**ĐẠI HỌC QUỐC GIA HÀ NỘI**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Trần Thị Ngân**

# TRÍCH CHỌN THÔNG TIN Y TẾ TIẾNG VIỆT CHO BÀI TOÁN TÌM KIẾM NGỮ NGHĨA

**KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Công nghệ thông tin**

***HÀ NỘI - 2009***

**ĐẠI HỌC QUỐC GIA HÀ NỘI**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Trần Thị Ngân**

# TRÍCH CHỌN THÔNG TIN Y TẾ TIẾNG VIỆT CHO BÀI TOÁN TÌM KIẾM NGỮ NGHĨA

**KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Công nghệ thông tin**

**Cán bộ hướng dẫn: PGS. TS. Hà Quang Thụy**
**Cán bộ đồng hướng dẫn: Th.S Nguyễn Cẩm Tú**

***HÀ NỘI - 2009***

# LỜI CẢM ƠN

Đầu tiên cho em gửi lời cảm ơn sâu sắc nhất đến PGS. TS. Hà Quang Thụy, Th.S Nguyễn Cẩm Tú đã tận tình chỉ bảo cho em trong suốt thời gian thực hiện khóa luận. Trong quá trình nghiên cứu em đã gặp phải nhiều khó khăn nhưng nhờ sự hướng dẫn tận tình của thầy và chị em đã dần vượt qua và hoàn thành được khóa luận.

Em xin bày tỏ lòng biết ơn đến các thầy cô trong trường Đại Học Công Nghệ đã giảng dạy và cho em những kiến thức quý báu, làm nền tảng để hoàn thành khóa luận cũng như thành công trong nghiên cứu, làm việc trong tương lai.

Em xin gởi lời cảm ơn tới các anh chị trong phòng Lab đã cho em những lời khuyên quý báu, bổ ích trong quá trình thực hiện quá luận.

Và em cũng xin lời cảm ơn tới những người bạn thân yêu, đặc biệt là các bạn trong phòng ký túc xá đã bên cạnh động viên trong để giúp em hoàn thành khóa luận cũng như vượt qua nhiều khó khăn trong cuộc sống.

Cuối cùng, cho con gửi lời cảm ơn sâu sắc tới gia đình, bố, mẹ, chị và em đã cho con nhiều tình thương cũng như sự động viên kịp thời để con vượt qua những khó khăn trong cuộc sống và hoàn thành được khóa luận.

# TÓM TẮT

Trích chọn thông tin y tế nhằm xây dựng được một tập dữ liệu tốt, đầy đủ để hỗ trợ việc tìm kiếm ngữ nghĩa đang là nhu cầu thiết yếu, nhận được sự quan tâm đặc biệt trong thời gian gần đây. Ontology là cách biểu diễn khái niệm, thuộc tính, quan hệ trong miền ứng dụng đảm bảo tính nhất quán và đủ phong phú. Xây dựng hệ thống trích chọn thông tin dựa trên một Ontology y tế Tiếng Việt cho phép tìm kiếm và khai phá loại dữ liệu thuộc miền ứng dụng hiệu quả hơn là một nhu cầu thiết yếu.

Khóa luận này đề cập tới việc xây dựng một hê thống trích chọn thông tin dựa trên một ontology trong lĩnh vực y tế tiếng Việt. Khóa luận đã phân tích một số phương pháp, công cụ xây dựng Ontology để lựa chọn một mô hình và xây dựng được một Ontology y tế tiếng Việt với 21 lớp thực thể,13 mối quan hệ và trên 500 thể hiện của các lớp thực thể. Khóa luận đã tiến hành chú thích cho 96 file dữ liệu với trên 1500 thể hiện. Hệ thống nhận diện thực thể thực nghiệm của khóa luận đã hoạt động có tính khả thi với độ đo F1 trung bình qua 10 lần thực nghiệm đạt khoảng 64%.

MỤC LỤC

# DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

# Lời mở đầu

Chăm sóc sức khỏe luôn là một nhu cầu thiết yếu của con người, vì thế tìm kiếm các thông tin về lĩnh vực y tế trên Internet luôn là một nhu cầu thiết yếu. Vấn đề này càng cần phải được quan tâm thích đáng khi con người đang phải đối mặt với nhiều dịch bệnh truyền nhiễm, ví dụ điển hình có thể kể tới dịch bệnh cúm A H1N1 đang phát triển và có chiều hướng gia tăng trong thời gian gần đây. Cùng với sự ra đời và phát triển không ngừng của các tài nguyên trực truyến, việc khai thác hiệu quả nguồn tài nguyên này để đưa tới nguồn tri thức hữu ích cho người dùng sẽ góp phần vào việc tuyên truyền và nâng cao sức khỏe cộng đồng.

Sự bùng nổ các tài nguyên y tế, đặc biệt là các thông tin trực tuyến liên quan đến lĩnh vực sức khỏe; nhiều trang web và thông tin thừa cũng như việc tổ chức thông tin một cách tự do (không hoặc bán cấu trúc) … làm cho người dùng khó có thể theo dõi cũng như nắm bắt những thông tin cập nhật nhất. Bên cạnh đó, công nghệ tìm kiếm thông tin truyền thống hoặc trả về kết quả ít do sự phong phú, phức tạp của việc diễn đạt ngôn ngữ tự nhiên; hoặc quá nhiều theo nghĩa người tìm tin chỉ muốn tìm kiếm những tri thức ẩn chứ không chỉ là các văn bản chứa từ khóa tìm kiếm. Do đó việc khai thác tối ưu nguồn tài nguyên phong phú này trở thành một đề tài quan trọng, thu hút nhiều nhà khoa học tham gia nghiên cứu trong hai thập niên gần đây, có nhiều công trình nhằm trích rút các thông tin có cấu trúc từ những tài nguyên này nhằm xây dựng các cơ sở tri thức cho việc tổ chức thông tin, tìm kiếm, truy vấn, quản lý và phân tích thông tin.

Nhiều bài toán đã được đặt ra trong lĩnh vực trích chọn thông tin y tế như BioCreative-I (nhận diện các tên genes và protein trong văn bản) [32], LLL05 (trích chọn thông tin về gene) [33], BioCreative-II (trích chọn quan hệ tương tác giữa các protein) [49], …Những bài toán được đưa ra nhằm đánh giá các chiến lược khai phá dữ liệu y tế và đặc biệt tập trung vào hai bài toán con: nhận diện thực thể và trích chọn quan hệ. Nhận diện thực thể đòi hỏi nhận biết các thành phần cơ bản như tên thuốc, tên bệnh, triệu chứng, gene, protein, … trong văn bản. Xác định quan hệ với một mẫu cho trước là nhận biết một trường hợp của quan hệ này trong văn bản. Ví dụ, xác định quan hệ <gây_ra> giữa một bệnh xác định và một virus xác định. Ontology là một trong những cách biểu diễn mẫu cho các khái niệm, quan hệ đó một cách nhất quán và phong phú nhất. Việc xây dựng một Ontology cho y tế trong

tiếng Việt sẽ là cơ sở cho phép tìm kiếm, khai phá loại thông tin này một cách hiệu quả.

Theo khảo sát dữ liệu cho thấy ở Việt Nam hiện nay các Ontology cho y tế tiếng Việt thì hầu như chưa có; tuy nhiên cũng có đã có một số nhóm nghiên cứu tập trung xây dựng Ontology với các miền cụ thể khác để phục vụ cho nhiều mục đích khác nhau. Đơn cử có thể kế tới Ontology VN–KIM [34] được phát triển tại Đại học Bách khoa, Đại Học Quốc gia TP.Hồ Chí Minh. Ontology này bao gồm 347 lớp thực thể và 114 quan hệ và thuộc tính. VN-KIM Ontology bao gồm các lớp thực thể có tên phổ biến như Con _người, Tổ_chức, tỉnh, Thành_phố,…, các quan hệ giữa các lớp thực thể và các thuộc tính của mỗi lớp thực thể .

Tồn tại nhiều phương pháp được đưa ra để xây dựng một hệ thống trích chọn thông tin cũnug như xây dựng mạng ngữ nghĩa và từ đó áp dụng cho bài toán tìm kiếm ngữ nghĩa. Khóa luận trình bày cách biểu diễn dựa trên Ontology - một trong số những phương pháp đang được sử dụng khá rộng rãi hiện nay. Khóa luận trình bày một số phương pháp xây dựng Ontology, mở rộng ontology một cách tự động, giới thiệu bài toán nhận dạng thực thể cũng như phân loại quan hệ dựa trên một số phương pháp khác nhau. Khóa luận cũng đã xây dựng được một dữ liệu cho y tế phục vụ cho việc nhận dạng thực thể và quan hệ được hiệu quả hơn.

# Chương 1

# TỔNG QUAN VỀ TÌM KIẾM NGỮ NGHĨA

## *1.1. Nhu cầu về tìm kiếm ngữ nghĩa*

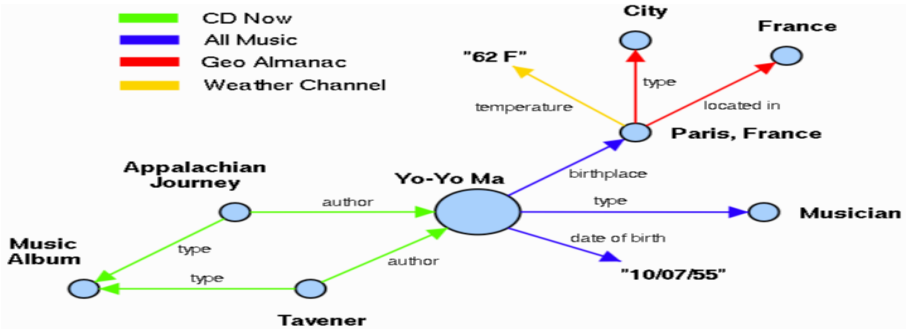
Sự bùng nổ các thông tin trực tuyến trên Internet và World Wide Web tạo ra một lượng thông tin khổng lồ đưa ra thách thức là làm thế nào để có thể khai phá hết được lượng thông tin này một cách hiệu quả nhằm phục vụ đời sống con người. Các máy tìm kiếm như Google, Yahoo… ra đời nhằm hỗ trợ người dùng trong quá trình tìm kiếm và sử dụng thông tin. Tuy kết quả trả về của các máy tìm kiếm này ngày càng được cải thiện về chất và lượng nhưng vẫn đơn thuần là danh sách các tài liệu chứa những từ xuất hiện trong câu truy vấn. Những thông tin từ các kết quả trả về này chỉ được hiểu bởi con người, máy tính không thể "hiểu" được, điều này gây những khó khăn cho quá trình tiếp theo xử lý thông tin tìm kiếm được. Thế hệ các máy tìm kiếm thực thể ra đời (hệ thống Cazoodle tại trang web http://www.cazoodle.com/, hệ thống Arnetminer tại trang web http://www.arnetminer.org/ **...**) đánh dấu một bước phát triển mới của các máy tìm kiếm. Thêm vào đó, với sự ra đời của máy tìm kiếm ngữ nghĩa **Wolfram**, được xây dựng và phát triển bởi dự án **Wolfram Research, Inc. Marketed** do **Stephen Wolfram** đề xuất [35], thì vấn đề tìm kiếm tri thức càng được quan tâm hơn nữa.

Sự ra đời của Web ngữ nghĩa (hay Semantic Web) do W3C (The World Wide Web Consortium) khởi xướng đã mở ra một bước tiến của công nghệ Web, những thông tin trong Web ngữ nghĩa có cấu trúc hoàn chỉnh và mang ngữ nghĩa mà máy tính có thể "hiểu" được. Những thông tin này, có thể được sử dụng lại mà không cần qua các bước tiền xử lý. Khi sử dụng các máy tìm kiếm thông thường (Google, Yahoo…), tìm kiếm thông tin trên Web ngữ nghĩa sẽ không tận dụng được những ưu điểm vượt trội của Web ngữ nghĩa, kết quả trả về không có sự cải tiến. Nói theo một cách khác thì với các máy tìm kiếm hiện tại thì Web ngữ nghĩa hay Web thông thường chỉ là một. Do vậy, cần thiết có một hệ thống tìm kiếm ngữ nghĩa (Semantic Search) tìm kiếm trên Web ngữ nghĩa hay trên một mạng tri thức mang ngữ nghĩa, kết quả trả về là các thông tin có cấu trúc hoàn chỉnh mà máy tính có thể "hiểu" được, nhờ đó việc sử dụng hay xử lý thông tin trở nên dễ dàng hơn [6][26][2]. Ngoài ra, việc xây dựng được một hệ thống tìm kiếm ngữ nghĩa cụ thể sẽ tạo tiền đề cho việc mở rộng xây dựng các hệ thống hỏi đáp tự động trên từng lĩnh vực cụ thể như : y tế, văn hóa … điều này mang một ý nghĩa thiết thực trong đời sống.

## *1.2. Nền tảng tìm kiếm ngữ nghĩa*

### 1.2.1.Web ngữ nghĩa

Web ngữ nghĩa hay còn gọi là Semantic Web theo Tim Berners-Lee là bước phát triển mở rộng của công nghệ Word Wide Web hiện tại, chứa các thông tin được định nghĩa rõ ràng để con người và máy tính làm việc với nhau hiệu quả hơn. Mục tiêu của Web ngữ nghĩa là phát triển dựa trên những chuẩn và công nghệ chung, cho phép máy tính có thể hiểu thông tin chứa trong các trang Web nhiều hơn nhằm hỗ trợ tốt con người trong khai phá dữ liệu, tổng hợp thông tin, hay trong việc xây dựng các hệ thống tự động khác… Không giống như công nghệ Web thông thường, nội dung chỉ bao hàm các tài nguyên văn bản, liên kết, hình ảnh, video mà Web ngữ nghĩa có thể bao gồm những tài nguyên thông tin trừu tượng hơn như: địa điểm, con người, tổ chức… thậm chí là một sự kiện trong cuộc sống. Ngoài ra, liên kết trong Web ngữ nghĩa không chỉ đơn thuần là các siêu liên kết (hyperlink) giữa các tài nguyên mà còn chứa nhiều loại liên kết, quan hệ khác. Những đặc điểm này khiến nội dung của Web ngữ nghĩa đa dạng hơn, chi tiết và đầy đủ hơn. Đồng thời, những thông tin chứa trong Web ngữ nghĩa có một mối liên hệ chặt chẽ với nhau. Với sự chặt chẽ này, người dùng dễ dàng hơn trong việc sử dụng, và tìm kiếm thông tin. Đây cũng là ưu điểm lớn nhất của Web ngữ nghĩa so với công nghệ Web thông thường [2].




**Hình 1. Ví dụ về Web ngữ nghĩa [6]**

Hình 1 là một ví dụ mô tả về một trang Web ngữ nghĩa chứa thông tin của một người tên là Yo-Yo Ma. Trang Web có cấu trúc như một đồ thị có hướng mang trọng số, trong đó mỗi đỉnh của đồ thị mô tả một kiểu tài nguyên chứa trong trang Web. Các cạnh của đồ thị thể hiện một kiểu liên kết (hay còn gọi là thuộc tính của tài nguyên) giữa các tài nguyên, trọng số của các liên kết đó thể hiện tên của liên kết [tên của thuộc tính] đó. Cụ thể ta thấy Yo-Yo Ma có thuộc tính ngày sinh là ***"10/07/55"*** có nơi sinh ở ***"Paris, France"***, ***"Paris, France"*** có nhiệt độ là ***"62 F"*** …

Như vậy, mỗi tài nguyên được mô tả trong Web ngữ nghĩa là một đối tượng. Đối tượng này có tên gọi, thuộc tính, giá trị của thuộc tính (giá trị có thể là một đối tượng khác) và liên kết với các tài nguyên (đối tượng) khác (nếu có). Để xây dựng được một trang Web ngữ nghĩa cần phải có tập dữ liệu đầy đủ, hay nói một cách khác là cần phải xây dựng một tập các đối tượng mô tả tài nguyên cho Web ngữ nghĩa. Các đối có quan hệ với nhau hình thành một mạng liên kết rộng, được gọi là *mạng ngữ nghĩa*.

Mạng ngữ nghĩa được chia sẻ rộng khắp do vậy các đối tượng trong một mạng ngữ nghĩa cần phải mô tả theo một chuẩn chung nhất. Ontology được sử dụng để mô tả về đối tượng, tài nguyên cho Web ngữ nghĩa [2].

### 1.2.2. Ontology

Có thể hiểu một cách đơn giản ontology là một mô hình dữ liệu trình bày một tập các khái niệm trong một miền và mối quan hệ giữa các khái niệm đó. Nó được sử dụng để lập luận (suy luận) về các đối tượng trong miền đó [12].

Ontology là một trong những cách biểu diễn mẫu cho các khái niệm, quan hệ đó một cách nhất quán và phong phú nhất, chính vì thế nó được sử dụng để xây dựng mạng ngữ nghĩa từ tập dữ liệu thô (không hoặc bán cấu trúc) tạo nền tảng xây dựng một máy tìm kiếm ngữ nghĩa một cách hiệu quả. Ontology sẽ được giới thiệu một cách cụ thể, kỹ lưỡng hơn trong chương 2 của khóa luận.

## *1.3. Kiến trúc của một máy tìm kiếm ngữ nghĩa*

Xét về cơ bản, một máy tìm kiếm ngữ nghĩa có cấu trúc tương tự với một máy tìm kiếm thông thường cũng bao gồm hai thành phần chính [2]:

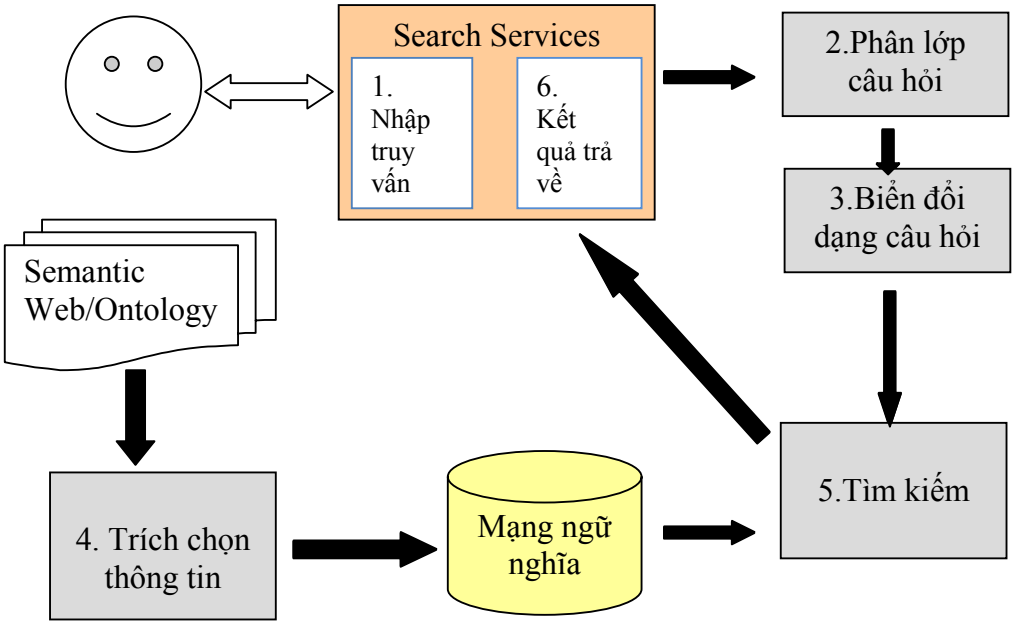
Phần giao diện người dùng (front end) có hai chức năng chính:

- Giao diện truy vấn: cho phép người dùng nhập câu hỏi, truy vấn.
- Hiển thị câu trả lời, kết quả.

Phần kiến trúc bên trong (back end) là phần hạt nhân của máy tìm kiếm bao gồm ba thành phần chính đó là:

- Phân tích câu hỏi
- Tìm kiếm kết quả cho truy vấn hay câu hỏi
- Tập tài liệu, dữ liệu tìm kiếm, mạng ngữ nghĩa.

Mô hình kiến trúc một máy tìm kiếm ngữ nghĩa được mô tả như Hình 2.




**Hình 2. Kiến trúc một máy tìm kiếm ngữ nghĩa [2]**

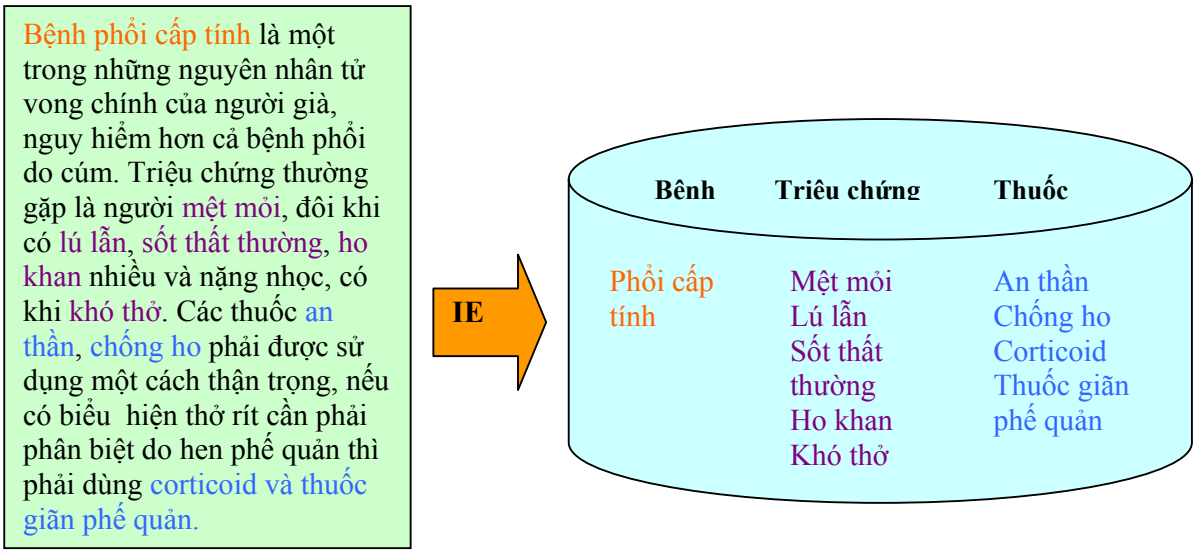
Có thể thấy rằng sự khác biệt trong cấu trúc của máy tìm kiếm ngữ nghĩa so với máy tìm kiếm thông thường nằm ở phần kiến trúc bên trong, cụ thể ở hai thành phần: phân tích câu hỏi và tập dữ liệu tìm kiếm.

Phân tích câu hỏi đã được đề cập chi tiết trong [2]. Tập dữ liệu tìm kiếm chính là web ngữ nghĩa và mạng ngữ nghĩa được xây dựng dựa trên ontology và hệ thống trích chọn thông tin. Khóa luận này tập trung nghiên cứu kỹ về xây dựng ontology, mở rộng tự động ontology nhờ trích chọn thông tin mà cụ thể là nhận dạng thực thể. Khóa luận cũng đề cập tới nhận dạng quan hệ ngữ nghĩa, phân loại câu chứa quan hệ nhằm mục đích như đã trình bày ở trên, đó là xây dựng được một tập dữ liệu tìm kiếm đầy đủ cho máy tím kiếm ngữ nghĩa trong tương lai.

### *1.4.Trích chọn thông tin*

Trích chọn thông tin là một lĩnh vực quan trọng trong khai phá dữ liệu văn bản, thực hiện việc trích rút các thông tin có cấu trúc từ các văn bản không có cấu trúc. Nói cách khác, một hệ thống trích chọn thông tin rút ra những thông tin đã được định nghĩa trước về các thực thể và mối quan hệ giữa các thực thể từ một văn bản dưới dạng ngôn ngữ tự nhiên và điền những thông tin này vào một văn bản ghi dữ liệu có cấu trúc hoặc một dạng mẫu được định nghĩa trước đó. Có nhiều mức độ trích chọn thông tin từ văn bản như xác định các thực thể (Element Extraction), xác định quan hệ giữa các thực thể (Relation Extraction), xác định và theo dõi các sự

kiện và các kịch bản (Event and Scenario Extraction and Tracking), xác định đồng tham chiếu (Co-reference Resolution)... Các kĩ thuật được sử dụng trong trích chọn thông tin gồm có: phân đoạn, phân lớp, kết hợp và phân cụm [1].




**Hình 3. Minh họa một hệ thống trích chọn thông tin**

Để có một hệ thống trích chọn thông tin đầu tiên chúng ta phải có một hệ thống nhận dạng thực thể và tiếp sau mới tính đến phân loại quan hệ. Bài toán nhận biết các loại thực thể là bài toán đơn giản nhất trong số các bài toán trích chọn thông tin, tuy vậy nó lại là bước cơ bản nhất trước khi tính đến việc giải quyết các bài toán phức tạp hơn trong lĩnh vực này. Ngoài ứng dụng trong hệ thống trích chọn thông tin, nó còn có thể được áp dụng trong tìm kiếm thông tin (Information Retrieval), dịch máy (machine translation) và hệ thống hỏi đáp (question answering).

Đã có rất nhiều bài toán được đặt ra trong lĩnh vực trích chọn thông tin y tế như BioCreative-I (nhận diện các tên genes và protein trong văn bản) [32], LLL05 (trích chọn thông tin về gene) [33], BioCreative-II (trích chọn quan hệ tương tác giữa các protein) [49], …Những bài toán được đưa ra nhằm đánh giá các chiến lược khai phá dữ liệu y tế và đặc biệt tập trung vào hai bài toán con: nhận diện thực thể và trích chọn quan hệ. Nhận diện thực thể đòi hỏi nhận biết các thành phần cơ bản như tên thuốc, tên bệnh, triệu chứng, gene, protein, … trong văn bản. Xác định quan hệ với một mẫu cho trước là nhận biết một trường hợp của quan hệ này trong văn bản. Ví dụ: xác định quan hệ <gây_ra> giữa một bệnh xác định và một virus

xác định. Ontology là một trong những cách biểu diễn mẫu cho các khái niệm, quan hệ đó một cách nhất quán và phong phú nhất. Việc xây dựng một ontology cho y tế trong tiếng Việt sẽ là cơ sở cho phép tìm kiếm, khai phá loại thông tin này một cách hiệu quả. Sau khi xây dựng ontology, công việc tiếp theo cũng rất quan trọng đó là mở rộng ontology một cách tự động. Việc có một hệ thống trích chọn thông tin (bao gồm nhận dạng thực thể và trích chọn quan hệ, …) là bước tiền để có thể mở rộng ontology một cách tự động.

# Chương 2

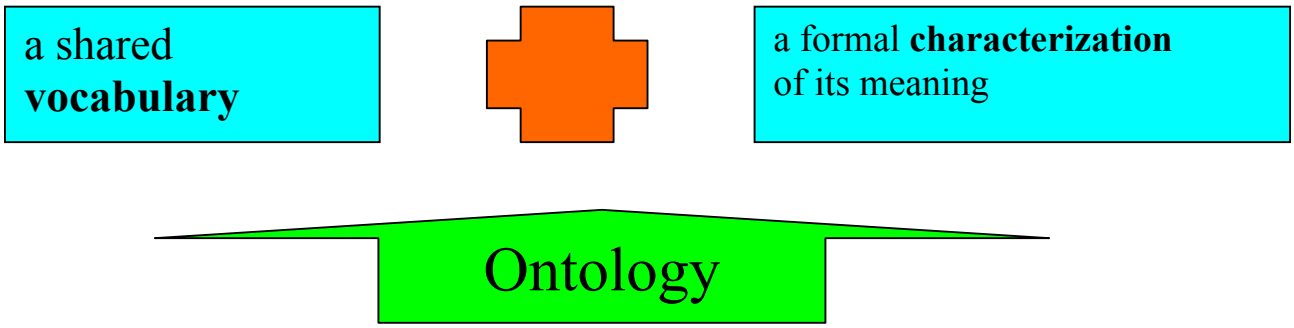
# XÂY DỰNG ONTOLOGY Y TẾ TIẾNG VIỆT

## *2.1. Giới thiệu Ontology*

### 2.1.1. Khái niệm Ontology

Trong những năm gần đây, thuật ngữ "Ontology" không chỉ được sử dụng ở trong các phòng thì nghiệm trên lĩnh vực trí tuệ nhân tạo mà đã trở nên phổ biến đối với nhiều miền lĩnh vực trong đời sống . Đứng trên quan điểm của ngành trí tuệ nhân tạo, một Ontology là sự môt tả về những khái niệm và những quan hệ của các khái niệm đó nhằm mục đích thể hiện một góc nhìn về thế giới. Trên miền ứng dụng khác của khoa học, một Ontology bao gồm tập các từ vựng cơ bản hay một tài nguyên trên một miền lĩnh vực cụ thể, nhờ đó những nhà nghiên cứu có thể lưu trữ, quản lý và trao đổi tri thức cho nhau theo một cách tiện lợi nhất [2].

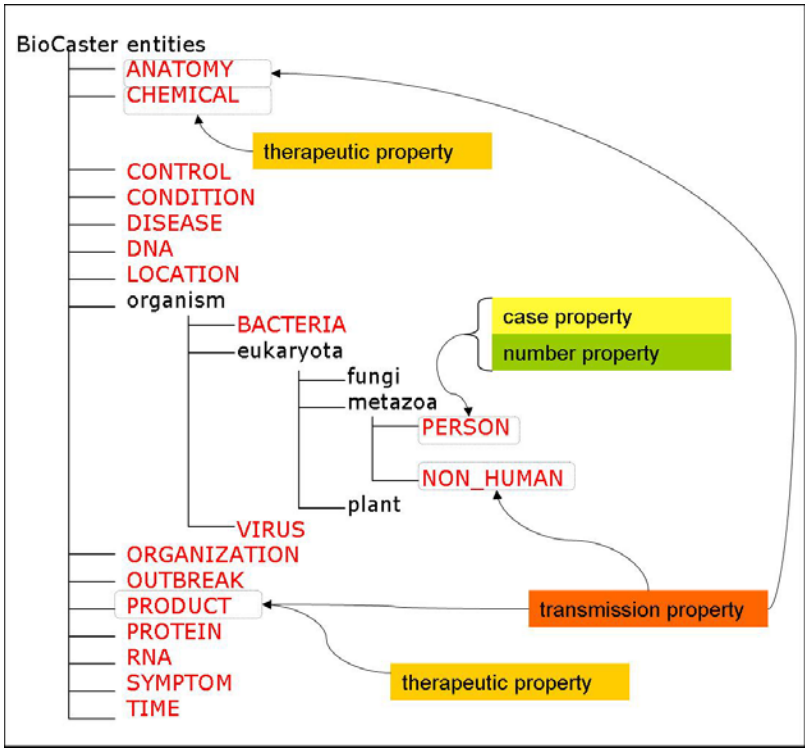
Hiện nay tồn tại nhiều khái niệm về Ontology, trong đó có nhiều khái niệm mâu thuẫn với các khác niệm khác, khóa luận này chỉ giới thiệu một định nghĩa mang tính khái quát và được sử dụng khá phổ biến được Kincho H. Law đưa ra: "Ontology là  biểu hiện một tập các khái niệm (đối tượng), trong một miền cụ thể và những mối quan hệ giữa các khái niệm này". Ontology chính là sự tổng hợp của một tập từ vựng chia sẻ và các miêu tả ý nghĩa của từ đó theo cách mà máy tính hiểu được.




**Hình 4.  Mô tả ý nghĩa của Ontology**

Hình 4 mô tả ý nghĩa của Ontology, trong đó tập từ vựng dùng chung (Vocabulary) chính là thể hiện của các lớp, quan hệ. Ví dụ, có thể có Vocabulary (...), Categories (Cat, White, Leg, Fish, Animal,…), Relations (Is-a, Part-of,

hasMother,…), Characterization (...) và các thể hiện quan hệ "A cat is an animal", "A cat has four legs"...




**Hình 5. Minh họa cấu trúc phân cấp của Ontology BioCaster [11]**

### 2.1.2. Các thành phần của Ontology

Các thành phần chính của Ontology là: Lớp (Class), thuộc tính (Property), thực thể (Individual).

Lớp (class) là một bộ những thực thể, các thực thể được mô tả logic để định nghĩa các đối tượng của lớp; lớp được xây dựng theo cấu trúc phân cấp cha con như là một sự phân loại các đối tượng. Thực thể được xem là thể hiện của một lớp, làm rõ hơn về lớp đó và có thể được hiểu là một đối tương nào đó trong tự nhiên (England, Manchester United, bệnh sởi, thủy đậu…).

Thuộc tính (Property) thể hiện quan hệ nhị phân của các thực thể (quan hệ giữa hai thực thể) như liên kết hai thực thể với nhau. Ví dụ thuộc tính ‘do_virus’ liên kết hai thực thể ‘bệnh’ và ‘virus’ với nhau.

Thuộc tính (property) có 4 loại (1) Functional: Một thực thể chỉ liên quan nhiều nhất đến một thực thể khác, ví dụ thuộc tính “có hương vị” đối với các thực thể lớp “thức_ăn”; (2) Inverse Functional: Thuộc tính đảo ngược của Functional,

thuộc tính "là hương vị của"; (3) Transitive: Thực thể a quan hệ với thực thể b, thực thể b quan hệ với thực thể c → thực thể a quan hệ với thực thể c; (4) Symmetric: Thực thể a quan hệ với thực thể b → thực thể b quan hệ với thực thể a.

Thuộc tính có 3 kiểu thể hiện (1) Object Property: Liên kết thực thể này với thực thể khác; (2) DataType Property: Liên kết thực thể với kiểu dữ liệu XML Schema, RDF literal; (3) Annotation Property: Thêm các thông tin metadata về lớp, thuộc tính hay thực thể khác thuộc 2 kiểu trên.

Để làm việc với ontology Web cần sử dụng ngôn ngữ ontology Web (The Web Ontology Language: OWL). OWL có thể có một kiểu thứ tư là Annotation propertie. Kiểu thuộc tính được sử dụng để thêm các thông tin (metadata – dữ liệu của dữ liệu) đối với các lớp, các thực thể hay các thuộc tính Object/ Datatype.

### 2.1.3 Một số công trình liên quan tới xây dựng Ontology

Ngày nay, Ontology được sử dụng rất nhiều trong các lĩnh vực liên quan đến ngữ nghĩa như trí tuệ nhân tạo (AI), semantic web, kĩ nghệ phần mềm, v.v… Vì những ứng dụng của Ontology nên không chỉ riêng Việt Nam, trên thế giới đã có nhiều dự án tập trung xây dựng Ontology đối với từng miền dữ liệu khác nhau và phục vụ cho nhiều mục đích đa dạng khác nhau. Đối với miền dữ liệu y tế có thể kể tới rất nhiều Ontology trong lĩnh vực y tế, sinh học đã được đưa ra bởi tổ chức The National Center for Biomedical Ontology [52]. Dự án này đã đưa ra được rất nhiều Ontology trong y tế cũng như trong sinh học, ví dụ như Ontology về cell type, Gene, FMA, Human disease…danh sách các Ontology đưa ra được hiển thị trong [41].

Ngoài ra có thể kể tới Disease Ontology [42] là một tập từ về y khoa được phát triển tại Bioinformatics Core Facility cùng với sự cộng tác của dự án NuGene Project tại trung tâm Center for Genetic Medicine. Ontology này được thiết kế với mục đích sắp xếp các bệnh và các điều kiện tương ứng đối với những code về y tế cụ thể như là ICD9CM, SNOMED và những cái khác….Disease Ontology cũng được sử dụng để liên kết những kiểu hình sinh vật mẫu đối với các bệnh của con người cũng như trong việc khai phá dữ liệu y học. Disease Ontology được thực hiện như là một đồ thị xoắn có hướng và sử dụng UMLS (Unified Medical Language System) là tập từ vựng để truy cập các Ontology về y tế khác như ICD9CM.

Một ontology tiếng Anh được đề cập rất nhiều trong lĩnh vực y tế trong thời gian gần đây đó là GENIA [43]. Mục đích chính mà ontology này hướng tới đó là

sự phản ứng lại của tế bào trong não người. Ontology này chủ yếu tập trung trong các lĩnh vực y tế và cũng được sử dụng trong các bài toán xử lý ngôn ngữ tự nhiên: truy hồi thông tin (Information Retrieval – IR), trích chọn thông tin, phân lớp và tóm tắt văn bản …Hình vẽ sau mô tả cấu trúc phân cấp của ontology GENIA.

Tồn tại nhiều Ontology về y tế hiện nay đã được xây dựng trên thế giới. Tuy nhiên ở Việt Nam hiện nay mặc dầu việc tìm kiếm ngữ nghĩa đang được tập trung nghiên cứu, nhưng các Ontology về y tế thì hầu như chưa có, cho nên việc tìm kiếm các trang web về thuốc, bệnh … của người dùng  chưa trả về các kết quả đầy đủ và đạt được hiệu quả. Tồn tại một Ontology đề cập đến các thuật ngữ y tế trong tiếng Việt, đó là Ontology Biocaster [44]. Đây là Ontology được nghiên cứu theo dự án Biocaster được phát triển tại Viện Tin học Quốc gia Nhật Bản với sự cộng tác của trường các trường đại học tại Nhật Bản, Thái Lan, Việt Nam... Đây là ontology viết cho nhiều ngôn ngữ như Nhật, Anh, Thái, Việt…

Ontology BioCaster [11] có các thuật ngữ của nhiều thứ tiếng trong đó có 371 thuật ngữ tiếng Việt, các thuật ngữ liên quan đến bệnh, virus, các triệu chứng của Việt Nam. Mặc dù Ontology này có xử lý trích chọn trong tiếng Việt, nhưng từ đó lại đưa ra các bài báo về y tế Việt Nam bằng tiếng Anh. Vì vậy, các thuật ngữ, thực thể, các bệnh hay virus được viết bằng tiếng Việt còn  các quan hệ được mô tả bằng tiếng Anh. Ví dụ, thuật ngữ Vietnamese_103, gán nhãn: vi rút gây bệnh thủy đậu, có hasLanguage: vi (Vietnamese), hasRootTerm : VIRUS_124…

## *2.2. Lý thuyết xây dựng Ontology*

### 2.1.1. Phương pháp xây dựng Ontology

Ngày nay, việc nghiên cứu quá trình xây dựng ontology ngày càng được quan tâm nhiều hơn. Có rất nhiều nhóm sau quá trình nghiên cứu đã đưa ra các phương pháp khác nhau nhằm xây dựng Ontology.

Phương pháp Ushold & King được xây dựng dựa trên việc phát triển Enterprise Ontology. Phương pháp này chủ yếu tập trung vào việc giúp người phát triển từ mục đích của ontology có thể có những hướng phát triển như thế nào, sau đó đánh giá và viết tài liệu cho ontology. Trong quá trình xây dựng, người dùng có thể tích hợp các ontology có sẵn vào ontology đang xây dựng. Ba cách tiếp cận sau được đưa ra nhằm định nghĩa các khái niệm chính trong ontology: cách tiếp cận top-down, bottom-up và middle-out. Phương pháp luận này được xây dựng không phụ thuộc vào ứng dụng, nghĩa là mục đích xây dựng ontology độc lập với quá

trình xây dựng chúng, không phụ thuộc vào nhau. Với bất kì ứng dụng nào, chúng ta đều có thể sử dụng chung phương pháp này [17].

Phương pháp luận tiếp theo được phát triển bởi Gruninger và Fox [16], được phát triển thông qua dự án ontology Toronto Virtual Enterprise (TOVE). Hệ thống này được xây dựng bắt nguồn từ tư tưởng về sự phát triển hệ thống dựa trên tri thức, sử dụng first order logic. Trong phương pháp này, các khái niệm nổi bật nhất được định nghĩa trước tiên, sau đó làm chi tiết và tổng quát hóa các khái niệm đó theo các hướng thích hợp. Như vậy, phương pháp này bắt đầu từ một số các khái niệm ở mức cao, đi rồi đến các khái niệm ở mức thấp và tổng quát ở các mức cao hơn. Phương pháp này sử dụng cách tiếp cận middle-out để định nghĩa các khái niệm và một phần phụ thuộc vào ứng dụng sau này của ontology, nghĩa là trước khi xây dựng ontology, người dùng cần quyết định mục đích sử dụng và tích hợp ontology vào ứng dụng gì.

METHONTOLOGY là một phương pháp xây dựng Ontology được phát triển từ phòng nghiên cứu trí tuệ nhân tạo của trường ĐH Polytechnic Madrid. Phương pháp này cho phép người sử dụng có thể xây dựng một ontology mới dựa trên bản mẫu thiết kế mới hoặc có thể sử dụng những ontology có sẵn. Bộ framework của METHONTOLOGY có thể giúp người dùng xây dựng cấu trúc ontology ở mức độ tri thức và bao gồm: định nghĩa quy trình phát triển ontology, một số kỹ thuật trong quá trình xây dựng quy trình trên (ví dụ quản lý và lập lịch, quản lý chất lượng, thu thập dữ liệu và tri thức, quản lý cấu hình, v.v.). Phương pháp luận này sử dụng chiến lược middle-out và không phụ thuộc vào ứng dụng.

### 2.1.2. Công cụ xây dựng Ontology

Bộ công cụ xây dựng và phát triển Ontology bao gồm các tool hỗ trợ và môi trường giúp người dùng có thể xây dựng một Ontology mới từ bản thiết kế mới hoặc sử dụng lại những Ontology mới có sẵn. Một số môi trường phát triển được xây dựng từ trước như Ontosaurus, Ontolingua và WebOnto. Những bộ công cụ mới được sử dụng nhiều gần đây bao gồm OntoEdit, OilED,WebODE, Chimera DAG-Edit và Protégé.

Ontoligua server [45] là bộ công cụ xây dựng ontology được phát triển từ những năm 1990 tại Phòng Thí nghiệm Hệ thống tri thức (Knowledge Systems Laboratory -KSL) của Trường ĐH Stanford (Mỹ). Các module chính của bộ công cụ bao gồm bộ biên tập ontology (ontology editor) và các module khác như Webster, OKBC (Open knowledge Based Connectivity) server.

Ontosaurus [46] được phát triển cùng trong khoảng thời gian đó bởi Viện Khoa học Thông tin ISI của Trường ĐH South Calfornia (Mỹ). OntoSaurus bao gồm 2 module chính: ontology server (sử dụng Loom) và một web browser cho Loom ontology. Ngoài ra, bộ công cụ còn hỗ trợ KIF, KRSS và C++, đồng thời OntoSaurus ontology cũng có thể được truy cập dựa trên protocol OKBC của Ontoligua server.

WebOnto là một ontology editor cho các Ontology OCML (Operational Conceptual Modelling Language), được phát triển bởi Viện Truyền thông Tri thức (KMI) tại Trường ĐH mở (Open University). Bộ công cụ này là sử dụng Java với webserver, cho phép người dùng có thể duyệt và thay đổi các mô hình tri thức thông qua Internet. Điểm mạnh chính của bộ công cụ này là có thể cho phép cộng tác giữa nhiều người nhằm thay đổi và hoàn thiện ontology [26].

Các bộ công cụ trên (Ontolingua server, Ontosaurus và WebOnto) được xây dựng đơn thuần nhằm hỗ trợ duyệt và biên tập các Ontology được viết bằng những ngôn ngữ riêng (Ontolingua, LOOM và OCML). Những bộ công cụ biên tập này hiện nay không còn đáp ứng đủ nhu cầu của người sử dụng. Thế hệ mới các bộ công cụ xây dựng Ontology có nhiều ưu việt cũng như tính năng hơn hẳn các bộ công cụ này, ví dụ như khả năng mở rộng, hệ thống kiến trúc các thành phần – giúp người dùng có thể cung cấp thêm các tính năng cho môi trường phát triển một cách dễ dàng.

WebODE [47] là một bộ công cụ có khả năng mở rộng được phát triển bởi nhóm Ontology của trường ĐH Technical Madrid (UPM), được xem như một thành công của ODE (Ontology Design Environment). WebODE được sử dụng như một Web server với giao diện web. Phần lõi chính của môi trường này là một dịch vụ (service) ontology, trong đó tất cả các dịch vụ và ứng dụng khác đều có thể sử dụng dịch vụ này. Phần soạn thảo Ontology cũng đồng thời cung cấp công cụ kiểm tra ràng buộc, tạo các luật tiên đề (axiom rule creation) và phân tích với WebODE Axiom Builder (WAB), tài liệu trong HTML, kết hợp ontology với các định dạng khác nhau [XML\RDF[s], OIL, DAML+OIL, CARIN, Flogic, Java và Jess].

OilED [48] là một bộ công cụ soạn thảo ontology cho phép người dùng có thể xây dựng Ontology bằng OIL và DAML+OIL, được xây dựng bởi Trường ĐH Manchester, Đại học Amsterdam và Interprice GmbH.

Protégé 2000 [51] là một trong những bộ công cụ được sử dụng rộng rãi nhất hiện nay, được phát triển bởi Trường ĐH Stanford. Bộ công cụ này được phát triển

dựa trên hai mục tiêu: có thể tương thích với các hệ thống khác, dễ dàng sử dụng và hỗ trợ các công cụ trích chọn thông tin. Phần chính của môi trường này là một biên tập ontology. Bên cạnh đó, Protégé còn bao gồm rất nhiều các plugin nhằm hỗ trợ chức năng như quản lý nhiều ontology, dịch vụ suy luận (inference service), hỗ trợ về vấn đề ngôn ngữ ontology (language importation/exportation).

### 2.1.3. Ngôn ngữ xây dựng Ontology

Hiện tại, các ngôn ngữ xây dựng ontology (ngôn ngữ ontology) điển hình bao gồm LOOM, LISP, Ontolingua, XML, SHOE, OIL, DAML+OIL và OWL.

Ngôn ngữ ontology được chia làm ba loại: định ngữ tập từ vựng sử dụng ngôn ngữ tự nhiên (object based-knowledge representation languages) như UML, và ngôn ngữ dựa trên lôgic vị từ bậc một (first order predicate logic) như logic mô tả (Description Logics). Ngôn ngữ ontology cần phải tương thích với những công cụ khác, tự nhiên và dễ học, tương thích với các chuẩn hiện tại của web như XML, XML Schema, RDF và UML. Dưới đây là một số các ngôn ngữ web-based.

EXtensible Markup Language [XML] là một chuẩn mở dùng để biểu diễn dữ liệu từ W3C, có tính mềm dẻo và mạnh hơn so với HTML. RDF (Resource Description Framework) được phát triển như một khung giúp mô tả và trao đổi các metadata [12].

SHOE (Simple HTML Ontology Extensions) được xây dựng vào năm 1996 tại Trường ĐH Maryland, như một mở rộng của HTML để có thể hợp nhất các tri thức ngữ nghĩa trên các văn bản web hiện tại thông qua việc chú thích các trang HTML [27].

OIL (Ontology Inference Layer) là mở rộng của RDF, được phát triển bởi dự án ON-To_Knowledge, là ngôn ngữ mô tả và trao đổi cho ontology. Ngôn ngữ này được kết hợp bởi ngôn ngữ dạng dựa trên frame (frame-based) với ngữ nghĩa hình thức (formal sematics) và dịch vụ suy luận từ logic mô tả (description logics). Ngôn ngữ được chia làm ba mức đối tượng lớp (các thực thể cụ thể), mức đầu tiên (first-meta, định nghĩa theo ontology) và mức thứ hai (second-meta, các mối quan hệ) [8].

DAML+OIL được phát triển dựa trên dự án DARPA năm 2000. Cả OIL và DAML+OIL đều cho phép mô tả các khái niệm, các phân cấp (taxonomy), các quan hệ nhị phân, chức năng và thực thể [9].

OWL là một ngôn ngữ ontology được sử dụng phổ biến hiện nay, được tối ưu hoá cho việc trao đổi dữ liệu và chia sẻ tri thức. Ngôn ngữ này được sử dụng khi thông tin chứa trong văn bản cần được xử lý bởi các ứng dụng. OWL l có thể được sử dụng để biểu diễn ngữ nghĩa các thuật ngữ trong tập từ vựng và mối quan hệ giữa những thuật ngữ này. OWL bao gồm OWL Lite, OWL DL [RDF] và OWL FULL.

### *2.3. Xây dựng Ontology y tế tiếng Việt*

Việc thiết kế và xây dựng một ontology bao gồm các bước sau:

- Định nghĩa các lớp trong ontology.
- Sắp xếp các lớp trong một  kiến trúc phân cấp (taxonomic hierarchy).
- Định nghĩa các thuộc tính (slot) và mô tả các giá trị cho phép cho những thuộc tính này.
- Điền giá trị của các thể hiện (instance) vào các slot.
- Sau đó, cơ sở tri thức được tạo ra bằng cách định nghĩa các thể hiện (instance) của những lớp này cùng với những giá trị của chúng.

Không có một phương pháp nào được gọi là phương pháp chuẩn xác cho việc xây dựng tất cả các Ontology [18]. Việc lựa chọn phương pháp xây dựng phù hợp nào được dựa trên mục đích và tính chất của từng Ontology. Qua quá trình khảo sát các dữ liệu về y tế và một số các phương pháp phát triển Ontology, chúng tôi lựa chọn môi trường Protégé OWL xây dựng một Ontology y tế bằng Tiếng Việt thử nghiệm.

Sau khi thu thập và khảo sát dữ liệu, chúng tôi liệt kê các thuật ngữ quan trọng nhằm có thể nêu định nghĩa cho người dùng với hướng nghiên cứu tiếp theo là tự động liên kết đến các định nghĩa có sẵn trên trang wikipedia. Từ các thuật ngữ trên, tiếp theo sẽ định nghĩa các thuộc tính của chúng. Việc xây dựng Ontology là một quá trình lặp lại được bắt đầu bằng việc định nghĩa các khái niệm trong hệ thống lớp và mô tả thuộc tính của các khái niệm đó.

# Chương 3

# NHẬN DẠNG THỰC THỂ

## *3.1. Giới thiệu bài toán nhận dạng thực thể*

### 3.1.1. Giới thiệu chung về nhận dạng thực thể

Nhận dạng thực thể có thể hiểu một cách đơn giản là phân loai các từ trong một văn bản thành các lớp thực thể đã được định nghĩa trước như người (PER), tổ chức (ORG), vị trí (LOC), bệnh (BENH), triệu chứng (TCHUNG), thuốc (THUOC). Nhận dạng thực thể cho chúng ta được một phân tích bề mặt, các thực thể sẽ  trả lời các câu hỏi quan trọng (có thể ứng dụng trong hệ thống hỏi đáp…).

Có rất nhiều phương pháp đã được dùng để giải quyết bài toán nhận dạng thực thể, từ các phương pháp thủ công đến các phương pháp học máy như các mô hình markov ẩn (Hidden Markov Models – HMM), các mô hình Markov cực đại hóa Entropy (Maximum Entropy Markov Models- MEMM), các mô hình miền phụ thuộc điều kiện (Conditional Random Field - CRF), phương pháp máy vector hỗ trợ (Support Vector Machine).

Tiêu biểu cho hướng tiếp cận thủ công là hệ thống nhận biết loại thực thể Proteus của đại học New York tham gia MUC-6. Hệ thống được viết bằng Lisp và được hỗ trợ bởi một số lượng lớn các luật, tuy nhiên hầu hết các luật đều còn tồn tại một số lượng lớn các trường hợp ngoại lệ, trong đó có những ngoại lệ chỉ xuất hiện khi hệ thống đưa vào sử dụng, mà ta khó có thể giải quyết hết. Dưới đây là một số ví dụ về các luật được sử dụng bởi Proteus cùng với các trường hợp ngoại lệ của chúng [1]:

Luật: Title Capitalized_Word => Title Person Name

→Trường hợp đúng : Mr. Johns, Gen. Schwarzkopf

→Trường hợp ngoại lệ: Mrs. Field’s Cookies (một công ty).

Luật: Month_name number_less_than_32 => Date

→Trường hợp đúng: February 28, July 15

→Trường hợp ngoại lệ: Long March 3 ( tên một tên lửa của Trung Quốc).

So với các phương pháp thủ công vừa tốn thời gian, công sức, mà kết quả đạt được lại không được như mong muốn, các phương pháp học máy hiện đang

được tập trung nghiên cứu nhiều hơn. Hầu hết các phương pháp đều có những ưu thế riêng đồng thời vẫn còn tồn tại một số hạn chế do đặc thù của mỗi mô hình. Tiêu biểu có thể kể đến các mô hình Markov ẩn HMM và các mô hình cải tiến của nó như MEMM, CRF; với các mô hình này ta có thể xem tương ứng mỗi trạng thái với một trong nhãn các nhãn thực thể và dữ liệu quan sát là các từ trong câu đang xét. Máy vector hỗ trợ (SVM) cũng là một trong những phương pháp học máy cho kết quả rất khả quan.

### 3.1.2. Một số kết quả nghiên cứu về nhận dạng thực thể

Trên thế giới bài toán nhận biết thực thể đã được quan tâm nghiên cứu từ lâu và đạt được những kết quả khá ấn tượng. Có rất nhiều phương pháp (từ các phương pháp thủ công đến các phương pháp học máy) đã được dùng để giải quyết bài toán này. Trong công trình nghiên cứu vào năm 2007 [5], David Nadeau đã đánh giá một số nghiên cứu tiêu biểu trước đó có liên quan đến bài toán nhận dạng thực thể. Nội dung các đánh giá của David Nadeau được trình bày như dưới đây.

Tiêu biểu cho hướng tiếp cận thủ công là hệ thống nhận biết loại thực thể Proteus của đại học New York tham gia MUC-6. Hệ thống được viết bằng Lisp và được hỗ trợ bởi một số lượng lớn các luật. Năm 1998, Radev công nghiên cứu nhận dạng những đoạn mô tả về thực thể được đưa ra, chẳng hạn như Bill Clinton sẽ được mô tả là “the President of the U.S.”, “the democratic presidential candidate” hay “an Arkansas native”… Hệ thống của Fung 1995 (và Huang 2005) giải quyết bài toán dịch các thực thể từ ngôn ngữ này sang ngôn ngữ khác (ví dụ như bản dịch tiếng Việt của thực thể “College of Technology” sẽ là “Trường Đại học Công nghệ”). Hệ thống này được đánh giá là gặp phải ít hơn 10% lỗi dịch. Tiếp theo đó, năm 2001, Charniak và cộng sự công bố kết quả nghiên cứu nhận dạng cấu trúc các phần trong tên người, ví dụ như cụm “Doctor Paul R. Smith” sẽ được chia thành cá thành phần chức danh, họ, đệm và tên). Nghiên cứu này là một bước tiền xử lý quan trọng trong bộ nhận dạng thực thể, để có thể xác định những trường hợp như “John F. Kennedy” và “President Kennedy” là cùng một người. Cũng trong năm 2001, hệ thống “Record linkage” của Cohen và Richman được xây dựng với mục đích tìm ra tất cả các dạng của cùng một thực thể trên toàn bộ cơ sở dữ liệu. Vào năm 2002, Dimitrov và cộng sự đã giải quyết vấn đề sử dụng các đại từ thay thế, ví dụ trong câu “Rabi finished reading the book and he replaced it in the library” đại từ “he” là đại từ thay thế cho “Rabi”. Nghiên cứu này có rất nhiều ứng dụng thực tế, ví dụ như trong hệ thống hỏi đáp tự động. Năm 2003, Mann và Yarowski xây dựng một hệ thống xóa bỏ các nhập nhằng về tên người, kỹ thuật này được sử dụng

để xây dựng tiểu sử - nền tảng của một số máy tìm kiếm như Zoominfo.com hay Spock.com. Năm 2005, Nadeau và Turney công bố kết quả nghiên cứu nhận dạng từ đầy đủ của các từ viết tắt trong một văn bản đang xét nào đó, ví dụ như "IBM" viết tắt của "International Business Machines" trong nhiều văn bản. Một nghiên cứu vào năm 2006 của Agbago nhằm xây dựng một hệ thống có khả năng phục hồi lại định dạng đúng của từ bao gồm việc bảo đảm cho ký tự đầu câu và đầu thực thể luôn được viết hoa là rất có ích trong dịch máy.

Cũng trong công trình nghiên cứu của mình [5], David Nadeau đã sử dụng tập nhãn thực thể ENAMEX theo mẫu của hội nghị MUC – 7 (Message Understanding Conference 7) và tiến hành huấn luyện - kiểm thử trên tập ngữ liệu Medstract Gold Standard Evaluation Corpus (Tập ngữ liệu này được xây dựng bởi Pustejovsky vào năm 2001). Tác giả sử dụng bộ công cụ Weka Machine Learning để kiểm thử nhiều thuật toán học có giám sát và đưa ra kết luận độ "tốt" của hệ thống phụ thuộc rất nhiều vào thuật toán được sử dụng và phương pháp học bán giám sát của mình cho kết quả khả quan nhất.

Tính đến nay, có khá nhiều hội nghị khoa học quốc tế lớn trao đổi về bài toán nhận dạng thực thể cũng như đánh giá đánh giá các hệ thống nhận dạng thực thể đã được xây dựng. Tiêu biểu có thể kể đến MUC (Message Understanding Conference, 1987-1997), MET (Multilingual Entity Task Conference, 1998), ACE (Automatic Content Extraction Program, 2000), HAREM (Evaluation contest for named entity recognizers in Portuguese, 2004-2006), IREX (Information Retrieval and Extraction Exercise, 1998-1999) …

### *3.2. Đặc điểm dữ liệu tiếng Việt*

Tiếng Việt thuộc ngôn ngữ đơn lập, tức là mỗi một tiếng (âm tiết) được phát âm tách rời nhau và được thể hiện bằng một chữ viết. Đặc điểm này thể hiện rõ rệt ở tất cả các mặt ngữ âm, từ vựng, ngữ pháp. Dưới đây trình bày một số đặc điểm của tiếng Việt theo các tác giả ở Trung tâm ngôn ngữ học Việt Nam đã trình bày. Việc nghiên cứu các đặc điểm dữ liệu tiếng Việt sẽ giúp em có cái nhìn tổng quan về các đặc trưng dữ liệu tiếng Việt. Hiểu rõ ràng hơn về dữ liệu sẽ giúp việc xây dựng Ontology và trích chọn thông tin được hiệu quả hơn.

#### 3.2.1. Đặc điểm ngữ âm

Tiếng Việt có một loại đơn vị đặc biệt gọi là "tiếng" mà về mặt ngữ âm thì mỗi tiếng là một âm tiết. Hệ thống âm vị tiếng Việt phong phú và có tính cân đối,

tạo ra tiềm năng của ngữ âm tiếng Việt trong việc thể hiện các đơn vị có nghĩa. Nhiều từ tượng hình, tượng thanh có giá trị gợi tả đặc sắc. Khi tạo câu, tạo lời, người Việt rất chú ý đến sự hài hoà về ngữ âm, đến nhạc điệu của câu văn.

### 3.2.2. Đặc điểm từ vựng

Nói chung, mỗi tiếng là một yếu tố có nghĩa. Tiếng là đơn vị cơ sở của hệ thống các đơn vị có nghĩa của tiếng Việt. Từ tiếng, người ta tạo ra các đơn vị từ vựng khác để định danh sự vật, hiện tượng..., chủ yếu nhờ phương thức ghép và phương thức láy.

Việc tạo ra các đơn vị từ vựng ở phương thức ghép luôn chịu sự chi phối của quy luật kết hợp ngữ nghĩa, ví dụ: đất nước, máy bay, nhà lầu xe hơi, nhà tan cửa nát... Hiện nay, đây là phương thức chủ yếu để sản sinh ra các đơn vị từ vựng. Theo phương thức này, tiếng Việt triệt để sử dụng các yếu tố cấu tạo từ thuần Việt hay vay mượn từ các ngôn ngữ khác để tạo ra các từ, ngữ mới, ví dụ như tiếp thị, karaoke, thư điện tử (e-mail), thư thoại (voice mail), phiên bản (version), xa lộ thông tin, siêu liên kết văn bản, truy cập ngẫu nhiên, v.v.

Việc tạo ra các đơn vị từ vựng ở phương thức láy thì quy luật phối hợp ngữ âm chi phối chủ yếu việc tạo ra các đơn vị từ vựng, chẳng hạn như chôm chỉa, chỏng chơ, đỏng đa đỏng đảnh, thơ thẩn, lúng lá lúng liếng, v.v.

Vốn từ vựng tối thiểu của tiếng Việt phần lớn là các từ đơn tiết [một âm tiết, một tiếng]. Sự linh hoạt trong sử dụng, việc tạo ra các từ ngữ mới một cách dễ dàng đã tạo điều kiện thuận lợi cho sự phát triển vốn từ, vừa phong phú về số lượng, vừa đa dạng trong hoạt động. Cùng một sự vật, hiện tượng, một hoạt động hay một đặc trưng, có thể có nhiều từ ngữ khác nhau biểu thị. Tiềm năng của vốn từ ngữ tiếng Việt được phát huy cao độ trong các phong cách chức năng ngôn ngữ, đặc biệt là trong phong cách ngôn ngữ nghệ thuật. Hiện nay, do sự phát triển vượt bậc của khoa học-kĩ thuật, đặc biệt là công nghệ thông tin, thì tiềm năng đó còn được phát huy mạnh mẽ hơn.

### 3.2.3. Đặc điểm ngữ pháp

Từ tiếng Việt không biến đổi hình thái. Đặc điểm này sẽ chi phối các đặc điểm ngữ pháp khác. Khi từ kết hợp từ thành các kết cấu như ngữ, câu, tiếng Việt rất coi trọng phương thức trật tự từ và hư từ.

Việc sắp xếp các từ theo một trật tự nhất định là cách chủ yếu để biểu thị các quan hệ cú pháp. Trong tiếng Việt khi nói “Anh ta lại đến” là khác với “Lại đến anh

ta”. Khi các từ cùng loại kết hợp với nhau theo quan hệ chính phụ thì từ đứng trước giữ vai trò chính, từ đứng sau giữ vai trò phụ. Nhờ trật tự kết hợp của từ mà "củ cải" khác với "cải củ", "tình cảm" khác với "cảm tình". Trật tự chủ ngữ đứng trước, vị ngữ đứng sau là trật tự phổ biến của kết cấu câu tiếng Việt.

Phương thức hư từ cũng là phương thức ngữ pháp chủ yếu của tiếng Việt. Nhờ hư từ mà tổ hợp “anh của em” khác với tổ hợp “anh và em”, “anh vì em”. Hư từ cùng với trật tự từ cho phép tiếng Việt tạo ra nhiều câu cùng có nội dung thông báo cơ bản như nhau nhưng khác nhau về sắc thái biểu cảm. Ví dụ, so sánh các câu sau đây:

- Ông ấy không hút thuốc.

- Thuốc, ông ấy không hút.

- Thuốc, ông ấy cũng không hút.

Ngoài trật tự từ và hư từ, tiếng Việt còn sử dụng phương thức ngữ điệu. Ngữ điệu giữ vai trò trong việc biểu hiện quan hệ cú pháp của các yếu tố trong câu, nhờ đó nhằm đưa ra nội dung muốn thông báo. Trên văn bản, ngữ điệu thường được biểu hiện bằng dấu câu. Sự khác nhau trong nội dung thông báo được nhận biệt khi so sánh hai câu sau:

- Đêm hôm qua, cầu gãy.

- Đêm hôm, qua cầu gãy.

Qua một số đặc điểm nổi bật vừa nêu trên đây, chúng ta có thể hình dung được phần nào bản sắc và tiềm năng của tiếng Việt cũng như khó khăn gặp phải trong việc nhận dạng thực thể cũng như trích chọn thông tin trong tiếng Việt.

### *3.3. Một số phương pháp nhận dạng thực thể*

Tồn tại nhiều phương pháp được đề cập tới trong bài toán nhận dạng thực thể. Tuy nhiên có thể tổng kết lại một số giai đoạn chính trong bài toán này như sau:

- Tiền xử lý: Loại bỏ HTML, tách câu, tách từ.
- Lựa chọn thuộc tính: Lựa chọn các nhãn thẻ (tag), mẫu ngữ cảnh (feature: viết hoa, viết thường, …).
- Giai đoạn huấn luyện, tự học: Sử dụng HMM, CRF, MEMM, SVM…
- Gán nhãn, khôi phục.

Tùy thuộc vào từng miền của bài toán nhận dạng thực thể thì sự lựa chọn các nhãn thẻ là khác nhau. Có thể đề cập tới bảy nhãn dạng cơ bản tổng quát nhất được lựa chọn đầu tiên: 7 dạng nhãn đầu tiên (theo Ralph & Beth, [5]): ORG (tổ chức), LOC (vị trí), PER (người), DATE,TIME,CUR (Biểu diễn tiền tệ), PCT (Phần trăm). Tập nhãn có thể được thay đổi, mở rộng tùy thuộc vào từng dự án. Dự án Biocaster [11] xây dựng 22 nhãn cho lĩnh vực y tế.

Mỗi một nhãn được gán bao gồm ba phần:

- Phần biên (boundary category): Xác định vị trí của từ hiện tại trong một thực thể.
- Phần thực thể (Entity category): Xác định kiểu thực thể.
- Tập đặc trưng (Feature set) : Xác định thông tin ngữ cảnh (mẫu ngữ cảnh).

Có nhiều cách để biểu diễn phần biên của các từ, trong đó cách biểu diễn thường được đề cập và dùng nhiều nhất có thể kể tới đó là: biểu diễn mỗi một nhãn gồm một tiếp đầu chữ B_ (bắt đầu một thực thể ), I_ (bên trong một thực thể), nhãn O (không phải thực thể). Lấy ví dụ: bệnh "viêm não nhật bản" có thể được gán nhãn như sau "B_DIS I_DIS I_DIS I_DIS".

Lựa chọn mẫu ngữ cảnh là bài toán quan trọng quyết định độ chính xác của nhận dạng thực thể. Mẫu ngữ cảnh tại vị trí quan sát bất kỳ cho ta thông tin ngữ cảnh. Bất kỳ một hệ thống nhận dạng thực thể hoàn thiện nào đều phải xây dựng được một tập các mẫu ngữ cảnh một cách chính xác và mô tả được từng lĩnh vực của bài toán nhận dạng. Bài toán nhận dạng thực thể chung: viết hoa, viết thường, ký tự % , chữ số, dấu chấm, phẩy…Bài toán tương tự trong y tế, đó là lựa chọn mẫu ngữ cảnh trong nhận dạng protein, gene, thuốc, tế bào .

Các loại mẫu ngữ cảnh [6]:

- Mẫu tiền định cơ bản (viết hoa, thường, chấm, phẩy): comma, dot, oneDigit, AllDigits
- Mẫu hình thái học: tiền tố, hậu tố (~virus, ~lipid, ~vitamin,…),
- Mẫu ngữ pháp: cụm động từ, cụm danh từ …
- Mẫu trigger ngữ nghĩa:

- Trigger danh từ chính: danh từ chính của một tổ hợp từ ( B Cell trong “activated human B cells”, bệnh trong “bệnh viêm xoang” ).
- Trigger động từ đặc biệt: nhiễm, lây, bao gồm, gây ra.

### 3.3.1. Phương pháp dựa trên luật, bán giám sát

Hệ thống dựa trên luật bao gồm một tập các luật cơ bản (Nếu-Thì), tập các sự vật (facts), bộ thông dịch (interpreter) sử dụng tập luật để sinh ra các sự vật. Sử dụng phương pháp dựa trên luật, đầu tiên chúng ta xây dựng một tập ban đầu các luật, các thực thể. Qua quá trình học dựa trên bán giám sát và kỹ thuật bootstrapping, chúng ta mở rộng tập thực thể cũng như tập luật ban đầu.

Học bán giám sát [28] được hiểu là phương pháp học máy sử dụng cả hai loại dữ liệu gán nhãn và chưa gán nhãn cho quá trình huấn luyên. Phương pháp này kết hợp được ưu điểm, giảm bớt những nhược điểm của phương pháp học có giám sát và học không giám sát. Các thuật toán bán giám sát có nhiệm vụ chính là mở rộng một tập dữ liệu huấn luyện nhỏ ban đầu  thành tập dữ liệu lớn hơn.

Một kỹ thuật chính của phương pháp học bán giám sát là bootstrapping. Kỹ thuật này bao gồm có giám sát ở mức độ nhỏ, từ một tập dữ liệu ban đầu (còn gọi là tập seed) bắt đầu quá trình huấn luyện. Ví dụ một hệ thống nhận dạng tên bệnh, lúc đầu yêu cầu một tập mẫu nhỏ các tên bệnh. Sau đó, hệ thống tìm kiếm các câu chứa các tên bệnh này và cố gắng tìm kiếm các thông tin ngữ cảnh chung cho một số tên bệnh trong tập này (ví dụ như có sự tương đồng về thông tin ngữ cảnh trong từng 5 mẫu tên bệnh). Sau đó từ các thông tin ngữ cảnh này, hệ thống sẽ tìm các thể hiện của tên bệnh xuất hiện trong các ngữ cảnh tương tự. Quá trình huấn luyện này sẽ được lặp đi lặp lại để tìm ra các ví dụ mới, cũng như khai thác được các thông tin ngữ cảnh mới có liên quan. Bằng cách lặp đi lặp lại quá trình này, một số lượng lớn các tên bệnh và một số lượng lớn các thông tin ngữ cảnh sẽ được thu thập lại.

### 3.3.2. Các phương pháp máy trạng thái hữu hạn

Các phương pháp máy trạng thái hữu hạn dùng một sơ đồ chung của máy trạng thái hữu hạn (finite state machine - FSM hoặc finite state automaton – FSA). Có thể coi máy trang thái hữu hạn là một máy trừu tượng được dùng trong các nghiên cứu về tính toán và ngôn ngữ với một số lượng hữu hạn, không đổi các trạng thái. Máy trạng thái hữu hạn được biểu diễn như một đồ thị có hướng, trong đó có hữu hạn cá nút (các trạng thái) và từ mỗi nút có không hoặc một số cung (bộ

chuyển) đi tới các nút khác. Một xâu đầu vào mà cần xác định dãy bộ chuyển phù hợp. Tồn tại một số kiểu máy trạng thái hữu hạn. Bộ nhận (Acceptor) cho câu trả lời "có hoặc không" tiếp nhận xâu đầu vào. Bộ đoán nhận (Recognizer) phân lớp đối với xâu đầu vào. Bộ biến đổi (Transducer) sinh ra một xâu kết quả ra tương ứng với xâu đầu vào. Mô hình máy trạng thái hữu hạn được ứng dụng trong trích chọn thông tin thuộc loại bộ biến đổi, trong đó với một xâu văn bản đầu vào, hệ thống đưa ra xâu các đặc trưng tương ứng với các từ khóa trong xâu văn bản đó. Theo một cách phân loại khác, thì có hai loại máy trạng thái hữu hạn là quyết định (Deterministic finite automaton- DFA) và không quyết định (Non-deterministic finite automaton – NFA).

Máy trạng thái hữu hạn bao gồm:

- Một bảng chữ $\Sigma$,
- Một tập các trạng thái S, trong đó
  - với DFA: có một trạng thái xuất phát và có từ không trở lên các trạng thái chấp nhận (dừng).
  - với NFA: có từ một trở lên các trạng thái được coi là trạng thái xuất phát và có từ không trở lên các trạng thái chấp nhận (dừng).
- Một hàm chuyển $T : S \times \Sigma \rightarrow S$.

Hoạt động máy trạng thái được mô tả như sau. Bắt đầu từ (tập) trạng thái xuất phát, lần lượt xem xét từng ký tự trong xâu đầu vào trong bảng chữ $\Sigma$, trên cơ sở hàm chuyển T để di chuyển tới trạng thái tiếp theo cho đến khi mọi ký tự của xâu đã được xem xét. Nếu gặp được trạng thái dừng là thành công. Trong trường hợp đó, xâu các trạng thái được gặp (xuất hiện) trong quá trình xử lý xâu đầu vào được coi là xâu kết quả, hay còn được gọi là xâu nhãn phù hợp với xâu đầu vào.

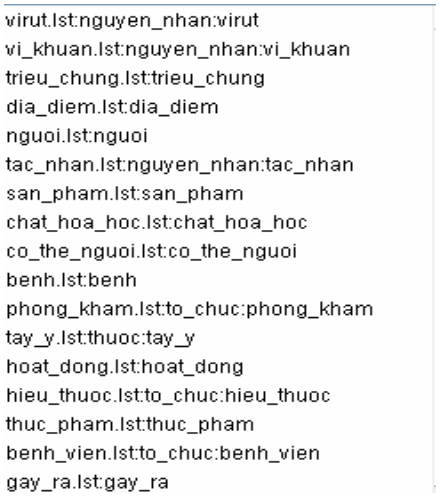
Mô hình máy trạng thái hữu hạn ứng dụng trong trích chọn thông tin được bổ sung thêm một số yếu tố, chủ yếu liên quan tới hàm chuyển T, thường T được mô tả như một quá trình Markov.

### 3.3.3. Phương pháp sử dụng Gazetteer

Từ điển Gazetteer (hay Gazetteer) được hiểu là một danh sách các thực thể như tên người, tổ chức, vị trí; hay riêng đối với lĩnh vực y tế là một danh sách các bệnh, tên thuốc, triệu chứng, nguyên nhân….Nếu có thể xây dựng được một tập dữ liệu gazetteer thật tốt, đầy đủ, chính xác thì sẽ tạo bước tiên quyết quan trọng đối

với hệ thống nhận dạng thực thể. Ngoài việc xây dựng Ontology sẽ đề cập tới công việc xây dựng một tập gazetteer ban đầu cho y tế tiếng Việt. Nhận dạng thực thể dựa trên tập Gazetteer này cho kết quả khả quan.

Các file gazetteer được biểu diễn theo định dạng sau: a.lst:b:c. Trong đó a.lst là file chứa các thể hiện của lớp thực thể a, b là kiểu major, c là kiểu minor. Có thể hiểu một cách đơn giản lớp thuộc kiểu minor là lớp con của lớp thuộc kiểu major. Ví dụ các file gazetteer biểu diễn nguyên nhân gây ra bệnh được biểu diễn như sau: "nguyen_nhan.lst:nguyen_nhan:vikhuan",
"nguyen_nhan.lst:nguyen_nhan:tac_nhan".




**Hình 6: Một số file Gazetteer được xây dựng phục vụ bài toán nhận dạng thực thể.**

Đã có khá nhiều bài báo đề cập tới viêc sử dụng tập dữ liệu để nhận dạng thực thể. Trong bài báo về xây dựng tập dữ liệu cho bài toán nhận dạng thực thể (được trình bày trong phần 3.4.1), nhóm tác giả đã đề cập tới tầm quan trọng của việc xây dựng một tập dữ liệu ban đầu cho quá trình nhận dạng thực thể. Bài báo đã sử dụng BioCaster NE để chú thích dữ liệu và sử dụng Yamcha để học mô hình SVM dựa trên các bài báo đã được chú thích [20].

## *3.4. Nhận dạng thực thể y tế tiếng Việt*

### 3.4.1. Nhận dạng thực thể tiếng Việt

Tồn tại một số công trình nghiên cứu đề cập tới viêc sử dụng tập dữ liệu để nhận dạng thực thể tiếng Việt. Nguyễn Cẩm Tú [1] xây dựng một hệ thống nhận diện thực thể nhận biết loại thực thể dựa trên mô hình trường ngẫu nhiên có điều

kiện (Conditional Random Fields - CRF) để xác định 8 loại thực thể, tương ứng với đó là 17 nhãn. Tác giả tiến hành thực nghiệm sử dụng công cụ FlexCRFs (công cụ mã nguồn mở được phát triển bởi Phan Xuân Hiếu và Nguyễn Lê Minh), sử dụng dữ liệu gồm 50 bài báo lĩnh vực kinh doanh (khoảng gần 1400 câu) lấy từ nguồn http://vnexpress.net.

Thao P.T.X. và cộng sự [21] đã đề cập tới việc khai thác các chiến lược bỏ phiếu (voting) bằng cách tổ hợp các bộ máy huấn luyện sử dụng phương pháp dựa trên từ (word-based). Ý tưởng chính của nhóm tác giả là đề cập tới đó là việc tổ hợp các máy huấn luyện sử dụng các thuật toán phân lớp khác nhau (SVM, CRF, TBL, Naïve Bayes) sẽ cho kết quả cao hơn khi sử dụng riêng rẽ mỗi thuật toán.

Trong [20], Thao P.T.X. và cộng sự đã đề cập tới tầm quan trọng của việc xây dựng một tập dữ liệu ban đầu cho quá trình nhận dạng thực thể. Các tác giả sử dụng BioCaster NE để chú thích dữ liệu và sử dụng Yamcha để học mô hình SVM dựa trên các công trình nghiên cứu liên quan. Nhóm tác giả dò tìm các bệnh truyền nhiễm thông qua các bài trực tuyến về y tế sức khỏe đã đề cập tới việc xây dựng tập dữ liệu cho bài toán nhận dạng thực thể đóng một vai trò rất quan trọng và đã đưa ra 22 nhãn thực thể để gán nhãn và chú thích dữ liệu.

Một nghiên cứu tiêu biểu có liên quan đến bài toán nhận dạng thực thể ở Việt Nam là công cụ VN-KIM IE [40] được xây dựng bởi nhóm nghiên cứu do phó giáo sư tiến sĩ Cao Hoàng Trụ đứng đầu, thuộc trường Đại học Bách khoa Thành phố Hồ Chí Minh. Chức năng của VN-KIM IE là nhận biết và chú thích lớp tự động cho các thực thể có tên trên các trang Web tiếng Việt.

### 3.4.2. Nhận dạng thực thể y tế tiếng Việt

Trên thế giới, một số nhà nghiên cứu (John McNaught[10], Sammy Wang [25], ...) đã lưu ý về một số vấn đề khó khăn trong xử lý dữ liệu y tế. Những khó khăn điển hình nhất là sự nhập nhằng và đa dạng của các từ, thực thể trong dữ liệu y tế có cấu trúc phức tạp, nguyên tắc hình thành đôi khi lại không giống như bình thường; hiện nay vẫn chưa có quy ước rõ ràng về tên các thực thể, vấn đề từ đồng nghĩa – từ trái nghĩa – từ viết tắt và trong nhiều trường hợp từ được sử dụng không mang nghĩa thường gặp của nó; nhiều từ cùng để chỉ một khái niệm và một từ có thể có nhiều nghĩa, ….

Đối với bài toán nhận dạng thực thể cho y tế tiếng Việt, ngoài những khó khăn chung của bài toán nhận dạng thực thể nói trên còn gặp một số trở ngại khác. Các văn bản tiếng Việt không có dữ liệu huấn luyện và các nguồn tài nguyên có thể

tra cứu (như Wordnet trong tiếng Anh), thiếu các thông tin ngữ pháp (POS) và các thông tin về cụm từ như cụm danh từ, cụm động từ cho tiếng Việt, trong khi các thông tin này giữ vai trò quan trọng trong việc nhận dạng thực thể; khoảng cách giữa các từ không rõ ràng, dễ gây nhập nhằng. Hơn nữa, đối với đặc trưng của dữ liệu y tế cũng gây ra không ít khó khăn cho bài toán nhận dạng thực thể: thông tin lưu trữ không hoặc bán cấu trúc (tên thuốc, virus), các kiểu viết tắt tên thực thể, kiểu tên thực thể dài, đa dạng, các cách viết khác nhau của cùng một thực thể. Riêng với thực thể bệnh tiếng Việt, có thể điểm qua một số đặc điểm gây khó khăn cho bài toán nhận dạng thực thể:

- Không tuân theo luật nào về ký tự viết hoa.
- Khó hạn chế số lượng từ vị: Có những tên bệnh chỉ gồm 01 từ (Như bệnh sởi, bệnh chẩn…), nhưng có những tên bệnh lại gồm rất nhiều từ như "chứng rối loạn tâm thần thể hoang tưởng", …
- Cấu trúc các từ tạo thành một thực thể có thể rất phức tạp: rối loạn chức phận não nhẹ ở trẻ em, …
- Có nhiều từ mượn, từ Hán Việt: Stress, bệnh paranoa, bệnh gout, bệnh thiên đầu thống …
- Cùng một bệnh đôi khi có nhiều cách viết không hoàn toàn giống nhau hay thậm chí khác hẳn nhau: thủy đậu hay trái rạ, bệnh gút hay gout hay còn gọi là thống phong, bệnh ung thư máu còn được gọi là bệnh máu trắng…
- Có nhiều từ viết tắt: AIDS (là viết tắt từ Acquired Immunodeficiency Syndrome hay từ Acquired Immune Deficiency Syndrome của tiếng Anh) trong nhiều tài liệu y tế tiếng Việt được dịch là "hội chứng suy giảm miễn dịch mắc phải", …
- Chứa những từ rất dễ bị "bỏ sót" vì cụm từ dù có hay không có các từ này vẫn có thể được tính là một thực thể, như mãn tính, cấp tính, nguyên phát, thứ phát

Bài toán nhận dạng thực thể đặc trưng cho dữ liệu sinh học và y tế cũng là một nội dung nghiên cứu rất được quan tâm. Các thực thể đặc trưng của dữ liệu sinh học – y tế thường được quan tâm đến nhiều nhất là: Bệnh, Thuốc, Gen, Sinh vật, Protein, Enzime, Các khối u ác tính (Malignancies), Fibrinogen [10] [23]…

Một trong những phương pháp đơn giản nhất được đề xuất cho bài toán nhận dạng thực thể trong dữ liệu y tế là sử dụng các từ điển hoặc tập từ vựng được định nghĩa trước. Đơn cử là sử dụng MeSH [23]. Đây là một bảng từ vựng y tế có kiểm

soát sử dụng để đánh chỉ mục. Thực chất nó là một danh sách các từ đã được xác nhận dùng để đánh chỉ mục và chỉ có các từ trong danh sách này được chấp nhận ở vai trò đó. Các từ trong MeSH được sắp xếp theo hệ thống có cấu trúc cây. Có tất cả 16 nhánh của cây MeSH, đây là những nhóm từ lớn nhất và đặc trưng nhất trong dữ liệu y tế, có thể kể đến nhánh A- Anatomy (giải phẫu học), nhánh B – Organisms (sinh vật), nhánh C – Dieases (bệnh), nhánh D – Chemicals and Drugs (hóa học và thuốc), nhánh G - Biological Sciences (sinh vật học) … Các nhánh lại chia làm các nhánh nhỏ, ví dụ nhánh A01 - Body Regions (bộ phận cơ thể), A02 – Sense Organs (các giác quan) …

Trong chuỗi hội nghị quốc tế BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology]: được tổ chức dưới dạng một cuộc thi, BioCreAtIvE I (2003-2004) tập trung vào chủ đề nhận dạng tên thực thể Gene và Protein, có thể điểm qua một vài kết quả tiêu biểu dưới đây [32]:

- Alexander Yeh và cộng sự sử dụng dữ liệu và phần mềm ước lượngcủa W. John Wilbur and Lorraine Tanabe cho kết quả F-measure khoảng 80-83%.

- Shuhei Kinoshita và cộng sự giải quyết vấn đề bằng cách coi bài toán nhận dạng thực thể như một dạng của bài toán gãn nhãn từ loại, thêm một nhãn GENE vào tập nhãn thông thường, các tác giả sử dụng phương pháp gán nhãn từ loại của Brill, sử dụng công cụ TnT – một công cụ dựa trên mô hình HMM, hệ thống không qua hậu xử lý cho kết quả độ chính xác là 68.0%, độ hồi tưởng là 77.2% và F-measure là 72.3%., nếu thêm một bước hậu xử lý (bằng một số luật để bắt lỗi) đạt độ chính xác là 80.3%, độ hồi tưởng 80.5% và F-measure là 80.4%; nếu sử dụng thêm một bước hậu xử lý dựa trên từ điển thì đạt được F-measure là 80.9%.

- Năm 2004, Yi-Feng Lin, Tzong-Han Tsai, Wen-Chi Chou, Kuen-Pin Wu, Ting-Yi Sung and Wen-Lian Hsu công bố nghiên cứu về áp dụng mô hình Markov cực đại hóa Entropy cho bài toán nhận dạng thực thể trong dữ liệu y tế. Kết quả được cho bởi độ chính xác P, độ hồi tưởng R và F-measure (2PR/(P+R)) là (0.512, 0.538, 0.525), sau khi hậu xử lý thì đạt được kết quả tương ứng là (0.729, 0.711, 0.72).

Năm 2004, Haochang Wang và cộng sự [7] đề xuất phương pháp nhận dạng thực thể cho dữ liệu y tế dựa trên bộ phân lớp kết hợp các phương pháp Generalized Winnow, Conditional Random Fields, Support Vector Machine và Maximum Entropy, các phương pháp này được phối hợp theo ba chiến lược khác

nhau. Hệ thống mà các tác giả xây dựng đạt được kết quả độ đo F khoảng 77.57%, là một kết quả khá tốt so với các nghiên cứu cùng thời điểm.

Năm 2007, Andreas Vlachos [3] so sánh hai phương pháp nhận dạng thực thể trong dữ liệu y tế dựa trên mô hình HMM và dựa trên mô hình CRF cùng với phân tích cú pháp. Hai bảng dưới đây chỉ ra kết quả thực nghiệm, bảng bên trái là kết quả thực nghiệm khi huấn luyện bằng một tập nhỏ dữ liệu đã được chú thích thực thể thủ công và kiểm thử trên toàn bộ tập huấn luyện, bảng bên phải là kết quả khi huấn luyện bằng một tập nhỏ dữ liệu nhiễu và kiểm thử trên toàn bộ tập huấn luyện

Gần đây nhất, vào tháng 3 năm 2009, Razvan C. Bunescu [45] khi trình bày về trích chọn quan hệ từ tập dữ liệu y tế đã lưu ý vấn đề nhận dạng thực thể đặc trưng trong dữ liệu y tế, các thực thể được quan tâm đến gồm có Bệnh, Gen và Protein. Sau khi đã nhận dạng được các thực thể này, tác giả tiến thêm một bước quan trọng là trích chọn ra quan hệ tương tác giữa chúng (ví dụ như Gen mã hóa một Protein, Protein hoàn thành chức năng của nó bằng cách tương tác với một Protein khác …).

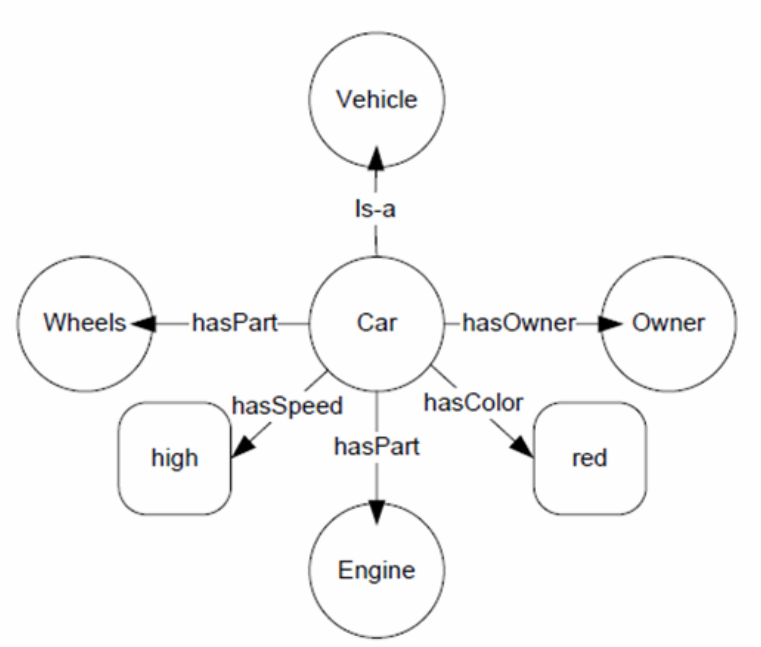
# Chương 4

# XÁC ĐỊNH QUAN HỆ NGỮ NGHĨA

## *4.1. Tổng quan về xác định quan hệ ngữ nghĩa*

### 4.1.1. Khái quát về quan hệ ngữ nghĩa

Như đã trình bày ở trên, sau khi có một tập lớp thực thể (qua bước nhận dạng thực thể) để có được một mạng ngữ nghĩa các thực thể, chúng ta cần thực hiện bước tiếp theo là bước trích chọn quan hệ ngữ nghĩa (semantic relation). Quan hệ ngữ nghĩa có thể được hiểu là mối quan hệ tiềm ẩn giữa hai khái niệm được biểu diễn bằng từ hoặc cụm từ [24]. Các mối quan hệ ngữ nghĩa đóng một vai trò quan trọng trong việc phân tích ngữ nghĩa từ vựng. Từ đó nó có thể ứng dụng vào nhiều bài toán khác: Xây dựng nền tảng tri thức ngữ nghĩa từ vựng, hệ thống hỏi đáp, tóm tắt văn bản,… Một số mối quan hệ ngữ nghĩa điển hình trong lĩnh vực y tế là IS_A (Cúm -- bệnh), PART_WHOLE (Virus – Nguyên nhân), CAUSE_EFFECT (virus – bệnh).



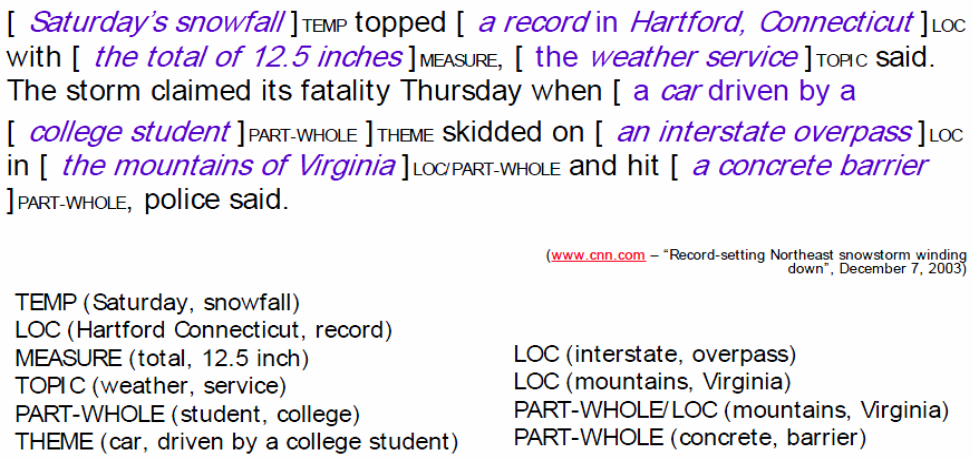

**Hình 7: Minh họa một quan hệ ngữ nghĩa cho thực thể car**

Tuy quan hệ ngữ nghĩa đóng một vai trò quan trọng trong phân tích ngữ nghĩa nhưng chúng thường tồn tại ở dạng ẩn gây khó khăn cho việc trích chọn các quan hệ này. Một câu hỏi đặt ra là làm thế nào chúng ta có thể khai thác được các

quan hệ ngữ nghĩa này một cách có hiệu quả từ tập dữ liệu thô (không hoặc bán cấu trúc). Trả lời cho câu hỏi này chính là mục tiêu chính của bài toán trích chọn quan hệ được đề cập nhiều trong thời gian gần đây.

### 4.1.2. Trích chọn quan hệ ngữ nghĩa

Mục đích của trích chọn quan hệ ngữ nghĩa là trích rút ra những quan hệ chuyên biệt, cụ thể nào đó giữa các thực thể trong nguồn ngữ liệu văn bản lớn. Thực chất nhiệm vụ của trich chọn quan hệ ngữ nghĩa là khi được cho một cặp thực thể x-y, phải xác định được ý nghĩa của cặp thực thể đó [24]. Lấy ví dụ từ câu "mất ngủ do căng thẳng, hồi hộp" chúng ta có thể suy ra quan hệ ngữ nghĩa: căng thẳng, hồi hộp là nguyên nhân của bệnh mất ngủ.



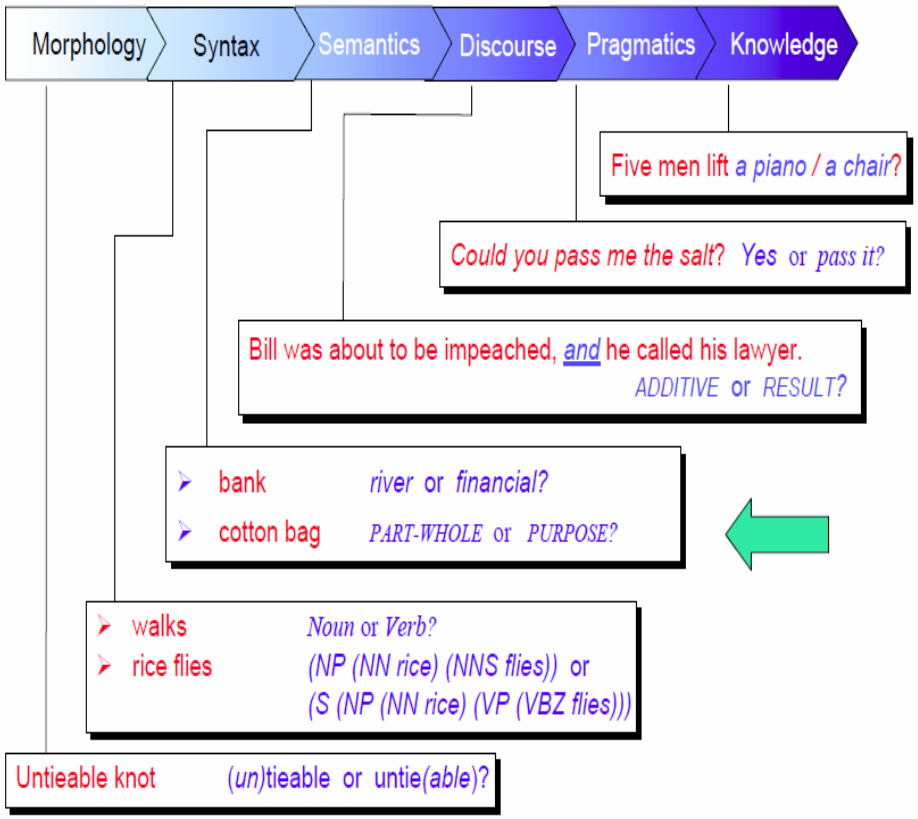

**Hình 8. Minh họa về trích chọn quan hệ ngữ nghĩa**

Các tài nguyên trich chọn quan hệ ngữ nghĩa bao gồm:

- Các tập dữ liệu: Dựa trên sự xuất hiện đồng thời và các phương pháp thống kê.
- Các tài nguyên sẵn có về các quan hệ ngữ nghĩa như WordNet và các bộ chuẩn mực.
- Sự đánh giá của con người.

Cũng như nhận dạng thực thể, nhận dạng quan hệ ngữ nghĩa cũng có một số khó khăn riêng như sau (1) chưa có được sự thống nhất về vấn đề số lượng các quan hệ ngữ nghĩa, các quan hệ ngữ nghĩa được ẩn giấu dưới các dạng khác nhau; (2) các sự kết hợp (danh từ - danh từ) không hoàn toàn tuân theo các quy tắc ràng buộc nhất định, các quan hệ ngữ nghĩa thường là ẩn, có thể có nhiều mối quan hệ giữa các cặp khái niệm, việc thông dịch có thể phụ thuộc nhiều vào ngữ cảnh, không có một tập đã được định nghĩa tốt về các quan hệ ngữ nghĩa.

Việc trích chọn quan hệ ngữ nghĩa là một phần của các dự án quan trọng mang tầm cỡ quốc tế trong lĩnh vực khai phá tri thức [24]. Ví dụ như ACE (Automatic Content Extraction). DARPA EELD (Evidence Extraction and Link Discovery), ARDA-AQUAINT (Question Answering for Intelligence), ARDA NIMD (Novel Intelligence from Massive Data), Global WordNet.



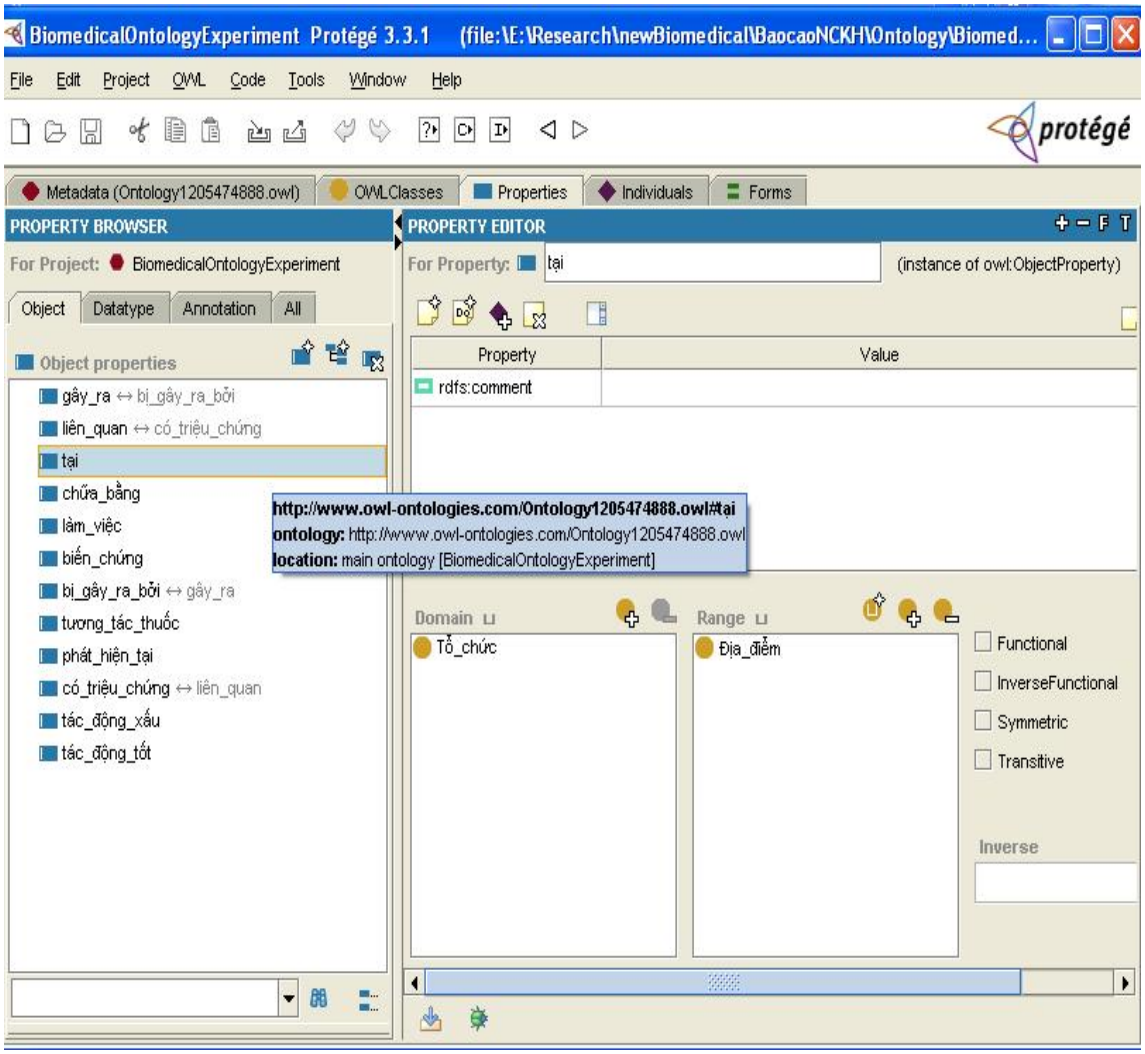

**Hình 9. Vị trí của khai phá quan hệ ngữ nghĩa trong xử lý ngôn ngữ tự nhiên**

Tùy thuộc vào từng miền, lĩnh vực mà chúng ta có các quan hệ ngữ nghĩa khác nhau. Bảng trong Hình 10 minh họa một số quan hệ ngữ nghĩa trong WordNet

| *Semantic relation* | *Description* | *Part-of-speech* | | | | *Example* |
|---|---|---|---|---|---|---|
| | | N | V | Adj | Adv | |
| *Synonym* | A concept that means exactly or nearly the same as another. *WordNet* considers immediate hypernyms to be synonyms. | × | × | × | × | *{ sofa, couch, lounge }* are all synonyms of one another. *{ seat }* is the immediate hypernym of the synset. |
| *Antonym* | A concept opposite in meaning to another. | × | × | × | × | *{ love }* is the antonym of *{ hate, detest }*. |
| *Hypernym* | A concept whose meaning denotes a superordinate. | × | × | | | A *{ feline, felid }* is a hypernym of *{ cat, true cat }*. |
| *Hyponym* | A concept whose meaning denotes a subordinate. | × | × | | | A *{ wildcat }* is a hyponym of *{ cat, true cat }*. |
| *Substance meronym* | A concept that is a substance of another concept. | × | | | | A *{ snowflake, flake }* is substance of *{ snow }*. |
| *Part meronym* | A concept that is a part of another concept. | × | | | | A *{ crystal, watch crystal, watch glass }* is a part of a *{ watch, ticker }*. |
| *Member meronym* | A concept that is a member of another concept. | × | | | | An *{ associate }* is a member of an *{ association }*. |
| *Substance of holonym* | A concept that has another concept as a substance. | × | | | | A *{ tear, teardrop }* has *{ water, H20 }* as a substance. |
| *Part of holonym* | A concept that has another concept as a part. | × | | | | A *{ school system }* has a *{ school, schoolhouse }* as a part. |

**Hình 10. Minh họa các quan hệ ngữ nghĩa được chỉ ra trong WordNet [37]**

Đối với miền dữ liệu y tế, qua khảo sát, chúng tôi thu thập được 12 loại quan hệ ngữ nghĩa, các quan hệ này sẽ được mô tả chi tiết trong Chương 5.




**Hình 11. Một số quan hệ ngữ nghĩa đã xây dựng được**

Hình 11 mô tả một số quan hệ ngữ nghĩa, ý nghĩa các quan hệ ngữ nghĩa này được mô tả trong bảng Bảng 1.

| Quan hệ | Ý nghĩa | Quan hệ đảo ngược |
|---|---|---|
| Gây_ra | Mô tả quan hệ nguyên_nhân gây ra bệnh | Bị_gây_ra_bởi |
| Có_triệu_chứng | Quan hệ bệnh có các triệu chứng | Liên_quan |
| Tại | Tổ_chức được đặt tại Địa_điểm | |
| Chữa_bằng | Bệnh được chữa bằng thuốc | Chữa |
| Làm_việc | Người làm việc ở tổ_chức | |
| Biến_chứng | Bệnh biến chứng sang bệnh khác | |
| Tương_tác_thuốc | Thuốc tương tác với thuốc | |
| Phát_hiện_tại | Bệnh được phát hiện tại Tổ_chức | |
| Tác_động_tốt | Thực_phẩm,Hoạt_động, Chất_hóa_học tác động tốt đến cơ_thể_người, bệnh | |
| Tác động xấu | Thực_phẩm, Hoạt_động, Chất_hóa_học tác động xấu đến cơ_thể_người, bệnh | |

**Bảng 1. Giải thích các mối quan hệ ngữ nghĩa**

### 4.1.3. Một số nghiên cứu liên quan đến xác định quan hệ ngữ nghĩa

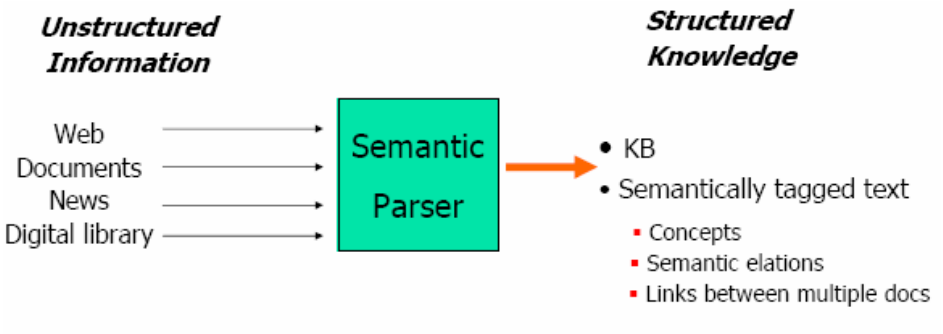
Tại Hội thảo SemEval 2007 [38], nhận dạng các mối quan hệ ngữ nghĩa giữa hai danh từ là một nội dung chính được đề cập.Ý nghĩa của 2 thực thể liên quan đến ý nghĩa của các từ khác trong ngữ cảnh, nhận dạng theo 1 kiểu quan hệ nào đó. Ví dụ: đi xe đạp và sự vui vẻ (quan hệ nhân quả)… Trích chọn quan hệ ngữ nghĩa dựa trên 7 mối quan hệ cơ bản là Cause- Effect, Instrument-Agency, Product-Producer,Origin-Entity, Theme-Tool, Part-Whole, and Content-Container.

Ngoài ra, có thể kể thêm một số phương pháp trích chọn quan hệ giữa hai khái niệm được mô tả như sau: **thuốc** là 1 cách điều trị của 1 **bệnh**, hay 1 **gene** là 1 nguyên nhân của 1 **bệnh**. Swanson [29] giới thiệu một mô hình để trích chọn các kiểu quan hệ trên trong cơ sở dữ liệu y sinh học từ đó mở ra một khái niệm thứ 3 (ví dụ 1 chức năng sinh lý) liên quan đến cả hai khái niệm **thuốc** và **bệnh**. Việc trích chọn loại khái niệm thứ 3 này cho phép một mối quan hệ giữa hai khái niệm chính (chứa tiềm ẩn trong một tài liệu nào đó) được hiển thị ra. Mô tả phương pháp trên một cách cụ thể hơn: X liên quan đến bệnh nào đó, Z liên quan đến thuốc, Y là một chức năng bệnh lý, sinh lý, triệu chứng…, X và Y, Y và Z thường được đề cập

cùng nhau, X và Z thì lại k cùng xuất hiện trong 1 tài liệu nghiên cứu. Từ đó ta có thể sử dụng khái niệm Y để vẽ 1 mối liên quan giữa hai khái niệm X và Z.

Đối với việc sử dụng Ontology, đã có nhiều nhóm tác giả đề cập tới việc học bán giám sát sử dụng Ontology như một hướng tiếp cận mới. Trong hướng tiếp cận đó, input là một tập các văn bản text (tên thực thể, tương ứng đối với các khái niệm trong ontology mà mới được xác định). Sử dụng các tập dữ liệu có sẵn như GENIA corpus [14], việc gán nhãn được thực hiện thủ công nhưng dữ liệu corpus có thể được tự động tạo ra sử dụng một hệ thống NER tương ứng. Output: Tập các mẫu bao gồm các cặp lớp và mối quan hệ trong ontology GENIA, (ví dụ template : virus infect cell).

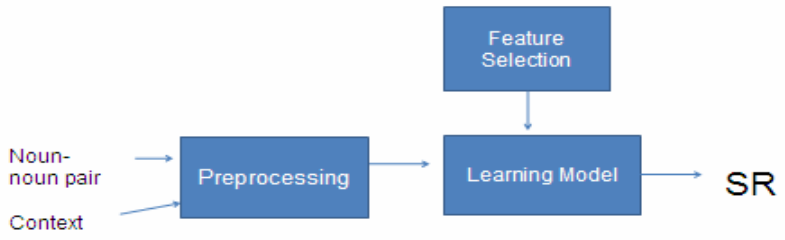
Có nhiều phương pháp được đưa ra để xác định quan hệ. Tuy nhiên nhiệm vụ chung của bài toán này đó là từ các văn bản thô như các trang Web, tài liệu, tin tức, …; qua bộ phân tích ngữ nghĩa (Semantic Parser) chúng ta có đầu ra là các cơ sở tri thức (Knowledge Base – KB), và các khái niệm, các mối quan hệ cũng như các liên kết giữa các văn bản [24]. Hình 12 mô tả nhiệm vụ chung của bài toán xác định thực thể.




**Hình 12. Nhiệm vụ chung của bài toán xác định quan hệ**

Bài toán xác định quan hệ cũng có thể hiểu là từ một cặp danh từ (thực thể) xác định được ý nghĩa của cặp danh từ đó [24]. Ý nghĩa đó được diễn đạt thông qua một danh sách các quan hệ, các cặp thực thể đã được nhận dạng và một số tài nguyên khác.

Đối với bộ phân tích ngữ nghĩa, như đã trình bày ở phần trên, đóng vai trò quan trọng trong việc trích rút các quan hệ ngữ nghĩa. Bộ phân tích ngữ nghĩa này bao gồm các thành phần được mô tả như trong Hình 13:




**Hình 13. Mô tả các bộ phận trong bộ phân tích ngữ nghĩa SR [24]**

- Preprocessing: Tokenizer, Part-of-speech tagger, Syntactic parser, Word sense disambiguation, Named entity recognition.

- Feature Selection: Xác định các tính chất, ràng buộc (hoặc ngữ cảnh) , sử dụng bộ phân lớp để phân biệt các mối quan hệ ngữ nghĩa.

- Learning Model: Phân loại các thể hiện (instance) input thành các mối quan hệ phù hợp

Bộ phân tích ngữ nghĩa (SR: Semantic Parsers) thực hiện hai nhiệm vụ chính:

- Labeling: Từ các mối quan hệ ngữ nghĩa được định nghĩa trước và cặp thực thể (danh từ - danh từ) ta gán nhãn mối quan hệ giữa hai thực thể đó. Ví dụ, Bánh xe ô tô – ô tô <Part_Whole>.

- Paraphrasing: Từ một cặp danh từ hay thực thể đưa ra được ý diến đạt của trong văn cảnh của danh từ đó. Ví dụ bệnh mất ngủ do căng thẳng, từ đó chúng ta có thể suy ra quan hệ **căng thẳng** là nguyên nhân của **mất ngủ**.

### *4.2. Gán nhãn ngữ nghĩa cho câu*

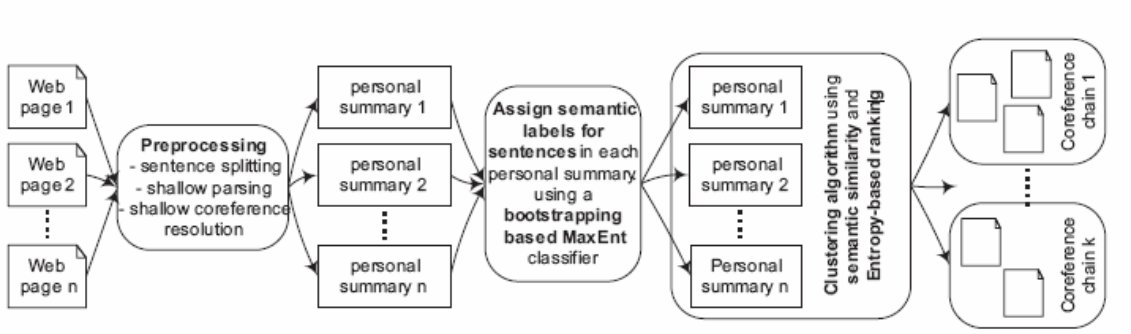
Trong [30], Xuan-Hieu Phan và cộng sự đã đề cập tới giải pháp ”khử nhập nhằng thực thể đa tài liệu” bằng cách gán nhãn ngữ nghĩa cho các câu trong văn bản. Khử nhập nhằng thực thể đa tài liệu là phân biệt các thực thể trùng thể hiện trong một tập tài liệu cho trước. Ví dụ, cho một tập các thực thể có cùng thể hiện là “Bill Clinton, ta phải xác định được tập con tài liệu thực sự nói về “Bill Clinton” – cựu tổng thống Mỹ, tập con tài liệu nào nói về “Bill Clinton” – cầu thủ golf hay tập nào nói về một “Bill Clinton” nào đó khác.

Gán nhãn ngữ nghĩa có thể được xem như là bài toán phân lớp các câu chứa quan hệ ngữ nghĩa. Bài báo đã sử dụng bộ phân lớp dựa trên Maxent lấy các câu từ tóm tắt cá nhân là các câu đầu vào và đầu ra với các nhãn ngữ nghĩa. Bộ phân lớp

dựa trên Maxent có ưu điểm là liên kết chặt chẽ giữa một số lượng rất lớn (lên tới hàng trăm nghìn hoặc triệu) của các đặc trưng chồng chéo, độc lập tại các mức độ khác nhau.

Các tác giả [30] cũng đề xuất một Framework cho việc khử nhập nhằng thực thể đa tài liệu gồm ba phần chính, và một phần không thể thiếu đó là gán nhãn ngữ nghĩa cho câu trong văn bản:

- Tiền xử lý: Sử dụng xử lý nông để một thu thập một tóm tắt bao gồm các câu liên quan tới thực thể được đề cập.
- Chỉ định các nhãn ngữ nghĩa đối với câu trong tóm tắt để đặt chúng vào các lớp khác nhau của sự vật. Sự chỉ định này được thực hiện bởi bộ phân lớp dựa trên Maxent có độ chính xác cao, trong đó dữ liệu được huấn luyện dựa trên phương pháp học bán giám sát.
- Sử dụng phương pháp phân cụm, độ tương đồng giữa các tóm tắt cá nhân của mỗi câu có cùng các nhãn ngữ nghĩa sẽ được đặt bằng nhau để tính toán độ gần ngữ nghĩa.




**Hình 14. Minh họa Framework giải quyết bài toán xác định tên riêng giữa các tài liệu.**

Hình vẽ 14 cho thấy gán nhãn ngữ nghĩa cho câu đóng một vai trò quan trọng trong bài toán xác định tên riêng giữa các tài liệu cũng như là cơ sở cho xác định quan hệ ngữ nghĩa.

Một số nhãn ngữ nghĩa cho câu được minh họa như trong Hình 15 sau đây

| Label | Notes |
|---|---|
| BirthInfo | birthdate or birthplace |
| Nationality | homeland |
| Parent | i.e., mother and father |
| Children | i.e., sons and daughters |
| Partner | e.g., darlings, husbands or wives |
| Education | e.g., school, university, degree, or certificate |
| WorkFor | work for company/organization |
| Position | e.g., leader, president, worker, etc. |
| Achivmt | i.e., a special prize, award, or achievement |
| Religion | e.g., Christianity, Judaism, etc. |
| Hobby | e.g., a kind of sports, music, etc. |
| OtherFact | i.e., other relevant relationships or events |

**Hình 15. Một số nhãn ngữ nghĩa được gán cho câu [30]**

Với các nhãn này, tóm tắt cá nhân của Bill Clinton sẽ được gán nhãn như Hình 16 dưới đây.

```
- [BirthInfo] W.J. Clinton was born on Aug. 19, 1946.
- [Religion] Religion: Baptist
- [OtherFact] As a delegate to Boys Nation while in
  high school, he met President John F. Kennedy ...
- [Education] After graduation, he attended Oxford
  University, and received a law degree at Yale ...
- [Partner] At Yale, Bill Clinton met Hillary Rodham,
  and they married in 1975.
- [Children] They have one daughter Chelsea, ...
- [WorkFor] Clinton taught law at the University ...
- [Position] He served two terms as the 42nd ...
- [OtherFact] As a result of the Monica Lewinsky
  scandal, he was acquitted by the Senate.
- ...
```

**Hình 16. Gán nhãn ngữ nghĩa cho các câu mô tả tổng thống Bill Clinton [30].**

Khóa luận đã gán nhãn thử nghiệm cho 1000 câu với các nhãn chứa quan hệ liên quan đến lĩnh vực y tế. Các nhãn và dữ liệu được gán nhãn sẽ được trình bày chih tiết trong Chương 5.

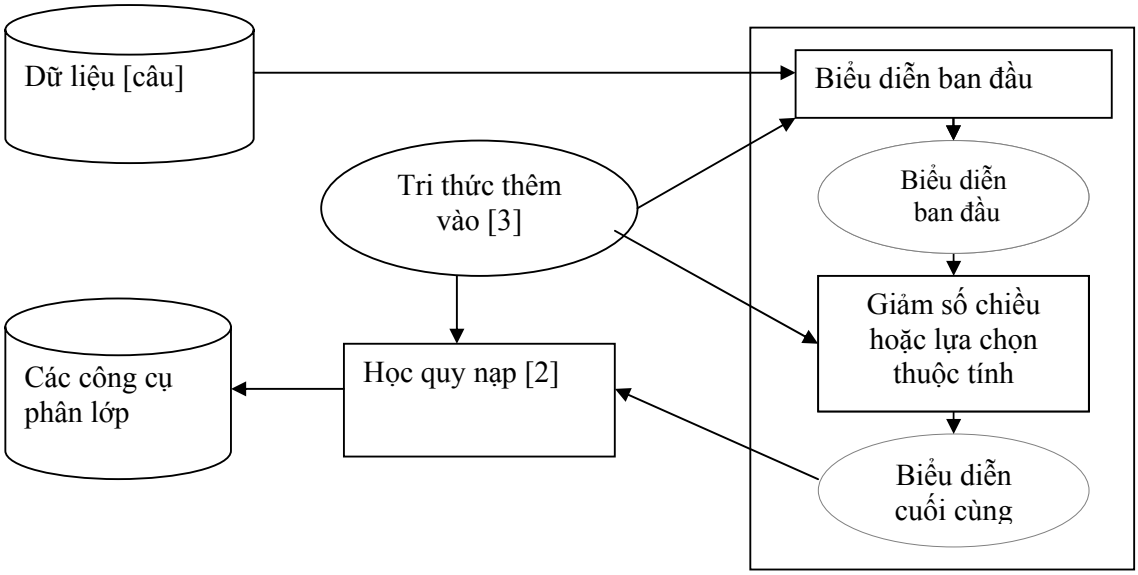
## *4.3. Phân lớp câu chứa quan hệ*

### 4.3.1. Phân lớp với xác định quan hệ, nhận dạng thực thể

Thực thể cần nhận dạng cũng như các mối quan hệ cần xác định tùy thuộc vào từng bài toán, từng miền ứng dụng (domain). Ví dụ tên thực thể có thể là tên người, tên tổ chức, địa danh, …(bài toán nhận dạng thực thể thông thường). Trong miền ứng dụng mà khóa luận thực hiện, tên thực thể có thể là tên bệnh, thuốc, triệu chứng, nguyên nhân, … Tuy nhiên đối với một số tên thực thể hay quan hệ, ví dụ tên bệnh, triệu chứng, nguyên nhân, quan hệ có_triệu_chứng và quan hệ có_biến_chứng thì việc nhận dạng và phân biệt chúng cũng là một bài toán phức

tạp. Có nhiều khi tên bệnh trùng với triệu chứng, nguyên nhân, ví dụ như : đau đầu, ho …có thể hiểu là bệnh, cũng có thể hiểu là nguyên nhân hay triệu chứng trong một số trường hợp ngữ cảnh khác nhau. Gắn liền nhận dạng thực thể, xác định quan hệ với vấn đề phân lớp. Các thực thể sau khi được nhận dạng ra cần được phân vào các lớp đúng. Hơn nữa, như đã trình bày ở phần trước về gán nhãn ngữ nghĩa cho câu bản chất cũng chính là dựa trên thuật toán phân lớp. Từ những lý do đó mà khóa luận đề cập tới bài toán phân lớp và các thuật toán phân lớp đã được nghiên cứu trong thời gian qua.

Hình 17 mô tả các giai đoạn trong quá trình phân lớp. Mô hình này bao gồm ba công đoạn chính: công đoạn đầu là biểu diễn dữ liệu, tức là chuyển các dữ liệu (các câu) thành một dạng có cấu trúc nào đó, tập hợp các mẫu cho trước thành một tập huấn luyện. Công đoạn thứ hai là việc sử dụng các kỹ thuật học máy để học trên các mẫu huấn luyện vừa biểu diễn. Như vậy là việc biểu diễn ở công đoạn một sẽ là đầu vào cho công đoạn thứ hai. Công đoạn thứ ba là việc bổ sung các kiến thức thêm vào do người dùng cung cấp để làm tăng độ chính xác trong biểu diễn văn bản hay trong quá trình học máy.



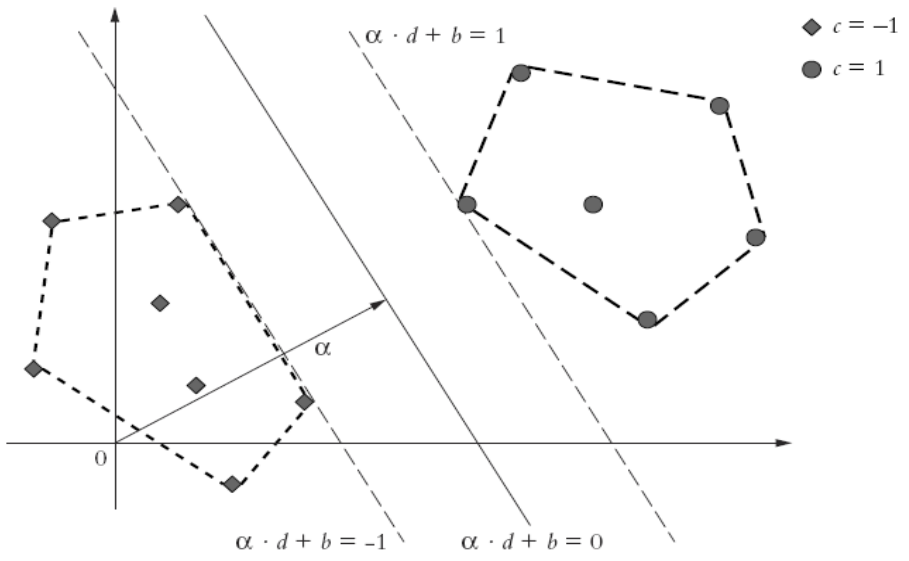

**Hình 17. Mô tả các giai đoạn trong quá trình phân lớp**

Trong nhiều năm gần đây đã có nhiều thuật toán được đưa ra để giải quyết bài toán phân lớp, ví dụ : SVM (Support Vector Machine), K – láng giềng gần nhất, phân lớp dựa vào cây quyết định, …Các thuật toán này đã được Nguyễn Minh Tuấn [2] mô tả khá chi tiết. Chúng tôi sử dụng phương pháp SVM để phân loại câu chứa quan hệ, trong các phần tiếp theo sẽ trình bày kỹ hơn về thuật toán này.

### 4.3.2. Thuật toán SVM (Support Vector Machine)

Thuật toán máy vector hỗ trợ (Support Vector Machine – SVM) được Corters và Vapnik giới thiệu vào năm 1995. SVM rất hiệu quả để giải quyết các bài toán với dữ liệu có số chiều lớn (như các vector biểu diễn văn bản).

Thuật toán SVM được thực hiện trên một tập dữ liệu học D= {(Xi,Ci), i=1,…n}.Trong đó Ci Є {-1,1} xác định dữ liệu dương hay âm. Mục đích của thuật toán là tìm một siêu phẳng $\alpha_{svm}.d + b$ phân chia dữ liệu thành hai miền. Phân lớp một tài liệu mới chính là xác định dấu của $f[d] = \alpha_{svm}.d + b$. Tài liệu sẽ thuộc lớp dương nếu $f(d) > 0$, thuộc lớp âm nếu $f(d) < 0$.




**Hình 18: Mô tả sự phân chia tài liệu theo dấu của hàm $f(d) = \alpha_{svm}.d + b$**

### 4.3.3 Phân lớp đa lớp với SVM

Bài toán phân lớp quan hệ yêu cầu một bộ phân lớp đa lớp do đó cần cải tiến SVM cơ bản (phân lớp nhị phân) thành bộ phân lớp đa lớp.

Một trong những phương pháp cải tiến đó là sử dụng thuật toán "one-against-all"[12]. Ý tưởng cơ bản như sau:

- Giả sử tập dữ liệu mẫu (x1,y1), … ,(xm,ym) với xi là một vector n chiều. và yi ∈ Y là nhãn lớp được gán cho vector xi .
- Chia tập Y thành m tập lớp con có cấu trúc như sau zi ={yi ,Y\$y_i$ } .
- Áp dụng SVM phân lớp nhị phân cơ bản với m tập Zi để xây dựng siêu phẳng cho phân lớp này.

Bộ phân lớp với sự kết hợp của m bộ phân lớp trên được gọi là bộ phân lớp đa lớp mở rộng với SVM.

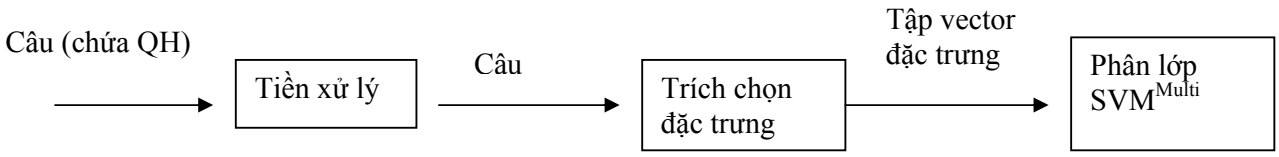
### 4.3.4. Áp dụng SVM vào phân loại quan hệ ngữ nghĩa trong lĩnh vực y tế tiếng Việt

Tuy mục tiêu ban đầu của SVM là dùng cho phân lớp nhị phân, nhưng hiện nay đã được cải tiến cho phân lớp đa lớp, có thể sử dụng cải tiến này để phân lớp các câu chứa quan hệ [2].

Hai quá trình chuẩn bị dữ liệu khi xây dựng được mô hình phân lớp quan hệ dựa trên SVM như sau:

- Thiết kế mô hình cây phân cấp (taxonomy) cho tập lớp quan hệ. Miền ứng dụng của quan hệ sẽ quyết định độ phức tạp (phân cấp) của taxonomy.
- Xây dựng tập dữ liệu mẫu (corpus) đã được gán nhãn cho từng lớp quan hệ. Trong bước này, cách lựa chọn đặc trưng để biểu diễn quan hệ có vai trò quan trọng. Phụ thuộc vào đặc điểm của từng ngôn ngữ mà tập các đặc trưng được lựa chọn khác nhau. Ví dụ với tiếng Anh thì tập đặc trưng của nó là các từ.

Sau khi xây dựng được tập các lớp câu hỏi cùng với tập dữ liệu sẽ tiến hành "học": Mô hình học như sau:




**Hình 19. Mô tả quá trình học của phân lớp câu chứa quan hệ [2]**

# Chương 5

# THỰC NGHIỆM

Việc xây dựng Ontology cho y tế tiếng Việt đồng thời mở rộng nó một cách tự động thông qua các bước của bài toán trích chọn thông tin: nhận dạng thực thể, xác định quan hệ…. sẽ làm tiền đề để khóa luận xây dựng một tập dữ liệu mang ngữ nghĩa (mạng ngữ nghĩa). Kết quả của công việc này đóng vai trò quan trọng trong nhiệm vụ xây dựng một máy tìm kiếm ngữ nghĩa trong tương lai.

## *5.1. Môi trường thực nghiệm*

### 5.1.1. Phần cứng

Chúng tôi sử dụng máy tính cá nhân với cấu hình phần cứng là Genuine Intel CPU T2050 1.60 GHz, CHIP 798 MHz, RAM 1Gb.

### 5.1.2 Phần mềm

Chúng tôi tích hợp các tiện ích trong các bộ công cụ Protégé, Gate để xây dựng ontology, chú thích dữ liệu và nhận dạng thực thể tiếng Việt đối với lĩnh vực y tế.

Protégé [13] là một công cụ xây dựng Ontology được xây dựng và phát triển tại Stanford Center for Biomedical Informatics Research của trường đại học Stanford University School of Medicine. Protégé có hai loại: Protégé Frame và Protégé OWL. Protégé Frame cung cấp một giao diện dùng đầy đủ và mô hình có sẵn để tạo, lưu trữ Ontology dưới dạng Frame. Còn Protégé OWL hỗ trợ về ngôn ngữ Web ontology, được chứng thực dựa vào web ngữ nghĩa hay W3C.

Gate [31] là một kiến trúc phần mềm để phát triển và triển khai các bộ phận phần mềm phục vụ công việc xử lý ngôn ngữ của con người. Gate giúp các nhà phát triển tiến hành công việc theo ba cách:

- Xác định một cấu trúc, kiến trúc tổ chức cho các phần mềm xử lý ngôn ngữ.
- Cung cấp một framework hay thư viện các lớp thực thể, thực hiện cấu trúc đã xác định và có thể được sử dụng cho các ứng dụng xử lý ngôn ngữ tự nhiên.
- Cung cấp một môi trường phát triển được xây dựng dựa trên framework của các công cụ đồ họa tiện lợi cho các thành phần phát triển.

Gate khai phá sự phát triển các phần mềm dựa trên bộ phận, hướng đối tượng và code lưu động, biến đổi nhanh. Framework và môi trường phát triển được viết bởi ngôn ngữ Java và là một phần mềm mã nguồn mở dưới sự cho phép của thư viện GNU. Gate sử dụng Unicode (Unicode Consortium 96) và được kiểm thử trên một só ngôn ngữ : Đức, Ấn Độ.

Gate bắt đầu được xây dựng và phát triển tại Trường ĐH Sheffield từ năm 1995 và từ đó được sử dụng trong nghiên cứu và các dự án. Phiên bản 1 được ra đời năm 1996 và được chứng nhận bởi hàng trăm tổ chức. Gate sử dụng một lượng lớn các ngữ cảnh từ phân tích ngôn ngữ vào trong nhiều thứ tiếng: Anh, Hy Lạp, Thụy Điển, Đức, Ý, Pháp… Các phiên bản tiếp sau được ra đời và ngày càng đáp ứng một cách hiệu quả trong nghiên cứu cũng như ứng dụng.

### 5.1.3 Dữ liệu thử nghiệm

Sau khi thu thập được hơn 500 trang web từ các web site http://suckhoedoisong.vn, chúng tôi đã loại bỏ, xử lý các văn bản nhiễu không giúp ích cho quá trình xây dựng Ontology cũng như nhận dạng thực thể. Sau khi xử lý đã thu thập được gần 400 trang web, tương ứng với trên 5000 câu để phục vụ cho việc xây dựng Ontology, nhận dạng thực thể và tạo nền tảng cho phân loại quan hệ câu.

Sử dụng công cụ tách từ JvnTextPro của Nguyễn Cẩm Tú [1] để loại bỏ HTML các trang Web cũng như tách câu, tách từ tập tài liệu này.

## *5.2 Xây dựng Ontology*

### 5.2.1. Phân cấp lớp thực thể

Với các dữ liệu về y tế thu thập được từ các trang web và ontology, chúng tôi liệt kê các thuật ngữ (term) quan trọng nhằm có thể nêu định nghĩa cho người dùng với hướng nghiên cứu tiếp theo là tự động liên kết đến các định nghĩa có sẵn trên trang wikipedia. Từ các thuật ngữ trên, tiếp theo sẽ định nghĩa các thuộc tính của chúng. Việc xây dựng Ontology là một quá trình lặp lại được bắt đầu bằng việc định nghĩa các khái niệm trong hệ thống lớp và mô tả thuộc tính của các khái niệm đó.

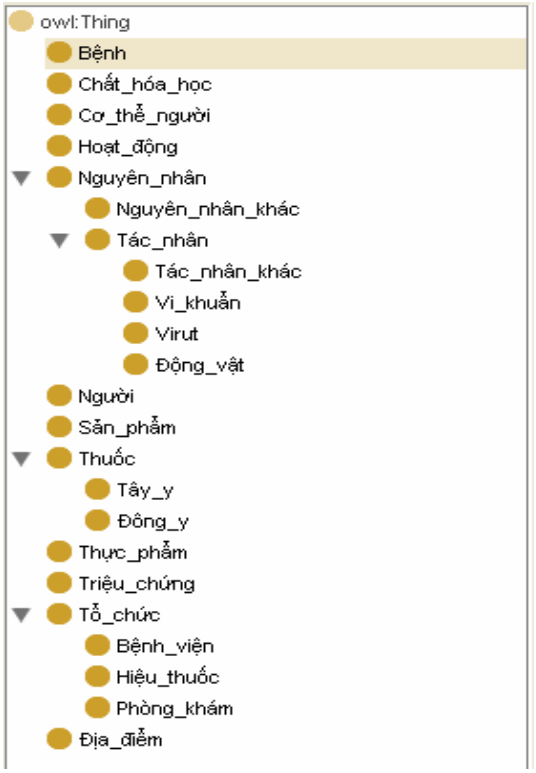
Qua khảo sát Ontology BioCaster với các thuật ngữ trong tiếng Việt, cùng với một số luợng lớn các trang Web về y tế hiện nay ở Việt Nam, chúng tôi tiến hành xây dựng nên một tập các thuật ngữ, các mối quan hệ cơ bản nhất để từ đó để xuất ra Ontology thử nghiệm ban đầu.

Sau đây là một số lớp thực thể do khóa luận đề xuất để xây dựng Ontology:

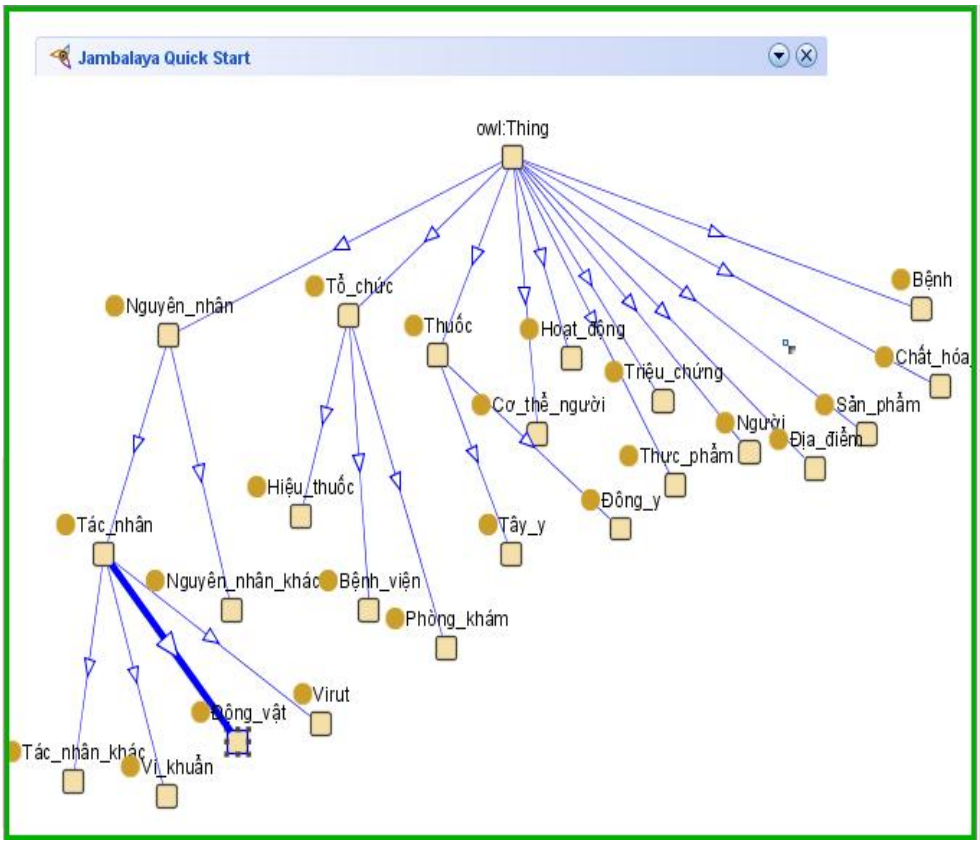
- Thuốc: Đông y, Tây y. Ví dụ như thuốc 5-Fluorouracil Ebewe chống ung thư (ung thư đại trực tràng, vú, thực quản, dạ dày), hay là thuốc Ciloxan sát trùng,

chống nhiễm khuẩn ở mắt. Thuốc đông y ngũ gia bì chữa bệnh phong thấp, tráng gân cốt …

- Bệnh, hội chứng: Các loại bệnh như cúm gà, viêm loét dạ dày, các hội chứng mất ngủ, suy tim …
- Triệu chứng: Ví dụ như triệu chứng của cúm H5N1 là sốt cao, nhức đầu, đau mỏi toàn thân,...
- Nguyên nhân: Tác nhân (virut, vi khuẩn..muỗi, gà, chim..), và các nguyên khác như là thiếu ngủ, lười tập thể dục, hút thuốc lá thụ động …
- Thực phẩm: Bao gồm các món ăn có lợi hoặc gây hại cho sức khỏe con người cũng như phù hợp với một số loại bệnh nào đó.
- Người: Bao gồm bác sỹ, giáo sư mà người bệnh có thể tìm kiếm để khám bênh, xin giúp đỡ khi mắc bệnh.
- Tổ chức: Bệnh viện, phòng khám, hiệu thuốc … là các địa điểm để bệnh nhân có thể tìm đến khi mắc bệnh.
- Địa điểm: Địa chỉ của một tổ chức nào đó mà bệnh nhân có thể tìm đến, các nơi dịch đang phát sinh và lan rộng.
- Cơ thể người: Là tất cả các bộ phận cơ thể người có thể thể bị nhiễm bệnh: mắt, mũi, gan, tim …
- Hoạt động: Chẩn trị, xét nghiệm, hồi cứu, hô hấp nhân tạo, phòng tránh, tiêm phòng ...
- Hóa chất: Vitamin, khoáng chất …gây tác động xấu, tốt đến cơ thể con người, ví dụ vitamin A có lợi cho mắt, Vitamin C, E làm giảm các nguy cơ bệnh tim…
- Hội chứng: hội chứng có thể xuất hiện của một bệnh [hội chứng sốc của bệnh sốt xuất huyết].
- Biến chứng: Từ một bệnh có thể biến chứng sang bệnh khác (bệnh quai bị biến chứng viêm màng não…).




**Hình 20: Minh họa các lớp trong Ontology đã xây dựng.**

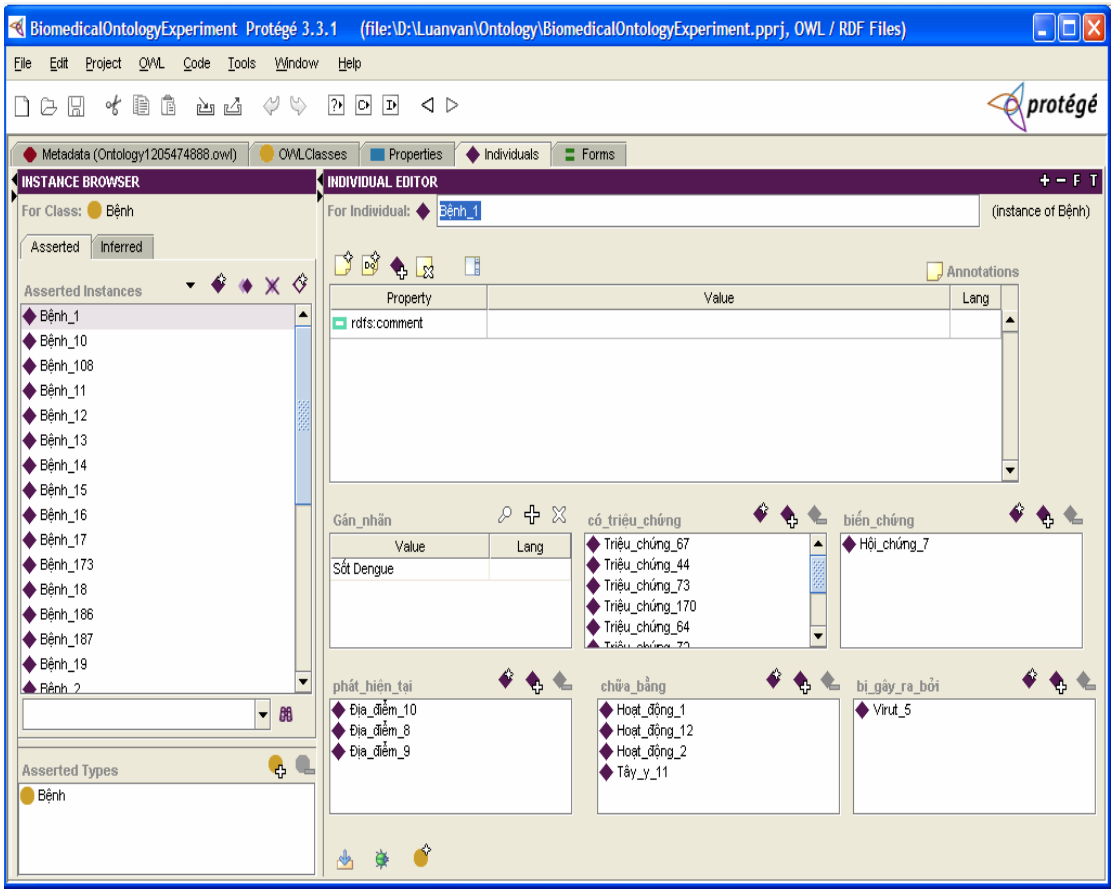



**Hình 21: Minh họa cấu trúc phân tầng của Ontology xây dựng được.**

### 5.2.2. Các mối quan hệ giữa các lớp thực thể

Khóa luận sử dụng một số quan hệ ngữ nghĩa dưới đây giữa các thực thể để xây dựng quan hệ ngữ nghĩa trong Ontology cũng như việc gán nhãn ngữ nghĩa cho câu:

- Sự tương tác thuốc – thuốc: Thuốc này có thể gây tác dụng phụ cho thuốc kia, hay có thể kết hợp các loại thuốc với nhau để chữa bệnh. Ví dụ thuốc chống ung thư Alexan không nên dùng chung với methotrexate hay 5-fluorouracil.
- Thực phẩm tác động xấu, tốt đến bệnh, cơ thể người. Ví dụ như uống xôđa nhiều có rủi ro mắc các bệnh rối loạn trao đổi chất, tăng vòng bụng, tăng huyết áp…
- Quan hệ bệnh – thuốc.
- Quan hệ nguyên nhân gây ra bệnh, hay bệnh có nguyên nhân.
- Quan hệ bệnh – triệu chứng.
- Quan hệ bệnh biến chứng thành bệnh khác.
- Các hoạt động tác động lên bệnh.
- Người làm việc trong một tổ chức tại địa điểm nào đó.
- Bệnh thuộc chuyên khoa của người.
- Bệnh được phát hiện, chữa trị ở tổ chức.
- Bệnh biến chứng sang bệnh khác.
- Quan hệ bệnh -- hội chứng.




**Hình 22. Minh họa các thể hiện của lớp thực thể và mối quan hệ giữa các thể hiện**

Hình 22 minh họa một mối quan hệ giữa các thể hiện của các lớp thực thể. Trên hình 22 là thể hiện "sốt Dengue" và các quan hệ với các thể hiện của lớp thực thể khác: Gán_nhãn, phát_hiện_tại, có_triệu_chứng, biến_chứng, chữa_bằng, bị_gây_ra_bởi.

Khóa luận đã xây dựng được một Ontology bao gồm 21 lớp thực thể, 13 mối quan hệ và trên 500 thể hiện của các lớp thực thể.

### *5.3. Chú thích dữ liệu*

Khóa luận tích hợp Ontology vào công cụ Gate (General Architecture for Text Mining) để chú thích dữ liệu.. Từ dữ liệu đã được thu thập và ontology đã xây dựng, quá  trình chú thích dữ liệu bao gồm các bước sau:

- Mở file chứa dữ liệu để chú thích, có thể dùng mở cả thư mục chứa nhiều file để chú thích. Sử dụng Data_Store của gate để lưu các dữ liệu được mở và sau khi được chú thích.

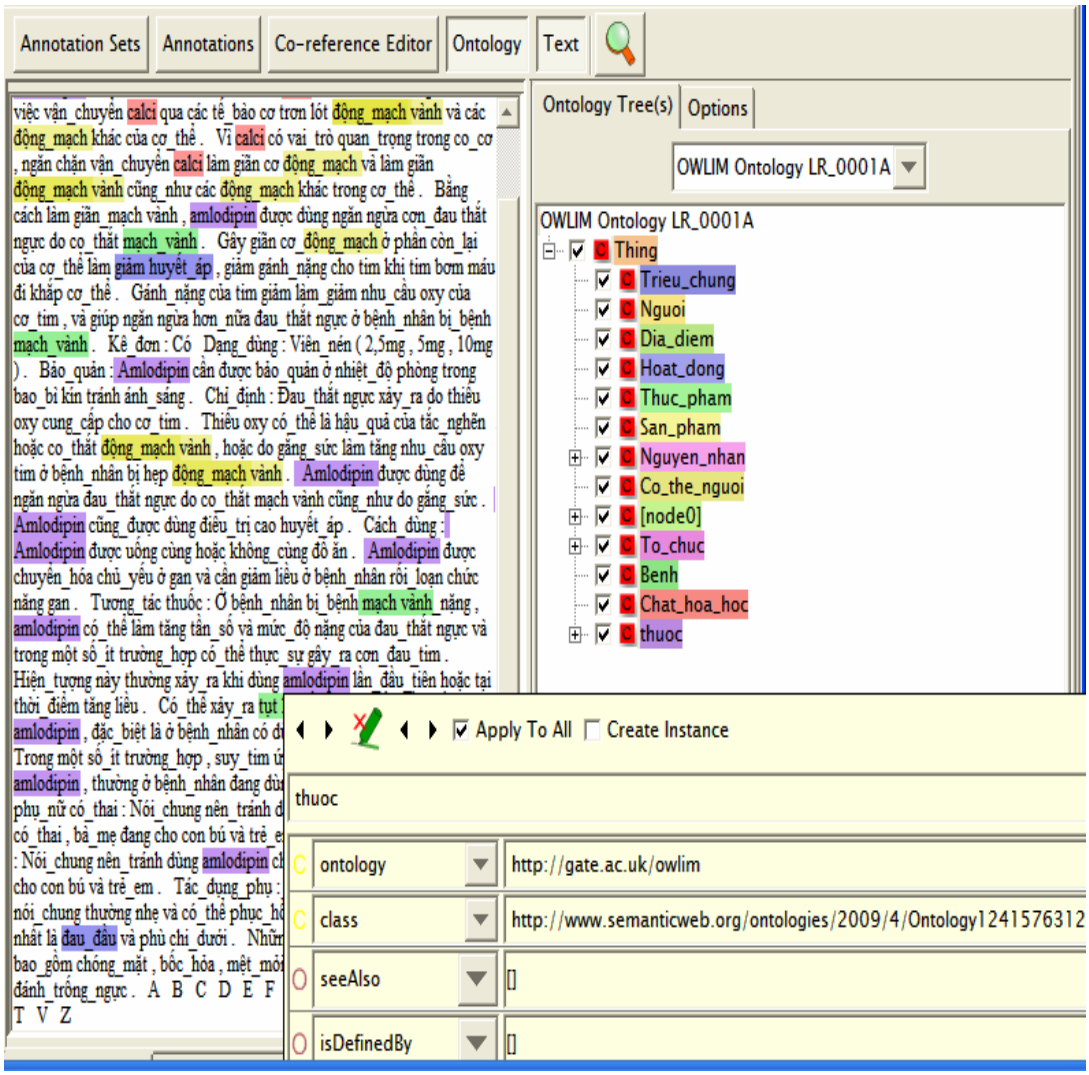
• Mở Ontology đã xây dựng được. Ontology có thể dùng công cụ Gate để chỉnh sửa lại các lớp, thuộc tính,…

• Thay đổi màu sắc chú thích các thực thể ở Ontology một cách phù hợp để có thể tiện phân biệt các thực thể một cách rõ ràng.

• Chọn thực thể cần chú thích và chọn tên lớp thực thể thuộc ontology để chú thích.

Kết quả sau quá trình chú thích, chúng ta có thể có một dữ liệu chứa các thực thể tương ứng với các lớp đã được xây dựng trên ontology. Chú thích dữ liệu giúp cho việc xây dựng tập corpus trên dữ liệu y tế một cách dễ dàng hơn, đồng thời góp phần vào việc tự động mở rộng các thực thể trên ontology.

Khóa luận đã chú thích được 96 file dữ liệu tương ứng với trên 1500 thể hiện.




**Hình 23: Minh họa một dữ liệu được chú thích bằng Ontology.**

## *5.4. Nhận dạng thực thể*

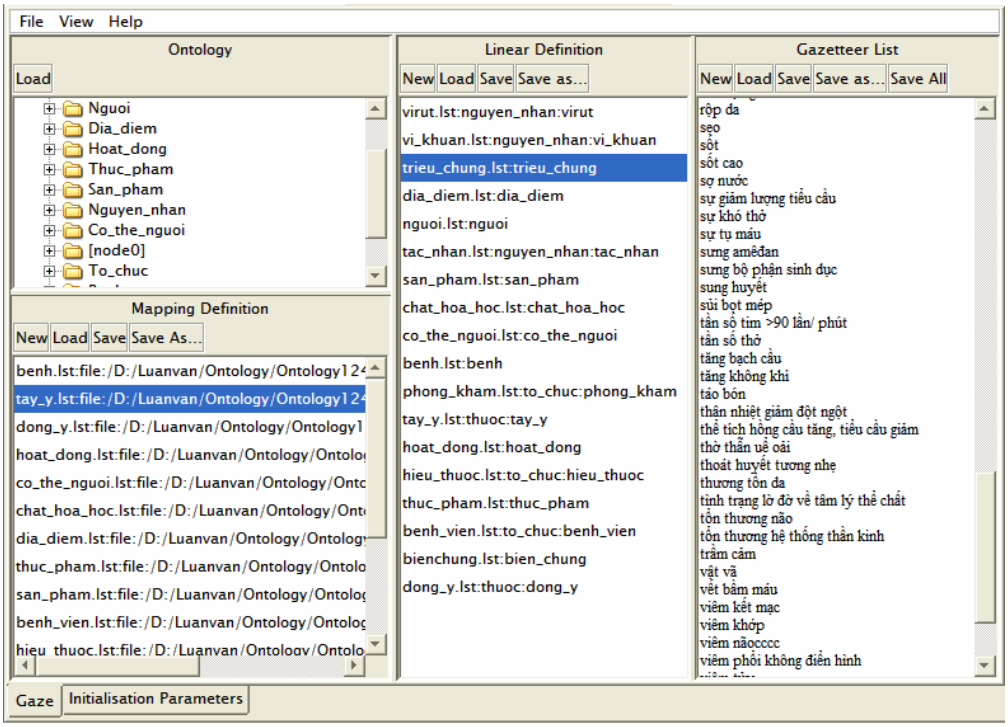
### 5.4.1. Xây dựng tập gazetteer

Sau khi chú thích dữ liệu, chúng ta có các file dữ liệu được chú thích với các lớp thực thể riêng biệt. Sau quá trình chú thích này, chúng ta có thể dựa trên các dữ liệu đã được chú thích để xây dựng môt tập dữ liệu tên các thực thể. Xây dựng được một tập dữ liệu tốt có thể giúp cho quá trình nhận dạng thực thể hiệu quả hơn. Khóa luận đã sử dụng Ontology cùng một mở rộng được tích hợp vào Gate là gazetteer để xây dựng. Ngoài việc xây dựng được một tập dữ liệu phục vụ cho nhiệm vụ trích chọn thực thể, dựa vào gazetteer chúng ta có thể liệt kê một số từ ngữ liên quan trực tiếp tới một số quan hệ, ví dụ như quan hệ gay_ra giữa thực thể "nguyên_nhân" và "bệnh" có các từ thường gặp như **gây**, **gây_ra**, **làm**, **làm_cho** …

Bảng 2 minh họa số lượng các thể hiện của các lớp thực thể trong tập dữ liệu gazetteer.

| Lớp thực thể | Số lượng |
|---|---|
| Bệnh | 232 |
| Triệu chứng | 246 |
| Cơ_thể_người | 78 |
| Virut | 53 |
| Vi_khuẩn | 38 |
| Phòng_khám | 27 |
| Bệnh_viện | 52 |
| Hiệu thuốc | 81 |
| Biến_chứng | 93 |
| Gây_ra | 15 |
| Thuốc (Đông y) | 212 |
| Thuốc (Tây y) | 151 |
| Thực phẩm | 145 |
| Chất_hóa_học | 122 |
| Hoạt_động | 147 |
| Tổng | 1692 |

**Bảng 2. Số lượng các thể hiện của các lớp thực thể trong tập dữ liệu gazetteer.**




**Hình 24. Minh họa các file chứa thực thể trong tập Gazetteer xây dựng được**

### 5.4.2.Đánh giá hệ thống nhận dạng thực thể

Các hệ thống nhận biết loại thực thể được đánh giá chất lượng thông qua ba độ đo: độ chính xác (precision), độ hồi tưởng (recall) và độ đo F (F-messure). Ba độ đo này được tính toán theo các công thức sau:

$$rec = \frac{correct}{correct+incorrect+missing}$$

$$pre = \frac{correct}{correct+incorrect+spurious}$$

$$F = \frac{2*pre*rec}{pre+rec}$$

Ý nghĩa của các giá trị correct, incorrect, missing và spurious được định nghĩa như Bảng 3 dưới đây.

| Giá trị | Ý nghĩa |
|---|---|
| Correct | Số trường hợp được gán đúng |
| Incorrect | Số trường hợp bị gán sai |
| Missing | Số trường hợp bị thiếu |
| Spurious | Số trường hợp thừa |

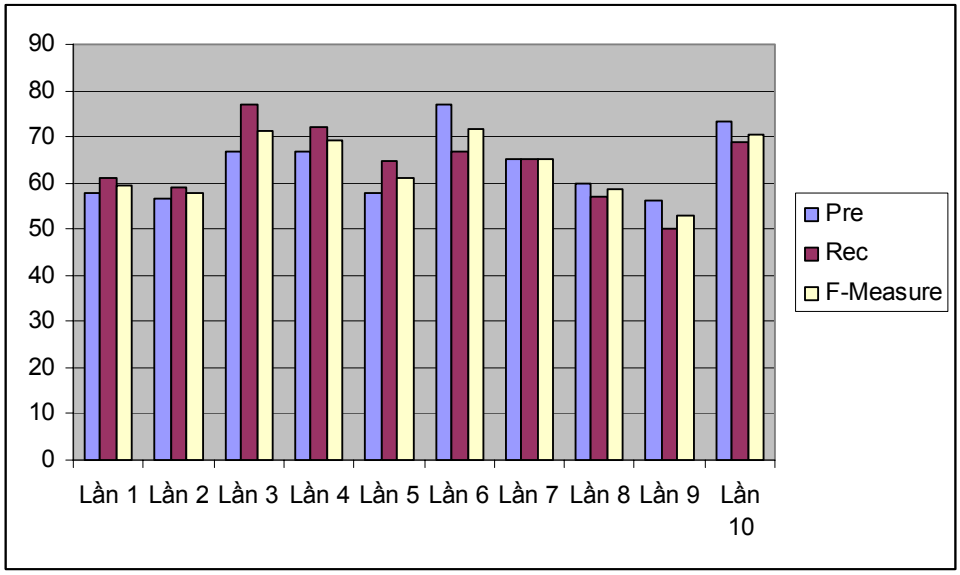
**Bảng 3. Các giá trị đánh gía một hệ thống nhận diện loại thực thể**

### 5.4.3. Kết quả đạt được

Kết quả sau 10 lần thực nghiệm nhận dạng thực thể các file đã được chú thích ngữ nghĩa được thể hiện dưới Bảng 4 dưới đây:

| Độ đo | Lần 1 | Lần 2 | Lần 3 | Lần 4 | Lần 5 | Lần 6 | Lần 7 | Lần 8 | Lần 9 | Lần 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Pre. [%] | 57.89 | 56.52 | 66.67 | 66.67 | 57.89 | 77.06 | 65.2 | 60 | 56.25 | 73.3 |
| Rec. [%] | 61.1 | 59.09 | 76.92 | 72.22 | 64.70 | 66.67 | 65.2 | 57.14 | 50 | 68.75 |
| F-Measure [%] | 59.45 | 57.77 | 71.42 | 69.33 | 61.10 | 71.49 | 65.2 | 58.53 | 52.94 | 70.45 |

**Bảng 4. Kết quả sau 10 lần thực nghiệm nhận dạng thực thể.**




**Hình 25. Kết quả 10 lần thực nghiệm nhận dạng thực thể**

### 5.4.4. Nhận xét và đánh giá

Nhận dạng thực thể sử dụng tập Gazetteer đưa ra kết quả khá cao (thấp nhất là 50% và cao nhất là 77.06 %). Sỡ dĩ sử dụng phương pháp gazetteer cho kết quả khả quan là do giữa các tài liệu huấn luyện và kiểm thử có sự tương đồng nhất định. Do đó các thực thể cần nhận dạng thường xuất hiện trong danh sách các gazetteer. Nếu tập dữ liệu kiểm thử được lấy từ một nguồn khác thì phương pháp này có thể không mang lại kết quả khả quan. Trong tương lai, chúng tôi sẽ sử dụng các đặc trưng dữ liệu, biểu thức chính quy,… để mang lại kết quả cao hơn cho bài toán nhận dạng thực thể.

### *5.5. Gán nhãn ngữ nghĩa cho câu*

Ontology đã mô tả được một số quan hệ giữa các lớp thực thể y tế tiếng Việt. Từ các quan hệ trong khóa luận, chúng tôi đã lược bỏ và sẽ chỉ sử dụng 6 loại quan hệ

- LÀ: Thực thể này là thực thể kia (cúm gà – cúm A H5N1).
- CÓ: Bệnh có các triệu chứng, biến chứng, hội chứng.
- GÂY_RA: Các nguyên nhân gây ra bệnh.
- LIÊN_QUAN: Triệu chứng liên quan đến bệnh nào đó.
- ĐIỀU_TRỊ: Các phương pháp điều trị bệnh.
- TÁC_ĐỘNG: Thực phẩm, hoạt động … tác động đến bệnh nào đó.

Từ tập dữ liệu thu thập được, chúng tôi đã gán nhãn dữ liệu cho 1000 câu để làm dữ liệu học. Do thời gian có hạn và tập dữ liệu xây dựng là quá lớn, khóa luận chỉ kịp xây dựng dữ liệu. Với tập dữ liệu được xây dựng, trong tương lai, chúng tôi sẽ sử dụng 500 câu để huấn luyện và 500 câu dùng để kiểm thử trong quá trình phân lớp câu chứa quan hệ sử dụng thuật toán SVM. Bảng 5 mô tả một số câu dữ liệu y tế được gán nhãn với các quan hệ vừa trình bày ở trên.

| |
|---|
| GÂY_RA Mắt hột là bệnh viêm kết mạc do vi khuẩn Chlamydia gây ra. |
| CÓ Bệnh có những đợt tái phát, viêm kết mạc, viêm biểu mô giác mạc. |
| CÓ Biểu hiện bệnh rất đa dạng, từ nhẹ không có triệu chứng gì đến những trường hợp bệnh nặng kéo dài, biến chứng nguy hiểm có thể dẫn đến mù lòa. |
| CÓ Những triệu chứng thường gặp là: cộm xốn mắt, vướng mắt như có hạt bụi trong mắt, ngứa mắt, hay mỏi mắt. |
| CÓ Tổn thương sẹo hóa của kết mạc dẫn đến cụp mi, lông siêu, lông quặm. |
| TÁC_ĐỘNG Phòng bệnh bằng cách: rửa mặt bằng khăn riêng sạch, nước rửa sạch, giữ tay sạch, không dụi bẩn lên mắt, không tắm ao hồ, tránh để nước bẩn bắn vào mắt, nên đeo kính khi đi đường, về nhà nên rửa mặt sạch sẽ; diệt ruồi nhặng. |
| ĐIỀU_TRỊ Đi khám bệnh ngay khi có những triệu chứng khó chịu ở mắt. Khi bị bệnh cần điều trị theo sự hướng dẫn của bác sĩ. |
| ĐIỀU_TRỊ Khi phát hiện thấy có những biểu hiện bất thường, bạn cần đi khám tại chuyên khoa mắt hay bệnh viện mắt để được tư vấn cách điều trị bệnh. |
| GÂY_RA Sau trận lụt lịch sử vừa qua, tại một số địa phương đã xuất hiện nhiều người mắc bệnh đau mắt đỏ. |
| GÂY_RA Đây là một bệnh dễ gặp ở các vùng bị ngập lụt do thiếu nước sạch sinh hoạt hoặc do tiếp xúc với hóa chất. |
| LÀ Đau mắt đỏ (ĐMĐ) còn gọi là viêm kết mạc. |

**Bảng 5. Ví dụ một số câu được gán nhãn quan hệ**

# PHỤ LỤC - MỘT SỐ THUẬT NGỮ ANH VIỆT

| Thuật ngữ | Giải thích |
|---|---|
| Assign sentence lable | Gán nhãn ngữ nghĩa cho câu |
| Classifier | Phân loại, phân lớp |
| Information Extraction | Trích chọn thông tin |
| Information Retrieval | Tìm kiếm thông tin |
| Machine Translation | Dịch máy |
| NE – Name Entity | Tên thực thể |
| NER-Name Entity Recognition | Nhận dạng tên thực thể |
| Semantic Relation | Quan hệ ngữ nghĩa |
| Semantic Search | Tìm kiếm ngữ nghĩa |
| Semi-Supervised | Học bán giám sát |

# KẾT LUẬN

Nhận biết được tầm quan trọng của việc sử dụng các tài nguyên trực tuyến trong lĩnh vực y tế nhằm phục vụ đời sống con người, khóa luận đã trình bày và thử nghiệm một số phương pháp khai phá nguồn dữ liệu y tế này nhằm mục đích đưa lại nguồn tri thức cho một số bài toán khác, ví dụ là bài toán tìm kiếm ngữ nghĩa. Khóa luận đã trình bày một số phương pháp, công cụ … xây dựng Ontology và xây dựng được một Ontology cho y tế tiếng việt. Ontology này mô tả tổng quát được các thực thể cơ bản rong dữ liệu y tế, làm tiền đề cho việc xây dựng mạng ngữ nghĩa cho bài toán tìm kiếm ngữ nghĩa. Khóa luận cũng trình bày một số phương pháp, công cụ để chú thích dữ liệu và xây dựng tập dữ liệu ban đầu cho quá trình nhận dạng thực thể cũng như mở rộng Ontology một cách tự động dùng Gazetteer. Kết quả thực nghiệm khi sử dụng tập dữ liệu tương đối khả quan (thấp nhất là 50% và cao nhất là 77.06%). Ngoài ra khóa luận cũng đề cập tới bài toán đang rất được quan tâm trong thời gian gần đây: xác định quan hệ. Đối với bài toán xác định quan hệ, chúng tôi đã trình bày khái quát về quan hệ, xác định quan hệ, gán nhãn ngữ nghĩa cho câu và phân lớp câu chứa quan hệ.

Hướng nghiên cứu trong tương lai, chúng tôi sẽ mở rộng Ontology một cách tự động, sử dụng phương pháp trích chọn đặc trưng, biểu thức chính quy và dựa trên hệ luật để có thể nâng cao hết quả của hệ thống nhận dạng thực thể. Khóa luận đã bước đầu thử nghiệm gán nhãn ngữ nghĩa cho câu với khoảng 1000 câu, các câu này sẽ được sử dụng thuật toán SVM để học và phân lớp quan hệ chứa ngữ nghĩa cho câu trong thời gian sắp tới.

# TÀI LIỆU THAM KHẢO

## Tiếng Việt

[1]. Nguyễn Cẩm Tú. Nhận biết các loại thực thể trong văn bản tiếng Việt nhằm hỗ trợ Web ngữ nghĩa và tìm kiếm hướng thực thể. Khóa luận tốt nghiệp ĐHCN 5/2005, tr. 3, tr.

[2]. Nguyễn Minh Tuấn. Phân lớp câu hỏi hướng tới tìm kiếm ngữ nghĩa tiếng Việt trong lĩnh vực y tế. Khóa luận tốt nghiệp ĐHCN 5/2008, tr. 2-26.

## Tiếng Anh

[3]. Andreas Vlachos. Evaluating and combining biomedical named entity recognition systems,Computer Laboratory ,University of Cambridge, 2007.

[4]. Brandon Beamer, Alla Rozovskaya, Roxana Girju. Automatic Semantic Relation Extraction with Multiple Boundary Generation. University of Illinois at Urbana-Champaign, 2008, tr. 3-4.

[5]. David Nadeau. Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision. Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for the PhD degree in Computer Science, 2007 tr. 15-16.

[6]. GuoDong Zhou, Jian Su. Named Entity Recognition using an HMM-based Chunk Tagger. Laboratories for Information Technology, Singapore, 2002, tr. 3-4.

[7]. Haochang Wang, Tiejun Zhao, Hongye Tan, Shu Zhang. Biomedical Named entity recognition based on classifiers ensemble. International Journal of Comput er Science and Applications, 2004; Vol. 5, No. 2 ,tr. 1-11.

[8]. I. Horrocks, D. Fensel, F. Harmelen, S. Decker, M. Erdmann, M. Klein, OIL in a Nutshell, ECAI00 Workshop on Application of Ontologies and PSMs, Berlin, 2000.

[9]. I. Horrocks, F. van Harmelen. Reference Description of the DAML þ OIL, Ontology Markup Language, Technical report, 2001.

[10]. John McNaught. Challenges for Terminology Management in Biomedicine. NaCTeM Associate, University of Manchester, 2005.

[11]. Kawazoe, A., and Collier, N. April. BioCaster Project Working Report on English Named Entity Annotation. National Institute of Informatics, Japan 2007 , tr. 4-6.

[12]. Lassila, R. Swick. Resource description framework (RDF) model and syntax specification, W3C Recommendation 1999, http://www.w3.org/TR/REC-rdf-syntax/.

[13]. LIU Yi, ZHENG Y F. One-against-all multi-Class SVM classification using reliability measures.Proceedings of the 2005 International Joint Conference on Neural Networks Montreal, Canada, 2005.

[14]. Massimiliano Ciaramita, Aldo Gangemi, Esther Ratsch Jasmin, Saric Isabel Rojas. Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology. Institute for Cognitive Science and Technology (CNR), Italy, 2005, tr 1-5.

[15]. M. Fernaandez-Loopez, A. Goomez-Peerez, A. Pazos-Sierra, J. Pazos-Sierra, Building a chemical ontology using METHONTOLOGY and the ontology design environment, IEEE Intelligent Systems & their applications 4 (1), 1999.

[16]. M. Gr€uuninger, M.S. Fox. Methodology for the design and evaluation of ontologies, Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, 1995.

[17]. M. Ushold, R M. Uschold, M. King. Towards a Methodology for Building Ontologies, IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, 1995

[18]. Noy, N.F., and McGuinness, D.L. Ontology Development 101: A Guide to Creating Your First Ontology SMI, Technical report SMI-2001-0880, Stanford University, 2001.

[19]. N. Guarino. Formal Ontology in Information Systems. Proceedings of FOIS'98:3-15, Trento, Italy, 6/1998. Amsterdam, IOS Press.

[20]. Thao Pham T. X., Tri T. Q., Ai Kawazoe, Dien Dinh, Nigel Collier. Construction of Vietnamese corpora for Named Entity Recognition.VNU of HCMC Vietnam, National Institute of Informatics, Tokyo, Japan, tr. 1-3.

[21]. Thao, P.T.X., Tri, T.Q., Dien, D., and Collier N., 2007. Named entity recognition in Vietnamese using classifier voting, ACM Trans. Asian. Lang. Inf. Process. 6, 4, Article 14 , 12/2007, tr. 2-3.

[22]. Tim Berners-Lee, "Semantic Web Road map", http://www.w3.org/DesignIssues/Semantic.html.

[23] Razvan C. Bunescu. Learning to Extract Relations from Biomedical Corpora. Electrical Engineering and Computer Science, Ohio University, Athens, OH, 3/2009.
[24] Roxana Girju. Semantic relation extraction and its applications, 20th European Summer School in Logic, Language and Information, 4/2008, tr. 2-10.
[25] Sammy Wang. Application of Data and Text Mining to Bioinformatics, 2008. University of Georgia.
[26] S.Cohen , Mamou, J., Kanza, Y., Sagiv, Y. Xsearch: A semantic search engine for xml. In: Proceedings of of the 29th VLDB Conference, Berlin, Germany, 2003.
[27] S. Luke, J. Heflin, SHOE 1.01. Proposed Specification, SHOE Project technical report, University of Maryland, 2000.
[28] Soumen Chakrabarti. Mining the web, Discovering Knowledge from Hypertext Data, Edition: 3, illustrated. Published by Morgan Kaufmann, 2003. Chapter Semi-supervised Learning.
[29] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med, 1986.
[30] Xuan-Hieu Phan, Le-Minh Nguyen, Susumu Horiguchi. Personal Name Resolution Crossover Documents by A semantics-Based Approach. in IEICE Trans Inf & Syst , 2006, tr. 1-5.
[31] http://gate.ac.uk/
[32]http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
[33] http://genome.jouy.inra.fr/texte/LLLchallenge/
[34] http://www.dit.hcmut.edu.vn/~tru/VN-KIM/products/vnkim-kb.htm.
[35] http://www.wolframalpha.com/
[36] http://www.w3.org/
[37]http://wordnet.princeton.edu/.
[38]http://nlp.cs.swarthmore.edu/semeval/
[39]http://www.nlm.nih.gov/mesh/-meshhome.html
[40]http://www.dit.hcmut.edu.vn/~tru/VN-KIM/products/vnkim-ie.htm.
[41 ]http://www.bioontology.org/ncbo/faces/pages/ontology_list.xhtml.
[42] http://diseaseontology.sourceforge.net/
[43 ]http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi
[44] http://biocaster.nii.ac.jp/
[45] http://www.ksl.stanford.edu/software/ontolingua/
[46] http://www.isi.edu/isd/ontosaurus.html
[47] http://www-sop.inria.fr/acacia/ekaw2000/ode.html

[48] http://www.xml.com/pub/r/861
[49]http://biocreative.sourceforge.net/
[50] http://www.owlseek.com/whatis.html
[51] http://protege.stanford.edu/
[52] http://www.bioontology.org/.