Tạp chí Tin học và Điều khiển học, T.24, S.2 (2008), 161–167

PHÉP DỊCH CÁC TRUY VẤN NGÔN NGỮ TỰ NHIÊN THÀNH CÁC TRUY VẤN SQL SỬ DỤNG VĂN PHẠM NGỮ NGHĨA

NGUYỄN KIM ANH

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

Abstract. For our natural language query systems, natural language queries are parsed syntactically and semantically using one semantic grammar. This paper presents a technique to translate the natural language queries into SQL queries and then, a relational DBMS is left to find all answers to the queries with its own special optimization and planning techniques.

Tóm tắt. Đối với các hệ truy vấn ngôn ngữ tự nhiên của chúng tôi, các truy vấn ngôn ngữ tự nhiên được phân tích cú pháp và ngữ nghĩa sử dụng một văn phạm ngữ nghĩa. Bài báo này trình bày một kỹ thuật cho phép dịch các truy vấn ngôn ngữ tự nhiên thành các truy vấn SQL và sau đó, một hệ quản trị cơ sở dữ liệu (DBMS) sẽ tìm tất cả các câu trả lời đối với câu hỏi với các kỹ thuật lập kế hoạch và tối ưu hoá đặc biệt riêng của nó.

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). 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 [2,5]. Lý do là các cơ sở dữ liệu (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 các hệ truy vấn ngôn ngữ tự nhiên của chúng tôi, các câu truy vấn ngôn ngữ tự nhiên được phân tích cú pháp và ngữ nghĩa sử dụng một văn phạm ngữ nghĩa. Bài báo này đề cập đến một kỹ thuật cho phép dịch các câu truy vấn ngôn ngữ tự nhiên thành các truy vấn SQL và sau đó, một DBMS sẽ tìm tất cả các câu trả lời đối với câu hỏi với các kỹ thuật lập kế hoạch và tối ưu hoá đặc biệt riêng của nó. Do vậy, các tính năng của một DBMS mạnh có thể được sử dụng khi trả lời câu hỏi và hệ thống có thể dễ dàng phát triển đối với các cơ sở dữ liệu rất lớn.

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. Phần 3 trình bày một kiến trúc phác thảo của hệ thống truy vấn ngôn ngữ tự nhiên làm cơ sở cho phép dịch các câu truy vấn ngôn ngữ tự nhiên thành các truy vấn SQL. Phần 4 trình bày phép dịch các

câu truy vấn ngôn ngữ tự nhiên thành các truy vấn SQL. Cuối cùng, phần 5 đưa ra một vài đánh giá và kết luận.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN

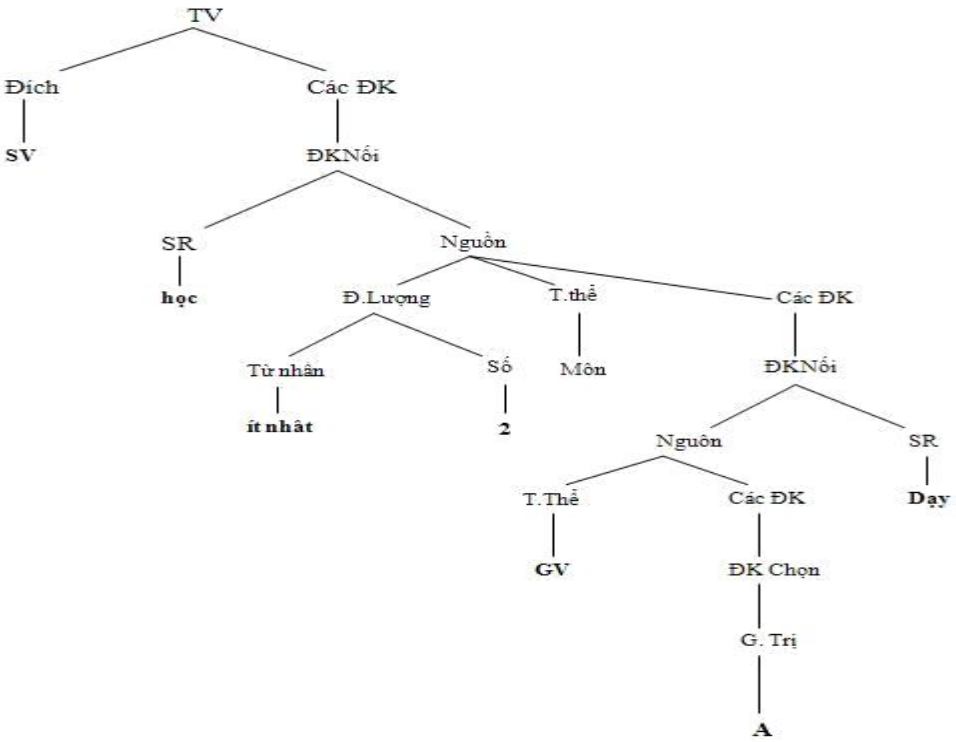
2.1. Sơ đồ thực thể-liên kết

Trong thực tế, khi thiết kế cơ sở dữ liệu (CSDL) 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 xí nghiệp 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. 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 phân tích 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. Đồng thời, sơ đồ thực thể-liên kết này cũng được sử dụng để ánh xạ vào mô hình dữ liệu quan hệ đối với xí nghiệp khi thiết kế cơ sở dữ liệu (CSDL) quan hệ cho xí nghiệp này. Để đơn giản, chúng tôi giả thiết tên của các quan hệ và tập thuộc tính của chúng được đặt trùng với các tên gọi tương ứng trong sơ đồ thực thể-liên kết được sử dụng khi thực hiện ánh xạ.

2.2. Văn phạm ngữ nghĩa (Semantic Grammar -SG)

Về mặt cú pháp, mỗi câu truy vấn của người sử dụng đều có một dạng văn phạm, mỗi vị trí của một từ hay cụm từ trong câu truy vấn có một phạm trù nhất định, phạm trù đó có thể là đối tượng mà người sử dụng cần hỏi, thuộc tính của đối tượng, giá trị mà người sử dụng đưa vào hay một liên kết giữa các thực thể. Việc xác định tập các luật cho phép sản sinh ra tập các mẫu truy vấn hạn chế là có thể và được gọi là văn phạm ngữ nghĩa. Trong [3, 4], các tác giả đã xây dựng một tập các mẫu truy vấn theo SG cố định và phụ thuộc miền ứng dụng. Ở đây, chúng tôi chỉ xây dựng tập luật SG cho một ngôn ngữ truy vấn tự nhiên hạn chế với các truy vấn có dạng ' [Hãy] đưa ra/tìm/liệt kê/cho biết/' hay có dạng 'Ai/Cái gì/?' có nghĩa là các truy vấn có dạng mệnh lệnh thức hay có từ để hỏi đặt ở đầu câu. Theo chúng tôi, ngôn ngữ truy vấn tự nhiên hạn chế này là đủ mạnh để bao phủ một lượng lớn các truy vấn thường gặp ở người dùng. Các kỹ thuật biến đổi các truy vấn tự nhiên không hạn chế thành các truy vấn tự nhiên hạn chế không thuộc phạm vi của bài báo này. SG là một văn phạm phi ngữ cảnh, do vậy, chúng tôi đã sử dụng thuật toán CYK- một thuật toán phân tích cú pháp cho văn phạm phi ngữ cảnh để phân tích cú pháp các câu truy vấn ngôn ngữ tự nhiên.

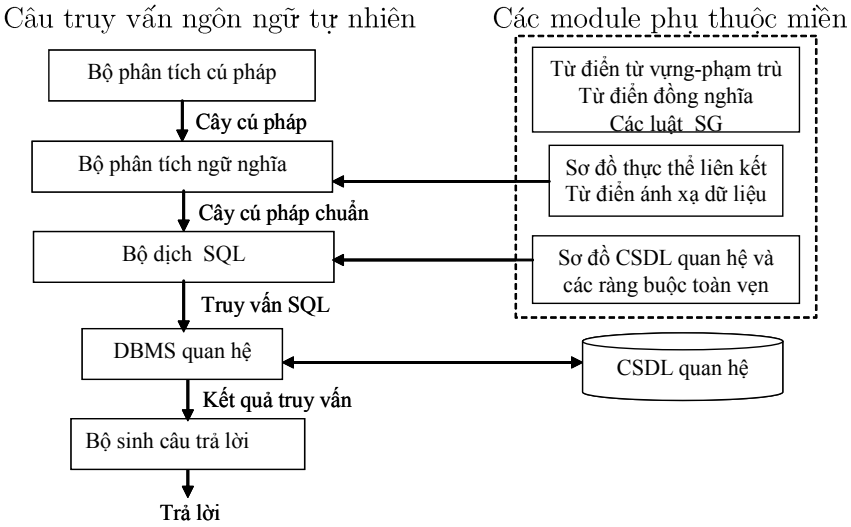
Ví dụ 1. Câu truy vấn 'Tìm các sinh viên học ít nhất 2 môn do giảng viên A dạy' có cây cú pháp như trong Hình 1.



Hình 1. Cây cú pháp của câu truy vấn ngôn ngữ tự nhiên

3. 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 các câu truy vấn SQL.



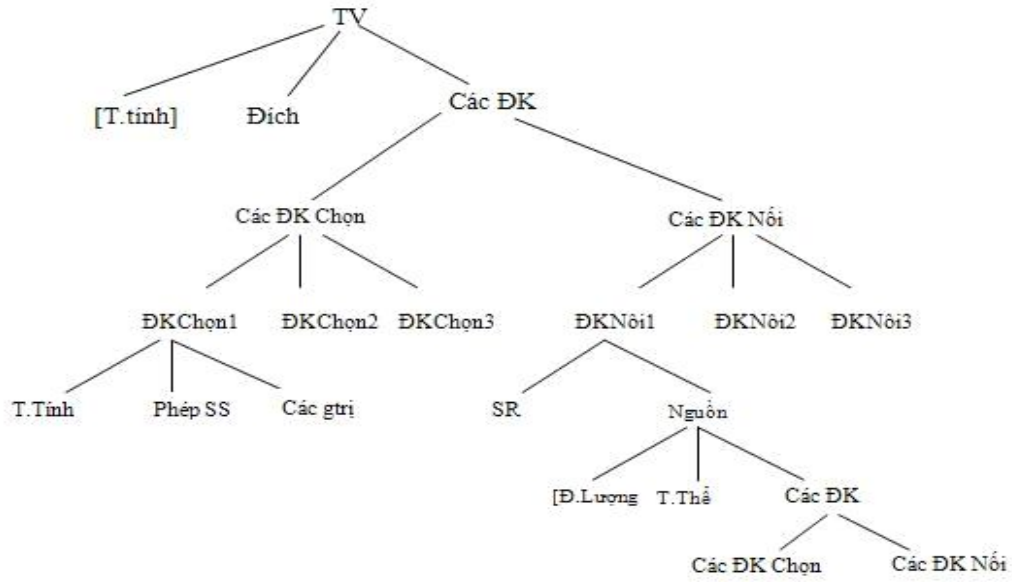
Hình 2. Kiến trúc hệ thống

Qua kiến trúc hệ thống trong Hình 2, có thể thấy rằng, bộ phân tích cú pháp, sử dụng các từ điển và SG, chỉ xác định được cấu trúc cú pháp của câu truy vấn đầu vào. Thực tế, các câu truy vấn tự nhiên thường thiếu thông tin hay không hoàn chỉnh về ngữ nghĩa. Sử dụng sơ đồ thực thể liên kết và từ điển ánh xạ dữ liệu- một từ điển cung cấp các thông tin về cấu trúc của CSDL với tên các bảng, các thuộc tính và các vai trò ngữ nghĩa (Semantic Role-SR), bộ phân tích ngữ nghĩa tiến hành bổ sung hoàn chỉnh và chuẩn hoá cây cú pháp thu được từ bộ phân tích cú pháp [1]. Cây cú pháp chuẩn kết quả của bộ phân tích ngữ nghĩa rất thuận lợi cho nhiệm vụ dích sang truy vấn SQL. Trong phần tiếp theo, chúng tôi chú trọng vào nhiệm vụ dịch các cây cú pháp chuẩn thành các 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 .

4. DỊCH CÂU TRUY VẤN NGÔN NGỮ TỰ NHIÊN THÀNH TRUY VẤN SQL

4.1. Cây cú pháp chuẩn đối với một câu truy vấn ngôn ngữ tự nhiên

Cây cú pháp chuẩn đối với một câu truy vấn ngôn ngữ tự nhiên thu được từ cây cú pháp sau khi đã bổ sung hoàn chỉnh các thông tin thiếu thông qua thao tác đẩy tất cả các ĐKChọn đối với một thực thể (Nguồn hoặc Đích) lên trước tất cả các ĐKNối đối với thực thể này. Cây cú pháp chuẩn như trong Hình 3 sẽ là đầu vào của bộ dịch SQL. Có thể thấy rằng, cây cú pháp chuẩn không những phản ảnh được cấu trúc cú pháp mà còn phản ảnh được ngữ nghĩa hay ý đồ tra cứu của người sử dụng trong câu truy vấn đầu vào.



Hình 3. Cây cú pháp chuẩn

4.2. Dịch cây cú pháp chuẩn thành truy vấn SQL

Để trả lời cho câu hỏi ngôn ngữ tự nhiên của người sử dụng, chúng tôi cần phái dịch các cây cú pháp chuẩn thành một truy vấn SQL. Chúng tôi xử lý một cách đơn giản mỗi một

nút Nguồn như một cây con, trong đó nút T.Thể và CácĐK sẽ được dịch thành một truy vấn SQL con. Phép dịch được thực hiện từ dưới lên đối với các nút Nguồn của cây. Cuối cùng, dịch nút gốc TV với Đích, chúng ta sẽ thu được một câu truy vấn SQL hoàn chỉnh đối với câu hỏi ngôn ngữ tự nhiên. Do vậy, nếu chúng tôi xây dựng được tập các luật dịch tổng quát đối với các cấu trúc có thể xuất hiện trong Nguồn và TV thì bất kỳ một cây cú pháp chuẩn nào cũng có thể được dịch thành một truy vấn SQL. Trước khi trình bày quá trình dịch, chúng tôi giả thiết:

• Mỗi quan hệ biểu diễn thực thể được bổ sung thêm một khoá đại diện. Ví dụ, trong cơ sở dữ liệu quản lý học tập, chúng tôi thêm các thuộc tính MãSV và MãGV như các khoá đại diện đối với quan hệ SV và GV tương ứng.

• Mỗi quan hệ biểu diễn liên kết được bổ sung thêm một khoá đại diện bao gồm các khoá đại diện của các quan hệ biểu diễn các thực thể được kéo theo bởi liên kết này. Ví dụ, khoá đại diện của HướngDẫn được hình thành bởi việc nhóm MãSV và MãGV thành (MãSV, MãGV).

• Thay các SR bằng các tên liên kết tương ứng SR' dựa vào từ điển ánh xạ dữ liệu.

Phép dịch các cấu trúc trong Nguồn và TV được kí hiệu bởi hàm SL. Các luật dịch tổng quát bao gồm:

1. $SL(S\ SR'\ R)$ = select K_S from SR' where $SR'.K_R$ in (select K_R from R)
2. $SL(S$'phủ định' $SR'R)$ = select K_S from S where K_S not in (select K_S from SR' where $SR'.K_R$ in (select K_R from R))
3. $SL(S$ 'ít nhất 1' $SR'R)$ = select K_S from SR' where $SR'.K_R$ in (select K_R from R)
4. $SL(S$'1'$SR'R)$ = select K_S from SR' where $SR'.K_R$ in (select K_R from R)
5. $SL(S$ 'ít nhất n' $SR'R)$ = select K_S from SR' group by K_S having count$(K_R) \geq n$
6. $SL(S$ 'ít nhất tất cả' $SR'R)$ = select K_S from SR' group by K_S having set(K_R) contains (select K_R from R)
7. $SL(S$ 'nhiều nhất n ' $SR'R)$ = select K_S from S where K_S not in (select K_S from SR' group by K_S having count$(K_R) > n$)
8. $SL(SSR'$ 'chỉ n' $R)$ = select K_S from SR' group by K_S having count$(K_R) = n$
9. $SL(S$ 'chỉ' SR' 'n' $R)$ = select K_S from SR' group by K_S having count$(K_R) = n$
10. $SL(E <$ Các ĐKChọn $><$ Các ĐKNối $>) = SL(SL(E <$Các ĐKChọn$>) <$ ĐKNối 1$>)$ intersect $SL(SL(E <$Các ĐKChọn$>) <$ ĐKNối 2$>)$ intersect...

intersect $SL(SL(E <$Các ĐKChọn$>) <$ ĐKNối $n >)$

(ở đây <Các ĐKNối>=< ĐKNối 1> 'và' < ĐKNối 2> 'và'...'và'< ĐKNối $n >$, các < ĐKNối $i >$ này được dịch theo các luật từ 1 đến 9 phụ thuộc vào cấu trúc tương ứng trong cây cú pháp).

11. SL(E <Các ĐKChọn>) = select K_E from E where ĐKChọn 1 and

ĐKChọn 2 and.....and ĐKChọn m

(ở đây <Các ĐKChọn>=< ĐKChọn 1>'và'< ĐKChọn 2>'và''và'< ĐKChọn $m >$)

12. $SL(<$Ttính$> S)$ = select Ttính from S

Ở đây, K_S, K_R là các khóa của bảng R và S, K_E là khóa của bảng biểu diễn thực thể E(E có thể là đích của truy vấn).

Mệnh đề 1. *Các luật dịch tổng quát ở trên đảm bảo bảo toàn ngữ nghĩa của các cấu trúc trong Nguồn và TV.*

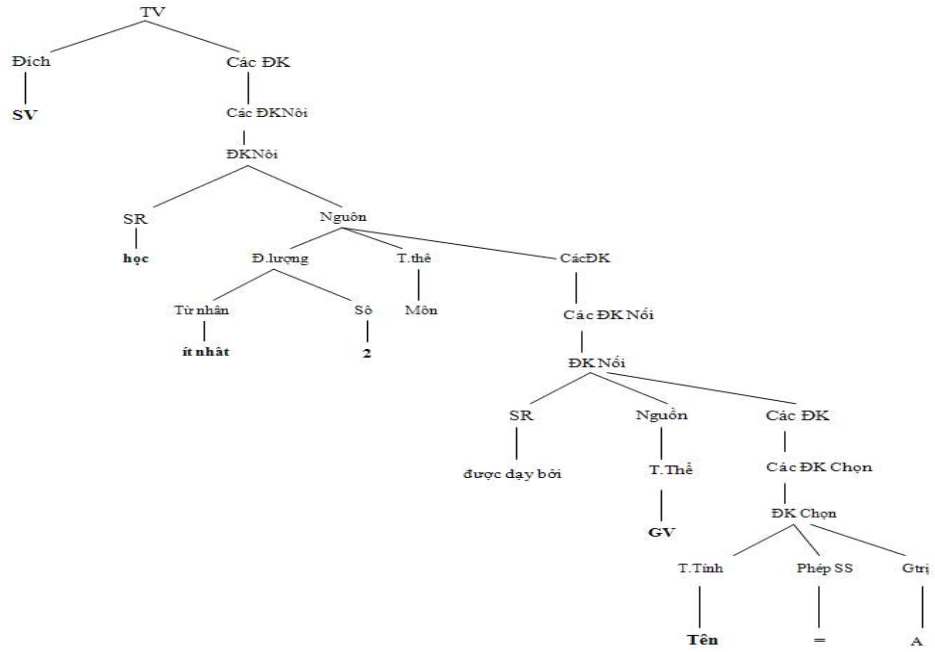
Chứng minh. Có thể dễ dàng kiểm tra ngữ nghĩa của các cấu trúc trong Nguồn và TV là trùng với kết quả của các lệnh truy vấn SQL tương ứng thông qua hàm dịch SL. Chẳng hạn, đối với luật số 1:

$$SL(SSR'R) = \text{select } K_S \text{ from } SR' \text{ where } SR'.K_R \text{ in (select } K_R \text{ from } R).$$

Ngữ nghĩa của cấu trúc trong vế trái là 'lấy ra các đối tượng $S(K_S)$ có liên kết SR với R' trùng với kết quả của lệnh truy vấn SQL ở vế phải 'select K_S from SR' where $SR'.K_R$ in (select K_R from $R)'$, ở đây R có thể là kết quả dịch của một cấu trúc lồng dưới đó.

Chú ý, ngôn ngữ SQL được hỗ trợ bởi các hệ quản trị cơ sở dữ liệu mạnh hiện nay cho phép các truy vấn con có thể đặt sau mệnh đề select, mệnh đề from và có thể lồng nhiều mức trong một truy vấn SQL. ■

Ví dụ 2. Cây cú pháp chuẩn của câu truy vấn trong Ví dụ 1 được cho trong Hình 4.



Hình 4. Cây cú pháp chuẩn đối với câu truy vấn trong Ví dụ 1

5. KẾT LUẬN VÀ ĐÁNH GIÁ

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. 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, hệ thống cài đặt đã đạt được các mục tiêu sau:

• Hệ thống đã cung cấp một giao diện khá thân thiện và thuận tiện cho người sử dụng khi truy vấn.

• Các truy vấn có lượng từ và phủ định rất khó biểu diễn bằng SQL đối với những người

dùng không chuyên nghiệp hoặc không thành thạo SQL đã được phát biểu bằng ngôn ngữ tự nhiên một cách khá tự nhiên và dễ dàng.

• Hệ thống có khả năng đưa ra những trợ giúp để giúp cho người sử dụng đặt câu hỏi đúng với mục đích.

• Với một số loại câu hỏi thì hệ thống có thể tự động sửa được.

• Hệ thống có tính khả chuyển cao. Khi thay đổi miền ứng dụng thì cần thay đổi các từ điển, sơ đồ thực thể liên kết và cơ sở dữ liệu quan hệ của 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] N. K. Anh, P. T. T. Hoài, Truy vấn ngôn ngữ tự nhiên không hoàn chỉnh đối với các cơ sở dữ liệu quan hệ, *Tuyển tập các bài báo khoa học của Hội nghị khoa học lần thứ 20 trường Đại học Bách khoa Hà Nội về Công nghệ thông tin*, Hà Nội, Việt Nam, 2006 (117–122).

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

[3] V. Owei, Enriching the conceptual basis for query formulation through relationship semantics in databases, *Inf.Syst.* **26** (6) (2001) 445–475.

[4] V. Owei, An intelligent approach to handling imperfect information in concept-based natural language queries, *ACM TOIS* **20** (3) (2002) 291–328.

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

Nhận bài ngày 15 - 8 - 2007