Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

NGHIÊN CỨU VỀ UPPER ONTOLOGY TRỰC TUYẾN

Huỳnh Nhứt Phát[1], Hoàng Hữu Hạnh[2] và Phan Công Vinh[3]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Khoa học Huế*
[2] *Ban hợp tác Quốc tế, Trường Đại học Huế*
[3] *Khoa Công nghệ Thông tin, Trường Đại học Nguyễn Tất Thành TP. HCM*

Thông tin chung:
Ngày nhận: 19/09/2015
Ngày chấp nhận: 10/10/2015

Title:
Research on an Online Upper Ontology

Từ khóa:
WordNet, SUMO, trình duyệt SUMO, ánh xạ, các tiên đề

Keywords:
WordNet, SUMO, SUMO Browser, Mapping, Axioms

ABSTRACT

We present a new approach for mapping WordNet to SUMO as well as the SUMO Browser-online tool that can be used for browsing SUMO(Suggested Upper Merged Ontology), and its connection to the WordNet lexicon. The Browser facilitates the process of getting familiar with SUMO contents. In this paper, we also present SUMO and WordNet briefly.

TÓM TẮT

Chúng tôi trình bày cách tiếp cận mới về ánh xạ giữa WordNet và SUMO (Suggested Upper Merged Ontology) cũng như công cụ trình duyệt SUMO trực tuyến có thể sử dụng để duyệt SUMO, và sự kết nối của nó với từ vựng WordNet. Trình duyệt tạo điều kiện thân thiện cho quá trình tiếp nhận nội dung SUMO. Trong bài báo này chúng tôi cũng trình bày ngắn gọn về SUMO và WordNet.

1 GIỚI THIỆU

Các tác vụ về mặt kỹ thuật như truy hồi thông tin, xử lý ngôn ngữ tự nhiên, biểu diễn tri thức, và khả năng tương tác dữ liệu đòi hỏi các nguồn tài nguyên mới. Các nguồn tài nguyên rất quan trọng và cần thiết trong việc thiết kế các hệ thống tri thức thuộc ontology hình thức với cấu trúc của cơ sở tri thức. Đặc biệt, lớp của các ontology được hình thành bởi các ontology ở tầng trên của các ontology với miền độc lập, nhằm tái sử dụng và mở rộng của các miền riêng để tạo thành một miền ontology chung. Các nguồn tài nguyên quan trọng khác là các từ điển điện tử. Các từ điển cung cấp cầu nối giữa kiến thức được diễn đạt trong các hệ thống tri thức và ngôn ngữ tự nhiên.

Bài báo này, chúng tôi trình bày một ontology cụ thể ở tầng trên là SUMO [9, 10] và một từ điển cụ thể đó là WordNet [6, 12]. Chúng tôi trình bày về các khả năng sử dụng từ điển WordNet với các tác vụ bao gồm việc xử lý tự động đối với ngôn ngữ tự nhiên. Phần tiếp theo, chúng tôi sẽ đưa ra ý tưởng mới của chúng tôi về việc liên kết giữa SUMO và WordNet. Phần cuối, chúng tôi giới thiệu trình duyệt SUMO [11] là công cụ trực tuyến để tiếp nhận nội dung của SUMO và WordNet với một hình thức thích hợp. Các phần tách biệt dành riêng cho tính năng diễn giải của trình duyệt, và ứng dụng mẫu của chúng tôi diễn tả chương trình truy cập đến nội dung SUMO và WordNet.

2 SUMO (SUGGESTED UPPER MERGED ONTOLOGY)

SUMO là ontology được nghiên cứu bởi nhóm SUO tại Teknowledge Corporation. SUMO được tạo ra bởi việc sáp nhập nội dung các ontology vào một cấu trúc thống nhất, toàn diện và gắn kết. SUMO là một tập của khoảng 1000 khái niệm đã được xác định và được chú thích rõ ràng, kết nối với nhau tạo thành mạng ngữ nghĩa và kèm theo một số tiên đề. Các khái niệm bao gồm những từ tổng quát như 'số lượng', và cụ thể như 'chim bồ câu'. Các tiên đề chủ yếu phản ánh những khái niệm với nghĩa phổ biến, chúng thường được chấp

nhận giữa các khái niệm. SUMO với miền độc lập, được sử dụng như là một thành phần hỗ trợ chính cho việc thiết kế các ontology miền.

Các tiên đề sẽ hỗ trợ sự thể hiện về ràng buộc của các khái niệm, và cung cấp các nguyên tắc cơ bản cho các hệ thống suy luận tự động để xử lý các cơ sở tri thức phù hợp với ontology SUMO. Ví dụ sau đây về tiên đề:

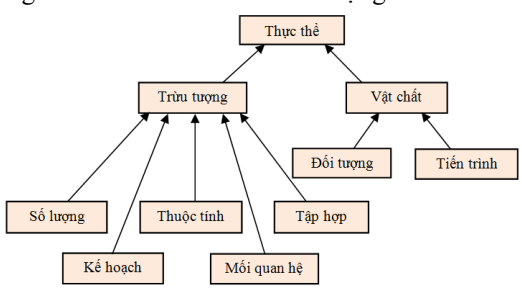
"Nếu c là một thể hiện của quá trình đốt cháy, thì tồn tại sự nung h và ánh sáng tỏa ra là l để cả hai h và l là quá trình con của c".

Điều này khá phức tạp, nhưng về mặt logic, câu nói đó là một quá trình của sự nung nóng và một quá trình phát ra ánh sáng đi kèm với quá trình đốt cháy (sự đốt). Hơn nữa, tiên đề này được mã hoá trong SUMO với một ngôn ngữ logic hình thức.

Các khái niệm trong SUMO được tổ chức thành một hệ thống phân cấp có nguồn gốc từ thực thể, đại diện cho các khái niệm chung nhất. Hai cấp độ đầu tiên của hệ thống phân cấp được mô tả trong Hình 1. Chúng ta có thể thấy, thực thể được chia thành vật chất (vật lý), và ý thức (trừu tượng). Mọi thứ vật chất được tiếp tục phân chia thành các đối tượng và các tiến trình,...

Các lớp con của một lớp thường là loại trừ lẫn nhau, nghĩa là chúng không chia sẻ các thể hiện chung. Ví dụ, không có thực thể nào vừa mang tính trừu tượng vừa mang tính vật lý và cũng không thể vừa là một đối tượng vừa là một tiến trình. Thuộc tính này được quy định rõ ràng trong SUMO. Tuy nhiên, một số lớp có thể có nhiều lớp cha. Ví dụ, con người vừa thuộc họ người (một thành viên của tầng lớp nhất định của loài động vật) vừa là tác nhân liên quan đến nhận thức (một thực thể có khả năng suy luận).

Một trong những hạn chế của SUMO là mức độ về những sự kiện được bao phủ tương đối thấp của nó và không cho phép sử dụng các ứng dụng về miền mở của mình. Nó cũng không có một kết nối giữa khái niệm của nó và các từ ngữ tự nhiên. Những hạn chế này đã được khắc phục một phần bằng cách kết nối SUMO với từ vựng WordNet.

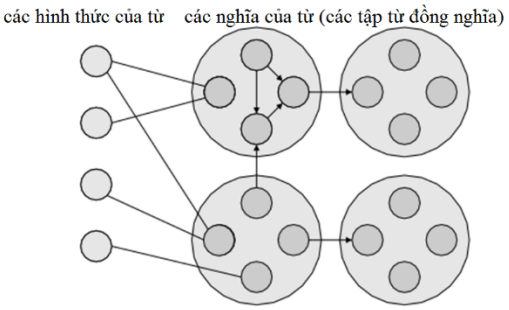


Hình 1: Các khái niệm ở mức cao trong SUMO

3 WORDNET - MỘT CƠ SỞ DỮ LIỆU VỀ TỪ VỰNG TRỰC TUYẾN

WordNet là một cơ sở dữ liệu trực tuyến về từ vựng có sẵn rộng rãi trên mạng. Các nhà ngôn ngữ học tại Đại học Princeton đã tạo ra nó với kết quả nghiên cứu thuộc ngôn ngữ tâm lý học. Tuy nhiên, trong thập kỷ qua WordNet được xác nhận là nguồn tài nguyên rất có giá trị để xử lý tự động đối với ngôn ngữ tự nhiên.

Về mặt kỹ thuật, WordNet là một từ điển điện tử, xác định tập hợp lớn về các nghĩa của từ vựng, được liên kết nhau thông qua các con trỏ ngữ nghĩa. Cấu trúc về logic của WordNet được thể hiện trong Hình 2.



Hình 2: Cấu trúc logic của WordNet

Các nghĩa của từ được kết hợp với các hình thức của từ và có thể biểu diễn chúng. Chúng ta có thể thấy hình minh hoạ về sự quan hệ giữa các hình thức của từ và các nghĩa của từ. Hình thức của từ có thể có nhiều nghĩa, và nhiều hình thức của từ có thể tham khảo đến nghĩa tương tự. Trường hợp đầu được gọi là đa nghĩa, trường hợp sau được gọi là đồng nghĩa. Việc xử lý với sự mơ hồ của ngôn ngữ tự nhiên là thách thức chính trong tiến trình xử lý tự động về ngôn ngữ tự nhiên.

Mỗi mục từ có nghĩa (còn gọi là tập từ đồng nghĩa), được đi kèm với định nghĩa ngắn gọn dễ hiểu (gọi là lời giải thích), và danh sách các hình thức của từ tham khảo đến nghĩa tương tự có thể đại diện cho tập từ đồng nghĩa trong ngôn ngữ nói hoặc viết. Tập từ đồng nghĩa được lưu giữ riêng biệt cho các từ loại khác nhau: như cơ sở dữ liệu của các danh từ (66.054 từ đồng nghĩa), động từ (12.156 từ đồng nghĩa), tính từ (17.944 từ đồng nghĩa) và trạng từ (3.604 từ đồng nghĩa). Cần lưu ý rằng các quan hệ ngữ nghĩa giữa các tập từ đồng nghĩa là khác nhau cho mỗi từ loại khác nhau. Ví dụ đối với các danh từ, mối quan hệ chính giữa các tập từ đồng nghĩa là mối quan hệ is-a, được biết đến từ mô hình dữ liệu. Trong WordNet, mối quan

hệ này được gọi là hypernymy/hyponymy (khái quát/khu biệt) .

Thoạt nhìn, các tập từ đồng nghĩa trong WordNet xây dựng được một mạng ngữ nghĩa lớn, chúng ta biết nó như là một mô hình biểu diễn tri thức của trí tuệ nhân tạo. Tuy nhiên, tầm nhìn gần hơn cho thấy mối quan hệ ngữ nghĩa trong WordNet đôi khi rất mơ hồ không logic, và không thể sử dụng được về suy luận logic. Các mối quan hệ được mã hóa bởi người biên soạn từ điển, có nghĩa là sự hiểu biết giống nhau của con người về các mối quan hệ giữa các nghĩa của từ. Hơn nữa, do kích thước rất lớn của mạng ngữ nghĩa, việc thiết kế về các mối quan hệ ngữ nghĩa là khá cục bộ, không chú ý đến cấu trúc tổng thể của toàn bộ mạng ngữ nghĩa.

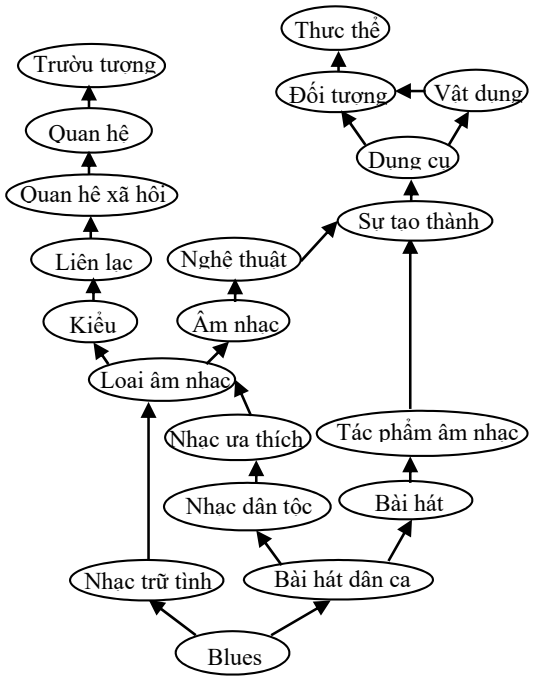
Chúng ta xét ví dụ về tập từ đồng nghĩa tương ứng với từ Blues. WordNet xác định Blues như là ‘một loại bài hát dân gian có nguồn gốc từ người Mỹ da đen ở đầu thế kỷ 20; có một âm thanh sầu thảm lặp đi lặp lại khi sử dụng các nốt nhạc xanh’.

Hệ thống phân cấp từ khái quát của tập từ đồng nghĩa này được thể hiện trong Hình 3. Từ hình này có thể nhận thấy rằng mạng ngữ nghĩa nhỏ này được tổ chức rất kém, do sự diễn giải ý nghĩa lỏng lẻo của khái niệm. Ví dụ, Blues vừa trừu tượng (không tồn tại về mặt nhận thức) vừa là thực thể (tồn tại về vật lý). Tương tự như vậy, khái niệm Bài hát dân ca có hai nghĩa: một nghĩa biểu thị một lớp của các Bài hát, và một nghĩa là loại Nhạc dân tộc.

Tuy nhiên, khái niệm này cũng có thể được hiểu như là một thuộc tính của các Bài hát, thuộc kiểu thể loại Nhạc dân tộc.

Với ngôn ngữ phổ biến hai khái niệm này thường không phân biệt rõ ràng, và do đó sự phân biệt này không được xử lý trong WordNet. Tình trạng tương tự xảy ra với nhiều từ khái quát của khái niệm về Loại âm nhạc, trong đó nó bỏ qua sự phân biệt giữa quá trình (Âm nhạc), và vai trò của nó (Quan hệ xã hội). Chúng ta thấy rằng, có nhiều khái niệm giống nhau làm cho việc xử lý trở nên khó khăn của nguồn tài nguyên WordNet khi xử lý ngôn ngữ tự nhiên tự động.

Những khác biệt về khái niệm là yếu tố quan trọng, để các nguồn tài nguyên như WordNet có thể tránh được sự khó khăn trong việc xử lý chúng. Vấn đề logic và ngôn ngữ là các lý do căn bản cho việc tổ chức các nghĩa của từ hoàn toàn khác nhau, đặc biệt là cho các khái niệm tổng quát, chúng gần với từ gốc của hệ thống phân cấp.



Hình 3: Hệ thống phân cấp khái quát của tập từ đồng nghĩa Blues

4 ÁNH XẠ GIỮA WORDNET VÀ SUMO

Ở các phần trước, cả SUMO và WordNet đều nói đến các khái niệm giống nhau, mặc dù tâm điểm là nói về các khái niệm khác nhau. Cả hai SUMO và WordNet định nghĩa các khái niệm (đơn giản) về thế giới thực. WordNet với mục đích chính là ánh xạ các khái niệm này thành những thuật ngữ về ngôn ngữ tự nhiên, và SUMO với mục đích là tổ chức chúng thành cấu trúc về logic. Do đó, nó thực hiện việc tạo một ánh xạ giữa hai nguồn tài nguyên này.

Chúng tôi đưa ra ý tưởng về ánh xạ giữa SUMO và WordNet. Việc ánh xạ làm phong phú thêm các tập tin cơ sở dữ liệu WordNet bằng cách gắn thẻ cho mỗi tập từ đồng nghĩa với khái niệm của SUMO tương ứng. Hơn nữa, mối quan hệ được trình bày giữa tập từ đồng nghĩa WordNet và khái niệm của SUMO, tập từ đồng nghĩa WordNet có thể được khai báo là tương đương với thuật ngữ của SUMO, như gộp vào nó, hoặc như là một thể hiện của nó. Ví dụ, tập từ đồng nghĩa {Animal (động vật), Beast (gia súc), Fauna (động vật hoang dã)} được đánh dấu là tương đương với khái niệm Animal của SUMO. Tập từ đồng nghĩa Scavenger (thú vật), có thể được khai báo gộp vào khái niệm Animal (động vật) (nếu có một lớp trong SUMO tương đương với tập từ đồng nghĩa này, nó sẽ là một lớp con của lớp Animal). Tương tự như vậy,

tập từ đồng nghĩa Pythagoras (Pi-ta-go) của WordNet được đánh dấu là một thể hiện của khái niệm Human (con người) của SUMO.

Những tập ánh xạ này cho phép ánh xạ các từ vựng ngôn ngữ tự nhiên thành những thuật ngữ SUMO, việc sử dụng các tập từ đồng nghĩa WordNet như là lớp trung gian.

Ví dụ về một mẫu tin trong cơ sở dữ liệu WordNet, hãy xét trường hợp sau đây:

00047131 04 n 02 accession 0 addition 0 001 @ 09536731 n 0000 | something added to what you have already; "the librarian shelved the new accessions"; "he was a new addition to the staff"

Phần đầu của mẫu tin (bản ghi) cho biết số 00047131 là định danh duy nhất của tập từ đồng nghĩa danh từ {accession, addition}. Phần giữa của mẫu tin ký hiệu "@" và ký hiệu "|" cho biết tập từ đồng nghĩa này được gộp trực tiếp bởi tập từ đồng nghĩa được định danh là 09536731. Tập từ đồng nghĩa thứ hai này tương ứng với nghĩa thu nhận. Thành phần cuối cùng của mẫu tin (bản ghi) ở ví dụ trên (văn bản sau ký hiệu "|") gồm chú giải của tập từ đồng nghĩa và một số ví dụ sử dụng.

4.1 Phương pháp ánh xạ

Chúng tôi chỉ xây dựng các ánh xạ cho các tập từ đồng nghĩa danh từ. Vì thực tế mỗi khái niệm SUMO có dạng của một danh từ và từ thực tế là các tập từ đồng nghĩa danh từ trong WordNet thường có nhiều quan hệ hơn và chứa nhiều thông tin hơn so với các tập từ đồng nghĩa của các loại khác.

Chúng tôi giải quyết trên các mối quan hệ được sử dụng để ánh xạ các các tập từ đồng nghĩa này đến các khái niệm của SUMO. Có ba quan hệ có thể quan tâm: **đồng nghĩa, khái quát và cụ thể**.

Một số ví dụ để làm rõ ba mối quan hệ này và cách thức chúng tôi sử dụng trong việc ánh xạ các khái niệm của SUMO cho các tập từ đồng nghĩa danh từ. Hãy xét ba mẫu tin sau trong cơ sở dữ liệu danh từ WordNet.

00008864 03 n 03 plant (thực vật) 0 flora (hệ thực vật) 0 plant_life (đời sống thực vật) 0 027 @ . . . | a living organism lacking the power of locomotion (sinh vật sống thiếu khả năng di chuyển)

Khi tập từ đồng nghĩa này là đồng nghĩa với khái niệm 'Plant' của SUMO, mẫu tin WordNet được thêm vào như sau:

00008864 03 n 03 plant 0 flora 0 plant_life 0 027 @ . . . | a living organism lacking the power of locomotion &%Plant=

Tiền tố '&%' cho biết thuật ngữ được lấy từ ontology SUMO, và hậu tố '=' cho biết quan hệ ánh xạ là đồng nghĩa.

Bây giờ chúng ta hãy xét trường hợp tập từ đồng nghĩa danh từ được ánh xạ tới khái niệm SUMO, khái niệm này có nghĩa rộng hơn (khái quát) so với tập từ đồng nghĩa. Hãy xét ví dụ mẫu tin sau đây trong tập tin danh từ WordNet.

04719796 09 n 01 Christian_Science (sự hiểu biết về Cơ Đốc giáo) 0 001 @ 04718274 n 0000 | religious system based on teachings of Mary Baker Eddy emphasizing spiritual healing (hệ thống tôn giáo dựa trên những lời giáo huấn của Mary Baker Eddy về phương pháp chữa trị bằng tâm linh)

Trong trường hợp này, không có thuật ngữ trong SUMO tương đương về nghĩa với 'Christian_Science'. Tuy nhiên, SUMO có chứa khái niệm tổng quát hơn 'Religious Organization' (tổ chức về tôn giáo). Theo đó, chúng tôi thêm chú thích '&% ReligiousOrganization+' ở cuối mẫu tin của WordNet đối với 'Christian_Science'. Lưu ý rằng hậu tố '+' cho biết khái niệm này là từ khái quát của tập từ đồng nghĩa có liên quan.

Cuối cùng, mối quan hệ ánh xạ được sử dụng trong trường hợp này thuộc loại cụ thể. Mối quan hệ này cho biết khái niệm được biểu thị bằng tập từ đồng nghĩa danh từ WordNet là một thành phần của lớp được biểu thị bởi khái niệm SUMO. Xét ví dụ mẫu tin sau trong cơ sở dữ liệu danh từ WordNet.

00034393 04 n 02 Underground_Railroad 0 Underground_Railway 0 001 @ 00032687 n 0000 | abolitionists secret aid to escaping slaves; preCivil War in US (những người chống chế độ nô lệ bí mật giúp đỡ để thoát khỏi cảnh nô lệ; trước cuộc nội chiến ở Mỹ).

Trong trường hợp này, khái niệm SUMO liên quan trực tiếp là 'Organization' (tổ chức). Tuy nhiên, mối quan hệ này không phải là một sự tương đương về nghĩa, cũng không phải là sự gộp về nghĩa. Underground_Railway là một tổ chức cụ thể. Từ thực tế này, chúng tôi bổ sung chú thích "&%Organization@" vào cuối của mẫu tin đối với 'Underground_Railway'.

4.2 Ví dụ về ánh xạ

Trong thực tế, hầu hết các khái niệm phức tạp ở tầng cao trong cơ sở dữ liệu danh từ tìm thấy sự tương đương có sẵn trong SUMO. Xét ví dụ, các mẫu tin danh từ sau đây được thêm vào:

00008019 03 n 06 animal 0 animate_being 0 beast 0 brute 0 creature 0 fauna 0 . . . | a living organism characterized by voluntary movement **&%Animal=**

00008864 03 n 03 plant 0 flora 0 plant_life 0 . . . | a living organism lacking the power of locomotion **&%Plant=**

00009457 03 n 02 object 0 physical_object 0 . . . | a physical (tangible and visible) entity; "it was full of rackets, balls and other objects" **&%Object=**

Tất cả ba trường hợp ánh xạ này rất đơn giản không có vấn đề gì khi danh từ được thêm vào. Tuy nhiên có một số trường hợp khó khăn cần phải kiểm tra chặt chẽ. Hãy xét ví dụ, tập từ đồng nghĩa WordNet đối với khái niệm 'Space'.

00015975 03 n 01 space 0 003 @ 00013018 n 0000 %p 00014887 n 0000 %p 06271859 n 0000 | the unlimited 3-dimensional expanse in which everything is located; "they tested his ability to locate objects in space" (mọi đối tượng đều nằm trong không gian 3 chiều; "chúng được kiểm tra để xác định vị trí ở đâu trong không gian").

Đây là vấn đề khó khăn khi khái niệm này liên quan đến SUMO, vì SUMO không có khái niệm "Space" và làm thế nào khái niệm như vậy sẽ có ích cho các tác vụ về công nghệ tri thức và mô hình dữ liệu. Vấn đề này được xử lý dễ dàng khi chúng tôi xét khái niệm song song với nó là khái niệm "Time" (thời gian), nó được diễn tả bởi khái niệm 'TimeMeasure' trong SUMO. Chúng ta thấy rằng, khái niệm về phép đo thời gian (time measure) có thể được dùng tương tự khi diễn tả khái niệm khoảng cách "Space". Khái niệm về phép đo khoảng cách 'LengthMeasure' của SUMO nói về khía cạnh định lượng của khái niệm "Space", tương tự như khái niệm 'TimeMeasure' nói về khía cạnh định lượng của "Time". Theo đó, chúng tôi đã thêm nó vào mẫu tin của tập từ đồng nghĩa trên với chú thích **"&% LengthMeasure**".

Trường hợp phát sinh khi một khái niệm duy nhất từ SUMO ánh xạ tới hơn một tập từ đồng nghĩa trong WordNet, hoặc ngược lại. Trong một số trường hợp, WordNet có sự phân biệt về ngôn ngữ và sự khác biệt về logic. Xét ví dụ, hai tập từ đồng nghĩa sau đây:

00002086 03 n 04 life_form (sinh vật) 0 organism (sinh vật) 0 being 0 living_thing (sinh vật sống) . . . | any living entity (mọi thực thể sống)

00002880 03 n 01 life (sự sống) 0 002 @ 00002086 n 0000 ~ 05988126 n 0000 | living things collectively (sinh vật sống chung); "the oceans are teeming with life" (các đại dương đầy ắp sự sống).

Hai tập từ đồng nghĩa này về cơ bản là giống nhau, nhưng trường hợp đầu tiên nhấn mạnh sinh vật là một thể hiện của lớp chung đối với những sinh vật sống, trong khi trường hợp thứ hai biểu thị đây là lớp trực tiếp của trường hợp đầu tiên. Mặc dù sự khác biệt này có thể có tầm quan trọng về ngôn ngữ, nhưng nó không có sự liên quan nào đến nhu cầu về công nghệ tri thức. Vì lý do này, cả hai tập từ đồng nghĩa chúng tôi gán chú thích **"&%Organism="**. Ví dụ về một tập từ đồng nghĩa ánh xạ với nhiều hơn một khái niệm của SUMO, xét mẫu tin sau trong WordNet:

00128951 04 n 02 substitution (sự thay thế) 0 exchange (sự trao đổi) 1 004 @ 00125689 n 0000 ~ 00129213 n 0000 ~ 00129804 n 0000 ~ 00129915 n 0000 | the act of putting one thing or person in the place of another: "he sent Smith in for Jones but the substitution came too late to help" (hành động đưa một vật hay một người vào vị trí khác: "ông ta đã đưa Smith vào vị trí của Jones nhưng sự thay thế để trợ giúp đã quá muộn").

Khái niệm của sự thay thế này liên quan đến việc loại bỏ một điều gì đó từ một vị trí cụ thể và đưa một cái gì khác vào cùng vị trí đó. Tuy nhiên, điều này rất khó khăn khi xây dựng các ràng buộc chính xác về thời gian và không gian đối với sự thay thế. Do đó, chúng tôi chỉ đơn giản là thêm vào mẫu tin ở trên với chú thích "&% Removing+&%Putting+".

4.3 Kết quả ánh xạ

Trong bài báo này, ta thấy có ba vấn đề quan trọng phát sinh trong sự kết nối với các ontology.

– Làm thế nào một ontology hình thức có thể sử dụng một cách hiệu quả bởi những người không thành thạo nhiều về logic và toán học?

– Làm thế nào một ontology có thể được sử dụng bởi các ứng dụng tự động (ví dụ như các ứng dụng truy hồi thông tin và xử lý ngôn ngữ tự nhiên) sao cho xử lý văn bản dễ dàng?

– Làm thế nào chúng ta có thể biết khi một ontology được bổ sung?

Các ánh xạ WordNet/SUMO sẽ giúp giải quyết những vấn đề này. Cụ thể, những ánh xạ này có chức năng như là một chỉ số của ngôn ngữ tự nhiên đối với các khái niệm trong ontology, như một cầu nối giữa những khái niệm có cấu trúc và văn bản đa dạng được xử lý bởi một số lượng ngày càng tăng của các ứng dụng, và được xem là "kiểm tra sự hoàn chỉnh" về nội dung của ontology.

Chúng tôi thảo luận lần lượt về ba vấn đề. Trước hết, các ánh xạ giữa WordNet và SUMO có thể được coi là một chỉ số ngôn ngữ tự nhiên đối với SUMO. Vì vậy, chúng tôi đã và đang trong quá trình phát triển một trình duyệt cho phép người dùng nhập vào các danh từ bằng tiếng Anh, và nó sẽ trả lại các khái niệm SUMO có liên quan với các danh từ đầu vào thông qua các tập từ đồng nghĩa WordNet. Bằng cách tương tác với trình duyệt, người dùng có thể xem tất cả các khái niệm SUMO có liên quan đến các thuật ngữ ngôn ngữ tự nhiên, và điều này làm cho nó dễ dàng hơn nhiều đối với các tác vụ về công nghệ tri thức và mô hình dữ liệu của ontology.

Bên cạnh sự thuận lợi cho việc tạo ra tri thức và các yếu tố dữ liệu theo chuẩn của SUMO, các ánh xạ cũng có thể là một nguồn tri thức quan trọng đối với các ứng dụng xử lý ngôn ngữ tự nhiên. Các ánh xạ có thể được sử dụng bởi các ứng dụng để gán các nghĩa có cấu trúc của SUMO đối với văn bản đa dạng. Cách đơn giản nhất để làm điều này là gán mỗi khái niệm SUMO cho mỗi từ có liên quan đến nó thông qua một tập từ đồng nghĩa WordNet. Cách tiếp cận phức tạp hơn có thể sử dụng một số thuật toán để định hướng khả năng phán đoán xác định khái niệm chính xác của SUMO trong một bối cảnh cụ thể. Trong cả hai trường hợp trên, các khái niệm của SUMO có thể sử dụng để tạo ra bản tóm lược tự động hoặc chúng có thể được sử dụng để tìm kiếm ngữ nghĩa.

5 TRÌNH DUYỆT SUMO

Chúng ta biết rằng, SUMO được tạo ra nhằm mục đích tạo cấu trúc của các khái niệm và các tiên đề được rõ ràng và dễ hiểu, nhưng phải mất một khoảng thời gian để cho người dùng có thể làm quen và hiểu biết về nguyên lý của nó. Để tạo thuận lợi cho quá trình này, chúng tôi trình bày một công cụ trực tuyến mới của chúng tôi, đó là trình duyệt SUMO tự thiết kế. Việc sử dụng công cụ này, người dùng có thể duyệt cả hệ thống phân cấp SUMO và WordNet, và điều hướng từ một khái niệm SUMO tương ứng với tập các từ đồng nghĩa WordNet và ngược lại.

Trình duyệt được thiết kế và trình bày nội dung của SUMO một cách thân thiện, vì vậy nó dễ hiểu ngay cả đối với những người dùng không chuyên. Một trong những tính năng đặc biệt của nó là diễn giải ngôn ngữ tự nhiên của các tiên đề, được mô tả chi tiết trong Phần 6. Hơn nữa, mặc dù việc thực hiện SUMO chỉ là một danh sách các tiên đề nội bộ, trình duyệt hiểu và giải thích một số các tiên đề và hiển thị chúng một cách thích hợp.

Trình duyệt có thể hiển thị thông tin về một khái niệm (lớp) tại một thời điểm. Nó hiển thị các thông tin sau:

– Phần ontology mà lớp phụ thuộc

– Danh sách các lớp con gần nhất

– Danh sách tất cả các lớp cha

– Danh sách các thể hiện

– Danh sách các thuật ngữ kết hợp (các lớp con khác của lớp cha gần nhất)

– Từ tương đương của các tập từ đồng nghĩa WordNet

– Mọi liên quan đến các tập từ đồng nghĩa WordNet (trên trang riêng)

– Danh sách các tiên đề có liên quan, cả về logic và diễn giải trong ngôn ngữ tự nhiên.

Ví dụ về khái niệm Animal (động vật) có thể được hiển thị trong Hình 4. Khi trỏ tới vị trí khái niệm thích hợp, khái niệm này được chuyển thành liên kết siêu văn bản để dẫn đến khái niệm SUMO hoặc tập từ đồng nghĩa WordNet khác.

Một tiên đề có liên quan đến khái niệm Animal (động vật) được hiển thị trên trang. Chúng ta có thể thấy các tiên đề được trình bày, vừa ngôn ngữ tự nhiên và vừa ký hiệu logic (Hình 4).

Ngoài ra, trình duyệt cho phép điều hướng thông qua hệ thống phân cấp lớp con và lớp cha của các khái niệm, tương tự như quản lý tập tin cho phép điều hướng một hệ thống phân cấp các thư mục trên một ổ đĩa. Chúng ta có thể thấy ví dụ về sự phân cấp xem trong Hình 5.

Để truy cập vào nội dung của SUMO hoặc WordNet bằng cách tìm một khái niệm cụ thể hoặc từ vựng tiếng Anh được cho trong văn bản, nếu một từ tiếng Anh được tìm kiếm, một danh sách tương ứng của tập các từ đồng nghĩa WordNet được hiển thị. Tập các từ đồng nghĩa này thể được điều hướng cùng các con trỏ ngữ nghĩa WordNet,

hoặc một khái niệm SUMO tương ứng có thể được hiển thị.

Các trình duyệt ontology trực tuyến khác, chẳng hạn như Ontolingua [2], Chimaera [5], hoặc trình duyệt SUMO nguyên bản, đều chọn sự đơn giản, đồng nhất về lôgíc và sự thân thiện đối với người dùng. Chúng cố gắng tránh sử dụng các mô hình tưởng tượng không theo thể thức nào cho các tiên đề khác biệt. Trình duyệt của chúng tôi, thường xuyên sử dụng các mô hình không theo thể thức như vậy. Việc làm như thế, thường tạo ở đầu ra sự dễ đọc và rõ ràng. Như vậy, nó đòi hỏi thời gian nhiều hơn trong việc thiết kế cũng như sự thực hiện trình duyệt của chúng tôi.

Một ví dụ về mô hình tưởng tượng không theo thể thức nào là sự tưởng tượng với loại ràng buộc của các chức năng hoặc các mối quan hệ. Ví dụ, thay vì hiển thị các tiên đề (domain ListOrderFn 1 List), (domain ListOrderFn 2 PositiveInteger) and (range ListOrderFn Entity) ở dạng thô của chúng, chúng được giải thích và hiển thị như là một biểu thức đơn:

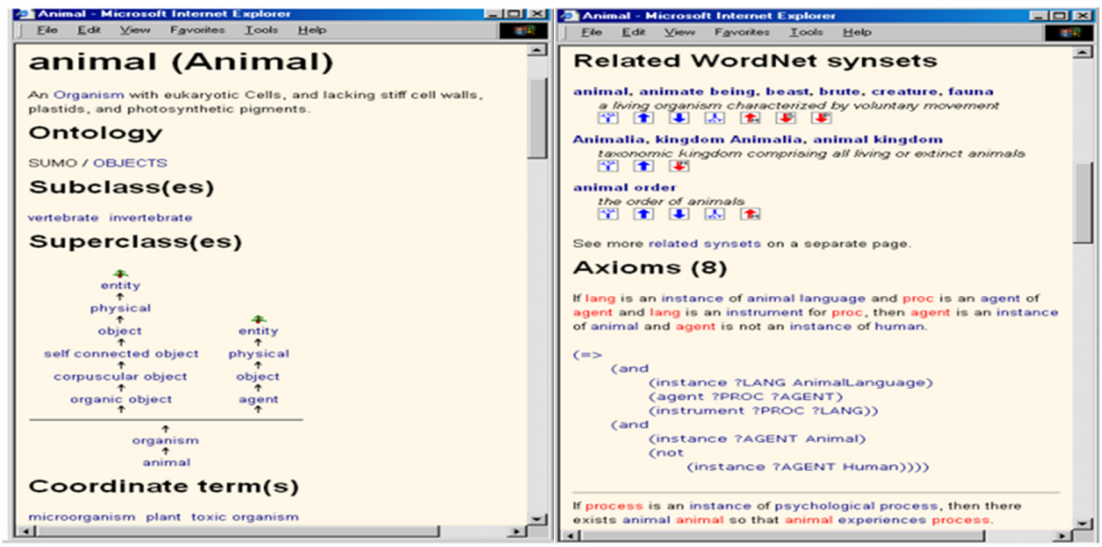
Entity ListOrderFn(List, PositiveInteger).

Mã nguồn của trình duyệt có sẵn đầy đủ. Điều này cho phép người sử dụng hoặc cài đặt nó trên máy tính, hoặc sử dụng lại mã nguồn để truy cập vào SUMO hoặc WordNet, như được mô tả trong phần 7.

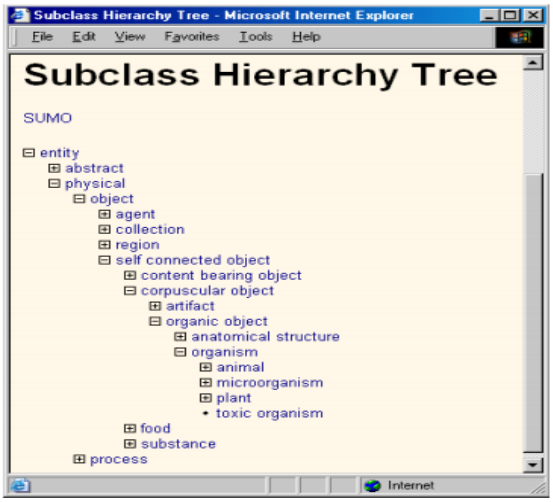
6 TIẾNG ANH DIỄN GIẢI CÁC TIÊN ĐỀ

Phần này mô tả ngắn gọn các tính năng diễn giải ngôn ngữ tự nhiên của trình duyệt. Mặc dù tính năng này chỉ đơn thuần là bản dịch các biểu thức logic mã hóa trong ngôn ngữ KIF [4] thành các câu như tiếng Anh, nó có thể trợ giúp rất nhiều cho người dùng mà không cần biết cú pháp của ngôn ngữ KIF. Các diễn giải cũng được phản hồi tốt đối với những người tạo ra các tiên đề.

Các tiên đề SUMO là các biểu thức logic vị từ. Các biểu thức này được mô tả như là cây nội tại, trong đó các nút nội tại là các liên từ và các lượng từ về logic, và các lá cây là các vị từ. Các thuật toán diễn giải là sự đệ quy đơn giản không phức tạp, nó áp dụng luật tốt nhất đối với mỗi nút để hình thành văn bản ngôn ngữ tự nhiên. Ví dụ, biểu thức E1 E2 được dịch là "N(E1) và N(E2)", trong đó N(x) là sự mô tả ngôn ngữ tự nhiên của biểu thức x, nó có thể áp dụng vào thuật toán tương tự. Bên cạnh các luật đơn giản, các kỹ thuật sau đây được sử dụng:



Hình 4: Ví dụ về trang trình duyệt SUMO với khái niệm Animal



Hình 5: Ví dụ nhận xét về hệ thống phân cấp

– Nếu biểu thức đầu vào có chứa một phủ định trong một số nút nội tại, sự phủ định là luôn luôn được phân phối cho nút lá bằng cách sử dụng luật chuyển đổi đơn giản; Ví dụ như, các biểu thức ¬(A∧B) và ¬(∀x:P(x)) được chuyển thành ¬A∨¬B và ∃x:¬P(x) tương ứng. Điều này cải thiện được mức độ rõ ràng của kết quả các câu tiếng Anh.

– Nếu cấu trúc biểu thức quá phức tạp, có thể dẫn đến những diễn giải không rõ ràng. Ví dụ, biểu thức (A∧B) ⇒ (C∨D) được dịch thành

- nếu A và B
- thì C hoặc D

Điều này tạo sự dễ đọc và rõ ràng ở đầu ra ngay cả đối với các biểu thức rất phức tạp.

– Mẫu biểu thức sau đây thường xuyên xảy ra:

∃x,y: instance(x,Person) ∧ instance(y,Animal) ∧ pet(x,y) được chuyển thành sự diễn giải ngắn gọn hơn “Tồn tại Person x và Animal y sao cho x có pet là y” (pet: vật cưng), chứ không phải “Tồn tại x và y sao cho x là một instance ...”.

Vấn đề còn lại là làm thế nào để diễn giải các nút lá của các biểu thức, tức là các vị từ và các hàm chức năng. Việc diễn giải thuật toán dựa trên sự hiện diện của các chuỗi định dạng cụ thể, chúng được nhúng vào trong ontology và xác định ngôn ngữ tự nhiên ở đầu ra của các vị từ và các hàm chức năng. Cụ thể, mối quan hệ định dạng kết hợp mỗi vị từ hay hàm chức năng với chuỗi định dạng của nó. Ví dụ, tiên đề kết hợp vị từ parent với định dạng đọc của nó “(format parent “%2 is %n a &%parent of %1”)”. Trong thời gian diễn giải, từ khóa %1 được thay thế bằng ngôn ngữ tự nhiên ở đầu ra với tham số đầu tiên của vị từ, %2 với tham số thứ hai, và từ khóa %n với từ ‘not’ nếu vị từ đang được kết xuất là từ phủ định, hoặc nếu không sẽ là một chuỗi rỗng. Từ khóa ‘&%’ có nghĩa là từ ‘paent’ được tạo thành một liên kết siêu văn bản dẫn đến khái niệm parent của SUMO.

Cách tiếp cận này đã được mở rộng để hỗ trợ diễn giải đa ngôn ngữ. Nó có thể kết hợp nhiều định dạng hơn cho vị từ đơn, cho mỗi ngôn ngữ khác nhau. Với sự hợp tác của những người tạo ra SUMO, và định dạng cho năm ngôn ngữ đã được phát triển: Tiếng Anh, Đức, Séc, Ý và Tiếng Hin-ddi.

7 SỰ THỂ HIỆN API CỦA TRÌNH DUYỆT SUMO

Như đã đề cập, trình duyệt SUMO phân tán mã nguồn của nó. Một phần của mã nguồn là những thư viện mà các file lập trình có thể truy cập SUMO và WordNet. Phần này trình bày một ứng dụng mẫu của chúng tôi cho việc sử dụng các thư viện này.

Một trong những vấn đề đối với cơ sở dữ liệu WordNet là nó rất lớn, vì nó có chứa các từ vựng ở tất cả các miền mà con người quan tâm. Người ta muốn loại bỏ các khái niệm từ những miền không quan tâm để tạo cơ sở dữ liệu dễ sử dụng hơn, bằng cách lược bớt hệ thống phân cấp WordNet. Việc ánh xạ SUMO-WordNet là một nguồn tài nguyên rất tốt cho tác vụ này, và API của trình duyệt SUMO cho phép chúng tôi thực hiện điều đó chỉ trong vài phút.

Giả sử rằng chúng ta không quan tâm đến các khái niệm về sinh học trong WordNet. Chúng ta có thể lược bỏ tất cả các tập từ đồng nghĩa có liên quan đến khái niệm sinh vật của SUMO hoặc một trong những lớp con của nó. Đoạn mã trong Hình. 6 trình bày cách thực hiện thủ tục này bằng cách sử dụng API của trình duyệt SUMO.

```
01 // tập các từ đồng nghĩa về danh từ của
      WordNET
02 set<SynSet> synsets;
03 // ontology SUMO
04 OntologyInfo ontology;
05
06 // đọc danh từ của cơ sở dữ liệu WordNET (đặc
      biệt, phiên bản được chú giải với các khái
      niệm SUMO)
07 SynSet::ReadDataFile(SynSets, ”noun.dat”);
```

```
08 // đọc ontology SUMO
09 ontology.ParseOntology(”merge.txt”);
10
11 set<SynSet>::iterator i, next;
12 // duyệt qua tất cả các tập từ đồng nghĩa của
      WordNET
13 i = synsets.begin();
14 while (i != synsets.end())
15 {
16 next = i; ++next;
17 // sửa cờ hiệu
18 bool prune = false;
19 // duyệt qua tất cả các thuật ngữ của SUMO kết
      hợp với tập từ đồng nghĩa hiện tại
20 for (list<SynSet::SumoTerm>::iterator k = i ->
      SumoTerms.begin();
21 !prune && k != i -> SumoTerms.end(); ++k)
22 {
23 // hiện tại thuật ngữ SUMO có kết hợp với lớp
   con của “Organism-Sinh vật” không?
24 if (ontology.IsSubclassOf(k -> Concept,
   ”Organism”))
25 // nếu như thế, đánh dấu nó để lược bỏ
26 prune = true;
27 }
28 // xoá tập từ đồng nghĩa nếu cờ hiệu chỉnh sửa
      được thiết lập
29 if (prune)
30 synsets.erase(i);
31 i = next;
32   }
33
34 // ghi các tập từ đồng nghĩa còn lại trở về tập tin
      cơ sở dữ liệu
35 SynSet::WriteDataFile(SynSet, ”noun.dat”);
```

Hình 6: Đoạn mã C++ về việc lược bỏ danh từ của cơ sở dữ liệu WordNet

Mã ở các dòng 1-9 tải danh từ của cơ sở dữ liệu ontology SUMO và WordNet vào trong bộ nhớ. Chu trình chính ở các dòng 14-32 duyệt mỗi danh từ của tập từ đồng nghĩa để lược bỏ. Tập từ đồng nghĩa được lược bỏ nếu bất kỳ các khái niệm SUMO liên quan với nó là một lớp con của khái niệm Organism (Sinh vật) của SUMO (dòng 20-27). Các tập từ đồng nghĩa còn lại được ghi lại vào đĩa ở dòng 35 trong định dạng được thiết kế cho WordNet.

8 KẾT LUẬN

SUMO, WordNet và trình duyệt SUMO đã được trình bày trong bài báo này. SUMO được xem là ontology có tên miền độc lập đối với việc thiết kế các ontology miền. Từ vựng WordNet cung cấp mối liên kết giữa nội dung hình thức được diễn đạt trong SUMO và ngôn ngữ tự nhiên. WordNet cũng được xác nhận tính hữu ích trong việc phát triển các ontology miền, được xây dựng ở tầng cao của SUMO, cũng như trong việc kiểm tra vùng giới hạn của SUMO hoặc SUMO tuân theo các ontology miền. Trình duyệt SUMO tạo điều kiện cho việc trình bày nội dung của SUMO, WordNet, và các ontology miền được dễ dàng sử dụng. Một trong những tính năng duy nhất của nó là việc diễn giải dòng chữ khó đọc về logic của các tiên đề sang ngôn ngữ tự nhiên. Mã nguồn của trình duyệt SUMO có chứa các thư viện để các tập tin lập trình có thể truy cập vào SUMO và WordNet. Chúng có thể sử dụng để thử nghiệm với các nguồn tài nguyên này.

TÀI LIỆU THAM KHẢO

1. Aynaz Taheri and Mehrnoush Shamsfard: Mapping FarsNet to SuggestedUpper Merged Ontology. Springer-Verlag Berlin Heidelberg 2011, 604–613
2. Farquhar, A., Fikes, R., Rice, J.: The Ontolingua Server: a Tool for Collaborative Ontology Construction; Proceedings of the Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop; Banff, Canada, 1996.
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
4. Genesereth, M.: Knowledge Interchange Format. Proceedings of the 2nd International Conference on the Principles of Knowledge Representation and Reasoning (KR-91), 1991
5. McGuinness, D. L., Fikes, R., Rice, J., Wilder, S.: The Chimaera Ontology Environment. Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000). Austin, Texas, 2000.
6. Miller, G.A., Beckwith, R., Fellbaum, C. Gross, D., Miller, K.J.: Introduction to WordNet: an on-line lexical database. In International Journal of Lexicography 3 (4), 1990, pp. 235 - 244.
7. ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps
8. Niles, I., Pease, A.: Toward a standard Upper Ontology. In: 2nd International Conference on Formal Ontology in Information Systems, Ogunquit, Maine (2001).

9. Niles, I., Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: International Conference on Information and Knowledge Engineering, Las Vegas (2003).
10. Pease, A., Niles, I., Li, J.: The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. To appear in Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web.
11. http://projects.teknowledge.com/AAAI-2002/Pease.ps
12. SUMO home page. http://ontology.teknowledge.com
13. SUMO Browser home page. http://virtual.cvut.cz/kifb/en/
14. Suggested Upper Merged Ontology (SUMO),
15. http://www.ontologyportal.org
16. WordNet home page. http://www.cogsci.princeton.edu/~wn/