**Mục lục**

**Danh mục các bảng sử dụng trong luận văn**

**Danh mục các hình sử dụng trong luận văn**



**Tổng quan**

Trong những năm gần đây các nghiên cứu trong lĩnh vực Xử lý ngôn ngữ tự nhiên (Naturual Language Processing) đã xác nhận sự cần thiết phải mở rộng và hoàn thành các Hệ Cơ sở tri thức từ vựng đó. Để đạt được các cấu trúc từ vựng hoặc cấu trúc ngữ nghĩa là một vấn đề khó và thường các phương pháp được sử dụng là tái sử dụng lại, trộn hoặc điều chỉnh lại các hệ thống đã có. Hiện nay, tiếng Anh đã có các hệ thống hoàn chỉnh như LDOCE của nhà xuất bản Longman[1], Alvey Lexicon của Grove 1993, COMLEX của Grishman 1994… Trong số đó, hệ thống phổ biến nhất hiện này là hệ WordNet (Miller, 1990). Đây là một mạng ngữ nghĩa đồ sộ với hơn 110.000 định nghĩa tiếng Anh và ngày càng được nâng cấp về số lượng và chất lượng. Tuy nhiên với các ngôn ngữ khác, một hệ thống như vậy vẫn chưa có nhiều. Điển hình là tiếng Việt, hiện nay chúng ta vẫn chưa có một hệ thống 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. Hiện nay có hai cách để tiếp cận vấn đề này. Cách thứ nhất: xây dựng hệ thống ngữ nghĩa được thực hiện bằng tay bởi một đội ngũ các nhà ngôn ngữ học, tâm lý học, tin học… Phương pháp này sẽ cho kết quả là một từ điển có

[1] Chúng tôi sẽ giới thiệu kỹ hơn về các hệ cơ sở tri thức này ở chương 1.

cấu trúc đáng tin cậy nhất nhưng nó đắt tiền và mất nhiều thời gian và công sức. Với lý do này nhiều nhà nghiên cứu đặt trọng tâm vào việc làm sao xây dựng từ điển có cấu trúc từ những cái đã có của hệ cơ sở tri thức từ vựng và thông tin ngữ nghĩa từ những nguồn tài nguyên từ vựng có cấu trúc càng tự động càng tốt.

Trong luận văn này, chúng tôi lựa chọn cách thứ hai để xây dựng nên một từ điển có cấu trúc cho tiếng Việt dựa trên  hệ thống WordNet của Miller. Hệ thống WordNet có thể xem như là một tập hợp các synset (synonym set)(chúng ta có thể xem đây là các nút trong một cây), mỗi synset sẽ có một số từ (cũng có thể là cụm từ, từ ghép…) để biểu thị cho ý nghĩa của cho synset này. Bên cạnh đó còn WordNet cung cấp một loạt các quan hệ giữa các synset này. Do đó, mục đích chính của luận văn chuyển các từ tiếng Anh trong mỗi synset sang tiếng Việt.

Tuy nhiên, sự ánh xạ giữa từ này và từ kia giữa các ngôn ngữ không phải là ánh xạ 1-1, nghĩa là một từ tiếng Anh không phải chỉ có mang một ý nghĩa trong ngôn ngữ khác. Nguyên nhân cốt lõi của vấn đề này là do bản chất của ngôn ngữ: ý nghĩa của ngôn ngữ không nằm ở hình vị mà nằm ở hình tượng mà con người gán cho nó. Do vậy, để xây dựng nên WordNet cho ngôn ngữ của mình các, các nhà tin học và ngôn ngữ học đều phải giải quyết vấn đề cốt lõi này. Có rất nhiều phương pháp được xây dựng để giải quyết vấn đề này:

Phương án đầu tiên sử dụng các từ điển song ngữ để chuyển các từ tiếng Anh trong WordNet sang ngôn ngữ tương ứng. Tất nhiên, sẽ có rất nhiều heuristic được sử dụng để nâng cao độ chính xác của phương pháp này như sử dụng công thức về độ tương đồng hình vị của hai từ, sử dụng xác suất xuất hiện của từ tiếng Anh trong WordNet và từ của ngôn ngữ tương ứng…

Phương án thứ hai đề xuất cách sử dụng chính cấu trúc của WordNet và một từ điển phân loại của vủa ngôn ngữ tương ứng để xây dựng nên WordNet. Phương án này đòi hỏi ngôn ngữ tương ứng phải có các từ điển phân loại đã có sẵn.

Phương án cuối cùng sử dụng các corpus song ngữ để xây dựng nên Wordnet của ngôn ngữ tương ứng.

Dựa trên các hướng tiếp cận trước, chúng tôi đã đề xuất một phương án khả dĩ có thể được áp dụng để 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.

Trước tiên, chúng tôi dịch từng synset qua tiếng Việt. Trong công đoạn này, chúng tôi 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 có giao nhau, synset có nhiều từ và tập các nghĩa tiếng Việt của các từ tiếng Anh trong các synset không giao nhau. Để khử các nhập nhằng phát sinh, chúng tôi đã sử dụng các thông tin về vector ngữ nghĩa được tính toán từ câu giải thích của synset và từ từ điển Việt-Việt.

Ở công đoạn thứ hai, chúng tôi gán nhãn synset cho từng từ tiếng Việt trong từ điển tiếng Việt. Trong công đoạn này, chúng tôi cũng 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 một 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. Để khử các nhập nhằng phát sinh, chúng tôi sử dụng chính cấu trúc WordNet để tính toán các hệ số, độ sâu ngữ nghĩa, khoảng cách ngữ nghĩa giữa các từ … nhằm xác định chính xác nhãn synset của từng từ tiếng Việt.

Bên cạnh đó, chúng tôi đã chứng minh các heuristic của mình để biến chúng thành các thuật toán. Điều này bảo đảm tính chắc chắn của mô hình.

Cuối cùng, để khẳng định tính thực tiễn của mô hình được đề xuất, chúng tôi đã cài đặt một chương trình để minh họa cho mô hình dịch WordNet phần danh từ (với động từ, tính từ, và phó từ thì áp dụng phương pháp tương tự) với độ chính xác tương đương với các ngôn ngữ cùng loại hình (tiếng Hoa, tiếng Thái …).

Luận văn này được chúng tôi trình bày thành 4 phần

Chương 1, Phần mở đầu: Chúng tôi giới thiệu về các hệ cơ sở tri thức hiện có, và các hướng tiếp cận của các ngôn ngữ khác khi giải quyết vấn đề này. Bên cạnh đó, chúng tôi hình thức hóa phát biểu bài toán và trình bày các trở ngại về mặt lý luận, kỹ thuật gặp phải khi giải quyết bài toán đối với tiếng Việt.

Chương 2, Phần cơ sở lý thuyết: Chúng tôi giới thiệu về các quan điểm về ngôn ngữ học làm tiền đề cho tính đúng đắng của mô hình. Đây là một phần rất quan trọng về mặt lý thuyết do, hiện nay, các nhà ngôn ngữ học tiếng Việt còn chưa thông nhất với nhau về các vấn đề về ngôn ngữ học tiếng Việt. Bên cạnh đó, chúng tôi giới thiệu về WordNet tiếng Anh trên các phương diện thống kê, cấu trúc, các cơ sở lý thuyết, các nguyên lý khi xây dựng hệ cơ sở tri thức này…

Chương 3, Phần mô hình: ở phần này, chúng tôi trình bày các bước để giải quyết vấn đề, bên cạnh đó, chúng tôi chứng minh tính đúng đắn của các mô hình này.

Chương 4, Phần cài đặt: ở phần này, chúng tôi trình bày một số các lưu đồ, các cấu trúc dữ liệu để giải quyết bài toán.

Chương 5, Phần đánh giá – kết luận: ở phần này, chúng tôi khảo sát các kết quả thu được, xác định nguyên nhân phát sinh lỗi. Bên cạnh đó, chúng tôi đưa ra các phương án khắc phục và nâng cao độ chính xác của mô hình.

# 1. Mở đầu

## 1.1. Bài toán

Bài toán có thể được khái quát như sau:

Xây dựng một cấu trúc mạng ngữ nghĩa các từ tiếng Việt (chú trọng về danh từ).

Từ đó chúng ta cần phải giải quyết hai vấn đề sau:

- Mạng ngữ nghĩa được tổ chức như thế nào?
- Mô hình nào được sử dụng để gắn các từ tiếng Việt vào mạng ngữ nghĩa đã được chọn ở trên.

### 1.1.1. Về vấn đề mạng ngữ nghĩa:

Để xử lý ngôn ngữ tự nhiên trên máy tính, chúng ta cần có các mô hình về ngữ nghĩa của ngôn ngữ. Thông thường các mô hình này là một từ điển phân loại của các từ hay nhóm từ, tức là mỗi từ sẽ được gán một hay nhiều nhãn ngữ nghĩa. Tuy nhiên, có nhiều mô hình còn đưa ra các 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, kế thừa… Cũng có mô hình chú trọng vào một số lĩnh vực hẹp hay phạm vi nhỏ. Nhưng các mô hình ngữ nghĩa là thành phần không thể thiếu được với một hệ thống xử lý ngôn ngữ tự nhiên. Hiện nay trên thế giới có rất nhiều hệ thống mạng ngữ nghĩa như:

*Với các ngôn ngữ khác:*

Từ điển Chinese Concept Dictionary (CCD) cũng có cách tổ chức tương tự như WordNet nhưng chặt chẽ hơn và thể hiện các đặc trưng riêng của tiếng Hoa.

KoreanWordNet, TamilWordNet: WordNet tiếng Hàn.

EuroWordNet: WordNet của các ngôn ngữ Châu Âu.

*Với tiếng Anh:*

Tuy nhiên do tính phổ biến của tiếng Anh, các hệ cơ sở dữ liệu này thường được xây dựng bằng tiếng Anh:

Hệ thống NETL chuyên sử dụng để dịch các ngữ danh từ trong tiếng Anh. Hiện nay, hệ thống này chỉ cho phép truy vấn cơ sở dữ liệu online và đã được nâng cấp lên version 2.0.

Hệ thống phân loại LLOCE (Longman Lexicon Of Contemporary English) (Arthur T.M 1997). Hệ thống này phân chia các 16.000 từ tiếng Anh thành một câu trúc phân lớp với 14 chủ đề, 129 nhóm, 2449 lớp ngữ nghĩa. Hệ thống này đã được dịch qua tiếng Việt.

Hệ thống LDOCE (Longman Dictionary Of Contemporary English). Hệ thống này với 45.000 từ tiếng Anh được phân bố thành 32 mã ngữ nghĩa, 100 chủ đề nguyên tố.

Hệ thống CORELEX gồm 39937 từ được chia thành các chủ đề.

Hệ thống FRAMENET: Hệ thống ngữ nghĩa này chú trọng vào phần động từ nên chúng tôi không trình bày kỹ ở đây.

…

Hệ thống mạng ngữ nghĩa WordNet: Hệ thống này được bắt đầu phát triển từ năm 1993. Hiện nay, hệ thống này đã phát triển lên version 2.0, bao gồm 152.059 từ được phân bố vào 115.424 ý niệm (synset) 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 được xem là một mạng ngữ nghĩa đầy đủ và hoàn thiện nhất.

Hiện nay, mỗi khi sử dụng các tri thức về thế giới thực, người ta thường sử dụng WordNet. Do đó, chúng tôi đã chọn tiêu chuẩn phân chia cấu trúc của WordNet để tổ chức mạng ngữ nghĩa cho tiếng Việt

### 1.1.2. Về vấn đề mô hình để gán nhãn ngữ nghĩa

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

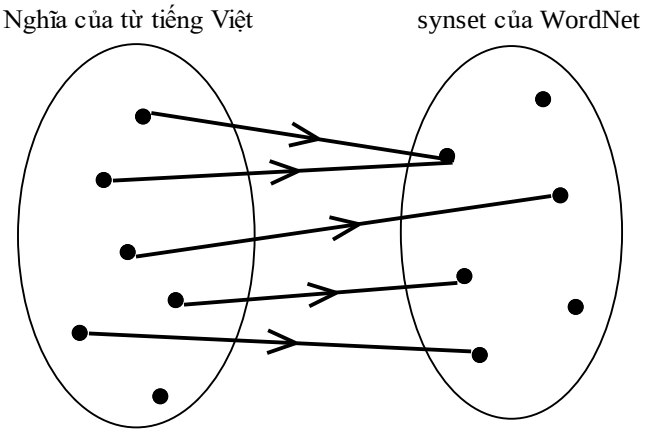
Chúng ta có thể đặc tả bài toán như sau:

$V$ : tập hợp các từ tiếng Việt

$\Omega$ : là tập các synset trong WordNet. Synset trong WordNet có thể được xem là hình vị hóa của ý niệm. Hay nói rõ hơn synset là một nhóm các từ có chung một ý niệm trong WordNet.

$\delta$ : Ánh xạ từ $V \to \Omega$

với $\delta(v) \in \Omega$ , $v \in V$




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

Gỉa thiết, chúng ta có tiên đề sau:

Ánh xạ từ $V \to \Omega$ là ánh xạ n – 1. 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 qui về là tìm ánh xạ $\delta$

## 1.2. Các hướng tiếp cập trước

Trên thế giới đã có nhiều cách tiếp cận để giải quyết bài toán này. Mỗi phương án được đề xuất đều xuất phát từ các 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.

Chúng tôi sẽ giới thiệu lần lượt các cách tiếp cận nhằm xây dựng một hệ thống WordNet của các ngôn ngữ khác theo cách phân loại như trên:

### 1.2.1. Xây dựng từ điển phân loại dựa trên từ điển MRD (Machine Readable Dictionary) đơn ngữ

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 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.

(Bruce R, Guthrie L 1992) đề xuất một phương pháp xây dựng từ điển phân loại ngữ nghĩa của danh từ một cách tự động sử dụng từ điển LDOCE kết hợp với thuật toán khử nhập nhằng để xác định đúng nhãn ngữ nghĩa của từ. Thuật toán sẽ chọn nhãn ngữ nghĩa đúng của từ căn cứ vào nhãn ngữ nghĩa (semantic category) của từ đầu mục từ trong từ điển. Nếu vẫn chưa chọn được nhãn thích hợp, thuật toán sẽ chọn các nhãn ngữ nghĩa lân cận. Cuối cùng, nếu vẫn chưa chọn được nhãn thích hợp, thuật toán sẽ chọn nhãn ngữ nghĩa dựa vào tần số sử dụng của từ. Tác giả đã đánh giá độ chính xác của mô hình này là 80%. (Gutinie, Louise, Brian Slator, Yorick Wilks, và Rebecca Bluce (1990))

Gần đây, (Rigau G., Rodriguez H., Agirre E (1998)) đã đưa ra một phương pháp học để gán nhãn ngữ nghĩa chính xác cho một từ điển phân loại từ một từ điển MRD chưa được phân loại dựa trên nghĩa gốc của từ. Sau đó, sử dụng một số bước xử lý để lọc ra các kết quả tốt nhất.

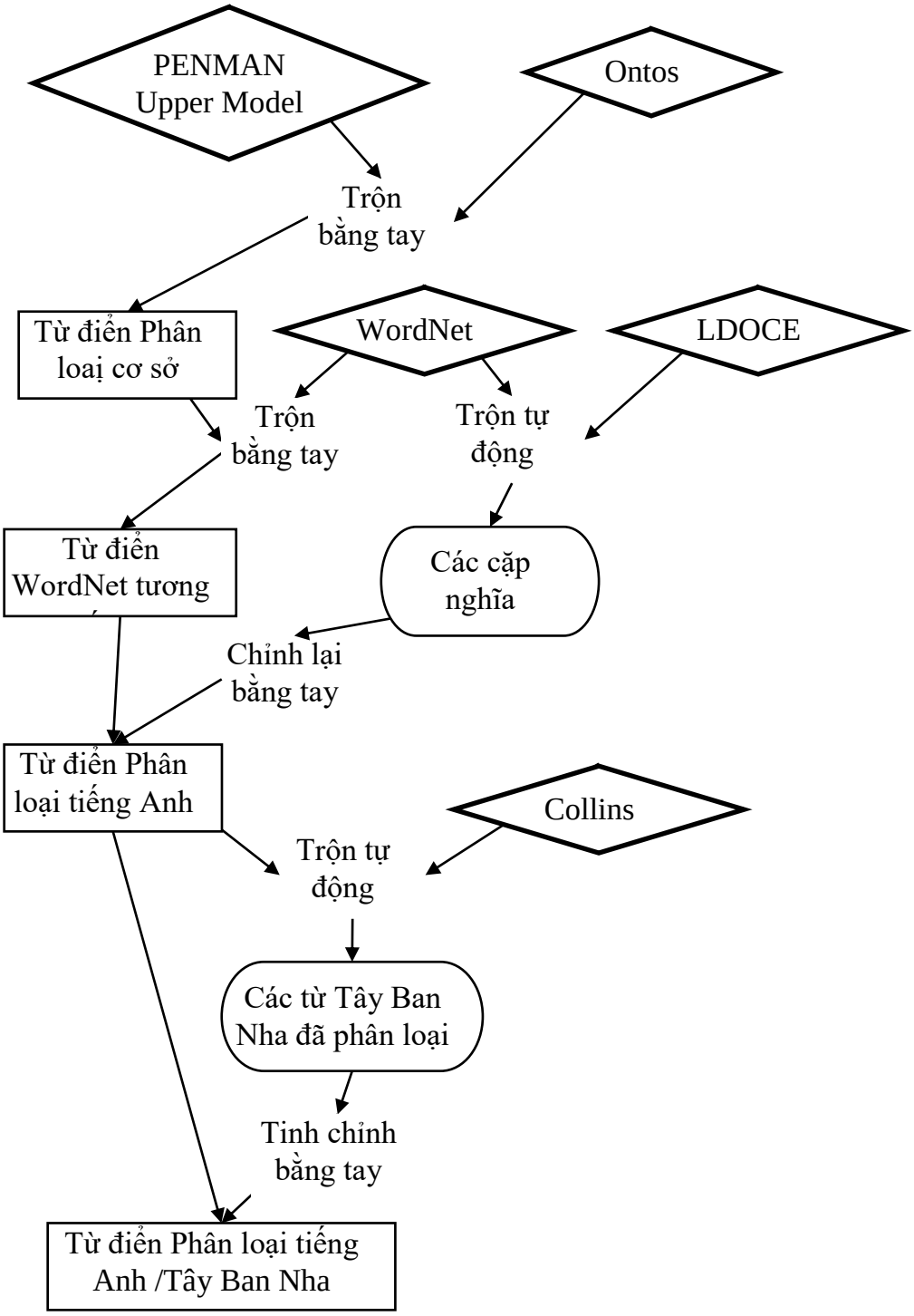
Hướng tiếp cận này có thể áp dụng cho mọi ngôn ngữ, do hầu như ngôn ngữ nào cũng có các từ điển đơn ngữ của ngôn ngữ mình. Tuy nhiên các phương pháp này không cho ra kết quả chính xác do chúng ta cần phải giải quyết các vấn đề của từ điển đơn ngữ như phân loại thiếu, phân loại không hợp lý và có rất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

### 1.2.2. Xây dựng từ điển phân loại sử dụng các liên kết trong các từ điển phân loại đã có

Các phương pháp thuộc cách tiếp cận dạng này sử dụng cho các ngôn ngữ đã có một từ điển đã được phân loại (taxonomy). Khi đó, chúng ta có thể sử dụng từ

điển này kết hợp với các phân loại khác nhau để tạo nên một cấu trúc ngữ nghĩa hoàn chỉnh đa ngôn ngữ.

Knight K. and Luk S. (1994) đề xuất một mô hình để xây dựng một cây ngữ nghĩa lớn để hỗ trợ các cơ sở tri thức của dịch máy. Cấu trúc (ontology) này được xây dựng dựa trên việc kết hợp một số từ điển điện tử trực tuyến, các mạng ngữ nghĩa, các ngữ liệu song ngữ … thông qua một phương pháp bán tự động: kết hợp các định nghĩa, các cấu trúc kế thừa, kết hợp các từ điển song ngữ.




*Hình 1-2: Mô hình của Kevin Kninght và Steve K.Luk*

Daude J., Padro L., Rigau G. (2000) đề xuất một cách tiếp cận, dựa trên các liên kết đã có giữa các từ vựng, các nghĩa được kế thừa. Họ đã sử dụng một thuật toán gán nhãn từ vựng dựa trên các ràng buộc về thống kê để chọn nhãn ngữ nghĩa

thích hợp trong một nhóm các nhãn ngữ nghĩa. Họ đã áp dụng thuật toán này để ánh xạ các từ trong Wordnet 1.5 và 1.6 và thu được kết quả khả quan. Tuy nhiên, khi áp dụng phương pháp này để tìm ánh xạ giữa hai ngôn ngữ khác nhau thì chúng ta sẽ không có được độ chính xác cao. Trước đó J. Daude, L. Padro & G. Rigau (1999) cũng đã áp dụng phương pháp này để xây dựng phân loại tiếng Tây Ban Nha và tiếng Anh nhưng kết quả thu được không khả quan nhiều.

### 1.2.3. Xây dựng từ điển phân loại dựa trên việc 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 cập này thu đuợc kết quả rất tốt nếu chúng ta sử dụng các quan hệ giữa các synset như đồng nghĩa, phản nghĩa, bao hàm …

Okumura A., Hovy E. (1994) mô tả một phương pháp bán tự động để phân bố một từ tiếng Nhật vào một từ điển đồng nghĩa tiếng Nhật bằng cách sử dụng từ điển song ngữ Nhật – Anh làm cấu nối. Bài báo đã sử dụng 3 thuật toán để liên kết từ tiếng Nhật với nghĩa thích hợp tương ứng trong từ điển đồng nghĩa: kết hợp các từ tương đương, kết hợp các giải thích, kết hợp các ví dụ. Tuy nhiên, họ không đưa ra một phương pháp kết hợp cả 3 phương pháp trên.

Rigau G., Agirre E. (1995) đã sử dụng hàm *mật độ quan niệm* để khử nhập nhằng về nghĩa trong 3 từ điển song ngữ (Pháp/Anh, Tây Ban Nha/Anh, Anh /Tây Ban Nha) bằng cách sử dụng WordNet.

### 1.2.4. Sử dụng các heuristic

#### 1.2.4.1. Thứ tự về nghĩa (sense ordering)

#### 1.2.4.2. Xác suất tiên nghiệm(prior probability)

#### 1.2.4.3. Độ tương tự lớn nhất (maximum similarity)

#### 1.2.4.4. Mối quan hệ IS_A (IS_A relation)

#### 1.2.4.5. Sự tương xứng của từ (word match)

#### 1.2.4.6. Đồng xuất hiện (co-occurernce)

## 1.3. Các vấn đề gặp phải khi giải quyết bài toán dịch tự động WordNet qua tiếng Việt

# 2. Cơ sở lý thuyết

## 2.1. 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 ngôn ngữ thiên về cách  làm sao cho máy tính dễ xử lý, và có nhiều trường phái lại rất khó áp dụng máy tính để xử lý.

Về cơ sở lý thuyết áp dụng cho luận văn này, tôi sử dụng và trích quan điểm về ngôn ngữ học tiếng Việt của Tiến Sĩ Đinh Điền (các quan điểm này được nêu rõ trong luận án tiến sĩ Ngôn ngữ học của tiến sĩ Đinh Điền).

### 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.

**Về giá trị ngữ âm**

Đứng về mặt ngữ âm thì hình vị thường trùng với âm tiết. Xét về ngữ âm, âm tiết là đơn vị ngữ âm rất dễ nhận diện, vì nó là một đơn vị phát âm tự nhiên ứng với sự căng lên và chùng 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, hoả, sơn; thực, khán, thính, toạ; mỹ; lạc, hỉ, 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 được. Chúng ta không chỉ vào nước mà nói rằng: đó là *thủy*, mà nói: đó là *nước*; chúng ta cũng không nói: *uống thủy* mà nói: *uống nước*. Nhưng loại tiếng này có thể được dùng để cấu tạo những đơn vị ở bậc 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 qui 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 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 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 gắn bó tương đối chặt chẽ về mặt nội dung và hình thức. Chẳng hạn: *nhà nước, xóm làng, quần áo; thợ sơn, hoa hồng, cá thu; quốc gia, giang sơn, huynh đệ; tàu thuỷ, bình thuỷ, lính thuỷ; dề dàng, gọn gàng, 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 tiếng kết với nhau thành từ. Ví dụ: *hợp tác xã, câu lạc bộ, cộng sản chủ nghĩa, chủ nghĩa xã hội*....

### 2.1.1.2. Từ

Chúng tôi sử dụng quan điểm của (Đinh Điền, 2004) xem "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, (Đinh Điền, 2004) 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 hoá 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.

Tác giả còn đưa ra cách hình thức hoá quan niệm từ tiếng Việt nói trên theo phương pháp như sau: Để lưu thông tin về ranh giới từ tiếng Việt, tác giả sử dụng khái niệm từ từ điển học kết hợp với nhãn hình thái trong từ điển. Nghĩa là khái niệm từ từ điển học ở đây không hoàn toàn như trong định nghĩa, mà phải là "những đơn vị mà căn cứ vào đặc điểm ý nghĩa của nó phải xếp riêng trong từ điển VÀ có đánh dấu đây là đơn vị từ của ngôn ngữ". Thật vậy, vì trong từ điển chúng tôi đưa vào nhiều đơn vị ngôn ngữ khác nhau (từ, ngữ, cụm từ cố định,…) nên chúng tôi cần phải có thêm nhãn hình thái để đánh dấu đây là đơn vị từ của ngôn ngữ. Việc chọn lựa những từ nào sẽ đưa vào từ điển là hoàn toàn do các nhà ngôn ngữ hay người xây dựng kho ngữ liệu quyết định, dựa theo quan điểm về từ đã nêu trên. Với cách hình thức hoá như vậy, chúng ta có thể thay đổi quan niệm về từ mà không phải thay đổi về cấu trúc dữ liệu cũng như mô hình xử lý của máy tính.

Tóm lại từ trong luận văn này được hiểu là "từ từ điển" + "nhãn đơn vị từ".

Giống như cách trình bày của WordNet, trong luận văn, chúng tôi sẽ dùng thêm ký 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 Không làm thay đổi từ loại

**Dẫn xuất**

Có 2 dạng 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 thay đổi từ loại của từ

### 2.1.3. Nghĩa của từ

Lưỡng phân ngôn ngữ, ta nhận ra hai mặt của nó: 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 có vỏ ngữ âm như ta đọc lên ([kej1]), 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 nếu ra từ lâu và cũng đã có nhiều cách hiểu, nhiều định nghĩa khác nhau. Tuy vậy, việc nêu lại và bình luận các quan điểm về nghĩa, chúng ta đành tạm gác sang một bên cho cách trình bày ở đây đỡ cồng kềnh, phức tạp.

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, mà anh ta quy chiếu, gán nó vào đúng sự vật có tên gọi là từ đó như cả cộng đồng xã hội vẫn gọi; đồng thời ít nhiều anh ta cũng biết được những đặc trưng bản chất của sự vật đó; và anh ta sử dụng từ đó trong giao tiếp đúng với các mẹo luật mà ngôn ngữ có từ đó cho phép; ta nói rằng anh ta 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 anh ta có thể:

- Quy chiếu, gắn được từ cây vào mọi cái cây bất kỳ 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 ngữ pháp tiếng Việt.

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

Cho tới nay, đa số các 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 kỳ 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... và 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 với những cái mà nó (từ) chỉ ra (những cái mà nó làm tín hiệu cho).

### 2.1.3.2. <u>Nghĩa của từ tồn tại ở đâu?</u>

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 không 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, trong bộ óc trí tuệ của con người chỉ tồn tại sự hiểu biết về nghĩa của từ chứ không phải là nghĩa của từ.

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 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 (significative meaning): Là liên hệ giữa từ với ý (hoặc ý nghĩa, ý niệm – signification – 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 nghĩa cấu trúc.

Nghĩa ngữ dụng (pragmatical meaning), còn được gọi là nghĩa biểu thái, nghĩa hàm chỉ (connotative meaning) là lối liên hệ giữa từ với thái độ chủ quan, cảm xúc của người nói.

Nghĩa cấu trúc (structural meaning) là mối quan hệ giữa từ với các từ khác trong hệ thống từ vựng. Quan hệ giữa từ này với từ khác thể hiện trên hai trục: trục đối vị (paradigmatial axis) và trục ngữ đoạn (syntagmatical axis). Quan hệ trên trục đối vị cho ta xác định được giá trị của từ, khu biệt từ 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.

Thật ra, những phân biệt như trên là cần thiết và hợp lí; nhưng không phải các thành phần nghĩa đó hiện diện trong mỗi từ bao giờ cũng đồng đều và rõ ràng như nhau. Vì thế, trong từ vựng-ngữ nghĩa học, nhiều khi người ta chỉ nhắc đến nghĩa ngữ dụng, nghĩa cấu trúc; thậm chí cả nghĩa biểu vật nữa, như những xác nhận về sự tồn tại của chúng hơn là phân tích, chứng minh cho thật minh bạch.

Đối với từ vựng-ngữ nghĩa học, cái quan trọng nổi lên hàng đầu là nghĩa biểu niệm. Cần phải hiểu mối liên hệ mà chúng ta nói tới trong quan niệm về nghĩa của từ ở đây chính là mối liên hệ chỉ xuất, mối liên hệ phản ánh; cho nên nghĩa biểu niệm cũng có thể hiểu là sự phản ánh sự vật-biểu vật (đúng hơn, là phản ánh các thuộc tính, các đặc trưng của chúng) trong ý thức con người, được tiến hành bằng 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 phải 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 một 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 tới chân lí khoa học; và có thể được diễn đạt bằng hàng loạt những 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 chugn, 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 lí 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 [hoá học] về nước là: Hợp chất của oxy và hidro mà trong thành phần của mỗi phân tử nước, có hai nguyên từ hydro với 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 thật 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 mắm, nước xốt, nước dứa, nước ép hoa quả*

*phở nước* *(đối lập với phở xào)*

*mỡ nước* *(đối lập với mỡ khổ)*

*nước gang (gang lỏng - Ví dụ: Đổ nước gang vào khuôn đúc)*

…

*nước dãi, nước bọt, nước mắt, nước giải, nước ối…*

Phân tích như trên đây chứng tỏ rằng nghĩa và khái niệm không đồng nhất. Đó là nói về các từ nói chung. Đối với nhiều thuật ngữ khoa học, sự phân biệt giữa nghĩa và khái niệm không cần đặt ra nữa: chúng đã tiệm cận đến giới hạn của nhau.

### 2.1.3.5. Từ, dạng thức và nguyên lý trình bày từ điển

Theo (John Lyons, 1995) Phần lớn các từ tiếng Anh không phải chỉ có hơn một dạng thức. Chúng còn có thể có hơn một nghĩa; và xét theo khía cạnh này thì tiếng Anh là tiêu biểu cho các ngôn ngữ tự nhiên. (Mặc dù tồn tại những ngôn ngữ tự nhiên trong đó mỗi từ có một và chỉ có một dạng thức, nhưng gần như chắc rằng không có, và chưa hề có ngôn ngữ tự nhiên nào trong đó mỗi từ có một và chỉ một nghĩa). Ví dụ, từ ‘food’ (bàn chân) có một số nghĩa. Nếu chúng ta muốn phân biệt các nghĩa này khi trình bày, chúng ta có thể đánh số chúng và ghi kèm các số này như là chú dẫn cho các ký hiệu biểu nghĩa, ví dụ: ‘food1’, ‘food2’,‘food3’... Khái quát hơn, giả định rằng X là dạng trích dẫn của một từ, chúng ta sẽ biểu thị từ đó là ‘X’ và nghĩa của nó (tức tập hợp một hoặc hơn một các nghĩa của nó) là “X”; và nếu nó có hơn một ý nghĩa thì chúng ta có thể phân biệt các ý nghĩa này như là “X1”, “X2”, “X3” v.v...

Tất nhiên, việc dùng các chú dẫn như vậy chỉ là một công cụ trình bày giản tiện, không nói lên điều gì cả về nghĩa của từ. Khi nào cần phải xác định các nghĩa khác nhau chứ không phải chỉ là biểu diễn ký hiệu như trên đây, chúng ta có thể dùng lối định nghĩa hoặc khúc giải. Ví dụ, đối với từ ‘foot’, chúng ta có thể nói rằng “foot1” là “phần kết thúc của chân”, rằng “foot2” là “phần thấp nhất của một ngọn đồi hay ngọn núi”...Do đó, chúng tôi giả định các từ điển mà chúng tôi sử dụng đều có các nghĩa của từ mang hai đặc điểm sau: (i) tách biệt, và (ii) có thể phân biệt với nhau.

### 2.1.4. So sánh từ tiếng Việt và tiếng Anh về mặt hình thái

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.

#### 2.1.4.1. 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. Chúng tôi minh họa sự khác biệt này bằng một ví dụ của (Đinh Điền, 2004)

Xét mục từ "*display*" và nghĩa tiếng Việt tương ứng của nó là "*hiển_thị*" thì đây là ánh xạ *1-1*, còn nếu mục từ "*display*" (mục từ khác), có nghĩa tiếng Việt tương ứng là "*màn hiển_thị*" thì đây là ánh xạ *1-n*. Nếu mục từ tiếng Anh là "*carry-out*" và nghĩa tiếng Việt là "*thực_hiện*" thì đây là ánh xạ *m-1*. Nếu mục từ tiếng Anh là "*call-up*" và nghĩa tiếng Việt là "*gọi điện_thoại*" thì đây là ánh xạ *m-n,...*

#### 2.1.4.2. Sự khác biệt về hình thái học

Bên cạnh sự về từ vựng, sự khác nhau về loại hình ngôn ngữ cũng tạo nên sự khác nhau về hình thái từ của 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.

( Đinh Điền, 2004) đưa ra phương pháp đối chiếu về mặt hình thái từ giữa tiếng Anh và từ 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 từ hư. Vì vậy, ứng với một từ trong tiếng Anh, khi chưa biến cách, ánh xạ của từ tiếng Việt tương ứng là 1-1 (nếu không tính yếu tố khác biệt về từ vựng hoá), 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 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 | book*s*; two student*s* | những/các +N; Φ | *những/các* cuốn-sách; hai sinh_viên. |
| 2 | Động từ ngôi 3 số ít | V + -s | He sleep*s* | Φ | Anh ấy ngủ |
| 3 | Sở hữu cách | X's Y | John*'s* book; teacher*s'* books | Y của X | cuốn-sách *của* John; các cuốn-sách *của* những giáo_viên. |
| 4 | Hiện phân từ | V-ing | Sleep*ing* | đang V | *đang* ngủ |
| 5 | Quá khứ | V-ed | Work*ed* | đã V | (*đã*) làm việc |
| 6 | Quá phân từ | V-en | Spok*en* | đã V | (*đã*) nói |
| 7 | So sánh hơn | Adj-er<br>Adv-er | short*er*<br>slow*er* | Adj - hơn | ngắn *hơn*<br>chậm *hơn* |
| 8 | So sánh nhất | Adj-est<br>Adv-est | short*est*<br>slow*est* | Adj - nhất | ngắn *nhất*<br>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 bằng 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 giữa từ 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 dẫn xuất 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_hoá, 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, ...*

Danh sách đối chiếu hình thái từ cho các hậu tố và tiền tố dẫn xuất trong tiếng Anh và từ tiếng Việt tương ứng đã được liệt kê chi tiết trong (Đinh Điền, 2004).

**Những khác biệt do đặc 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). (Đinh Điền, 2004) xem 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.

### 2.1.5. So sánh từ tiếng Việt và 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 từ loại như: danh từ ( noun, nom); động từ (verb, verbe); tính từ( adjective, adjectif); đại từ( pronoun, pronom...).

#### 2.1.5.1. 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).

#### 2.1.5.2. 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 đa số các nhà ngôn ngữ học chấp nhận đó là chia từ loại tiếng Việt thành 2 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, càm từ và từ chỉ hướng.

#### 2.1.5.3. Đối chiếu nhãn từ loại của 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 chỉ 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 khác 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 giữa từ loại tiếng Anh và từ loại tiếng Việt không là ánh xạ 1-1, nghĩa là từ X trong tiếng Anh có nghĩa Y thì không chắc từ loại X là là từ loại của Y. (Đinh Điền, 2004) đưa ra các bản đối chiếu từ loại giữa hai ngôn ngữ như sau:

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,..*<br>Danh từ riêng (NP): *John, Hanoi,…*<br>Danh từ (NN): *attention, help,..*<br>Trạng từ (RB): *above,below,here,..* | Danh từ (N): *bàn, người*, …<br>Danh từ riêng (Nn): *Tuấn, Hà-Nội*,…<br>Động từ (V): *chú_ý, giúp_đỡ*, …<br>Danh từ vị trí (Np): *trên, dưới, đây,*… |
| Động từ (VB): *eat, learn*,… | Động từ (V): *ăn, học,* … |
| Tính từ (JJ): *big, good,..*<br>Tính từ (JJ): *every, each,..*<br>Tính từ (JJ): *electric, national,…* | Tính từ (J): *lớn, tốt*, …<br>Phó từ (R): *mọi, từng,.*<br>Danh từ (N): *điện, quốc_gia*, … |
| Đại từ (PP): *I, you, he,..* | Đại từ (P): *tôi, anh, anh ấy,..* |
| Trạng từ (RB): *strongly, slowly,...*<br>Trạng từ (RB): *still, just,..;already,..*<br>Trạng từ(RB): *perhaps, of course,...*<br>Trạng từ (RB): *even, ...*<br>Trợ động từ (MD): *can, may, will,...* | Tính từ (J): *mạnh_mẽ, chậm_chạp,..*<br>Phó từ (R): *vẫn, vừa,..;đã, đang,sẽ,..*<br>Phó từ (R): *có_ lẽ, tất_nhiên, ...*<br>Trợ từ (M): *cả, chính,...*<br>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, ..*<br>Tính từ (JJ): *few, several, some,..* | Số từ (Q): *một, hai*<br>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 tiếng Anh và tiếng Việt*

| | **Ý nghĩa ngữ pháp** | **Từ pháp tiếng Anh** | **Từ pháp tiếng Việt** |
|---|---|---|---|
| 1 | Danh từ số nhiều | book*s*/NNS;<br>two/CD student*s*/NNS | *những*/Q cuốn/L-sách/N;<br>hai/Q sinh_viên/N |
| 2 | Động từ ngôi 3 số ít | He/PP sleep*s*/VBZ | Anh/P-ấy/P ngủ/V |
| 3 | Sở hữu cách | John/NP *'s*/POS book/NN;<br>teacher*s*/NNS *'*/POS book*s*/NNS | cuốn/L-sách/N *của*/I John/Nn;<br>các/Q cuốn/L-sách/N *của*/I những/Q giáo_viên/N. |
| 4 | Hiện phân từ | sleep*ing*/VBG | *đang*/R ngủ/V |
| 5 | Quá khứ | work*ed*/VBD | (*đã*/R) làm_việc/V |
| 6 | Quá phân từ | spok*en*/VBN | (*đã*/R) nói/V |
| 7 | So sánh hơn | short*er*/JJR<br>slow*er*/RBR | ngắn/J *hơn*/J<br>chậm/J *hơn*/J |
| 8 | So sánh nhất | short*est*/JJS<br>slow*est*/RBS | ngắn/J *nhất*/J<br>chậm/J *nhất*/J |

Với từ dẫn xuất

Như đã 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. Chi tiết của các luật này đã được trình bày ở (Đinh Điền, 2004).

### 2.1.6. 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, nhưng trong một ngữ cảnh cụ thể, thì chúng sẽ mang một nghĩa nhất định nào đó. Chẳng hạn, danh từ "bank" trong tiếng Anh có thể là "ngân hàng", hoặc "bờ sông" hoặc *"dãy*"; danh từ "*đường"* trong tiếng Việt có thể có nghĩa là *"đường ăn*" (sugar) hay *"đường đi*" (line),… Để dễ phân biệt các nghĩa từ vựng khác nhau, các nhà ngữ nghĩa học, từ vựng học và tâm lý học – ngôn ngữ đã phân chia toàn bộ các ý nghĩa từ vựng có thể có thành hệ thống các ý niệm (cây ý niệm) và mỗi ý niệm như vậy được coi như là một nhãn ngữ nghĩa của từ.

Chẳng hạn, với danh từ "*bank*" nói trên, các nghĩa tương ứng của chúng sẽ là: "*ngân hàng*" thuộc về ý niệm "*công trình xây dựng nhân tạo*" (nhãn HOU); "*bờ sông*" sẽ thuộc về ý niệm "*công trình thiên tạo*" (nhãn NAT); "dãy" sẽ thuộc về ý niệm "*sự sắp xếp tổ chức*" (nhãn GRP). Tương tự cho danh từ *"đường"* trong tiếng Việt, nghĩa "*đường ăn*" sẽ được xếp vào ý niệm "*hoá chất*" (nhãn CHM); còn nghĩa "*đường đi*" sẽ được xếp vào ý niệm "*đường nét, dấu vết*" (nhãn LIN);…

(Theo Đinh Điền, 2004) Các nhà ngôn ngữ học – tâm lý đã chứng minh bằng thực nghiệm là: với một từ kích thích "*aunt*" cho nhiều người khác nhau, thì đa số đều cho biết trong đầu họ nghĩ đến từ "*uncle*" trước nhất, điều này chứng tỏ rằng: ngay "lời nói bên trong" của con người chúng ta, thì từ "*uncle*" và "*aunt*" đã có quan hệ với nhau. 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 của tư duy, mà tư duy là sự 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, trong tiếng Anh là HUSBAND, tiếng Hoa là /fu/ ), nhưng cái được biểu đạt thì như nhau. Vì ý niệm và từ không trùng nhau, nên hệ thống ý niệm này đảm bảo sử dụng được 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 tư 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 hoá. Các nhà nghiên cứu chia phổ quát ngôn ngữ thành 2 dạng sau:

- Các phổ quát về thực thể: là những nét chung về sự tổ chức các thực thể ngôn ngữ. Chẳng hạn, mọi ngôn ngữ đều tồn tại các phạm trù danh từ và động từ, nó là cơ sở để biểu hiện cấu trúc chìm của câu trong mọi ngôn ngữ.
- Các phổ quát về dạng thức: chẳng hạn, ngữ pháp tạo sinh coi rằng bộ phận cơ sở của cú pháp trong mọi ngôn ngữ thì giống nhau.

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

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

(Đinh Điền,2004) đưa ra phương pháp đối chiếu nhãn ngữ nghĩa của tiếng Anh và tiếng Việt như sau:

#### 2.1.6.1. 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"; "help,NN" và "giúp đỡ,V"*): Khi đó, chuyển từ 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.

#### 2.1.6.2. Với liên kết 1-n

Với trường hợp này, một từ tiếng Anh được dịch ra bởi nhiều từ tiếng Việt. Khi đó, vấn đề là làm thể nào để chọn đúng nhãn ngữ nghĩa của chúng các từ này. Trong trường hợp này, (Đinh Điền, 2004) 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 vào 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_toán_hoá/V"* có ánh xạ hợp lệ là "*computerization/NN*" → "*điện_toán_hoá/V*")

#### 2.1.6.3. 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 thể nào để chọn đúng nhãn ngữ nghĩa của chúng các từ này. Trong trường hợp này, (Đinh Điền, 2004) đưa ra 2 trường hợp xem ánh xạ m-

1 là m ánh xạ 1-1 giữa các m từ tiếng Anh và một từ tiếng Việt và xem xét các ánh xạ nào là ánh xạ ánh xạ chính.

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"* →*"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 cửa 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 "*brother/NN*" → "*anh/V*")

#### 2.1.6.4. <u>Với liên kết m-n</u>

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 thể nào để chọn đúng nhãn ngữ nghĩa của chúng các từ này. (Đinh Điền, 2004) xem 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 ánh xạ chính và đưa về một trong ba trường hợp trên.

### 2.2. <u>Wordnet</u>

Năm 1980, Miller và cộng sự tại trường Đại học Princeton (Mỹ) đã xây dựng nên WordNet. **WordNet** là một cơ sở dữ liệu tri thức từ vựng học 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. Trong từ trong Wordnet được phân loại thành danh từ, động từ, tính từ, và phó 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. Hiện nay WordNet đã phát triển lên đến version 2.0 bao gồm hơn 110.000 synset 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 nguyên cứu. Wordnet là một kho tàng tri thức ngữ nghĩa từ vựng khổng lồ và đã được rất nhiều các nhà ngôn ngữ học và ngôn

ngữ học – máy tính khai thác, 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, ..).

Trong tiếng Anh, thông thường một từ có thể có nhiều nghĩa (*word meaning*) do đó mỗi nghĩa của nó sẽ thuộc vào những tập đồng nghĩa (synset) khác nhau. Ngược lại mỗi tập đồng nghĩa lại có thể chứa một hoặc nhiều hơn một từ khác nhau. Để dễ dàng hơn ta xét ví dụ sau.

*Ví dụ* : Khi tìm từ letter trong WordNet ta sẽ được kết quả như sau:

*+ the noun letter has 5 senses* :

1. **letter**, missive : *a written message addressed to a person or organization; "wrote an indignant letter to the editor"*.
2. **letter**, letter of the alphabet, alphabetic character : *the conventional characters of the alphabet used to represent speech; "his grandmother taught him his letters"*.
3. **letter** : *a strictly literal interpretation (as distinct from the intention); "he followed instructions to the letter"; "he obeyed the letter of the law"*.
4. **letter**, varsity letter : *an award earned by participation in a school sport; "he won letters in three sports"*.
5. **Letter,** *owner who lets another person use something (housing usually) for hire*

*Giải thích ví dụ* : Trong WordNet danh từ **letter** có 4 nghĩa thuộc vào 4 tập đồng nghĩa

1. Tập đồng nghĩa thứ nhất gồm: **letter**, **missive** với nghĩa tiếng Việt tương ứng là “lá thư”, “thư tín”.
2. Tập đồng nghĩa thứ hai gồm: **letter**, **letter of the alphabet**, **alphabetic character** với nghĩa tiếng Việt tương ứng “ký tự”, “chữ” hay “chữ cái”.
3. Tập thứ ba chỉ gồm một từ: **letter** với nghĩa tiếng Việt là “nghĩa chật hẹp”, “nghĩa mặt chữ”.
4. Tập thứ tư gồm hai từ: **letter**, **varsity letter** với nghĩa tiếng Việt tương ứng là “huy hiệu”, “danh hiệu” tặng cho những sinh viên có thành tích thể thao đặc biệt ở trường.
5. Tập cuối cùng: chỉ một từ với nghĩa tương ứng là “người cho thuê”.

Như vừa trình bày trong phần trên, các từ trong WordNet được sắp xếp vào thành các tập đồng nghĩa. Các tập đồng nghĩa này có mối quan hệ ngữ nghĩa với nhau. Các mối quan hệ đó bao gồm;

### 2.2.1. Số lượng từ , synset trong WordNet

*Bảng 2-4: 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 |

### 2.2.2. Thông tin về tính đa nghĩa

*Bảng 2-5: Số lượng từ và nghĩa của WordNet 2.0*

| Từ loại | Đơn nghĩa | Đa nghĩa | |
|---|---|---|---|
| | Số lượng từ và nghĩa | Số lượng từ | Số lượng nghĩa |
| Danh từ | 99524 | 15124 | 42325 |
| Động từ | 6256 | 5050 | 18522 |
| Tính từ | 16103 | 5333 | 14979 |
| Phó từ | 3901 | 768 | 1913 |
| Tổng cộng | 125784 | 26275 | 77739 |

*Bảng 2-6: Bảng trung bình từ / nghĩa*

| Từ loại | Trung bình từ /nghĩa | |
|---|---|---|
| | Kể cả các từ đơn nghĩa | Không kể các từ đơn nghĩa |
| Danh từ | 1.23 | 2.79 |
| Động từ | 2.17 | 3.66 |
| Tính từ | 1.44 | 2.80 |
| Phó từ | 1.24 | 2.49 |

### 2.2.3. Hình thức hóa WordNet

Trong các ứng dụng ngôn ngữ học, sự phân cấp khái niệm sự phân cấp khái niệm có thể được xem như khái niệm "hypernym" trong WordNet, các quan hệ khác như "meronymy" và "antonymy" có thể được sử dụng phụ trợ. Sự phân tích về mặt lý thuyết này thể hiện sự phụ thuộc của các khái niệm, sự kế thừa của các khái niệm nhỏ với các khái niệm lớn. Sự phân tích này không cung cấp một hệ thống hoàn chỉnh các tiên đề của các mối quan hệ ngữ nghĩa, nhưng nó có thể tạo điều kiện để tìn ra các đặc tính logic của các khái niệm này. Ví dụ: nó không trả lời câu hỏi rằng "meronymy" có đc chuyển tiếp hay không? Nhưng nó xác định các điều kiện chuyển tiếp bề mặt để xác định các đặc tính mà quan hệ "meronymy" không thể có được.

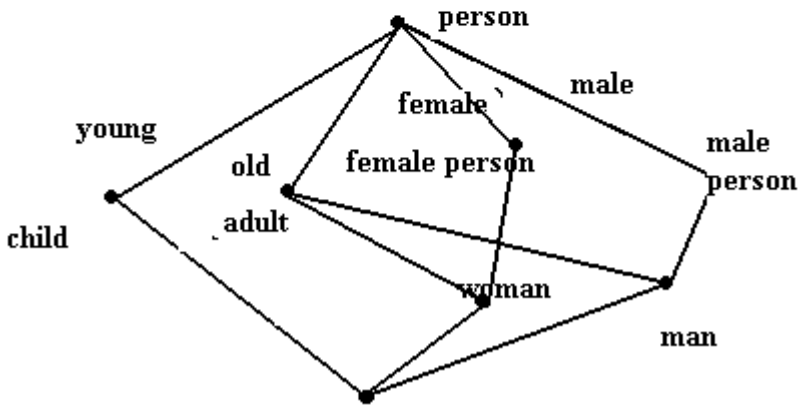
Cấu trúc khái niệm được mô hình bởi một mạng phân cấp được hình thức hóa dưới dạng lý thuyết dàn (trong toán học). Điều này cho phép chúng ta có thể sử dụng các ký tự hình vị để biểu diễn ngữ nghĩa. Một vài khái niệm về dàn được từ vựng hóa, bởi vì chúng có thể được đinh nghĩa bằng từ. Các khái niệm còn lại chúng ta không thể biểu diễn bằng từ được (cách chuỗi rỗng về dàn) có thể được mô tả bởi các quan hệ ngữ nghĩa như hypernym, thuộc tính hay các mối quan hệ giữa chúng, nhưng chúng không định nghĩa bởi từ. Hai hình thức quan trọng của ngữ cảnh cần được nguyên cứu về mối quan hệ ngữ nghĩa. *Ngữ cảnh biểu thị (denotative context)*, mà nó chứa các ngữ nghĩa cần biểu đạt của các dạng từ và các thứ tự ngữ cảnh. *Ngữ cảnh từ vựng (lexical context)* sử dụng các từ làm thành tố. Từ luôn được sử dụng một cách không nhập nhằng bằng cách sử dụng các con số

(ví dụ các *sense number* trong WordNet) để tránh các vấn đề về đa nghĩa và đồng âm. Ngữ cảnh biểu thị thường không đầy đủ bởi vì nó không thể liệt kê thành danh mục tất cả các trường hợp biểu thị của một ngôn ngữ. Tuy nhiên, vì các mối quan hệ về nghĩa ám chỉ các mối quan hệ giữa các trường hợp cần diễn đạt nên chúng không thể xác định chính xác dựa vào từ mà không xem xét các trườhg hợp diễn đạt trong ngữ cảnh. Các từ điển đồng nghĩa hay các từ điển có thể xem đây là ngữ cảnh từ vựng. Từ là tên cho khái niệm trong *ngữ cảnh biểu thị* và là đối tượng hình thức trong *ngữ cảnh từ vựng*. Do đó, cần phải tìm các mối quan hệ về nghĩa có cùng đặc tính trong cả hai ngữ cảnh. WordNet được hình thức như là một ngữ cảnh từ vựng, nhưng chỉ trong trường hợp các synset danh từ.

#### 2.2.3.1. Khái niệm hình thức

Trước tiên, khái niệm hình thức định nghĩa ngữ cảnh hình thức *K* là một bộ ba $(G, M, I)$ bao gồm 2 tập *G* và *M* và một quan hệ I giữa G và M $(I \subseteq G \times M)$. Các thành phần của *G* và *M* được gọi là đối tượng hình thức và thuộc tính hình thức tương ứng. Mối quan hệ được viết *gIm* hoặc $(g, m) \in I$ và được đọc là "đối tượng hình thức *g* có thuộc tính hình thức *m"*. Một ngữ cảnh hình thức có thể được biểu diễn bởi một bảng chéo, bảng chéo này có các hàng là các đối tượng hình thức *g*, còn cột là các thuộc tính hình thức *m*. Giao của hàng *g* và cột *m* là *gIm*. Đồ thị Hình 2-1 à mọt ví dụ của ngữ cảnh hình thức . Nó có *"person","adult","female person","male person","child","woman"* .. là các đối tượng hình thức, *"young","old","female","male"* là các thuộc tính hình thức. Điều này cho thấy cách dùng này của thuật ngữ *khái niệm* phải được xem là hơi khác với *khái niệm* trong ngôn ngữ học.

| | Young | Old | Female | Male |
|---|---|---|---|---|
| Person | | | | |
| Adult | | X | | |
| Female person | | | X | |
| Male person | | | | X |
| Child | X | | | |
| Woman | | X | X | |
| Man | | | X | X |




*Hình 2-2: Ngữ cảnh hình thức và lược đồ của của nó trong dàn khái niệm*

Trong ngữ cảnh *(G,M,I)* tập hợp các thuộc tính hình thức phổ biến của tập $A \subseteq G$ của đối tượng hình thức được định nghĩa bởi $tA := \{m \in M \mid gIm, \forall g \in A\}$ và tương tự tập hợp của các đối tượng hình thức phổ biến của tập $B \in M$ của các thuộc tính hình thức là $\varepsilon B := \{g \in G \mid gIm, \forall m \in B\}$. Ví dụ: trong ngữ cảnh hình thức của đồ thị Hình 2-3 $t\{\text{man}\}=\{\text{old,male}\}$ và $\varepsilon\{\text{old}\}=\{\text{adult, woman, man}\}$.

Một cặp (A,B) được xem là khái niệm hình thức của ngữ cảnh hình thức *(G,M,I)* nếu $A \subseteq G, B \subseteq M, A = \varepsilon B,$ B=tA. Khái niệm hình thức còn được ký hiệu bởi $c, c_1, ..c_i$ . Với khái niệm hình thức $c := (A, B)$, $A$ được gọi là phạm vi (được định nghĩa là *Ext(c)* ) và B được gọi là mục đích (được định nghĩa là *Int(c)*) của khái niệm hình thức. Ví dụ Hình 2-4 cho biết (*{adult, woman, man},{old}*) là khái niệm

hình thức bởi vì t{adult,woman,man}={old} và $\varepsilon${old}={adult, woman, man}. Tập hợp của tất cả các khái niệm hình thức của (*G,M,I*) được định nghĩa bởi $K(G,M,I)$. Cấu trúc quan trọng nhất trên *K(G,M,I)* được cho bởi mối quan hệ khái niệm hình thức lớn, khái niệm hình thức nhỏ. Định nghĩa về khái niệm hình thức lớn và khái niệm hình thức nhỏ như sau:

Khái niệm hình thức $c_1$ được gọi là khái niệm hình thức nhỏ (formal subconcept) của khái niệm hình thức $c_2$ *(ký hiệu* $c_1 \le c_2$*)* nếu $Ext(c_1) \subseteq Ext(c_2)$, tương đương với $Int(c_2) \subseteq Int(c_1)$ ; $c_2$ được xem là khái niệm hình thức lớn của $c_1$. *(ký hiệu* $c_1 \ge c_2$*)*. Ví dụ: (*{adult, woman, man},{old}*) là khái niệm hình thức lớn của (*{ woman},{old,female}*) do (*{adult, woman, man},{old}*) có nhiều đối tượng hình thức nhưng ít thuộc tính hình thức hơn (*{ woman},{old,female}*). Theo cách định nghĩa này, mỗi khái niệm hình thức là một khái niệm hình thức lớn của chính nó. Điều này hơi khác so với ngôn ngữ tự nhiên, khi đó, khái niệm không thể là khái niệm lớn của chính nó. Mối quan hệ $\le$ là một quan hệ thứ tự trong toán học và được gọi là thứ tự khái niệm hình thức của *B(G,M,I)* với các tập hợp của các khái niệm hình thức của một dàn toán học được ký hiệu bởi *B(G,M,I).*

Về phương diện đồ họa, dàn toán học có thể được xem như là các lược đồ đường thẳng biểu thị một khái niệm hình thức bởi một vòng tròn nhỏ. Với mỗi đối tượng hình thức g, khái niệm hình thức nhỏ nhất mà phạm vi thuộc về nó được được ký hiệu bởi $\gamma g$ ; và đối với mỗi thuộc tính hình thức *m*, khái niệm hình thức lớn nhất của nó mà mục đích *m* thuộc về được định nghĩa là $\mu m$. Khái niệm $\gamma g$ và $\mu m$ được gọi là đối tượng khái niệm (object concept) và thuộc tính khái niệm (attribute concept). Trong lược đồ đường thẳng, chúng ta không cần thiết phải viết đầy đủ mục đích và phạm vi của mỗi khái niệm. Thay vào đó, tên (cách nói) của mỗi đối tượng hình thức *g* được viết nhẹ dưới vòng tròn của $\gamma g$ và tên của mỗi thuộc tính *m* được viết nhẹ dưới vòng tròn của $\mu m$. Nửa dưới của hình 2-5 trình bày biểu đồ đường thẳng về dàn khái niệm gồm tất cả các ngữ cảnh hình thức được mở ra bằng việc bắt đầu với khái niệm hình thức và sau đó bằng việc tập hợp tất cả các ngữ

cảnh hình thức được viết trong các khái niệm hình thức nhỏ của các khái niệm hình thức đó. Tương tự, intent được rút ra bằng việc tập hợp các thuộc tình hình thức được viết trong các khái niệm hình thức lớn của khái niệm hình thức .

### 2.2.3.2. WordNet là một ngữ cảnh hình thức

Chúng ta hình thức hóa WordNet theo 2 ngữ cảnh: Trong ngữ cảnh biểu thị $K_D := (D, A_D, I_D)$ biểu thị $d \subseteq D$ là đối tượng hình thức. Tập hợp $A_D$ của các thuộc tính hình thức bao gồm các thuộc tính của biểu thị . Khái niệm có thể được bổ sung được gọi tên bởi một từ không nhập nhằng $w \in W$ thông qua hàm $dnt : W \to B(K_D)$. Bởi vì từ không nhập nhằng, *dnt* là một hàm số thực. Một cấu trúc quan hệ bao gồm một ngữ cảnh biểu thị $K_D$, một tập hợp $W$ của từ, hàm *dnt*, và các liên hệ khác trên biểu thị , thuộc tính, hay khái niệm được gọi là cấu trúc biểu thị và được ký hiệu $\varphi_D$ . Một ngữ cảnh từ vựng $K_L := (W, A_L, L_L)$ bao gồm một tập hợp $W$ các từ không nhập nhằng như là các đối tượng hình thức. một tập $A_L$ của các thuộc tính, và một quan hệ $I_L$. Thuộc tính trên $A_L$ có thể ghi nhận tính chất của các từ của các biểu hiện của các từ, các thuộc tính hàm ý, hoặc các thuộc tính hình thức ,như "has four letters". Trong nhiều ứng dụng, các thuộc tính của một từ trong $K_L$ là tính chất của ý nghĩa trong một ý cơ sở $K_D$ ; điều này có nghĩa là $I_L$ được định nghĩa bởi $wI_L m :\Leftrightarrow (dnt(w) \le \mu m \text{ in } B(K_D))$ và do đó $A_L$=$A_D$. Trong các ứng dụng khác, đạec biệt trong phân tích thành phần, chúng sử dụng nhiều dấu hiệu và các thuộc tính hàm ý, ngữ cảnh từ vựng không có $A_D$ như là tập hợp của các thuộc tính. Một biểu diễn tương tự của $B(K_D)$ là $B(K_D^*)$ với $K_D^* := (D \cup W, A_D, I_D^*)$, với những từ được ở đây là những từ của tập thứ hai của đối tượng kết nối với tập biểu thị. Quan hệ $I_D^*$ được định nghĩa $I_D \cup I_L$. Khi $K_L$ (với tập thuộc tính $A_D$) chứa trong $K_D^*, B(K_L)$ là đa hình của dàn $B(K_D^*)(\cong B(K_D))$

Ta xem WordNet là một ngữ cảnh từ vựng $K_{WN} := (W, S, I_{WN})$ với các tự không nhập nhằng $w \in W$ là đối tượng. Giống như ở trên, không tập thuộc tính $A_D$ nào được cho bởi WordNet có các từ phân biệt bởi giải thích của từ. Một hàm tương

tự, *synonymy (SYN)* được định nghĩa trên tập *W* của các từ không nhập nhằng thông qua $w_1 SYN w_2 :\Leftrightarrow syn(w_1) = syn(w_2)$ với synset của từ *w* được biểu diễn bởi *syn(w)*. Với các mối quan hệ thứ tự, hyponymy (HYP) được định nghĩa trên tập S của synset S và khái niệm dàn được tính toán theo. Do đó, một cách hình thức, mỗi synset được thể hiện thông qua các thuộc tính. Ví dụ: synset {dog} được thể hiện qua thuộc tính TO-BE-A-DOG. Quan hệ $I_{WN}$ được định nghĩa bởi $w\ I_{WN}\ syn\ (w_1) :\Leftrightarrow syn(w)\ HYN\ syn\ (w_1)$. Nó chỉ ra rằng mục đích của khái niệm bao gồm tất cả các từ thuộc các synset của khái niệm đó hoặc khái niệm nhỏ hơn. Ý định của khái niệm bao gồm tất cả các từ thuộc các synset của khái niệm đó hoặc khái niệm lớn hơn. Mỗi khái niệm có thể là một đối tượng khái niệm cho tối thiểu một synset. Một câu hỏi mở cho WordNet là liệu tập thuộc tính hình thức S có thể thay thế tập thuộc tính biểu thị $A_D$ hoặc có thể thay thế thứ tự hypornym của ssynset. Các thuộc tính thông thường không thể dùng như $A_D$ bởi vì chúng không kế thừa bởi các khái niệm con, ví dụ: không phải tất cả cac loại chim đều bay được. Ngược lại, cơ bản ngữ cảnh $K_D$ không được xem như là cơ sở để tạo ra $K_{WN}$. Tuy nhiên, trường hợp này có thể chấp nhận nếu giả sử đây là tiên đề . $K_{WN}$ lấy thông tin từ các trường hợp có thể của $K_D$.

### 2.2.3.3. Meronymy

Mối quan hệ ngữ nghĩa, như meronymy, synonymy và hyponymy, là các thuật ngữ về quan hệ của WordNet dùng để mô tả mối quan hệ cuả synset. Chúng thể hiện các mối quan hệ về từ vựng, như antonymy (đây là mối quan hệ giữa các từ, không phải là quan hệ trên synset). Trong WordNet, quan hệ $s \subseteq W \times W$ trên các từ không nhập nhằng được gọi là quan hệ ngữ nghĩa (semantic relations ) nếu

$$dnt(w_1) = dnt(w_2) \Rightarrow \forall_{w \in W}(w_1 sw \Leftrightarrow w_2 sw)\ and\ (w\ sw_1 \Leftrightarrow w\ sw_2)$$

được thỏa mãn. Quan hệ $s \subseteq W \times W$ không thỏa mãn điều kiện trên sẽ được gọi là quan hệ từ vựng (lexical relations). Quan hệ merynymy lad quan hệ ngữ nghĩa. Quan hệ antonymy là quan hệ từ vựng, nhưng quan hệ indirect antonymy là một quan hệ ngữ nghĩa, nó biểu thị khái niệm “antonymous”.

Mặc dù, quan hệ meronymy có tính kế thừa, nhưng nó không thể mô hình hóa bởi mô hình toán học dàn. Một lý do hiển nhiên cho việc không thể mô hình hóa quan hệ meronymy như một dàn khái niệm là việc sử dụng cái được biểu thị của đối tượng hình thức và thuộc tính hình thức và meronymy như là các quan hệ ngữ nghĩa giữa chúng, ví dụ: thuộc tính hình thức "ketchup" và "pizza" cùng chia sẻ đối tượng hình thức "sugar" và "salt" như là một phần. Do đó, một khái niệm hình thức "salt, sugar" sẽ mở, nhưng "salt, sugar" thường chỉ là một pha trộn và không là một khái niệm được biểu thị bằng từ trong tiếng Anh. Do đó, một dàn khái niệm có thể được cung cấp như gắn vào quan hệ meronymy, nhưng không phải tất cả có cách dùng tương tự.Giải pháp tốt hơn cho vấn đề này là sử dụng mối quan hệ bộ phần- toàn thể (part-whole) như một thuộc tính, ví dụ: HAS_HANDLE_AS_PART, ví dụ cho quan hệ này là sự khác nhau giữa "cup" và "glass". Sự lựa chọn thứ ba là hiểu quan hệ meronymy như là một mối quan hệ cộng thêm bên cạnh mối quan hệ dựa vào thứ tự. Chúng ta xét định nghĩa sau:

Trong một cấu trúc bao hàm quan hệ $\varphi_D$ ngữ nghĩa meronymy được định nghĩa như sau: hai từ không nhập nhằng trong mối quan hệ ngữ nghĩa meronymy nếu từ biểu thị khái niệm ở trong quan hệ $R^m_{(Q^4:Q^2)}$ với m là một quan hệ meronymy trong biểu thị đó là:

$$w_1 \; MER^m_{(Q^4:Q^2)} w_2 :\Leftrightarrow dnt(w_1) R^m_{(Q^4:Q^2)} dnt(w_2)$$

và mối quan hệ ngữ nghĩa meronymy là mối quan hệ không phản chiếu, phản đối xứng, và không tuần hoàn.

Ngược lại của quan hệ antonymy không trực tiếp, những loại của chúng phân biệt bởi các thành phần quan hệ, nhiều loại của quan hệ meronymy khác với số lượng các tag của chúng, các tag này được sử dụng để phân cách lớp thô. Ví dụ: có 4 loại quan hệ meronymy bao gồm các sự kết hợp của 3 loại số lượng của chúng $\| \geq 1 \|, \| \geq 0 \|$ và $\| \text{all} \|$:

1. $MER_0^m$: không bắt buộc-không bắt buộc; ví dụ: một em bé có thể là thành viên của một câu lạc bộ tennis, nhưng không phải tất cả em bé là thành viên của câu lạc bộ tennis, cũng không phải tất cả các câu lạc bộ tennis có trẻ em là thành viên.

2. $MER_{(\geq 0;\geq 1)}^m$: chính tắc-không bắt buộc; ví dụ: tất cả các tay cầm cửa là một phần của cửa, nhưng không phải tất cả các cửa đều có tay cầm.

3. $MER_{(\geq 1;\geq 0)}^m$: không bắt buộc- chính tắc; ví dụ: cục đá ở tủ ướp lạnh bao gồm nước, nhưng tất cả nước đều chưa chắc ở dạng đông đá.

4. $MER_{(\geq 1;\geq 1)}^m$: chính tắc - chính tắc; ví dụ: mỗi lông chim là một phần của con chim, mỗi con chim đều có lông .

### 2.2.3.4. Hình thức hóa mối quan hệ Hyponymy và Synonymy

Nếu quan hệ *r* là quan hệ bình đẳng, Quan hệ khái niệm thích hợp $R^{=}$ trùng với thứ tự khái niệm của dàn. Trong các ứng dụng ngôn ngữ học những quan hệ này được gọi tên riêng

Định nghĩa: Trong cấu trúc biểu hiện $\varphi_D$ mối quan hệ ngữ nghĩa được định nghĩa như sau:

Một từ không nhập nhằng là một hyponym của từ khác nếu khái niệm nó biểu thị là một khái niệm nhỏ của khái niệm mà từ khác biểu thị :

$$w_1\ HYN\ w_2 :\Leftrightarrow dnt(w_1) \leq dnt(w_2)(\Leftrightarrow dnt(w_1) R^{=}_{(\geq 0;\geq 1)} dnt(w_2))$$

Quan hệ ngược lại của hyponymy gọi là hypernymy.

Hai từ không nhập nhằng là không tách rời (disjoint) với nhau nếu chúng có chung một đối tượng chung trong mục đích của chúng:

$$w_1 \neg DIST\ w_2 :\Leftrightarrow dnt(w_1) R_0^{=} dnt(w_2)$$

Hai từ không nhập nhằng là đồng nghĩa (synonyms) với nhau nếu chúng biểu thị cùng một khái niệm:

$$w_1\ SYN\ w_2 :\Leftrightarrow dnt(w_1) = dnt(w_2)(\Leftrightarrow dnt(w_1)R^{=}_{(\geq 1;\geq 1)}dnt(w_2))$$

Nếu một ngữ cảnh biểu thị ẩn dụ được sử dụng trong WordNet, một từ là hyponym của từ khác nếu cái được biểu thị của nó là một tập con của cái được biểu thị của từ khác . Synonymy được định nghĩa giữa hai từ nếu khái niệm của chúng có cùng phạm vi (do đó cũng là có cùng mục đích) trong $K_D$. Điều này cho thấy những định nghĩa về hypornym hay synonym phụ thuộc vào ngữ cảnh $K_D$. Chúng ta có thể áp dụng những định nghĩa cho nhiều loại ngữ cảnh, nhưng nếu ngữ cảnh ẩn dụ không thuộc $K_D$ , những định nghĩa này có thể có một nghĩa hoàn toàn khác.

Tập trung vào mối kết hợp giữa hyponymy và meronymy, chúng ta thấy đây là mối quan hệ có tính kế thừa: $MER^m_{(\geq 0;\geq 1)}$ được kế thừa bởi quan hệ hypernym của tất cả và hyponyms của một phần. $MER^m_{(\geq 1;\geq 0)}$ được kể thừa bởi quan hệ hypornym của tất cả và hypernym của một phần. $MER^m_0$ được kế thừa bởi hypernym của một phấn và toàn thể. Có thể cả 2 phần của một quan hệ meronymy và một quan hệ hyponymy giữa hai khái niệm. Ví dụ, *"ice" là một loại "water" và nó cũng thành phần của "water", "musical string" là "musical supplies" và là một phần của "musical supplies"* . Tuy nhiên, thông thường có thể thêm các điều kiện vào các khái niệm với mối quan hệ đơn. Ví dụ: "water molecules" và "musical instruments" có thể được thêm vào bởi vì "ice" bao gồm "water molecules" và nó là một loại của "water" và "musical string" là một phần của "musical instruments", nó là một loại của "musical supplies".

## 3. Mô hình

Trên cơ sở khảo sát các phương pháp dịch tự động WordNet qua tiếng Việt, chúng tôi đã đưa ra một mô hình khả thi để có thể dịch tự động WordNet qua tiếng Việt.

Chúng tôi đề xuất sử dụng cả 2 giai đoạn để giải quyết vấn đề trên.

## 3.1. Dịch từ WordNet

### 3.1.1. Đặt vấn đề

Gọi

$S$ : là synset cần dịch

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

$V_i^{jk}$ : là từ thứ j trong dòng nghĩa thứ khóa của từ $E_i$ trong từ điển Anh-Việt

Trong đó

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

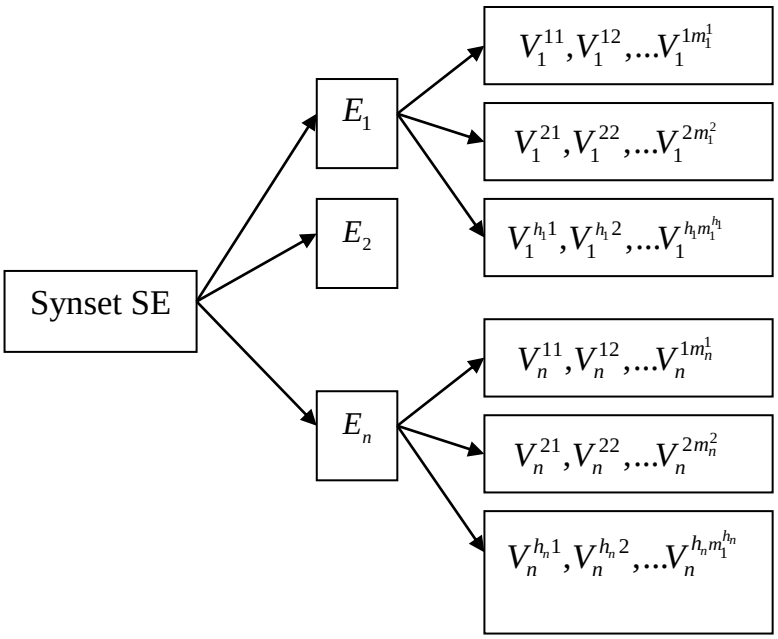
$0 \leq j \leq h_i$ :với $h_i$ là số lượng dòng nghĩa của từ $E_i$ trong từ điển Anh – Việt.

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

Chúng tôi định nghĩa thêm

$V_i$ : tập hợp các nghĩa tiếng Việt cuả từ $E_i$

$V_i^j$ : tập hợp các nghĩa tiếng Việt cuả từ $E_i^j$




*Hình 3-1: Mô hình diễn giải các ký 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*.

### 3.1.2. Hướng giải quyết

Chúng ta có cách trường hợp sau

#### 3.1.2.1. Trường hợp 1

Trường hợp này, synset *SE* chỉ có một từ tiếng Anh và từ tiếng Anh này chỉ có một dòng nghĩa 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 $V_1$ .

**Ví dụ**

Trong WordNet, chúng ta cần dịch synset

**poisoning**

2. poisoning -- (the act of giving poison to a person or animal with the intent to kill)

- Trong từ điển Anh-Việt

  @poisoning: (noun) sự đầu độc

- Do đó, chúng ta có thể sử dụng (cụm) từ *sự đầu độc* để biểu thị synset

  poisoning -- (the act of giving poison to a person or animal with the intent to kill)

**Chứng minh:**

Như đã đề cập đến ở phần cơ sở lý thuyết, mỗi synset sẽ mang một ý niệm nào đó.

(1) Trong trường hợp này ý niệm *SE* được biểu diễn bởi duy nhất hình vị *E* (tiếng Anh) (giả thiết $n=1$).

(2) Trong khi đó, hình vị *E* (tiếng Anh) lại tương ứng về nghĩa với duy nhất (giả thiết $n_i^j=1$) một hình vị $V_1^{11}$ (tiếng Việt) (giả thiết là từ *E* sẽ được dịch thành từ $V_1^{11}$).

(1),(2) cho chúng ta kết luận ý niệm *SE* sẽ được biểu thị bởi từ $V_1^{11}$.

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê sau

### 3.1.2.2. Trường hợp 2

Trường hợp này, synset *SE* chỉ có một từ tiếng Anh và từ tiếng Anh này có một nhiều dòng nghĩa tiếng Việt ta gọi là $V_i$. Do đó, việc chon lựa hình vị tiếng Việt cho synset *SE* sẽ được chọn lựa từ các dòng nghĩa tiếng Việt trên.

**Đặc tả**

Nếu $n=1$ và $n_i^j \geq 1$ thì synset *SE* sẽ có từ biểu thị sẽ được chọn từ tập ứng viên $V_i$ và căn cứ vào độ tương đồng giữa câu giải thích synset tiếng Anh và câu giải thích từ tiếng Việt [2].

Độ tương đồng của synset và từ tiếng Việt ứng viên được chúng tôi tính toán dựa trên công thức sau:

---

[2] Chúng tôi cũng đã thử nghiệm sử dụng các thông tin về hypernym và hyponym của synset để tính toán ra cách chọn nghĩa $V_i$ thích hợp, tuy nhiên, mức độ chính xác kém xa phương án mà chúng tôi trình bày ở đây.

$$SemanticValue(S,V) = \frac{|V_S \cap V_V|}{|V_S \cup V_V|}$$

với

$V_S$ là tập các từ tiếng Việt được suy ra từ câu giải thích của synset

$V_V$ là tập các từ tiếng Việt được tách từ từ câu giải thích của từ tiếng Việt trong từ điển Việt-Việt.

**Ví dụ**

Trong WordNet, chúng ta cần dịch synset

**chimney**

1. (9) chimney -- (a vertical flue that provides a path through which smoke from a fire is carried away through the wall or roof of a building)

Trong từ điển Anh-Việt

@chimney:

- (noun) ống khói
- (noun) lò sưởi,
- (noun) thông phong đèn, bóng đèn,
- (noun) miệng,
- (noun) khe núi, hẻm,

Trong từ điển Việt-Việt

@ống khói:

- (noun) *ống để dẫn cho khói, bụi thoát lên cao.*

@lò sưởi:

- (noun) *khí cụ đốt nóng để sưởi ấm.*

@thông phong:

- (noun) *bóng đèn dầu hoả.*

@miệng:

- (noun) *bộ phận trên mặt người hay ở phần trước của đầu động vật, dùng để ăn, và để nói; thường được coi là biểu tượng của việc ăn uống hay nói năng của con người.*
- (noun) *miệng ăn .*
- (noun) *lời nói trực tiếp, không phải viết.*
- (noun) *phần trên cùng, chỗ mở ra thông với bên ngoài của vật có chiều sâu.*

@hẻm:

- (noun) *lối đi hẹp hai bên có vách núi cao.*
- (noun) *ngõ hẻm .*

- (adjective) *hẹp, khó đi, hai bên thường có tường vách.*

Các từ tiếng Việt được rút trích từ câu giải thích nghĩa của synset

- vertical: *thẳng đứng,đứng,ở điểm cao nhất,ở cực điểm,đỉnh đầu,ở đỉnh đầu,thiên đỉnh,ở thiên đỉnh,đường thẳng đứng,mặt phẳng thẳng đứng*
- flue: *lưới đánh cá ba lớp mắt,nùi bông,nạm bông,ống khói,ống hơi,đầu càng mỏ neo,đầu đinh ba,thuỳ đuôi cá voi,đuôi cá voi,bệnh cúm,loe,mở rộng*
- provides : *chuẩn bị đầy đủ,dự phòng,cung cấp,chu cấp,lo cho cái ăn cái mặc cho,lo liệu cho,cung cấp,kiếm cho,quy định,chỉ định,bổ nhiệm,*
- path: *đường mòn,đường nhỏ,con đường,đường đi,đường lối.*
- through: *qua,xuyên qua,suốt,do,vì,nhờ,bởi,tại,qua,xuyên qua,suốt,từ đầu đến cuối,đến cùng,hết,hoàn toàn,đã nói chuyện được,đã nói xong,suốt,thẳng,*
- smoke: *khói,hơi thuốc,điếu thuốc lá,điếu xì gà,bốc khói,lên khói,toả khói,bốc hơi,hút thuốc,làm ám khói,làm đen,làm có mùi khói,hun,hút thuốc,nhận thấy,cảm thấy,ngờ ngợ,khám phá,phát hiện,chế giễu,*
- fire: *lửa,ánh lửa,sự cháy,hoả hoạn,sự cháy nhà,ánh sáng,sự bắn hoả lực,lò sưởi,sự tra tấn bằng lửa,sự sốt,cơn sốt,ngọn lửa,sự hăng hái,nhiệt tình,sự sốt sắng,sự vui vẻ hoạt bát,sự xúc động mạnh mẽ,nguồn cảm hứng,óc tưởng tượng linh hoạt,đốt cháy,đốt,làm nổ,bắn,làm đỏ,nung,sấy,đốt,khử trùng,thải,đuổi,sa thải,kích thích,khuyến khích,làm phấn khởi,bắt lửa,cháy,bốc cháy,nóng lên,rực đỏ,nổ,nổ súng,bắn,chạy,*
- carried: *tư thế cầm gươm chào,sự bồng vũ khí,tầm súng,tầm bắn xa,sự khiêng thuyền xuống,nơi kéo thuyền lên khỏi mặt nước,mang,vác,khuân,chở,ẵm,đem theo,đeo,mang theo,tích trữ,nhớ được,mang lại,kèm theo,chứa đựng,dẫn,đưa,truyền,chống,chống đỡ,có tầm,đạt tới,tầm xa,tới,đi xa,vọng xa,đăng,mang,sang,nhớ,làm dài ra,kéo cao lên,tiếp nối,thắng,lấy được,chiếm được,đoạt được,thuyết phục được,vượt qua,được thông qua,được chấp nhận,giành được thắng lợi cho ta,có dáng dấp,đi theo kiểu,giữ theo kiểu,có thái độ,xử sự,cư xử,ăn ở,*
- away: *xa,xa cách,rời xa,xa ra,đi,biến đi,mất đi,hết đi,không ngừng liên tục,không chậm trễ,ngay lập tức,*
- wall: *tường,vách,thành,thành luỹ,thành quách,lối đi sát tường nhà trên hè đường,rặng cây ăn quả dựa vào tường,bức tường có cây ăn quả dựa vào,vách ngoài vỉa,thành,xây tường bao quanh,xây thành bao quanh,*
- roof: *rễ,cây con cả rễ,các cây có củ,chăn,gốc,căn nguyên,gốc rễ,nguồn gốc,căn bản,thực chất,căn,nghiệm,gốc từ,nốt cơ bản,con cháu,làm bén*

*rễ,làm bắt rễ,làm ăn sâu vào,làm cắm chặt vào,nhổ bật rễ,trừ tận gốc,làm tiệt nọc,bén rễ,ăn sâu vào,*

- building: *kiến trúc,sự xây dựng,công trình kiến trúc,công trình xây dựng,toà nhà,*

Chúng tôi tính toán mức độ tương đồng của 2 câu giải thích Anh Việt và rút ra kết quả như sau

SematicValue(ống khói,chimney_1) = 0.019048
SematicValue(lò sưởi,chimney_1) = 0.009709
SematicValue(miệng,chimney_1) = 0.000000

Do đó, chúng ta có thể sử dụng (cụm) từ *ống khói* để biểu thị synset

Chimney_1 -- (a vertical flue that provides a path through which smoke from a fire is carried away through the wall or roof of a building)

**Chứng minh:**

(1) Trong trường hợp này ý niệm *SE* được biểu diễn bởi một hình vị $E_1$ (tiếng Anh).

(2) Trong khi đó, hình vị $E_1$ (tiếng Anh) lại tương ứng về nghĩa với các tập hình vị $\{V_1^1, V_1^2 ...\}, \{V_2^1, V_2^2 ...\}$, (tiếng Việt) (Chúng tôi gọi tắt là các tập $V_1, V_2, ...$)Nguyên nhân của trường hợp này là do hình vị $E_1$ được dùng để biểu thị nhiều ý niệm khác nhau tương ứng với các tập hình vị $V_1, V_2 ...$ . Dĩ nhiên nghĩa của synset *SE* tương ứng với nghĩa của một tập $V_i$ nào đó

(1),(2) cho chúng ta kết luận ý niệm *SE* được biểu diễn bởi một số hình vị trong tập $V_i$ .

Do đó, vấn đề của trường hợp này là làm thế nào để nâng cao chất lượng của công đoạn chọn hình vị biểu diễn thích hợp $V_i^j$ trong tập $V_i$ . Và chúng tôi đã khảo sát và chọn ra cách sử dụng vector tương đồng ngữ nghĩa của các câu giải thích synset và từ tiếng Việt

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê sau

### 3.1.2.3. 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, chúng tôi 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 $\bigcap_{i=1}^{n}\{\bigcup_{j=1}^{n_i}\{V_i^{j1}, V_i^{j2}...\}\} \neq \varnothing$ thì synset *SE* được biểu thị bởi tập:

$$\bigcap_{i=1}^{n}\{\bigcup_{j=1}^{n_i}\{V_i^{j1}, V_i^{j2}...\}\}$$

**Ví dụ:**

Trong WordNet, chúng ta cần dịch synset

**{organism, being}**

2. (7) organism, being -- (a living thing that has (or can develop) the ability to act or function independently)

Trong từ điển Anh-Việt

@organism

- (noun) cơ thể,**sinh vật,**
- (noun) cơ quan,tổ chức,

@being

- (noun**) sinh vật**,con người,
- (noun) sự tồn tại,sự sống,
- (noun) bản chất,thể chất,
- (adjective) hiện tại,hiện nay,này,

- Do đó, chúng ta có thể sử dụng (cụm) từ *sinh vật* để biểu thị synset

organism, being -- (a living thing that has (or can develop) the ability to act or function independently)

**Chứng minh**

(1) Trong trường hợp này ý niệm *SE* được biểu diễn bởi một số hình vị $E_1, E_2, E_3...$ (tiếng Anh).Tất nhiên là các hình vị $E_1, E_2, E_3...$ sẽ biểu thị ý niệm gần nhau (vì cùng thuộc một synset).

(2) Trong khi đó, mỗi hình vị $E_i$ (tiếng Anh) lại tương ứng về nghĩa với các hình vị thuộc tập $V_i$ , (tiếng Việt).

(3) Cần chú ý, trong tập hình vị $V_i$ có thể biểu thị nhiều ý niệm khác nhau (như đã giải thích ở trường hợp 2). Tuy nhiên, trong mỗi tập $V_i$ sẽ phải có từ biểu thị ý niệm *SE*

(1),(2),(3) cho chúng ta kết luận ý niệm *SE* được biểu bởi các hình vị trong tập $\bigcap_{i=1}^{n}\{\bigcup_{j=1}^{n_i}\{V_i^{j1},V_i^{j2}...\}\}$ (do giả thiết phần giao này khác rỗng).

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê như sau:

### 3.1.2.4. 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 chúng tôi căn cứ vào một số đặc điểm để chọn hình vị biểu thị cho synset *SE*.

**Đặc tả**

Nếu $n>1$ và $\bigcap_{i=1}^{n}\{\bigcup_{j=1}^{n_i}\{V_i^{j1},V_i^{j2}...\}\}=\varnothing$ thì synset *SE* được biểu thị bởi tập:

**Ví dụ:**

Trong WordNet, chúng ta cần dịch synset

**tempo, pace**

2. tempo, pace -- (the rate of some repeating event)

Trong từ điển Anh-Việt

@pace

- bước chân,bước,
- bước đi,nhịp đi,tốc độ đi,tốc độ chạy,
- nước đi,cách đi,
- nước kiệu,
- nhịp độ tiến triển,tốc độ tiến triển,

@tempo

- độ nhanh,
- nhịp,nhịp độ,

Tương tự như trường hợp 2 đã nói ở trên, Chúng tôi tính toán mức độ tương đồng của 2 câu giải thích Anh Việt và rút ra kết quả như sau

SematicValue(bước,tempo_pace) = 0.027778
SematicValue(bước đi, tempo_pace) = 0.000000
SematicValue(nước kiệu, tempo_pace) = 0.025000
SematicValue(xin lỗi,tempo_pace) = 0.000000
SematicValue(nhịp độ, tempo_pace) = 0.034483

Do đó, chúng ta có thể sử dụng (cụm) từ *nhịp độ* để biểu thị synset

tempo, pace -- (the rate of some repeating event)

**Chứng minh**

Tương tự trường hợp 2

(1) Trong trường hợp này ý niệm *SE* được biểu diễn bởi một số hình vị $E_1, E_2, E_3 ...$ (tiếng Anh).Tất nhiên là các hình vị $E_1, E_2, E_3 ...$ sẽ biểu thị ý niệm gần nhau (vì cùng thuộc một synset).

(2) Trong khi đó, mỗi hình vị $E_i$ (tiếng Anh) lại tương ứng về nghĩa với các hình vị thuộc tập $V_i$ , (tiếng Việt).

(3) Cần chú ý, trong tập hình vị $V_i$ có thể biểu thị nhiều ý niệm khác nhau. Tuy nhiên, trong mỗi tập $V_i$ sẽ phải có từ biểu thị ý niệm liên quan đến *SE*

(1),(2),(3) cho chúng ta kết luận ý niệm SE được biểu bởi các hình vị trong tập $\bigcap_{i=1}^{n}\{\bigcup_{j=1}^{n_i}\{V_i^{j1}, V_i^{j2}...\}\}$. Tuy nhiên , trong trường hợp này tập $\bigcap_{i=1}^{n}\{\bigcup_{j=1}^{n_i}\{V_i^{j1}, V_i^{j2}...\}\} = \varnothing$. Nguyên nhân dẫn tới điều này là do các tập ý niệm $V_i^j$ này không giao nhau mà bao lẫn nhau (quan hệ hypernym _ hyponym).

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê như sau:

## 3.2. Dịch từ từ điển tiếng Việt

### 3.2.1. Đặt vấn đề

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.

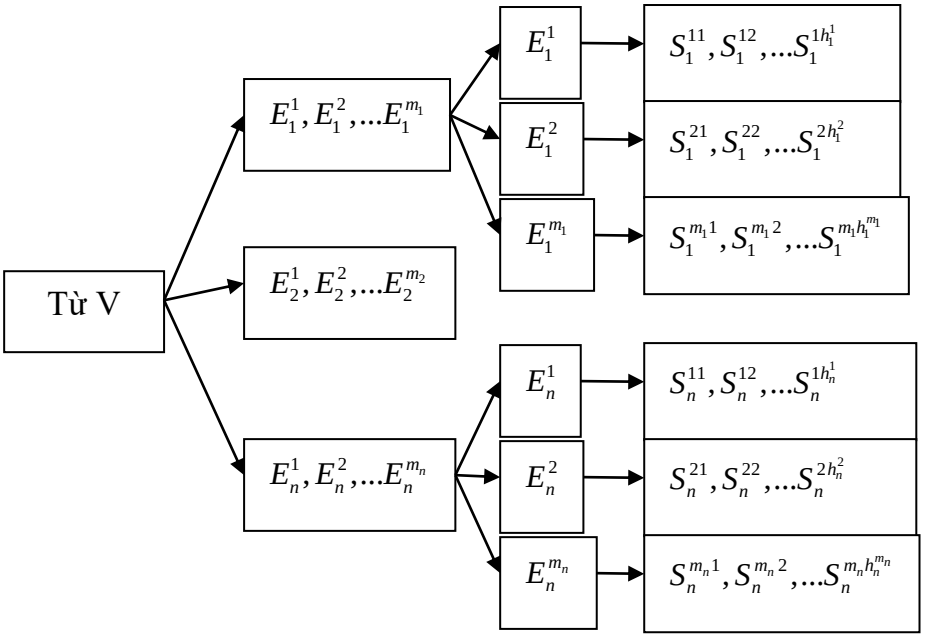
$0 \le j \le m_i$ :với $m_i$ là số lượng từ trong dòng nghĩa thứ i của từ V trong từ điển Việt Anh.

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

Chúng tôi định nghĩa thêm

$E_i$ : 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 3-2: Mô hình diễn giải các ký hiệu của mô hình gán nhãn synset cho các từ tiếng Việt*

Với mô hình này, vấn đề của chúng ta là 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 có thể 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 chọn synset cho từ V sẽ không phụ thuộc vào các dòng nghĩa khác của từ V.

Do đó, bài toán này trở thành bài toán làm thể nào để gán nhãn synset cho mỗi tập hợp $E_i$ $(0 \le i \le n)$

### 3.2.2. Hướng giải quyết

#### 3.2.2.1. Trường hợp 1

Trường hợp này, dòng nghĩa tiếng Anh chỉ có 1 từ và từ này chỉ thuộc 1 synset, chúng tôi sẽ lấy synset này làm nhãn synset cho tập $E_i$

**Đặc tả**

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

**Ví dụ**

Trong từ điển Việt Anh

**thực thể** : (noun) entity

- Trong WordNet

entity

The noun entity has 1 sense (first 1 from tagged texts)

(11) entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

1. Do đó, chúng ta có thể đoán ngay từ *thực thể* sẽ thuộc synset

entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

.

**Chứng minh:**

Như đã đề cập đến ở phần cơ sở lý thuyết, mỗi synset sẽ mang một ý niệm nào đó.

(1) Trong trường hợp này ý niệm *SV* được biểu diễn bởi hình vị *V* (tiếng Việt).

(2) Trong khi đó, hình vị *V* (tiếng Việt) lại tương ứng về nghĩa với duy nhất (giả thiết $n_i^j = 1$) một hình vị $E_i^1$ (tiếng Anh) (giả thiết là từ *V* sẽ được dịch thành từ $E_i^1$).

(3) Hơn thế nữa, hình vị $E_i^1$ chỉ biểu thị một ý niệm là $S_i^{j1}$.

(1),(2),(3) cho chúng ta kết luận ý niệm *SV* sẽ được gán nhãn là $S_i^{j1}$.

**Kết quả**

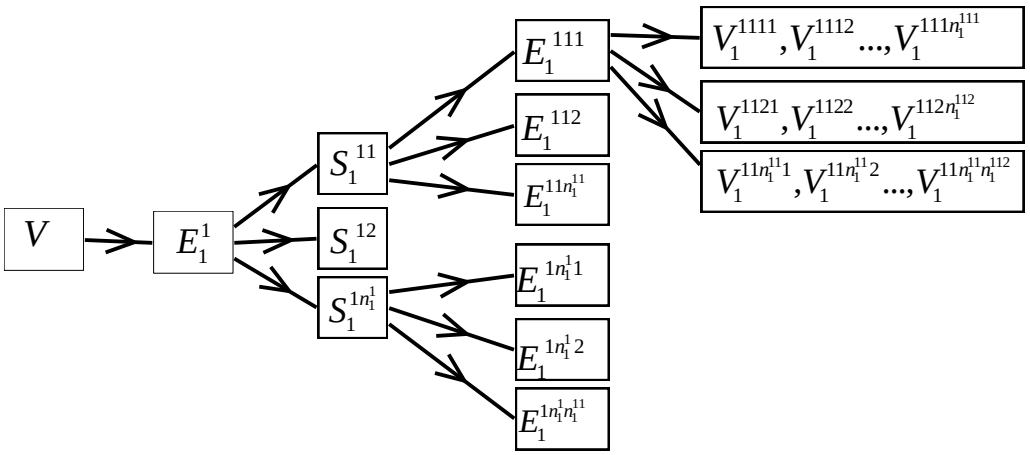
Qua khảo sát, chúng tôi đã thu được bảng thống kê sau

### 3.2.2.2. Trường hợp 2

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

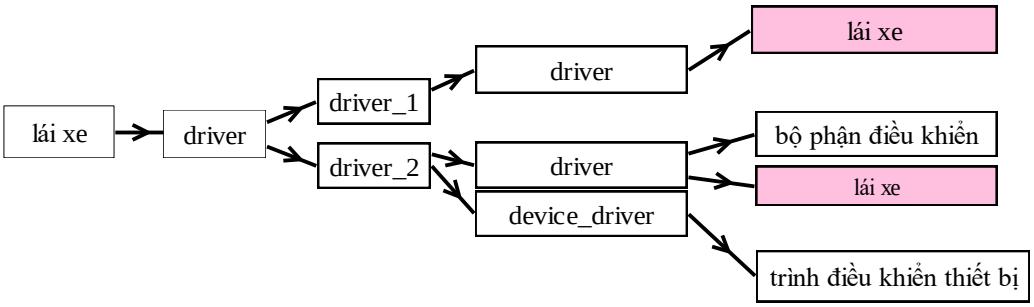
**Đặc tả**

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




*Hình 3-3: Mô hình diễn giải của trường hợp 2*

**Ví dụ**



Cụ thể, chúng tôi đã khảo sát 2 độ đo hình vị là hệ số Dice[3] và hệ số Jaccard[4] được đề xuất bởi P. Bhattacharyya and Narayan Unny (2002) và chúng tôi tự đề xuất thêm một mô hình chuỗi chung dài nhất[5] để tìm độ tương đồng của hai hình vị khi tính toán độ tương ứng hình vị trong trường hợp này. Qua khảo sát chúng tôi chọn độ đo (hệ số) *Dice* để chọn synset thích hợp.

**Chứng minh**

(1) Trong trường hợp này ý niệm SV được biểu diễn bởi hình vị V (tiếng Việt).

(2) Trong khi đó, hình vị V (tiếng Việt) lại tương ứng về nghĩa với duy nhất (giả thiết $n_i^j = 1$) một hình vị $E_i^1$ (tiếng Anh) (giả thiết là từ V sẽ được dịch thành từ $E_i^1$).

(3) Tuy nhiên, hình vị $E_i^1$ lại biểu thị nhiều ý niệm $S_i^{j1}, S_i^{j2}$...

---

[3] Hệ số Dice được định nghĩa như sau

$$Dice(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

với

$|X \cap Y|$: là số lượng từ chung giữa hai dãy văn bản X và Y

$|X|$: là số lượng từ của văn bản X

$|Y|$: là số lượng từ của văn bản Y

[4] Hệ số Jaccard

$$\text{Jaccard(X,Y)} = \frac{|X \cap Y|}{|X \cup Y|}$$

với

$|X \cap Y|$: là số lượng từ chung giữa hai dãy văn bản X và Y

$|X \cup Y|$: là số lượng từ có ở hai dãy văn bản X và Y

[5] Chúng tôi sử dụng tính chất chuỗi chung dài nhất để tính độ tương đồng giữa hai văn bản dựa trên thuật toán qui hoạch động

(1),(2),(3) cho chúng ta kết luận ý niệm $SV \in \{S_i^{j1}, S_i^{j2}...\}$

Do đó, vấn đề của trường hợp 2 là làm thế nào để nâng cao chất lượng của công đoạn chọn ý nhiệm nào trong tập ý niệm $S_i^{j1}, S_i^{j2}...$

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê như sau:

### 3.2.2.3. Trường hợp 3

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

**Đặc tả**

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

**Ví dụ:**

Trong từ điển Việt Anh

**ý kiến** : view, opinion

Trong WordNet

**opinion**

The noun opinion has 6 senses (first 5 from tagged texts)

1. (408) opinion, sentiment, persuasion, view, thought -- (a personal belief or judgment that is not founded on proof or certainty; "my opinion differs from yours"; "what are your thoughts on Haiti?")
2. (76) public opinion, popular opinion, opinion, vox populi -- (a belief or sentiment shared by most people; the voice of the people; "he asked for a poll of public opinion")
3. (51) opinion, view -- (a message expressing a belief about something; the expression of a belief that is held with confidence but not substantiated by positive knowledge or proof; "his opinions appeared frequently on the editorial page")
4. (34) opinion, legal opinion, judgment, judgement -- (the legal document stating the reasons for a judicial decision; "opinions are usually written by a single judge")
5. (10) opinion, ruling -- (the reason for a court's judgment (as opposed to the decision itself))

6. impression, feeling, belief, notion, opinion -- (a vague idea in which some confidence is placed; "his impression of her was favorable"; "what are your feelings about the crisis?"; "it strengthened my belief in his sincerity"; "I had a feeling that she was lying")

**view**

The noun view has 10 senses (first 8 from tagged texts)

1. (36) position, view, perspective -- (a way of regarding situations or topics etc.; "consider what follows from the positivist view")
2. (12) view, aspect, prospect, scene, vista, panorama -- (the visual percept of a region; "the most desirable feature of the park are the beautiful views")
3. (9) view, survey, sight -- (the act of looking or seeing or observing; "he tried to get a better view of it"; "his survey of the battlefield was limited")
4. (7) view, eyeshot -- (the range of the eye; "they were soon out of view")
5. (6) opinion, sentiment, persuasion, view, thought -- (a personal belief or judgment that is not founded on proof or certainty; "my opinion differs from yours"; "what are your thoughts on Haiti?")
6. (4) opinion, view -- (a message expressing a belief about something; the expression of a belief that is held with confidence but not substantiated by positive knowledge or proof; "his opinions appeared frequently on the editorial page")
7. (4) view -- (purpose; the phrase `with a view to' means `with the intention of' or `for the purpose of'; "he took the computer with a view to pawning it")
8. (3) scene, view -- (graphic art consisting of the graphic or photographic representation of a visual percept; "he painted scenes from everyday life"; "figure 2 shows photographic and schematic views of the equipment")
9. horizon, view, purview -- (the range of interest or activity that can be anticipated; "It is beyond the horizon of present knowledge")
10. view -- (outward appearance; "they look the same in outward view")

Do đó, chúng ta có thể đoán ngay từ *ý kiến* sẽ thuộc synset:

opinion, sentiment, persuasion, view, thought -- (a personal belief or judgment that is not founded on proof or certainty; "my opinion differs from yours"; "what are your thoughts on Haiti?")

opinion, view -- (a message expressing a belief about something; the expression of a belief that is held with confidence but not substantiated by positive knowledge or proof; "his opinions appeared frequently on the editorial page")

**Chứng minh**

(1) Trong trường hợp này ý niệm SV được biểu diễn bởi hình vị V (tiếng Việt).

(2) Trong khi đó, hình vị V (tiếng Việt) lại tương ứng về nghĩa với nhiều hình vị (giả thiết $n_i^j > 1$) $E_i^1, E_i^2, E_i^3 \ldots$, (tiếng Anh) (giả thiết là từ V sẽ được dịch thành các từ $E_i^1, E_i^2, E_i^3 \ldots$). Tất nhiên là các hình vị $E_i^1, E_i^2, E_i^3 \ldots$ sẽ biểu thị ý niệm gần nhau.

(3) Tuy nhiên, mỗi hình vị $E_i^j$ lại biểu thị nhiều ý niệm $S_i^{j1}, S_i^{j2} \ldots$

(1),(2),(3) cho chúng ta kết luận ý niệm $SV \subseteq \bigcap_{j=1}^{n_i} \{S_i^{j1}, S_i^{j2} \ldots\}$ (do giả thiết phần giao này khác rỗng).

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê như sau:

### 3.2.2.4. 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, chúng tôi sẽ căn cứ vào cấu trúc phân cấp của WordNet để chọn ra nhãn ngữ nghĩa thích hợp cho tập $E_i$

**Đặc tả**

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

Ở đây chúng tôi sử dụng 3 tiêu chuẩn:

1. **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 (hypernym)). Khi đó synset $\{E_1^1, E_1^2, ...\}$ được chọn là các synset anh em này.

Tức là

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

Ký hiệu

P is_hyper S: P là cấp cha của S

**2. 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 $\{E_1^1, E_1^2, ...\}$ được chọn là các synset anh em này.

Tức là

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

Ký hiệu

*P* is_hyper *S*: *P* là cấp cha của *S*

**3. 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 $\{E_1^1, E_1^2, ...\}$ được chọn là các synset cấp dưới này.

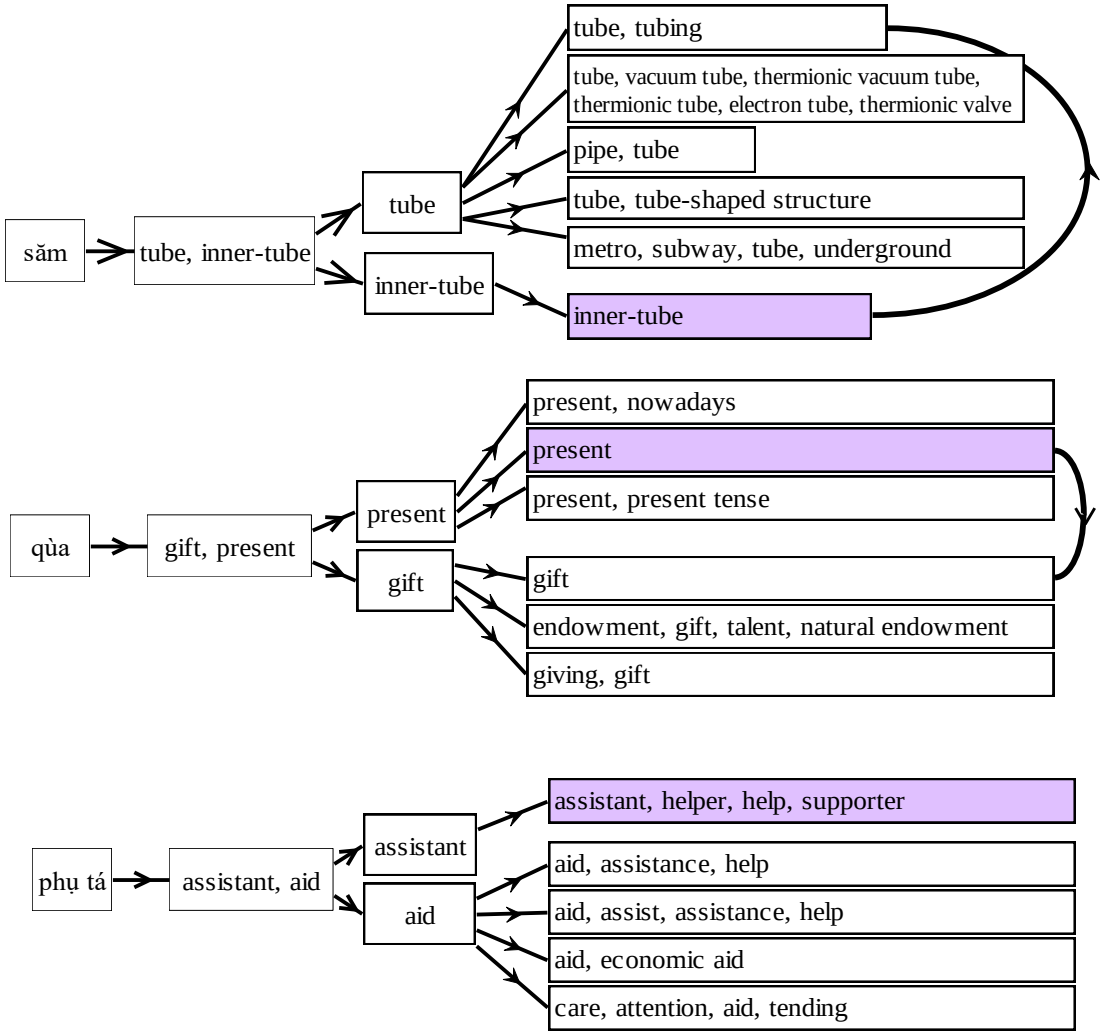
Tức là

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

Ký hiệu

*P* is_dist_hyper *S*: *P* là cấp trên của *S*

**Ví dụ**




**Chứng minh**

(1) Trong trường hợp này ý niệm SV được biểu diễn bởi hình vị V (tiếng Việt).

(2) Trong khi đó, hình vị V (tiếng Việt) lại tương ứng về nghĩa với nhiều hình vị (giả thiết $n_i^j > 1$) $E_i^1, E_i^2, E_i^3 \ldots$, (tiếng Anh) (giả thiết là từ V sẽ

được dịch thành các từ $E_i^1, E_i^2, E_i^3 ...$). Tất nhiên là các hình vị $E_i^1, E_i^2, E_i^3 ...$ sẽ biểu thị ý niệm gần nhau.

(3) Tuy nhiên, mỗi hình vị $E_i^j$ lại biểu thị nhiều ý niệm $S_i^{j1}, S_i^{j2} ...$

(1),(2),(3) cho chúng ta kết luận ý niệm $SV \subseteq \bigcap_{j=1}^{n_i} \{S_i^{j1}, S_i^{j2} ...\}$. Tuy nhiên theo giả thiết $\bigcap_{j=1}^{n_i} \{S_i^{j1}, S_i^{j2} ...\} = \varnothing$ . Nguyên nhân dẫn tới điều này là do các tập ý niệm này không giao nhau mà bao lẫn nhau (quan hệ hypernym _ hyponym). Do đó chúng tôi sử dụng 3 tiêu chuẩn để tính toán lại $\bigcap_{j=1}^{n_i} \{S_i^{j1}, S_i^{j2} ...\}$ sao cho phần giao của các ý niệm này khác rỗng bằng cách thay một số ý niệm bằng các ý niệm có nghĩa nhỏ hơn (thay các synset tổng quát bằng các synset rõ hơn).

**Kết quả**

Qua khảo sát, chúng tôi đã thu được bảng thống kê như sau:

### 3.2.2.5. Tổng hợp

### 1. Trường hợp dịch V_S

*Bảng 3-1: Tổng hợp kết quả các trường hợp V_S*

| | Số lượng Từ điển | Danh từ tiếng Anh | Danh từ tiếng Việt | Synset | Kết nối |
|---|---|---|---|---|---|
| | WordNet | 114648 | - | 79689 | |
| | Từ điển Việt-Anh | | 13706 | | |
| | Từ điển Anh-Việt | | | | |
| | Từ điển gom chung | | | | |
| | Mức độ bao phủ | | | | |
| 1 | Số lượng | 6556 | 2571 | 1990 | 2677 |
| | so với WordNet | 5.72 | 18.76 | 2.5 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| 2 | Số lượng | 5571 | 3920 | 1727 | 4286 |
| | so với WordNet | 4.86 | 28.6 | 2.17 | |

| | | | | | |
|---|---|---|---|---|---|
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| 3 | Số lượng | 3659 | 769 | 686 | 938 |
| | so với WordNet | 3.19 | 5.61 | 0.86 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| 4 | Số lượng | 2818 | 534 | 925 | 1189 |
| | so với WordNet | 2.46 | 3.9 | 1.16 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| Tổng cộng | Số lượng | 18604 | 7794 | 5328 | 9090 |
| | so với WordNet | 16.23 | 56.87 | 6.69 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |

**2. Trường hợp dịch V_S**

*Bảng 3-2: Tổng hợp kết quả các trường hợp S_V*

| | Số lượng<br>Từ điển | Danh từ tiếng Anh | Danh từ tiếng Việt | Synset | Kết nối |
|---|---|---|---|---|---|
| | WordNet | 114648 | - | 79689 | |
| | Từ điển Việt-Anh | | 13706 | | |
| | Từ điển Anh-Việt | | | | |
| | Từ điển gom chung | | | | |
| | Mức độ bao phủ | | | | |
| 1 | Số lượng | 9314 | | 5909 | |
| | so với WordNet | 8.12 | 0 | 7.42 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| 2 | Số lượng | 6529 | | 4561 | |

| | | | | | |
|---|---|---|---|---|---|
| | so với WordNet | 5.69 | 0 | 5.72 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| 3 | Số lượng | 6354 | | 2422 | |
| | so với WordNet | 5.54 | 0 | 3.04 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| 4 | Số lượng | 7303 | | 2086 | |
| | so với WordNet | 6.37 | 0 | 2.62 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| Danh từ riêng | Số lượng | 29725 | | 15203 | |
| | so với WordNet | 25.93 | 0 | 19.08 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |
| Tổng cộng | Số lượng | 59225 | 0 | 30181 | 0 |
| | so với WordNet | 51.65 | 0 | 37.88 | |
| | so với Từ điển Việt-Anh | | | | |
| | so với Từ điển Anh-Việt | | | | |
| | so với Từ điển gom chung | | | | |
| | so với Mức độ bao phủ | | | | |

### 3.3. Cách làm giàu từ điển song ngữ (các trường hợp không có trong từ điển)

Một trong những hạn chế của kết quả mà chúng tôi thu được ở trên là do sự không đầy đủ của từ điển song ngữ. Thật vậy, các synset trong WordNet không chỉ được biểu thị bởi các từ mà còn được biểu thị bởi các từ ghép (collations). Tuy nhiên, các từ ghép này lại không được giải thích trong từ điển Anh –Việt. Hay ở công đoạn gán nhãn ngữ nghĩa cho một từ ghép tiếng Việt, trong khi từ ghép này không có trong từ điển Việt_Anh. Bên cạnh đó, vấn đề hình thái của từ tiếng Anh

(số nhiều, sở hữu cách … ) và tiếng Việt (vấn đề chuyển đổi từ loại của từ sau khi thêm các từ "sự", "việc", "cuộc"… ) khiến cho kết quả của mô hình còn nhiều hạn chế. Để giải quyết vấn đề này, chúng tôi sử dụng thêm hai mô hình sau:

### 3.3.1. Với các tiếng Anh

Với tiếng Anh, chúng tôi giải quyết hai vấn đề

#### 3.3.1.1. Về vấn đề hình thái

Tiếng Anh là một ngôn ngữ biến hình, điều này sẽ gây khó khăn cho cách tra từ điển theo kiểu truyền thống. Nói rõ hơn, một từ trong tiếng Anh có thể có hình vị khác nhau nếu từ này ở trong ngữ cảnh khác nhau. Tuy nhiên, các từ điển chỉ lưu trạng thái gốc của từ. Ví dụ:Trong WordNet chỉ có từ c*hild* chứ không có từ c*hildrent.* Điều này sẽ gây trở ngại rất lớn cho vấn đề tìm kiếm. Để giải quyết vấn đề này, chúng tôi sử dụng thêm phần phân tích hình thái cho từ tiếng Anh.

Thật vậy, nếu từ (cụm từ) tiếng Anh cần tìm không tìm thấy trong từ điển, chúng tôi sử dụng các cách phân tích để tìm các hình thái khác nhau của từ này và tra lại trong từ điển các biến thể hình thái của từ này. Hiện nay, vấn đề phân tích hình thái của từ tiếng Anh được giải quyết thông qua các luật về hình vị. Ví dụ, từ số nhiều sẽ được thêm *s* hay *es* vào cuối. … Tuy nhiên, mô hình luật lại cho ra kết quả có độ chính xác chưa cao. Bên cạnh đó, mô hình luật này còn gặp phải vấn đề về sự nhập nhằng. Do đó, chúng tôi sử dụng luôn mô hình phân tích hình thái từ của WordNet. Mô hình này của WordNet không chỉ dựa trên luật mà còn sử dụng một từ điển hình thái lớn để lưu lại các biến cách của từ tiếng Anh. Do đó, độ chính xác và vấn đề nhập nhằng được giải quyết thỏa đáng.

#### 3.3.1.2. Về vấn đề từ ghép

Tiếng Anh chỉ có gần 60.000 từ đơn, tuy nhiên số lượng từ tiếng Anh trong WordNet lên đến 152.059 từ. Điều này chứng tỏ số lượng từ tiếng Anh trong WordNet là từ ghép và cụm từ chiếm tỉ lệ đáng kể. Đây cũng là một trong những trở ngại mà chúng tôi gặp phải khi tìm cách tra từ ghép này trong từ điển Anh _ Việt,

thông thường từ điển Anh Việt không lưu các từ ghép hay cụm từ (đặc biệt là cụm danh từ )…

Chúng tôi giải quyết trên cả hai phương diện: chủ động và thụ động

**Phương án chủ động**

Với phương án này, chúng tôi tìm cách nâng cao vốn từ cho từ điển Anh-Việt bằng cách sử dụng thêm các từ mẫu, hay ví dụ ngắn trong từ điển như một từ bình thường (entry) của từ điển.

*Ví dụ:*

Trong từ điển Anh Việt

**fat**

tính từ
▪ được vỗ béo (để giết thịt)
▪ béo, mập, béo phì, mũm mĩm
…
▪ màu mỡ, tốt
ví dụ ◦ fat lands đất màu mỡ
▪ béo bở, có lợi, có lãi
ví dụ ◦ fat job việc làm béo bở
▪ đầy áp
ví dụ ◦ a fat purse túi tiền đầy ắp, túi tiền dầy cộm

Chúng tôi đã rút trích các ví dụ để tạo thêm các mục từ mới trong từ điển Anh Việt

**fat**

tính từ
▪ được vỗ béo (để giết thịt)
▪ béo, mập, béo phì, mũm mĩm
…
▪ màu mỡ, tốt
▪ béo bở, có lợi, có lãi
▪ đầy áp

**fat job**

▪ việc làm béo bở

**fat lands**

▪ đất màu mỡ

**a fat purse**

▪ túi tiền đầy ắp, túi tiền dầy cộm

Vì việc thêm các mục từ này được thực hiện ở giai đoạn trước khi thực hiện chương trình nên chúng tôi đặt tên cho phương án này là chủ động.

**Phương án thụ động**

Với phương án này được sử dụng đối với các từ ghép tiếng Anh không được tìm thấy trong từ điển Anh Việt đã mở rộng theo cách chủ động ở trên. Trong trường hợp này chúng tôi chuyển các từ tiếng Anh này sang hình vị gốc (lemma) và tra lại trong từ điển, nếu vẫn không tìm thấy, chúng tôi tìm cách dịch từng hình vị trong từ ghép này để tìm nghĩa tiếng Việt thích hợp. Nếu có nhiều tổ hợp được tạo ra bởi cách dịch này, tuy nhiên, điều này sẽ không ảnh hưởng lớn đến độ chính xác của mô hình do chúng tôi chỉ sử dụng các tổ hợp này trong việc ra quyết định chọn nhãn synset của từ trong trường hợp 3.2.2.

### 3.3.2. Với các tiếng Việt

Với tiếng Việt, chúng tôi giải quyết ba vấn đề

#### 3.3.2.1. Về vấn đề hình thái

Tương tự như tiếng Anh, vấn đề hình thái cũng gây khó khăn cho việc tra từ điển để tìm các từ tiếng Việt tương ứng. Tuy nhiên, khác với tiếng Anh, Tiếng Việt là một ngôn ngữ đơn lập, điều này sẽ mang đến cho chúng ta nhiều thuận lợi khi phân tích hình thái của từ (cụm từ) tiếng Việt. Điều này là do các danh từ trong tiếng Việt đi kèm với các phó danh yừ

Thật vậy, nếu từ (cụm từ) tiếng Anh cần tìm không tìm thấy trong từ điển, chúng tôi sử dụng các cách phân tích để tìm các hình thái khác nhau của từ này và tra lại trong từ điển các biến thể hình thái của từ này. Hiện nay, vấn đề phân tích hình thái của từ tiếng Anh được giải quyết thông qua các luật về hình vị. Ví dụ, từ số nhiều sẽ được thêm *s* hay *es* vào cuối. …  Tuy nhiên, mô hình luật lại cho ra kết quả có độ chính xác chưa cao. Bên cạnh đó, mô hình luật này còn gặp phải vấn đề về sự nhập nhằng. Do đó, chúng tôi sử dụng luôn mô hình phân tích hình thái từ của WordNet. Mô hình này của WordNet không chỉ dựa trên luật mà còn sử dụng một từ điển hình thái lớn để lưu lại các biến cách của từ tiếng Anh. Do đó, độ chính xác và vấn đề nhập nhằng được giải quyết thỏa đáng.

### 3.3.2.2. Về vấn đề từ ghép

Tiếng Anh chỉ có gần 60.000 từ đơn, tuy nhiên số lượng từ tiếng Anh trong WordNet lên đến 152.059 từ. Điều này chứng tỏ số lượng từ tiếng Anh trong WordNet là từ ghép và cụm từ chiếm tỉ lệ đáng kể. Đây cũng là một trong những trở ngại mà chúng tôi gặp phải khi tìm cách tra từ ghép này trong từ điển Anh _ Việt, thông thường từ điển Anh Việt không lưu các từ ghép hay cụm từ (đặc biệt là cụm danh từ )…

Chúng tôi giải quyết trên cả hai phương diện: chủ động và thụ động

**Phương án chủ động**

Với phương án này, chúng tôi tìm cách nâng cao vốn từ cho từ điển Anh-Việt bằng cách sử dụng thêm các từ mẫu, hay ví dụ ngắn trong từ điển như một từ bình thường (entry) của từ điển.

*Ví dụ:*

Trong từ điển Anh Việt

**fat**

tính từ
- được vỗ béo (để giết thịt)
- béo, mập, béo phì, mũm mĩm

…
- màu mỡ, tốt

ví dụ ◦ fat lands đất màu mỡ
- béo bở, có lợi, có lãi

ví dụ ◦ fat job việc làm béo bở
- đầy áp

ví dụ ◦ a fat purse túi tiền đầy ắp, túi tiền dầy cộm

Chúng tôi đã rút trích các ví dụ để tạo thêm các mục từ mới trong từ điển Anh Việt

**fat**

tính từ
- được vỗ béo (để giết thịt)
- béo, mập, béo phì, mũm mĩm

…
- màu mỡ, tốt
- béo bở, có lợi, có lãi
- đầy áp

**fat job**
- việc làm béo bở

**fat lands**
- đất màu mỡ

**a fat purse**

- túi tiền đầy ắp, túi tiền dầy cộm

Vì việc thêm các mục từ này được thực hiện ở giai đoạn trước khi thực hiện chương trình nên chúng tôi đặt tên cho phương án này là chủ động.

**Phương án thụ động**

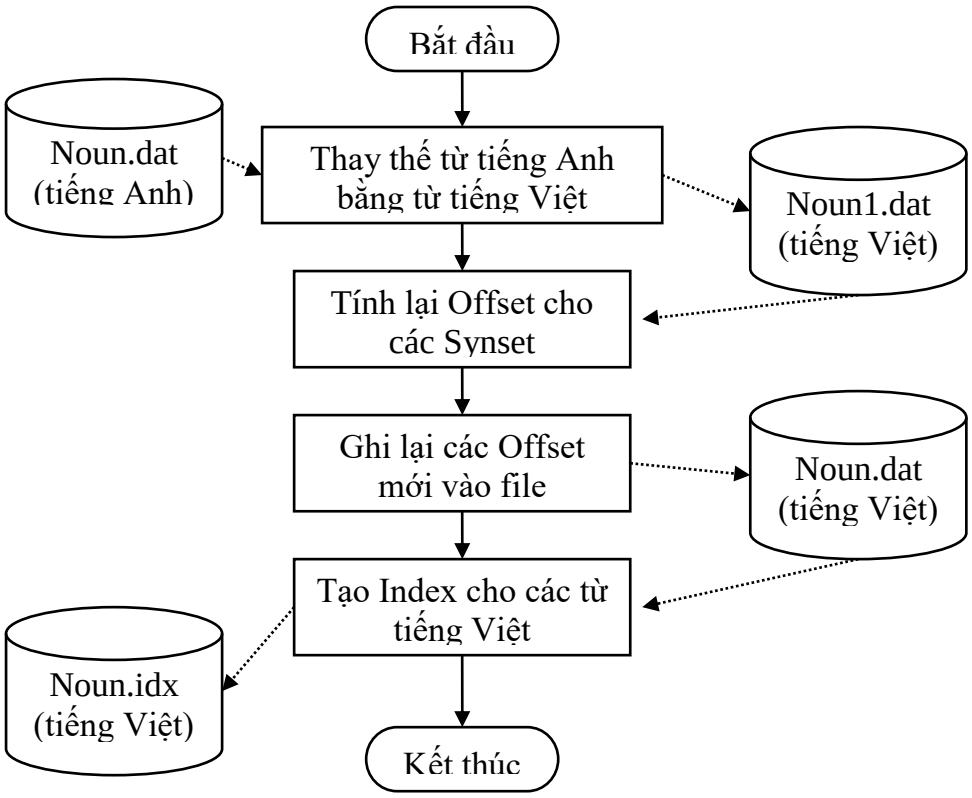
Với phương án này được sử dụng đối với các từ ghép tiếng Anh không được tìm thấy trong từ điển Anh Việt đã mở rộng theo cách chủ động ở trên. Trong trường hợp này chúng tôi chuyển các từ tiếng Anh này sang hình vị gốc (lemma) và tra lại trong từ điển, nếu vẫn không tìm thấy, chúng tôi tìm cách dịch từng hình vị trong từ ghép này để tìm nghĩa tiếng Việt thích hợp. Nếu có nhiều tổ hợp được tạo ra bởi cách dịch này, tuy nhiên, điều này sẽ không ảnh hưởng lớn đến độ chính xác của mô hình do chúng tôi chỉ sử dụng các tổ hợp này trong việc ra quyết định chọn nhãn synset của từ trong trường hợp 3.2.2.

## 3.4. 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 tôi 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, chúng tôi đã sử dụng ngay cách tổ chức WordNet của tiếng Anh để lưu cây WordNet tiếng Việt.

Chúng tôi đã sử dụng mô hình sau để tạo lại cấu trúc WordNet sau khi đã dịch xong:




*Hình 3-4: Mô hình quá trình tổ chức dữ liệu cho WordNet tiếng Việt*

***Ví dụ:***

Dữ liệu trong file noun.dat (tiếng Anh)

00003226 03 n 02 organism 0 being 0 055 @ 00003009 n 0000 … ~ 00004358 n 0000 ~ 00004483 n 0000 … | a living thing that has (or can develop) the ability to act or function independently

00004358 03 n 01 benthos 0 001 @ 00003226 n 0000 | organisms (plants and animals) that live at or near the bottom of a sea

Dữ liệu thu được sau khi thực hiện 2 công đoạn dịch WordNet

@00004358

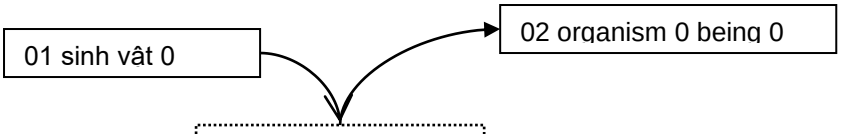
-sinh vật đáy,

@00003226

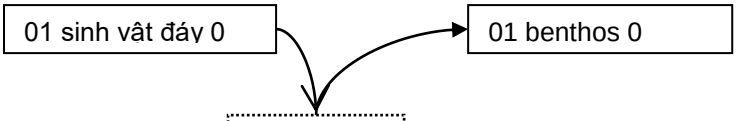
-sinh vật,

Khi đó chúng ta minh họa từng bước như sau

**Bước 1**

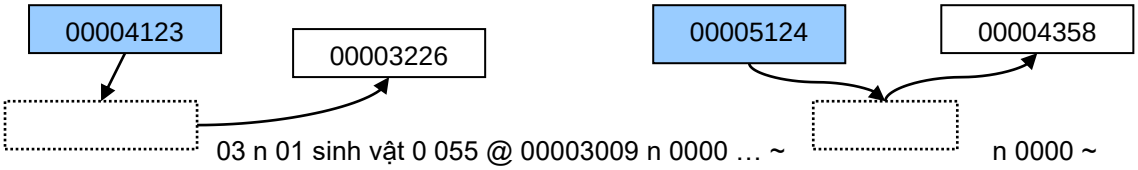



00003226 03 n [ ] 055 @ 00003009 n 0000 … ~ 00004358 n 0000 ~ 00004483 n 0000 … | a living thing that has (or can develop) the ability to act or function independently

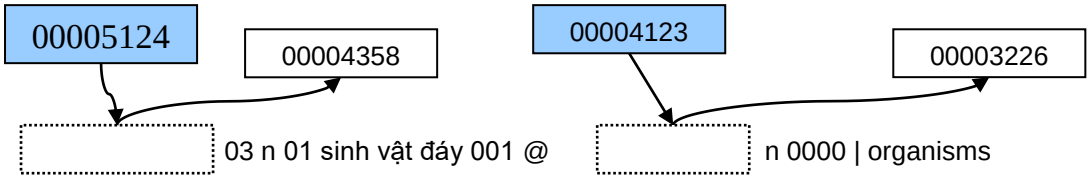



00004358 03 n [ ] 001 @ 00003226 n 0000 | organisms (plants and animals) that live at or near the bottom of a sea

**Bước 2-3**

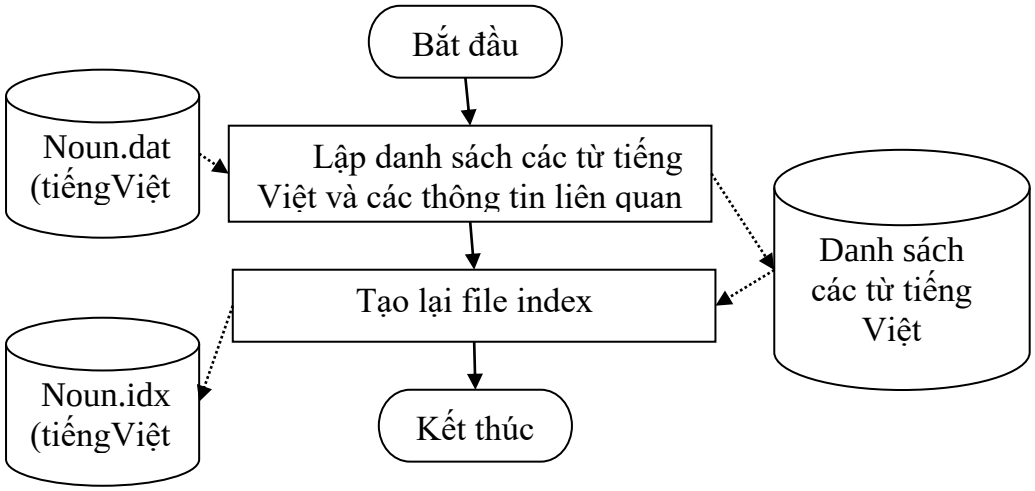



[ ] 03 n 01 sinh vật 0 055 @ 00003009 n 0000 … ~ [ ] n 0000 ~ 00004483 n 0000 … | a living thing that has (or can develop) the ability to act or function independently




[ ] 03 n 01 sinh vật đáy 001 @ [ ] n 0000 | organisms (plants and animals) that live at or near the bottom of a sea

**Bước 4**




Trong file noun.dat

> 00005124 03 n 01 sinh vật đáy 001 @ 00004123 n 0000 | organisms (plants and animals) that live at or near the bottom of a sea

Sẽ tạo ra file noun.idx như sau

sinh_vật_đáy n 2 1 @ 1 0 00005124

Khi đó chúng ta có thể tận dụng chép đè hai file noun.dat và noun.idx tiếng Việt lên 2 file tương ứng trong tiếng Anh. Khi đó, chúng ta có thể sử dụng các bộ duyệt (browser) của WordNet tiếng Anh cho tiếng Việt. Điều quan trọng hơn là chúng ta có thể sử dụng các hàm API của WordNet tiếng Anh cho tiếng Việt mà không cần bất cứ sự can thiệp nào.

## 4. Cài đặt

Dựa trên cơ sở lý thuyết và mô hình đã đề xuất, chúng tôi đã xây dựng thử nghiệm chương trình dịch và tổ chức dữ liệu tự động từ WordNet tiếng Anh sang WordNet tiếng Việt.

## 4.1. Chuẩn hóa các từ điển

Như đã trình bày ở phần trên, mô hình dịch WordNet tiếng Anh sang tiếng Việt phải sử dụng 4 từ điển: Từ điển Chính tả Tiếng Việt (gọi tắt là Vdic, từ điển này liệt kê tất cả các từ chính tả tiếng Việt), Từ Điển Tiếng Việt (gọi tắt là VVDic, từ điển này được chúng tôi lấy từ cuốn Từ Điển Tiếng Việt của Hội Ngôn Ngữ Học), Từ điển Anh Việt, từ điển Việt Anh.

### 4.1.1. Lý do

Không may mắn, các từ điển trên đều là từ điển dành cho người. Điều này gây rất nhiều khó  khăn cho độ chính xác và khả năng của chương trình.

*Ví dụ:*

Xét từ điển Anh-Việt

Từ "*fan*"(danh từ): "*cái quạt*".

Tuy nhiên, với từ điển dành cho máy tính (MRD: Machine Readable Dictionary) thì khái niệm này phải là

Từ "*fan*"(danh từ): "*quạt*".

Nguyên nhân của một loạt vấn đề này và các vấn đề tương tự là do sự khác nhau về loại hình giữa hai ngôn ngữ tiếng Anh và tiếng Việt. Với các ngôn ngữ thuộc loại hình đơn lập (như tiếng Việt), các từ thường đi kèm với các phó danh từ (như sự, việc, cuộc)…  và điều này sẽ gây khó khăn cho vấn đề so sánh, đối chiếu giữa các từ với nhau.

### 4.1.2. Giải quyết

Chúng tôi đã xây dựng một mô hình tự động tạo một từ điển MRD từ từ điển dành cho người[6]. Cuối cùng , chúng tôi dã xây dựng được 4 từ điển cần thiết với các thông số sau:

Từ điển Chính tả Tiếng Việt có 37454 từ trong đó có

[6] Mô hình này sẽ được chúng tôi công bố trong một công trình khác.

| Từ loại | Số từ |
|---|---|
| Danh từ | 18776 |
| Động từ | 11931 |
| Tính từ | 8410 |
| Phó từ | 928 |
| Tổng cộng | 37454 |

Từ điển Việt-Anh có 11364 từ trong đó có

| Từ loại | Số từ |
|---|---|
| Danh từ | 4933 |
| Động từ | 3965 |
| Tính từ | 2336 |
| Phó từ | 854 |
| Tổng cộng | 11364 |

## 5. Đánh giá – Kết luận

## 6. Tài liệu tham khảo


- Arthur T.M (1997). Longman Lexicon Of Contemporary English. Bản dịch tiếng Việt của Trần Tất Thắng. NXB Giáo Dục.
- Atserias J., Climent S., Farreras J., Rigau G., Rodriguez H. (1997) .*Combining Multiple Methods for the Automatic Construction of Multilingual WordNets.* In Proceeding of the Conference on Recent Advances on NLP. xem
- Bruce R, Guthrie L (1992). *Genus Disambiguation: A Study in Weighted Preference.* Proceedings of COLING'92. xem
- Đinh Điền 2004. *Xây dựng và khai thác kho ngữ liệu song ngữ Anh Việt điện tử.* Luận án Tiến Sĩ Ngôn Ngữ Học. ĐH KHXH&NV Tp HCM. xem
- Gutinie, Louise, Brian Slator, Yorick Wilks, và Rebecca Bluce (1990). *Is there content in Empty tleads*? Proceedings of the 13th International

Conference on Computational Linguistics (COLING-90), Helsinki, Finland, 3,pp.138-143.

- J. Daude, L. Padro & G. Rigau (2000) *Mapping WordNets Using Structural Information*. Proceeding of 38th Annual Meeting of the ACL.xem
- J. Daude, L. Padro & G. Rigau (1999) *Mapping Multilingual Hierarchies Using Relaxation Labeling.* In Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99). xem
- Knight K. and Luk S. (1994) *Building a Large-scale Knowledge Base for Machine Translatio*n. Proceeding of the American Association for Articial Intelligence. xem
- Okumura A., Hovy E. (1994). *Building Japanese-English Dictionary based on Ontology for Machine Translation.* Proceedings of ARPA Workshop on Human Language Technology.
- P. Bhattacharyya and Narayan Unny (2002), *Word Sense Disambiguation and Text Similarity Measurement Using WordNet*, chapter in Real World Semanic Web Applications, IOS Press, Amsterdam, 2002. Vipual Kashyap and Leon Shklar (ed), ISBN: 1 58603 306 9 xem
- Rigau G., Rodriguez H., Agirre E (1998). *Building Accurate Semantic Taxonomies from Monolingual MRDs*. In Proceedings of COLING-ACL'98. xem
- Rigau G., Agirre E. (1995). *Disambiguating Bilingual Nominal Entries against WordNet*. Proceedings of Third Workshop on Very Large Corpora. xem
- Hoàng Phê (2001). Từ điển Tiếng Việt. Hội Ngôn Ngữ Học - Nhà Xuất Bản Đà Nẵng.
- Mai Ngọc Chừ – Vũ Đức Nghiệu – Hoàng Trọng Phiến (1997). *Cơ sở ngôn ngữ học và tiếng Việt*. Nxb Giáo dục, H., 1997, tr. 166 - 171.

- John Lyons (1995). Linguistic semantics - An introduction. Cambridge University Press, 1995. (Người dịch  Nguyễn Văn Hiệp)