Tạp chí Tin học và Điều khiển học, T.21, S.3 (2005), 271–279

MỘT CÁCH TIẾP CẬN ĐỐI VỚI PHÉP DỊCH CÁC CÂU TRUY VẤN NGÔN NGỮ TỰ NHIÊN THÀNH DẠNG LOGIC

NGUYỄN KIM ANH

Khoa Công nghệ thông tin, Trường Đại học Bách khoa Hà Nội

Abstract. For most natual language processing tasks, an interpreter that translates the natural language sentences into a semantic representation is significantly more useful than a parser that simply recognizes syntactically well-formed strings. In this paper, we focus on the task of translating database queries directly into an executable logical form.

Tóm tắt. Đối với hầu hết các nhiệm vụ xử lý ngôn ngữ tự nhiên, một bộ thông dịch nhằm dịch các câu ngôn ngữ tự nhiên thành một biểu diễn ngữ nghĩa có ích hơn một bộ phân tích cú pháp chỉ đơn giản nhận biết về cú pháp các xâu đầu vào được thiết lập tốt. Trong bài báo này, chúng tôi tập trung vào nhiệm vụ dịch các câu truy vấn cơ sở dữ liệu thành một dạng logic có thể thực hiện được.

1. GIỚI THIỆU

Giúp máy tính dễ sử dụng hơn, gần gũi với con người hơn là điều mà các nhà lập trình và nghiên cứu máy tính đã, đang và sẽ tiếp tục cố gắng thực hiện. Ngôn ngữ nói là một trong những cách giao tiếp thông dụng và tự nhiên nhất của con người. Để giúp máy tính giao tiếp được với con người thông qua ngôn ngữ nói, chúng ta cần có các thành phần xử lý ngôn ngữ tự nhiên (NLP). Mục đích chính của các thành phần NLP này là từ một câu tiếng Việt (hoặc một ngôn ngữ bất kỳ) sẽ được thông dịch bởi máy tính và một hành động tương ứng sẽ được thực hiện. Do tính mập mờ, đa nghĩa trong ngôn ngữ nói nên cho đến nay, các hệ thống NLP xây dựng được đều bị giới hạn trong một miền nhỏ và chỉ thông dịch được một số loại câu nhất định.

Một lĩnh vực mà các hệ thống NLP có thể áp dụng hiệu quả là các hệ truy vấn cơ sở dữ liệu (CSDL). Lý do là các CSDL thường phủ một miền đủ nhỏ nên những câu truy vấn tiếng Việt về dữ liệu có thể phân tích được bởi một hệ thống NLP. Đối với hầu hết các nhiệm vụ xử lý ngôn ngữ tự nhiên, một bộ thông dịch nhằm dịch các câu đầu vào thành một biểu diễn ngữ nghĩa có ích hơn một bộ phân tích cú pháp chỉ đơn giản nhận biết về cú pháp các xâu đầu vào được thiết lập tốt. Tuy nhiên, theo chúng tôi, một số cách tiếp cận dịch các câu truy vấn ngôn ngữ tự nhiên thành dạng logic trong [2] và [7] chưa khai thác tốt các tri thức miền, cụ thể là các ngữ nghĩa được biết về CSDL mà chúng ta đang xem xét khi giải quyết các nhập nhằng và kiểm tra tính nhất quán của các câu truy vấn đầu vào. Bài báo này đề cập đến một cách tiếp cận cho phép dịch các câu truy vấn ngôn ngữ tự nhiên tiếng Việt thành một dạng logic có thể thực hiện được. Nội dung bài báo được trình bày như sau: phần 2 mở đầu với một số khái niệm cơ bản liên quan đến việc xác định và biểu diễn ngữ nghĩa của

CSDL quan hệ. Mục 3 trình bày một kiến trúc phác thảo của hệ thống làm cơ sở cho phép dịch các câu truy vấn tự nhiên. Mục 4 trình bày các bước chính để dịch các câu truy vấn tự nhiên tiếng Việt thành dạng logic có thể thực hiện được. Cuối cùng, Mục 5 trình bày một số ví dụ minh họa và Mục 6 đưa ra một vài đánh giá và kết luận.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN

2.1. Sơ đồ thực thể - liên kết

Trong thực tế, khi thiết kế cơ sở dữ liệu quan hệ cho một xí nghiệp, chúng ta thường sử dụng một sơ đồ thực thể - liên kết biểu diễn cấu trúc logic tổng thể của CSDL đối với đối tượng này. Các thành phần cơ bản của một sơ đồ thực thể - liên kết là các thực thể, các thuộc tính và các liên kết. Một tập thực thể (gọi đơn giản là thực thể) ký hiệu một tập các xí nghiệp có các tính chất chung và được gán một tên gọi là một danh từ. Các tập thực thể được xác định thông qua một tập các tính chất, được gọi là các thuộc tính, để phản ánh các đặc trưng của tập thực thể. Mỗi một thuộc tính được gán một tên gọi cũng là một danh từ. Một tập liên kết (gọi đơn giản là liên kết) ký hiệu một tập các bộ mà mỗi bộ biểu diễn một sự kết hợp giữa các thực thể được kéo theo bởi liên kết này. Mỗi liên kết được gán một tên gọi là một động từ. Thông thường, ngữ nghĩa của các thực thể, các thuộc tính và các liên kết đã phần nào được phản ánh thông qua tên gọi của chúng. Do vậy, sơ đồ thực thể - liên kết đối với một xí nghiệp có một ý nghĩa quan trọng nhất định đối với bộ phân tích cú pháp cũng như bộ thông dịch ngữ nghĩa để hiểu nghĩa của các câu truy vấn đối với CSDL của xí nghiệp này và đối với chúng tôi, sơ đồ thực thể - liên kết đối với một xí nghiệp có thể được xem như là những tri thức về ngữ nghĩa đã được biết về CSDL mà chúng ta đang xem xét.

2.2. Logic mô tả CIFR [4]

Các logic mô tả là các hệ hình thức cho phép biểu diễn và lập luận trên các lớp đối tượng phức tạp (được gọi là các khái niệm) và các mối quan hệ giữa chúng (thường được biểu diễn bởi các quan hệ hai ngôi và cũng còn được gọi là các vai trò).

Một cơ sở tri thức của logic mô tả gồm có hai thành phần:

+ TBoxes chứa một tập các mô tả khái niệm và biểu diễn cho sơ đồ chung mô hình hóa miền quan tâm.

+ ABoxes là một sự thể hiện bộ phận của sơ đồ này bao gồm một tập các khẳng định liên quan đến các cá thể của các lớp hay các cá thể có quan hệ với nhau thông qua các mối quan hệ giữa chúng.

CIFR là một sự mở rộng khá tự nhiên của CIF với mục đích biểu diễn trực tiếp các quan hệ n - ngôi mà đặc biệt có ý nghĩa trong ngữ cảnh của chúng tôi, biểu diễn các truy vấn đối với một cơ sở dữ liệu quan hệ.

Giả sử chúng ta có một tập hữu hạn các khái niệm nguyên tố ký hiệu bởi A, các vai trò nguyên tố ký hiệu bởi P và các quan hệ n-ngôi ký hiệu bởi $\boldsymbol{R}$. Chúng tôi sử dụng R ký hiệu các vai trò tùy ý, C ký hiệu các khái niệm tùy ý và T là khái niệm đỉnh, $\perp$ là khái niệm đáy, $\sqcap$ là phép giao và $\sqcup$ là phép hợp. Các khái niệm và vai trò được xây dựng phù hợp với cú pháp sau:

$C \leftarrow T|\perp|A|C_1 \prod C_2|C_1 \coprod C_2|\neg C|\forall R.C|\exists R.C|(\leqslant 1P)|(\leqslant 1P^{-}),$

$|(\leqslant 1R[U])\forall R[U].T_1 : C_1, ..., T_m : C_m|\exists R[U].T_1 : C_1, ..., T_m : C_m,$

$R \leftarrow P|R[U,U']|R_1 \coprod R_2|R_1 \circ R_2|R^-|id(C)$.

Ngữ nghĩa của CIFR, như thông thường, được cho thông qua hàm diễn dịch $I = (\Delta^I, \cdot^I)$. Đặc biệt, nếu R là một quan hệ n-ngôi mà tập các r-vai trò của nó là rol $(R) = \{U_1, ..., U_n\}$ thì R^I là một tập các bộ được gán nhãn có dạng $\langle U_1 : d_1, ..., U_n : d_n\rangle$, ở đây $d_1, ..., d_n \in \Delta^I$. Chúng ta viết $r[U]$ ký hiệu giá trị được kết hợp với U-thành phần của bộ r.

Các cấu trúc mới được diễn dịch như sau:

$R[U]^I = \{(d, r) \in \Delta^I \times R^I | d = r[U]\}$

$R[U, U']^I = \{(d, d) \in \Delta^I \times \Delta^I | \exists r \in R^I . d = r[U] \wedge d' = r[U']\}$

$(\leqslant 1R[U])^I = \{d \in \Delta^I |$ton tai nhieu nhat mot $r \in R^I |$sao cho $r[U] = d\}$,

$$(\forall R[U])T_1 : C_1, ..., T_m : C_m)^I = \{d \in \Delta^I | \forall r \in R^I . r[U] = d \rightarrow (r[T_1] \in C_1^I, ..., r[T_m] \in C_m^I)\},$$

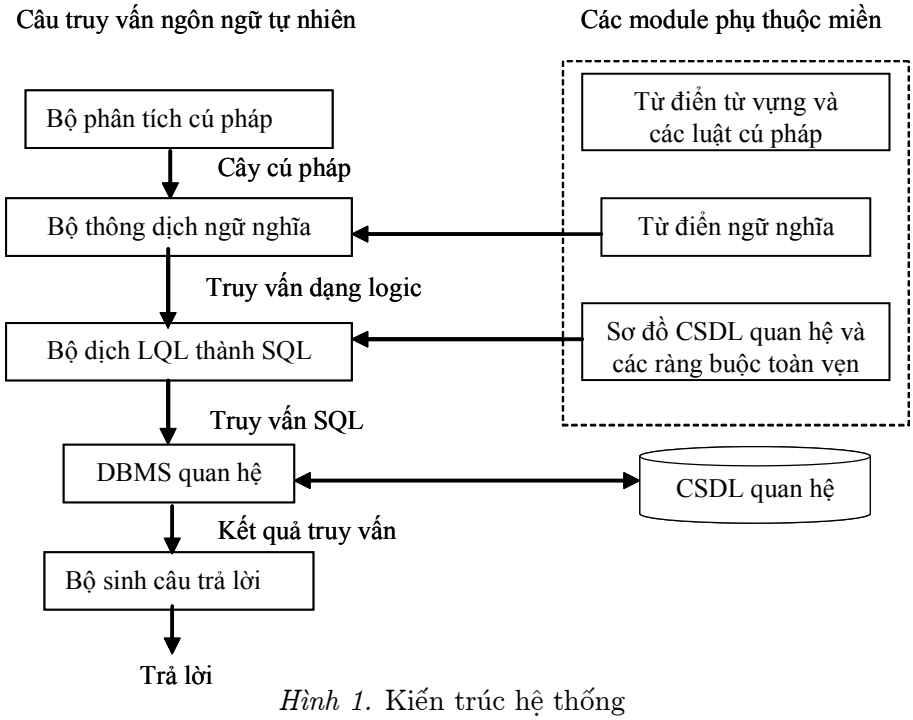
$$(\exists R[U])T_1 : C_1, ..., T_m : C_m)^I = \{d \in \Delta^I | \exists r \in R^I . r[U] = d \wedge r[T_1] \in C_1^I \wedge, ..., \wedge r[T_m] \in C_m^I)\}.$$

CIFR-TBoxes được định nghĩa là một tập hữu hạn các khẳng định bao hàm $C_1 \subseteq C_2$, ở đây C_1, C_2 là các khái niệm tùy ý của CIFR.

CIFR-ABoxes được định nghĩa là một tập hữu hạn các khẳng định $A(a)$ với a là một thể hiện của khái niệm nguyên tố A, các khẳng định $P(a, b)$ với (a, b) là một thể hiện của vai trò nguyên tố P và các khẳng định $R(U_1 : d_1, ..., U_n : d_n)$ với $\langle U_1 : d_1, ..., U_n : d_n.\rangle$ là một thể hiện của quan hệ n-ngôi R.

Tính thỏa mãn của khái niệm cũng như phép kéo theo logic trong CIFR-TBoxes được định nghĩa như thông thường nên chúng tôi không đề cập đến trong phần này nữa.

3. KIẾN TRÚC HỆ THỐNG



Hình 1. Kiến trúc hệ thống

Trong phần này, chúng tôi sẽ trình bày một kiến trúc phác thảo đối với hệ truy vấn ngôn ngữ tự nhiên và một số phân tích liên quan đến phép dịch các câu truy vấn ngôn ngữ tự nhiên thành dạng logic.

Theo kiến trúc trong Hình 1, câu truy vấn ngôn ngữ tự nhiên trước tiên được phân tích bởi bộ phân tích cú pháp. Bộ phân tích cú pháp tham chiếu đến từ điển từ vựng để phân tách các từ có nghĩa trong câu truy vấn tự nhiên, xác định loại từ và từng bước tạo nên cây cú pháp đối với câu truy vấn thông qua một tập các luật cú pháp. Tiếp sau đó, cây phân tích cú pháp kết quả được xử lý bởi bộ thông dịch ngữ nghĩa để hiểu nghĩa của câu truy vấn và sinh ra câu truy vấn dạng logic. Ngôn ngữ được lựa chọn để biểu diễn các câu truy vấn logic phải có khả năng mô tả hay định nghĩa được các tính chất hay các điều kiện trích rút được từ câu truy vấn đầu vào. Chúng tôi đề nghị sử dụng một logic mô tả như một ngôn ngữ trung gian để biểu diễn các truy vấn logic dưới dạng một biểu thức logic mô tả. Tiếp theo, câu truy vấn logic này sẽ được dịch thành một truy vấn SQL mà có thể được thực hiện bởi một phần mềm hệ quản trị CSDL quan hệ nào đó có hỗ trợ SQL. Bộ sinh câu trả lời sử dụng các kết quả của truy vấn SQL để đưa ra câu trả lời cho người sử dụng.

Để giúp cho bộ phân tích cú pháp có thể tạo ra được cây phân tích cú pháp đối với câu truy vấn đầu vào, từ điển từ vựng của hệ cần phải liệt kê đầy đủ tất cả các dạng từ có thể xuất hiện trong các câu truy vấn của người sử dụng. Do vậy, từ điển từ vựng phải chứa tất cả các thông tin được biết về CSDL quan hệ như tên các thực thể, các thuộc tính của thực thể và các liên kết giữa các thực thể được lưu trữ trong CSDL này. Một số những thông tin cần thiết khác như các từ đồng nghĩa của một từ vựng trong từ điển và tất cả các giá trị miền đối với các thuộc tính cũng cần được lưu trữ trong từ điển từ vựng. Ngoài ra, một số các phép toán xuất hiện trong các câu truy vấn như các phép toán so sánh (lớn hơn, nhỏ hơn,....) và các phép toán logic (và, hoặc/hay, không) cũng cần được đưa vào trong từ điển từ vựng. Đặc biệt, từ điển này không chỉ đơn giản lưu trữ thông tin về loại từ (danh từ, tính từ, động từ,...) mà lưu cả các thông tin về tổ hợp các tri thức liên quan đến từ vựng. Nói cách khác, từ điển không chỉ lưu thông tin một cách đơn giản về từ loại mà cả nghĩa loại và các thông tin liên quan khác khi cần thiết. Đây là nguồn dữ liệu chủ yếu về thông tin thuộc tính của từ vựng phục vụ cho mục đích phân tích cú pháp các dạng câu truy vấn có thể từ người sử dụng và cây phân tích cú pháp kết quả cần chứa đầy đủ những thông tin cần thiết giúp cho bộ thông dịch ngữ nghĩa có thể hiểu được nghĩa của câu truy vấn đầu vào.

Từ điển ngữ nghĩa của hệ giúp cho bộ thông dịch ngữ nghĩa có thể hiểu được câu truy vấn đầu vào. Do vậy, ý nghĩa của mỗi từ liệt kê trong từ điển từ vựng cần được mô tả ý nghĩa trong từ điển ngữ nghĩa dưới dạng một vị từ logic. Chúng tôi cũng sử dụng cùng một logic mô tả được lựa chọn biểu diễn câu truy vấn logic để mô tả các từ vựng trong từ điển từ vựng. Bộ thông dịch ngữ nghĩa sử dụng các vị từ logic này để hình thành nên câu truy vấn logic biểu thị ý nghĩa của truy vấn đầu vào. Hai từ điển này và sơ đồ CSDL quan hệ là các module phụ thuộc miền và cần phải được xây dựng mỗi khi hệ thống được cài đặt cho một CSDL mới hay một miền tri thức mới. Tuy nhiên, về cơ bản, hai từ điển này có thể được tạo lập một cách bán tự động với sự can thiệp nhất định từ người quản trị hệ thống.

Qua phần trình bày trên, có thể thấy rằng, trong các nhiệm vụ cần thực hiện khi thiết kế và cài đặt một giao diện truy vấn ngôn ngữ tự nhiên đối với các CSDL, nhiệm vụ thông dịch ngữ nghĩa nhằm dịch các câu truy vấn đầu vào thành một biểu diễn ngữ nghĩa câu truy

vấn là quan trọng hơn cả nhất là khi các truy vấn tự nhiên được phát ra khá tự do từ những người sử dụng không chuyên về tin học, không có những hiểu biết đầy đủ và nhất định về CSDL cần tra cứu. Do vậy, trong phần tiếp theo, chúng tôi chú trọng vào nhiệm vụ dịch các câu truy vấn tự nhiên thành dạng logic có thể thực hiện được.

4. DỊCH CÂU TRUY VẤN TỰ NHIÊN THÀNH DẠNG LOGIC

Để biểu diễn ý nghĩa của các từ vựng trong từ điển từ vựng, chúng tôi sử dụng một logic mô tả đặc biệt, CIFR, đã được giới thiệu trong Mục 2.2.

4.1. Dịch sơ đồ thực thể - liên kết thành CIFR-TBoxes

Trong phần này, chúng tôi sẽ chỉ ra rằng, các ngữ nghĩa được phản ánh trong sơ đồ thực thể - liên kết có thể được nắm bắt trong CIFR thông qua một phép dịch từ sơ đồ thực thể - liên kết thành CIFR-TBoxes.

Cơ sở tri thức CIFR-TBoxes được suy ra từ một sơ đồ thực thể - liên kết S được xác định như sau:

+ Cơ sở tri thức này chứa một khái niệm nguyên tố A đối với mỗi miền giá trị thuộc tính hay mỗi thực thể A, một vai trò nguyên tố P đối với mỗi thuộc tính P và một quan hệ $n+m$-ngôi R đối với mỗi liên kết R n-ngôi (kéo theo n thực thể) có m thuộc tính liên kết. Tập các khẳng định bao hàm của cơ sở tri thức được xác định như sau:

+ Với mỗi cặp các thực thể E, F sao cho E là-một F trong S, chúng ta có khẳng định: $E \subseteq F$ với E và F là các khái niệm nguyên tố ứng với các thực thể E và F.

+ Với mỗi thực thể E có các thuộc tính $A_1, A_2, ..., A_k$ với các miền $D_1, D_2, ..., D_k$ tương ứng, chúng ta có khẳng định:

$$E \subseteq \forall A_1.D_1 \Pi ..... \Pi \forall A_k.D_k \Pi (\leqslant 1 A_1) \Pi ..... \Pi (\leqslant 1 A_k).$$

+ Với mỗi liên kết n-ngôi R giữa n thực thể $E_1, ..., E_n$ có m thuộc tính liên kết $T_1, ..., T_m$ với các miền $D_1, D_2, ..., D_m$, tương ứng, chúng ta có khẳng định:

$$E_i \subseteq \forall R[E_i].T_1 : D_1, ..., T_m : D_m.$$

4.2. Dịch nội dung của CSDL quan hệ thành CIFR-ABoxes

Với mỗi miền D chứa các giá trị rời rạc và hữu hạn, chúng ta có khẳng định: $D(d)$ với d là một thể hiện của D.

4.3. Phân tích cú pháp câu truy vấn ngôn ngữ tự nhiên

Trước tiên, bộ phân tích cú pháp tra cứu từ điển từ vựng để phân tách các từ có nghĩa trong câu truy vấn đồng thời xác định loại từ cũng như những thông tin cần thiết khác có liên quan đến từ được lưu trữ trong từ điển từ vựng. Dựa trên những thông tin này và một tập các luật cú pháp, bộ phân tích cú pháp sẽ xác định cấu trúc cú pháp của câu truy vấn đầu vào và đặc biệt, xác định dạng của câu truy vấn (câu có từ để hỏi, câu nghi vấn, câu mệnh lệnh thức với hàm ý yêu cầu,...). Trong ngữ cảnh của chúng tôi, để bộ phân tích cú pháp có thể phân tích câu truy vấn đầu vào một cách hiệu quả và xác định được dạng câu truy vấn, một tập các luật cú pháp được xây dựng dựa trên việc liệt kê các dạng câu truy vấn có thể từ người sử dụng cần được bổ sụng vào tập các luật cú pháp của hệ thống. Cấu

trúc cú pháp kết quả của câu truy vấn đầu vào sẽ được xử lý tiếp bởi bộ thông dịch ngữ nghĩa.

4.4. Dịch cây phân tích cú pháp thành câu truy vấn logic

Để thông dịch ngữ nghĩa các câu truy vấn tự nhiên đối với một CSDL quan hệ, chúng tôi bổ sung vào tập các khái niệm nguyên tố của CIFR một tập đếm được các kiểu cơ sở , được ký hiệu là B với $B = \{\text{Int, Real, String, Bool, Date}, d_1, d_2\}$, ở đây, các d_k ký hiệu các phần tử của Int $\cup$ Real $\cup$ String $\cup$ Bool $\cup$ Date. Ngoài ra, để thông dịch được các phép so sánh, chúng tôi bổ sung vào tập các vai trò nguyên tố của CIFR một vai trò nguyên tố LớnHơn và định nghĩa các vai trò sau: Bằng = id(Int $\cup$ Real), LớnHơnHoặcBằng = LớnHơn $\coprod$ Bằng, NhỏHơn = LớnHơn, NhỏHơnHoặcBằng = NhỏHơn $\coprod$ Bằng.

Mục đích của bộ thông dịch ngữ nghĩa, cũng như các giao diện truy vấn ngôn ngữ tự nhiên khác, cần phải trích ra những thông tin cần thiết từ câu truy vấn tự nhiên của người sử dụng để có thể phát biểu được một câu truy vấn có thể thực hiện được đối với một CSDL quan hệ. Do vậy, đối với mỗi câu truy vấn tự nhiên, chúng ta cần xác định được ba phần: danh sách các thuộc tính cần đưa ra, các thực thể có liên quan và các điều kiện ràng buộc của truy vấn. Sau khi phân tích cú pháp câu truy vấn, chúng ta đã xác định được dạng câu truy vấn và cấu trúc cú pháp của câu. Tuy nhiên, để hiểu được câu truy vấn này, chúng ta cần biết ý nghĩa của mỗi từ hay cụm từ có nghĩa trong câu truy vấn đầu vào. Mỗi từ hay cụm từ này có thể là tên của một thực thể, một thuộc tính, một liên kết, một giá trị thuộc tính, một phép so sánh hay phép toán logic. Khó khăn của nhiệm vụ thông dịch là những ý định của người sử dụng có thể biểu thị một cách không trực tiếp và tiềm tàng nhiều nhập nhằng trong câu truy vấn đưa vào. Vì vậy, những thông tin trích rút ra được từ câu truy vấn đưa vào có thể là không đầy đủ và không rõ ràng.

Thông thường, việc xác định ý nghĩa của các từ hay cụm từ có nghĩa không phải luôn luôn tầm thường. Trong quá trình thông dịch, bộ thông dịch phải tham chiếu đến từ điển ngữ nghĩa hay sử dụng các tri thức của hệ để xác định ý nghĩa phù hợp nhất đối với cho các từ hay cụm từ .

Danh sách thuộc tính cần đưa ra đối với truy vấn của người sử dụng phụ thuộc vào những thuộc tính nào người sử dụng muốn đưa ra trong câu trả lời. Hệ thống của chúng tôi xác định hai loại truy vấn vào: loại thứ nhất có danh sách thuộc tính đưa ra tường minh và loại thứ hai không có danh sách thuộc tính đưa ra.

1. Các dạng câu truy vấn thuộc loại thứ nhất:

+ Câu truy vấn với động từ hàm ý 'yêu cầu' đứng đầu. Các động từ hàm ý 'yêu cầu' như: liệt kê, cho biết, đưa ra,... luôn luôn được sử dụng để phát ra một yêu cầu tra cứu thông tin trong một CSDL. Các động từ này không chuyển hành động. Do vậy, chúng ta không cần kết hợp nó với một vị từ biểu diễn liên kết nào cả. Tuy nhiên, những danh từ đứng ngay sau những động từ này cho phép xác định được những thuộc tính hay thực thể cần đưa ra trong câu trả lời cho người sử dụng.

+ Câu truy vấn có từ để hỏi. Bộ thông dịch xác định các thuộc tính được nhắc tới bởi người sử dụng, cụ thể đó là những thuộc tính có xuất hiện trong câu đưa vào nhưng không tham gia vào việc xác định các điều kiện truy vấn. Những thuộc tính này sẽ được đưa vào danh sách thuộc tính cần đưa ra một cách tự động. Chúng tôi cho rằng, các thuộc tính được

sử dụng trong các điều kiện truy vấn đã có giá trị và những thuộc tính không thuộc điều kiện truy vấn cần phải tra cứu và cần được đưa ra.

2. Các dạng câu truy vấn thuộc loại thứ hai: Đối với các câu truy vấn thuộc loại này, hệ sẽ đưa ra các thuộc tính ngầm định đối với thực thể được nhắc tới bởi người sử dụng. Có thể thấy rằng, phần thông tin cần thiết và quan trọng nhất cần trích rút từ câu truy vấn đầu vào là các điều kiện ràng buộc của truy vấn do các thực thể có liên quan đã được xác định tiềm ẩn khi biểu diễn điều kiện ràng buộc của truy vấn.

Động từ đóng một vai trò rất quan trọng trong việc thông dịch ngữ nghĩa câu truy vấn. Do vậy, bộ thông dịch ngữ nghĩa của chúng tôi lấy động từ làm thành phần trung tâm của câu để xác định các chủ thể và đối tượng có liên quan đến động từ hay xác định các vai trò ngữ nghĩa của động từ. Tiếp theo, chúng ta sẽ lần lượt xét các động từ xuất hiện trong cây phân tích cú pháp để xác định điều kiện ràng buộc của truy vấn.

a. Động từ đang xét tương ứng với một quan hệ n-ngôi trong cơ sở tri thức của hệ.

+ Dịch động từ thành quan hệ n-ngôi tương ứng đồng thời xác định các cụm danh từ hay mệnh đề đóng vai trò chủ thể và đối tượng của động từ tương ứng với các thực thể tham gia vào liên kết được biểu diễn bởi quan hệ n - ngôi này. Ngoài ra, đoạn trạng ngữ của câu sẽ được dịch thành thuộc tính của liên kết đã xác định.

+ Dịch các danh từ đứng đầu cụm danh từ thành các khái niệm tương ứng với các thực thể đã xác định. Thông thường, các cụm danh từ này có thể có một số định ngữ bổ nghĩa cho danh từ đứng đầu. Các định ngữ này có thể là các danh từ, tính từ hay một mệnh đề. Tra cứu từ điển ngữ nghĩa đối với các danh từ và tính từ, chúng ta có thể xác định được các định ngữ này tương ứng với thuộc tính và/hay giá trị thuộc tính của thực thể đã xác định. Dịch các định ngữ này thành các ràng buộc giá trị đối với các vai trò tương ứng.

Trường hợp bổ ngữ của động từ hay định ngữ của danh từ là mệnh đề sẽ được xét tiếp sau đây.

b. Động từ đang xét không tương ứng với một quan hệ n-ngôi trong cơ sở tri thức của hệ. Trong trường hợp này, động từ đang xét không chuyển hành động mà chỉ mô tả một thuộc tính của thực thể đóng vai trò chủ thể của động từ và bổ ngữ của động từ có thể danh từ/tính từ hay một cụm danh từ.

+ Bổ ngữ là danh từ/tính từ: Tra cứu từ điển ngữ nghĩa để xác định danh từ/tính từ này là giá trị của thuộc tính nào đối với thực thể đã xác định. Dịch bổ ngữ này thành ràng buộc giá trị đối với các vai trò tương ứng

+ Bổ ngữ là cụm danh từ: Dịch danh từ đứng đầu thành vai trò tương ứng với một thuộc tính của thực thể đã xác định và dịch định ngữ của danh từ đứng đầu thành ràng buộc giá trị đối với vai trò đã xác định (có thể thông qua một phép toán so sánh).

c. Dịch các từ chỉ quan hệ 'và', 'hoặc/hay' thành các phép toán $\prod, \coprod$ trong CIFR tương ứng.

d. Dịch từ chỉ phủ định 'không' thành phép $\neg$ trong CIFR tương ứng.

e. Các dạng truy vấn đặc biệt:

+ Câu truy vấn ở dạng nghi vấn với động từ 'là': Động từ 'là' là động từ liên hệ thường được sử dụng để mô tả chủ thể của động từ. Các động từ này không chuyển hành động. Do vậy, chúng ta không cần kết hợp nó với một vị từ biểu diễn liên kết nào cả.

+ Câu truy vấn ở dạng sở hữu: chú ý rằng, một số liên kết có thể được truy vấn dưới dạng sở hữu như 'hãy liệt kê các A của B'. Trong trường hợp này, không có động từ nào

xuất hiện trong câu truy vấn tự nhiên và do vậy không thể xác định được các liên kết cần thiết từ câu đầu vào để dịch câu truy vấn này thành dạng logic có thể thực hiện được. Đối với các câu truy vấn dạng này, trong quá trình dịch, trước tiên, hệ sẽ dịch các cụm danh từ (A và B) thành các khái niệm/vai trò tương ứng. Tiếp sau đó, dựa trên cơ sở tri thức của hệ, hệ sẽ suy diễn để thiết lập các đường dẫn có thể giữa các khái niệm/vai trò này và cho phép mô tả chi tiết câu truy vấn logic có thể thực hiện được. Trong trường hợp tồn tại nhiều đường dẫn có thể, người sử dụng sẽ được hỏi để khẳng định một đường dẫn phù hợp với ý định của người sử dụng.

5. MỘT SỐ VÍ DỤ MINH HỌA

Câu 1. Hãy đưa ra tên các sinh viên ở Hà Nội và sinh sau năm 85: TênSV là thuộc tính cần đưa ra và biểu thức logic mô tả là: $\exists$ ĐịaChỉ$\circ$HàNội $\prod \exists$ NămSinh$\circ$LớnHơn$\circ$85.

Câu 2. Cho biết các giảng viên chỉ dạy môn Cơ sở dữ liệu hay môn Hệ quản trịCSDL:

$\forall$ Dạy[GV, MônHọc]$\circ$TênMôn$\circ$(Cơ sở dữ liệu$\coprod$Hệ quản trị CSDL).

Câu 3. Cho biết các sinh viên của giảng viên A:

Hệ xác định được 2 đường dẫn giữa SV và GV là:

1. SV ⟨Học⟩ MônHọc ⟨Dạy⟩ GV và
2. SV ⟨HướngDẫn⟩ GV.

Người sử dụng sẽ được hỏi để lựa chọn đường dẫn phù hợp:

1. $\exists$ Học[SV, MônHọc]$\circ$Dạy[MônHọc, GV]$\circ$TênGV$\circ$A.
2. $\exists$ HướngDẫn [SV, GV]$\circ$ TênGV$\circ$ A

6. ĐÁNH GIÁ VÀ KẾT LUẬN

Chúng tôi đã tiến hành cài đặt thử nghiệm một hệ truy vấn ngôn ngữ tự nhiên tiếng Việt đối với CSDL Quản lý học tập một khoá của trường Đại học Bách khoa Hà Nội. Hệ thống cài đặt đã đáp ứng được các yêu cầu và mục tiêu đề ra đối với một hệ thống truy vấn ngôn ngữ tự nhiên. Tuy nhiên, hiệu quả của hệ thống phụ thuộc rất nhiều vào vốn từ vựng mà ta đưa vào. Đây chính là khó khăn lớn nhất và cũng là vấn đề cơ bản của bất kỳ hệ thống xử lý ngôn ngữ tự nhiên nào - sự hiểu biết của nó về CSDL cụ thể.

Theo đánh giá của chúng tôi, cách tiếp cận dịch các câu truy vấn tự nhiên tiếng Việt được giới thiệu trong bài này thành một biểu thức logic mô tả là rất có triển vọng. Câu truy vấn ở dạng logic này khá tự nhiên và rất gần với câu truy vấn tự nhiên. Hơn nữa, sử dụng khả năng lập luận của hệ logic mô tả, chúng ta có thể dịch được các truy vấn không đầy đủ thông tin, không rõ ràng, kiểm tra tính nhất quán của câu truy vấn đầu vào và đặc biệt có thể áp dụng các kỹ thuật tối ưu hoá về ngữ nghĩa đối với các câu truy vấn phức tạp. Cách tiếp cận này đặc biệt phù hợp với các truy vấn tra cứu thông tin về một khái niệm - một dạng truy vấn phổ biến đối với các hệ CSDL quan hệ.

Cuối cùng, chúng tôi hy vọng rằng hệ thống cài đặt sẽ được cải tiến và phát triển hoàn thiện hơn nữa để đáp ứng đầy đủ các yêu cầu của một hệ truy vấn ngôn ngữ tự nhiên tiếng Việt và thực sự cho phép những người sử dụng không được đào tạo về Tin học có thể khai thác tốt các CSDL.

TÀI LIỆU THAM KHẢO

[1] S. Abiteboul, R. Hull, IFO, A formal semantic database model, *ACM TODS* **12** (4) (1987) 525–565.

[2] I. Androutsopoulos, "Interfacing a natural language front-end to relational database", Tech. Paper no.11, Dept.of AI, Univ. of Edingburgh, 1993.

[3] D. Calvanese, M. Lenzerini, D. Nardi, *Logics for Databases and Information Systems*, Kluwer, 1998.

[4] G. D. Giacomo, M. Lenzerini, Description logic with inverse roles, functional restrictions, and n-ary relations, *Proc. of the 4th European Workshop on Logic in AI* (1994) 332–346.

[5] G. G. Hendrix, et.all, Developing a nature language interface to complex data, *ACM TODS* **3** (3) (1978) 105–147.

[6] J. S. Kaplan, Designing a portable nature language database query system, *ACM TODS* **9** (1) (1984) 1–19.

[7] C. A. Thompson, R. S. Mooney, and L. R. Tang, Learning to parse natural language database queries into logical form, Workshop on Automata Induction, Grammatical Inference and Language Acquisition, 1997.

[8] D. L. Waltz, An English language question answering system for a large relational database, *Comm. ACM* **21** (7) (1978) 526–539.

Nhận bài ngày 18 - 5 - 2005
Nhận lại sau sửa ngày 17 - 10 -2005