tục cho đến cuối dòng. Các Gloss có thể chứa một định nghĩa, ví dụ một hoặc nhiều câu, hoặc cả hai.

2.2.4 Số lượng từ, synset trong WordNet

Bảng 2-7: Số lượng từ, synset trong WordNet 2.0

Từ loại	Số từ	Số synset	Tổng số mục từ
Danh từ	114648	79689	141690
Động từ	11306	13508	24632
Tính từ	21436	18563	31015
Phó từ	4669	3664	5808
Tổng cộng	152059	115424	203145

Chương 3: XÂY DỰNG MÔ HÌNH VÀ THỰC NGHIỆM

Hiện nay để giải quyết vấn đề có cơ sở lưu trữ từ vựng giống WordNet. Chúng ta cần giải quyết vấn đề dịch các từ tiếng Anh trong synset ra tiếng Việt để tạo nên WordNet tiếng Việt trên nền tảng tận dụng tất cả những tài nguyên (từ điển) hiện đã có của tiếng Việt, có hai cách để tiếp cận vấn đề này.

- **Cách thứ nhất:** cách thức rút trích (bán) tự động mối liên hệ ngữ nghĩa trong WordNet tiếng Anh và thông qua một số từ điển song ngữ xây dựng một mạng từ vựng tiếng Việt phần danh từ.

- **Cách thứ hai:** Xây dựng hệ thống ngữ nghĩa được thực hiện bởi một đội ngũ các nhà ngôn ngữ học, tâm lý học và tin học..

3.1 Phương pháp dịch tự động WordNet qua tiếng Việt

(Tham khảo phương pháp này của Nguyễn văn Toàn ĐH KH-TN ĐHQG Tp.HCM)

3.1.1 Dịch từ WordNet

Gọi

S: là synset cần dịch

E_i : là tiếng Anh thứ i tong một synset ($n \geq 1$)

V_i^{jk}: là từ thứ j trong dòng nghĩa thứ k của từ E_i trong tư điển Anh Việt.

Do đó,

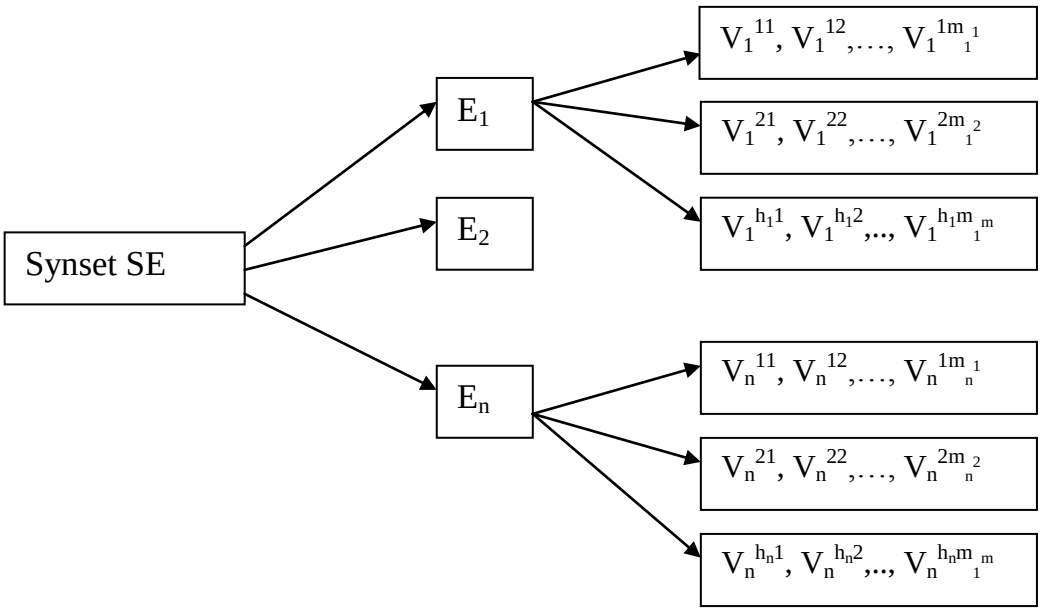
$0 \leq i \leq n$: với n là số lượng từ tiếng Anh của 1 synset.

$0 \leq j \leq h$: với h_i là số lượng dòng nghĩa của từ E_i trong từ điển Anh-Việt.

$0 \leq k \leq m_i^j$ với m_i^j là số lượng từ trong dòng nghĩa thứ j của từ E_i trong từ điển Anh-Việt.

V_i: tập hợp các nghĩa tiếng Việt của E_i

V_i^j:tập hợp các nghĩa tiếng Việt của E_i^j



Hình 2: Mô hình diễn giải các kí hiệu của mô hình dịch các synset trong WordNet

Với mô hình này, vấn đề của chúng ta là chọn nghĩa V_i^{j1},V_i^{j2},…nào cho synset SE. Để giải quyết vấn đề này chúng ta cần giải quyết các trường hợp sau:

a)Trường hợp 1

Trường hợp này, synset SE chỉ có một từ tiếng Anh và từ tiếng Anh này chỉ có một dòng tiếng Việt. Do đó, synset SE sẽ được biểu thị trong tiếng Việt bằng từ tiếng Việt trên.

Đặc tả

Nếu n=1 và n_i^j =1 thì synset S sẽ có từ biểu thị là V_1

b)Trường hợp 2

Trường hợp này, synset SE chỉ có một từ tiếng Anh và từ tiếng Việt này có một nhiều dòng nghĩa tiếng Việt ta gọi là V_i. Vấn đề được đặt ra là chúng ta sẽ chọn dòng nghĩa V_i nào tương ứng.

Đây là một trong hai trường hợp phổ biến nhất trong cả bốn trường hợp (một trường hợp phổ biến là trường hợp thứ 4 cũng có cách xử lý tương tự). Phương án để chọn ra một mô hình khả dĩ có thể chọn được nghĩa tiếng Việt đúng cho synset: mô hình hệ thống dựa trên lớp ngữ nghĩa.

Đặc tả

Nếu n=1 và $n_i^j \geq 1$ thì synset SE sẽ có tiếng Việt biểu thị sẽ được chọn từ tập ứng viên V_i, việc lựa chọn sẽ căn cứ vào xác suất của các hình vị V_i trong lớp từ.

Gọi E_i là từ tiếng Anh cần dịch trong Wordnet. Xác suất của cách dịch nó qua tiếng Việt được đặt tên là P (C/ E_i).Do đó, cách dịch tốt nhất là V^* với

$$V^*(E_i) = \underset{C \in T(E)}{\text{agr max}}\ P(V/E_i) \quad (1)$$

Với T(X) là tập hợp các cách dịch cảu từ thứ X trong từ điển Anh Việt

$$P(V|\ E_i) \tilde{=} P(V|g) \quad (2)$$

Với g= g(E_i) là lớp ngữ nghĩa chứa E_i.

Xác suất của P (V|g) có thể được ước lượng bằng cách sử dụng thuật toán EM (Estimation Maximization).

Bước 1: $P(V|\ E_i) = \frac{1}{m}, m = |\ T(E)|, V \in T(E)$ (3)

Bước 2: $P(V|g) = \frac{\sum_{E,k,i} P(V_k|E_i) I(V = V_k) I(E_i \in g)}{\sum_{E,k,i} P(V_k|E_i) I(E_i \in g)}$ (4)

Với V_k = cách dịch thứ k của E_i trong T(E_i)

I(x)=1 nếu x đúng và ngược lại

Bước 3: $P'(V|\ E_i) = P(V|g)$ (5)

Với g=g(E_i) là lớp chứa E_i.

Bước 4: $P(V|\ E_i) = \frac{P'(V|E_i)}{\sum_{D \in T(Ei)} P'(V|E_i)}$ (6)

Lặp lại bước 2 cho đến khi P(V| E_i) hội tụ

Để tránh vấn đề phân tán của dữ liệu

Chúng ta tính lại công thức EM như sau:

Bước 2:

$$P_u(u|g) = \frac{\sum_{E,k,i,j} \frac{1}{m} I(E_i \in g) I(u = u_{k,j}) P(u_{k,j}|E_i)}{\sum_{E,k,i,j} \frac{1}{m} I(E_i \in g) P(u_{k,j}|E_k)} \quad (4a)$$

Với $u_{k,j}$=unigram thứ j của cách dịch thứ k trong T(E_i)

M= số kí tự trong cách dịch thứ k trong T(E_i)

$$P_b(b|g) = \frac{\sum_{E,k,i,j} \frac{1}{m-1} I(E_i \in g) I(b = b_{k,j)} P(b_{k,j}|E_i)}{\sum_{E,k,i,j} \frac{1}{m-1} I(E_i \in g) P(b_{k,j}|E_i)} \quad (4b)$$

Với $b_{i,j}$=bigram chồng lên nhau của cách dịch thứ I trong $T(E_k)$

Bước 3:

$$P'(V|E_i)=P(V|g) = \sum_{k=1}^{m} \frac{P_u(u_k|g)}{m} \text{ (unigram)} \quad (5a)$$

$$P'(V|E_i)=P(V|g) = \sum_{k=1}^{m} \frac{P_u(u_k|g)}{2m} + \sum_{k=1}^{m-1} \frac{P_b(b_k|g)}{2(m-1)} + (bigram) \quad (5b)$$

Với u_k là một unigram, b_k là một bigram chồng lên của V và m số kí tự của V_d

c)Trường hợp 3

Trường hợp này, synset S có nhiều từ tiếng Anh. Các từ tiếng Anh này có nhiều nghĩa tiếng Việt (thuộc nhiều dòng nghĩa khác nhau), do đó, sẽ lấy phần giao của các $\{V_1^{11}, V_1^{12},..\}$, $\{V_1^{21}, V_1^{22},..\}$… để biểu thị cho synset SE.

Đặc tả

Nếu n>1 và $\bigcap_{i=1}^{n}\left\{\bigcup_{j=1}^{n_i} V_i^{j1}, V_i^{j2},...\right\}^{-}_{-} \neq Ø$ thì synset SE được biểu thị bởi tập:

và $\bigcap_{i=1}^{n}\left\{\bigcup_{j=1}^{n_i} V_i^{j1}, V_i^{j2},...\right\}^{-}_{-}$

d)Trường hợp 4:

Trường hợp này, synset S có nhiều từ tiếng Anh. Các từ tiếng Anh này có nhiều nghĩa tiếng Việt (thuộc nhiều dòng nghĩa khác nhau). Tuy nhiên, không giống trường hợp 3, các dòng nghĩa của các từ tiếng Anh không giao nhau nên đưa trường hợp này về trường hợp 2.

Đặc tả

Nếu n>1 và $\bigcap_{i=1}^{n}\left\{\bigcup_{j=1}^{n_i} V_i^{j1}, V_i^{j2},...\right\}^{-}_{-} \neq Ø$ thì synset SE được biểu thị bởi tập:

3.1.2 Dịch từ từ điển tiếng Việt

Gọi

V: là từ tiếng Việt cần gán nhãn synset

E_i^j: là nghĩa tiếng Anh thứ j của dòng nghĩa thứ i trong từ điển Việt-Anh

S_i^{jk}: là synset thứ k của từ E_i^j trong WordNet

Trong đó

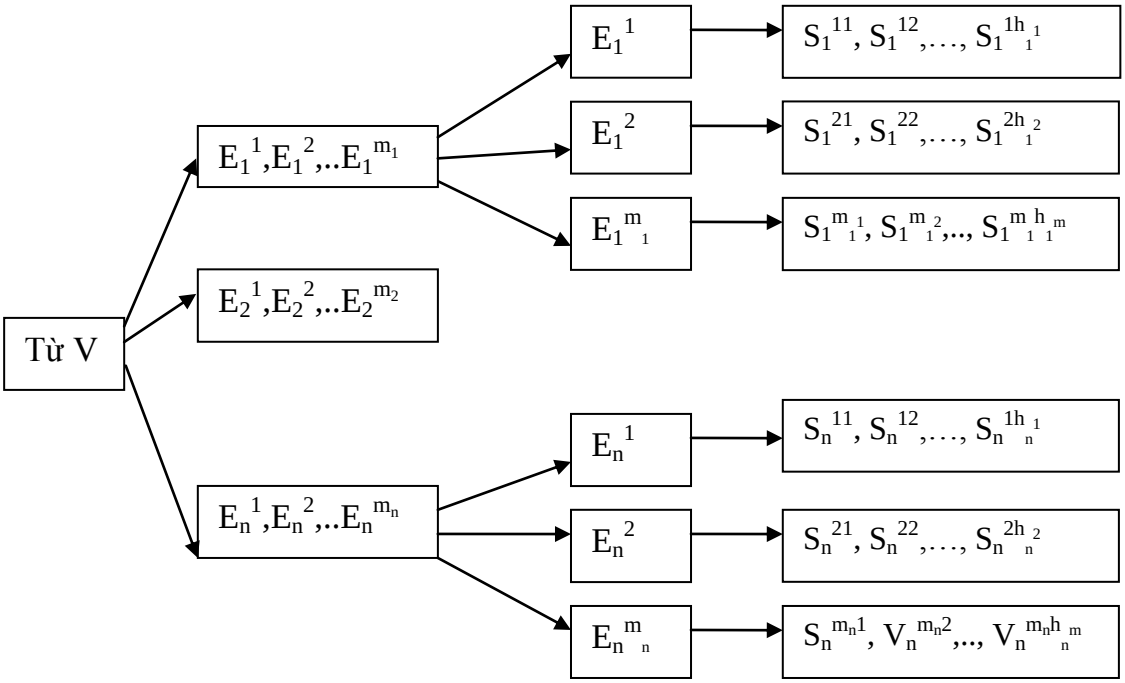
$0 \le i \le n$: với n là số lượng dòng nghĩa của từ V trong từ điển Việt-Anh.

$0 \le j \le m_i$: với m_i là số lượng từ trong dòng nghĩa thứ i của từ V trong từ điển Việt-Anh.

$0 \le k \le h_i^j$: với h_i^j là số lượng synset mà từ E_i^j thuộc.

E_i: tập hợp các từ E_i^j ($\forall$ j, $0 \le j \le m_i$)

S_i^j: tập hợp các synset S_i^{jk} ($\forall$ k, $0 \le k \le h_i^j$)



Hình 3: Mô hình diễn giải các kí hiệu của mô hình gán nhãn synset cho các từ tiếng Việt

Với mô hình này, vấn đề của chúng ta chọn nhãn synset S_i^{jk} nào cho từ V.

Dĩ nhiên, mỗi từ V có thể có nhiều nghĩa khác nhau, tương ứng với nghĩa của các tập E_i, E_j…do đó, khi chọn nhãn synset cho từ V chúng ta chọn nhiều synset.

Hơn thế nữa, do mỗi nghĩa của từ V tương ứng với nghĩa của tập E_i ($0 \le i \le n$) và các tập này rời rạc nhau nên việc chon synset cho từ V sẽ không phụ thuộc vào các dòng nghĩa khác nhau của từ V.

Do đó, bài toán này trở thành bài toàn làm thế nào để gán nhãn synset cho mỗi tập E_i ($0 \le i \le n$). Để giải quyết vấn đề này chúng ta cần giải quyết các trường hợp sau:

a)Trường hợp 1

Trong trường hợp này, dòng nghĩa tiếng Anh chỉ có một từ và từ này chỉ thuộc một synset, sẽ lấy synset này làm nhãn synset cho tập E_i

Đặc tả

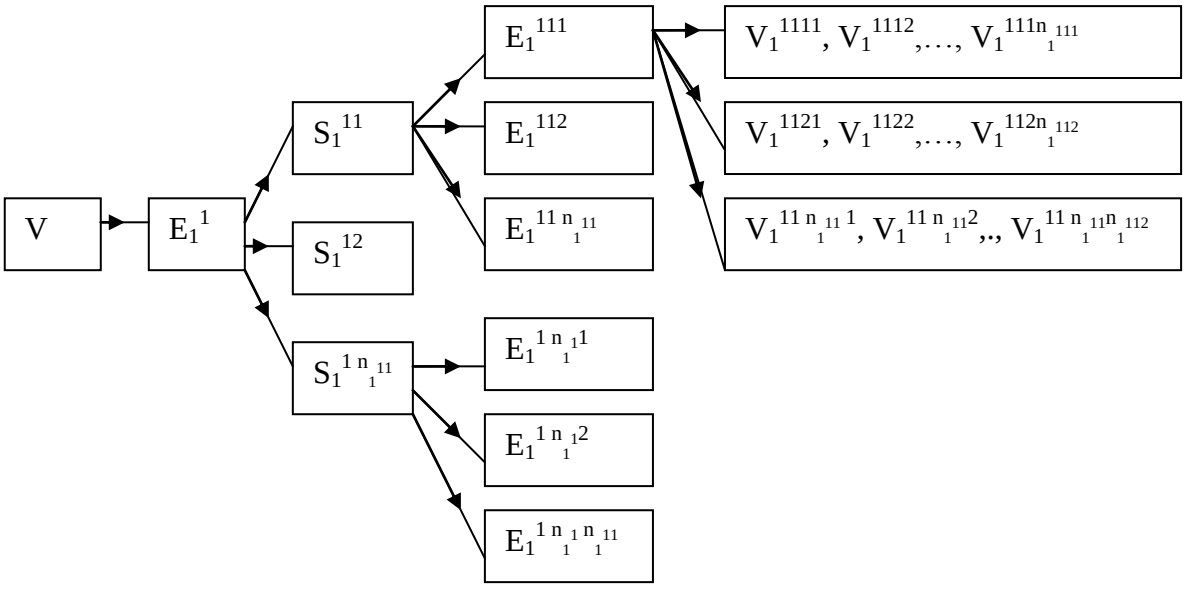
Nếu $n_i=1$ và $h_i^j=1$ (tức $|\{S_i^{j1},S_i^{j2},\ldots\}|=1$) thì synset của $\{E_1^1,E_1^2,\ldots\}$ chính là S_i^{j1}

b)Trường hợp 2

Với trường hợp này, dòng nghĩa tiếng Anh chỉ có một từ và từ này thuộc nhiều synset, khi đó căn cứ vào nghĩa tiếng Việt của các synset này để chọn ra nhãn synset cho $\{E_1^1, E_1^2, \ldots\}$

Đặc tả

Nếu $n_i=1$ và $h_i^j > 1$ (tức $|\{S_i^{j1}, S_i^{j2}, \ldots\}|>1$) thì synset của $\{E_1^1, E_1^2, \ldots\}$ được chúng tôi lựa chọn bằng cách sử dụng thêm từ điển Anh-Việt.



Hình 4: Mô hình diễn giải trường hợp 2

c)Trường hợp 3

Trường hợp này dòng nghĩa tiếng Anh có nhiều từ. Các từ này có nghĩa (thuộc nhiều synset khác nhau), do đó, sẽ lấy phần giao của các $\{ S_1^{11}, S_1^{12}, \ldots\}$, $\{ S_1^{21}, S_1^{22}, \ldots\}$…để gán nhãn ngữ nghĩa cho tập E_i

Đặc tả

Nếu $n_i > 1$ và $\bigcap_{j=1}^{m_i} \{ S_1^{j1}, S_1^{j2}, \ldots\} \neq \emptyset$ thì synset của $\{E_1^1, E_1^2, \ldots\}$ là

$$\bigcap_{j=1}^{m_i} \{ S_1^{j1}, S_1^{j2}, \ldots\}$$

d)Trường hợp 4

Trường hợp này, dòng nghĩa tiếng Anh có nhiều từ . Các từ này có nhiều nghĩa (thuộc nhiều synset khác nhau), tuy nhiên, khác với trường hợp 3 các tập synset này không giao nhau. Do đó, căn cứ vào cấu trúc của WordNet để chọn nhãn ngữ nghĩa thích hợp cho tập E_i.

Đặc tả

Nếu $n_i > 1$ và $\bigcap_{j=1}^{n_i}\{ S_1^{j1}, S_1^{j2},\ldots\} \neq Ø$ thì synset của $\{E_1^1,E_1^2,\ldots\}$ là sẽ được chọn lựa qua các mối liên hệ giữa các $\{ S_1^{11}, S_1^{12},\ldots\}$, $\{ S_1^{21}, S_1^{22},\ldots\}\ldots$

Ở đây sử dụng 3 tiêu chuẩn:

Tiêu chuẩn Anh em

Tiêu chuẩn này được áp dụng khi các tập synset S_i^j đều có các synset là anh em với nhau (có cùng synset cha (hypernymy)). Khi đo synset $\{E_1^1,E_1^2,\ldots\}$ được chọn là các synset Anh em này.

Tức là:

$SV= \{S_i^{jk} / S_i^{jk} \in S_i^j\ (\forall j : 0 \le j \le n_i^j) : \exists S_p : (S_p is\_hyper\ S_i^{jk}))\}$

Kí hiệu:

P is_hyper S: P là cấp cha của S:

Tiêu chuẩn cha con

Tiêu chuẩn này được áp dụng khi trong các tập synset S_i^j có một synset là cha của các synset còn lại (chỉ cần mỗi tập synset còn lại có một synset là con của synset cha nói trên). Khi đó synset $\{E_1^1,E_1^2,\ldots\}$ được chọn là các synset Anh em này.

Tức là:

$SV=\{S_i^{jk} / \exists S_p \in S_i^h\ (h \in [1..n_i^j]), S_i^{jk} \in S_i^j\ (\forall j : 0 \le j \le n_i^j, j \neq h) : (S_p is\_hyper\ S_i^{jk})\}$

Kí hiệu:

P is_hyper S: P là cấp cha của S:

Tiêu chuẩn ông cháu

Tiêu chuẩn này được áp dụng khi trong các tập synset S_i^j có một synset là cấp trên của các synset còn lại (chỉ cần mỗi tập synset còn lại có một synset là cấp dưới của synset cấp trên nói trên). Khi đó synset $\{E_1^1,E_1^2,\ldots\}$ được chọn là các synset cấp dưới này.

Tức là:

$SV==\{S_i^{jk} / \exists S_g \in S_i^h\ (h \in [1..n_i^j]), S_i^{jk} \in S_i^j\ (\forall j : 0 \le j \le n_i^j, j \neq h) : (S_g is\_dist\_hyper\ S_i^{jk})\}$

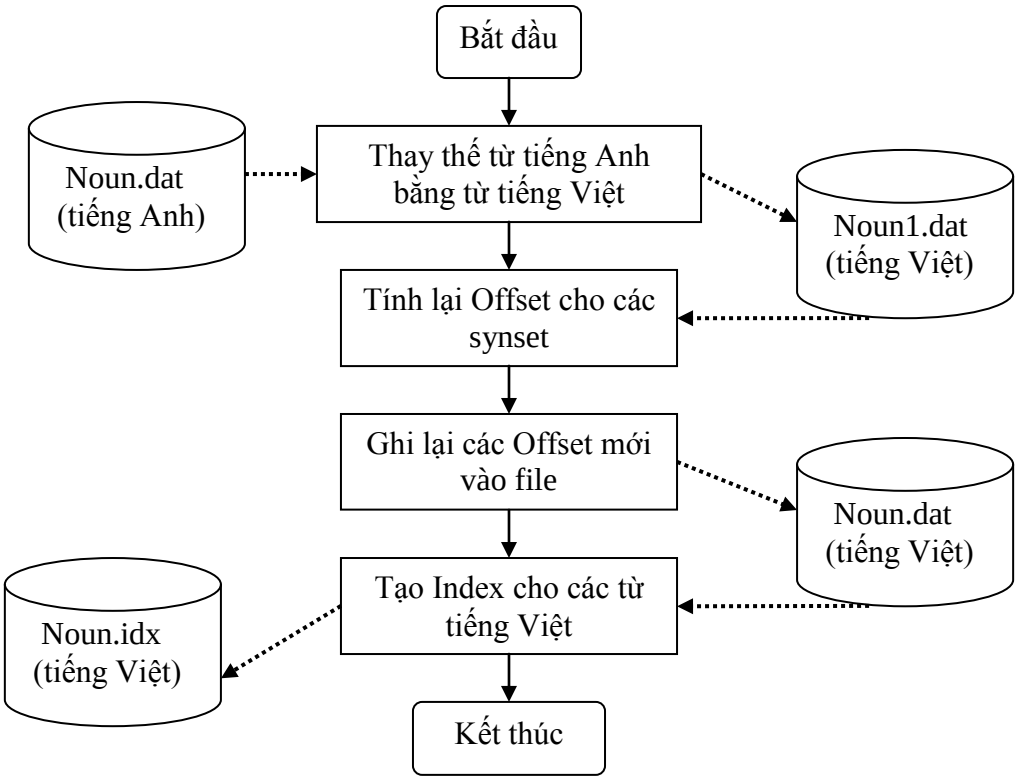
Kí hiệu:

P is_dist_hyper S: P là cấp trên của S:

3.1.3 Tổ chức dữ liệu

Sau khi đã xác định được từ (cụm) từ tiếng Việt tương ứng cho mỗi synset, công việc kế tiếp của chúng ta là tổ chức cơ sở tri thức WordNet tiếng Việt có hiệu quả và hợp chuẩn.

Để thuận tiện cho vấn đề chuẩn hóa, trao đổi giữa các cơ sở tri thức WordNet của các ngôn ngữ khác nhau, sử dụng ngay cách tổ chức WordNet của tiếng Anh để lưu cây WordNet tiếng Việt sau khi đã dịch xong:



Hình 5: Mô hình quá trình tổ chức dữ liệu cho WordNet tiếng Việt

3.2 Phương pháp làm thủ công

WordNet là một hệ thống từ vựng khổng lồ, đây là một hệ thống từ điển mã nguồn đóng nên việc sử dụng lại file data là rất khó khăn. Do thời gian làm đồ án có hạn, đây chỉ là chương trình thực nghiệm nên em xây dựng một số từ demo để khẳng định phương pháp xây dựng từ điển dựa trên cơ sở lưu trữ từ vựng của Wordnet. Phương pháp này sẽ cho kết quả là một từ điển có cấu trúc đáng tin cậy nhất nhưng đắt tiền, mất nhiều thời gian và công sức. Còn phương pháp tự động dịch nhanh nhưng độ chính xác không cao, nảy sinh nhiều vấn đề về ngôn ngữ mà máy tính không thể giải quyết được. Do vậy, để có một từ điển có độ chính xác cao, đơn giản, dễ thực hiện em đã chọn phương pháp thứ hai để xây dựng từ điển danh từ tiếng Việt dựa theo WordNet.

Trước tiên, dịch các synset ra tiếng Việt. Trong công đoạn này, đã giải quyết bốn trường hợp : synset có một từ và từ tiếng Anh có một nghĩa tiếng Việt, synset có một từ và từ tiếng Anh có nhiều nghĩa tiếng Việt, synset có nhiều từ và tập các nghĩa tiếng Việt của các từ tiếng Anh trong các synset không giao nhau. Sử dụng mô hình phân lớp ngữ nghĩa (semantic class-base translation model) để khử các nhập nhằng phát sinh.

Ở công đoạn thứ hai, gán nhãn synset cho từng từ tiếng Việt trong từ điển tiếng Việt. Trong công đoạn này, giải quyết bốn trường hợp : từ tiếng Việt có một nghĩa tiếng Anh và nghĩa tiếng Anh này chỉ thuộc một synset, từ tiếng Việt có một nghĩa tiếng Anh và nghĩa tiếng Anh này thuộc nhiều synset, từ tiếng Việt có nhiều nghĩa tiếng Anh và tập nhãn synset của các nghĩa tiếng Anh này có giao nhau, từ tiếng Việt có nhiều nghĩa tiếng Anh và tập nhãn synset của các nghĩa tiếng Anh này không giao nhau.

Cuối cùng, để mô phỏng kết quả của mô hình trên, Em đã cài đặt một chương trình để minh họa cho mô hình WordNet phần danh từ tiếng Việt.

3.3 Chương trình thực nghiệm

KẾT LUẬN

Qua quá trình tìm hiểu, nghiên cứu và hoàn thành đề tài, em được hiểu biết thêm về ngữ pháp tiếng Việt và cơ sở lưu trữ từ vựng của WordNet. Dựa vào công cụ hỗ trợ em đã xây dựng được từ điển cho phần danh từ tiếng Việt dựa theo WordNet.

Sau khoảng thời gian tìm hiểu và nghiên cứu đề tài em gặp phải một số vấn đề:

- Với Tiếng Việt, để tiến hành xử lý ngữ nghĩa trên máy tính, chúng ta cần phải có một cơ sở tri thức ngữ nghĩa từ vựng Tiếng Việt khá lớn mà thời gian có hạn nên việc xây dựng hoàn thiện cơ sở dữ liệu ngữ nghĩa cho phần danh từ là không thể.
- Với WordNet, WordNet là hệ thống mã nguồn đóng nên việc sử dụng lại cơ sở dữ liệu là rất khó khăn.

Dựa vào mô hình xây dựng và công cụ hỗ trợ để xây dựng từ điển danh từ tiếng Việt dựa theo cơ sở lưu trữ từ vựng của WordNet. Từ mô hình này chúng ta có khả năng áp dụng cho các từ loại khác như tính từ, động từ, trạng từ để hoàn thiện cho bộ từ điển Tiếng Việt theo WordNet . Em hi vọng, trong tương lai gần, sẽ hoàn thành việc xây dựng một hệ cơ sở tri thức ngữ nghĩa từ vựng Tiếng Việt tương đối hoàn chỉnh. Đây cũng là tiền đề để đẩy mạnh công việc xử lý Tiếng Việt trên máy tính.

Tài liệu tham khảo

[1] Diệp Quang Ban và Hoàng Văn Thung, *Ngữ Pháp tiếng Việt*. Tập 1 . NXB Giáo Dục

[2] Diệp Quang Ban và Hoàng Văn Thung, *Ngữ Pháp tiếng Việt*. Tập 2 . NXB Giáo Dục

[3] Nguyễn Thiện Giáp (chủ biên), Đoàn Thiện Thuật, Nguyễn Minh Thuyết, *Dẫn luận ngôn ngữ học* . NXB Giáo Dục

[4] Đinh Điền (2004), *Luận án Tiến Sĩ ngữ văn chuyên ngành Ngôn Ngữ học so sánh. ĐH XH&NV Tp.HCM.*

[5]Mai Ngọc Chừ, Vũ Đức Nghiệu, Hoàng Trọng Phiến (1997). *Cơ sở ngôn ngữ học và tiếng Việt*.NXB Giáo dục

[6] Đỗ Hữu Châu, *Từ vựng ngữ nghĩa tiếng Việt*. NXB GD (1997)

[7] http://wordnet.princeton.edu/

[8] http://en.wikipedia.org/wiki/WordNet

[9] http://www.cl.ut.ee/yllitised/viderorav.html

[10] George Miller, Richard Beckwith, Christiane Fellbaum, Dereck Gross, and Katherine Miller (Revised August 1993)- *Introduction to WordNet : an on-line lexical database.*

[11] Xavier Farreres, German Rigau, Horacio Rodriguez, *Using WordNet buiding WordNets.*

[12] Vũ Xuân Lương và Nguyễn Thị Minh Huyền, *Nghiên cứu và xây dựng từ điển Tiếng Việt cho Máy tính (Buiding a Vietnamese Computational Lexicon).*

[13] J.Daude, L.Padro & G.Rigau (1999) *Mapping WordNets Using Structural Information.*

[14] Jonh Lyons (1971), *Nhập môn ngôn ngữ học lý thuyết* (Bản dịch năm 1977). NXB GD Hà Nội.

[15] Hoàng Phê , *Từ điển tiếng Việt*.Hội ngôn ngữ học, NXB Đà nẵng.

[16] J. Daudé, L. Padró, G. Rigau, Mapping WordNets using structural information, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics.

[17] Automatic WordNet Mapping Using Word Sense Disambiguation