Một giải pháp suy diễn câu trả lời trong hệ thống hỏi đáp thông tin

A Solution for Answer Reasoning in QA System

Phan Thị Tươi, Nguyễn Chánh Thành

Abstract: *Question Answering (QA) is an important research field that research groups focus on to develop useful QA system for many languages in the world (English, Chinese, Japanese …) but just few for Vietnamese. One of most difficult processings in QA is the answer reasoning applying to natural language question, especially Vietnamese. The paper introduces an approach of reasoning answers for Vietnamese question based on Graph Theory and Artificial Intelligence (AI). The experimentation, which is done for Vietnamese questions in initial phase, shows that the proposed approach is feasible for Vietnamese and it can be upgraded more for other languages in future.*

Keywords*: reasoning, CG, question answering, QA.*

I. GIỚI THIỆU

Hỏi đáp thông tin (*Question Answering, QA*) là một trong những nhu cầu thiết thực của mọi người dùng trên thế giới. Nhiều phương pháp của các nhóm nghiên cứu trên thế giới về vấn đề này đã và đang được tiến hành với một số kết quả minh chứng tính khả thi của những phương pháp đó.

Một trong những hướng nghiên cứu về QA được thực hiện trong lĩnh vực trí tuệ nhân tạo liên quan đến hệ chuyên gia với một số thành công nhất định.

Ngoài ra, với mong muốn cải tiến các hệ thống QA nhằm tăng khả năng xử lý và chất lượng nội dung trả lời, các nghiên cứu về QA được phát triển theo hướng kết hợp với ngôn ngữ tự nhiên. Hướng nghiên cứu này thường kết hợp với một số lĩnh vực khác như truy xuất thông tin (trực tuyến hoặc không trực tuyến), hệ chuyên gia.

Trong những năm 1960, một số hệ thống hỏi đáp đầu tiên bằng ngôn ngữ tự nhiên được xây dựng như Elisa, Lunar, Baseball. Các hệ thống QA liên tục được hoàn thiện và phát triển cùng với ngành ngôn ngữ học tính toán (*Computational Linguistics*) và hiểu văn bản (*Text Comprehension*) trong thập niên 1970 kế tiếp. *TREC* (trec.nist.gov) chính thức đưa ra các đề tài và các hoạt động nghiên cứu trong những năm 1990 về hệ thống QA. Hiện nay đã tồn tại một số hệ thống QA bằng tiếng Anh, như *AquaLog* [1], *START* [2].

Trong những nghiên cứu về QA, các giải pháp về suy diễn trả lời câu hỏi luôn là một vấn đề khó khăn và phức tạp. Một số nghiên cứu giới hạn lại phạm vi xử lý chủ yếu chỉ dựa trên cơ chế so khớp cấu trúc cú pháp của câu truy vấn với một số mẫu cấu trúc có sẵn, như [1], [3], [4], [5], [6], [7] và [9]. Một số dự án như S-CREAM [11] và MnM [12] hay AquaLog [1] dùng nhiều kỹ thuật học máy để rút trích quan hệ giữa các đối tượng, tuy nhiên chỉ làm được bán tự động. Công trình của IBM Watson giới thiệu môt hướng tiếp cận khác dựa trên ngôn ngữ tự nhiên [10]. Nhóm tác giả [13] và [14] đã dùng đồ thị ý niệm để biểu diễn ngữ nghĩa các quan hệ giữa các thực thể để từ đó xác định các phương án trả lời cho nội dung câu hỏi.

Ngoài ra, một số phần mềm tập trung xử lý suy diễn cho một vấn đề tổng quát[1], từ đó có thể ứng dụng vào hệ thống QA để hỗ trợ suy diễn câu trả lời cho nội dung hỏi của người dùng, như bảng 1, với (*) là dạng mã nguồn mở và (**) không miễn phí.

Hướng nghiên cứu về QA có hỗ trợ truy vấn dạng

[1] *Liệt kê chi tiết trong en.wikipedia.org/wiki/Semantic_reasoner*

ngôn ngữ tự nhiên tiếng Việt, công trình [19] giới thiệu một phương pháp khai thác thông tin trong cơ sở dữ liệu thông qua giao diện người dùng có hỗ trợ truy vấn dạng ngôn ngữ tự nhiên.Tuy nghiên cứu này tập trung về việc phân tích truy vấn theo dạng văn phạm ngữ nghĩa với kết quả chính xác 91.91% và tạo ra cây cú pháp hướng ngữ nghĩa (*syntactics-semantic tree*), nhưng đây cũng là một trong các hướng tiếp cận ảnh hưởng đến kỹ thuật phân tích câu hỏi trong tiếng Việt.

Ngoài ra, kết quả của [20] trình bày một cách tiếp cận dựa trên bản thể học để sinh câu trả lời với kết quả trả lời đúng 95% trên tập 60 câu hỏi tiếng Việt. Phương pháp sinh nội dung ở đây tập trung xử lý việc suy diễn dựa trên các quan hệ ngữ nghĩa có trong bản thể học để đối sánh các ý niệm liên quan đến nội dung nhóm từ nghi vấn.

Công trình [26] giới thiệu một nghiên cứu chuyên sâu về mô hình hệ thống QA hỗ trợ tiếng Việt dựa trên bản thể học. Trong đó giải pháp suy diễn câu trả lời tập trung theo hướng sử dụng quan hệ ngữ nghĩa *is-a* để thực hiện việc tìm kiếm kết quả phù hợp. Kết quả thực nghiệm khả quan của họ cho thấy triển vọng của hướng tiếp cận này trong tương lai. Tuy nhiên, việc suy diễn tự động có hỗ trợ xử lý bắc cầu cho các câu hỏi phức tạp là mục tiêu không những của [26] mà còn là một vấn đề được giới thiệu trong bài báo này.

Cách tiếp cận của [27], tuy không triển khai cho câu hỏi tiếng Việt nhưng hướng đến việc phân tích và biểu diễn truy vấn với cốt lõi là cụm từ theo dạng đồ thị ý niệm, từ đó thực hiện việc suy diễn nội dung trả lời và đặc biệt có hỗ trợ suy luận bắc cầu, đem lại kết quả ở mức 93% và 98% cho độ chính xác và độ bao phủ trên tập thử nghiệm 346 truy vấn. Hướng tiếp cận này có ảnh hưởng lớn đến nghiên cứu của chúng tôi, tuy nhiên việc sinh tự động toàn bộ chuỗi các suy luận nội dung trả lời theo phương án hợp lý cũng là một vấn đề khó mà chúng tôi trình bày hướng giải quyết trong bài báo này.

Công trình [21] đề xuất một phương pháp phi cú pháp phục vụ cho việc hiểu truy vấn ngôn ngữ tự nhiên để hình thành câu hỏi không chuẩn mực. Phương pháp này thực hiện khai thác kiến thức trong một bản thể học để nhận biết các thực thể và xác định mối quan hệ của họ trong một truy vấn, từ đó đơn giản hóa quy tắc chuyển đổi từ các truy vấn ngôn ngữ tự nhiên dựa trên việc liên kết chặt chẽ giữa ngôn ngữ tự nhiên và đồ thị ý niệm.

Các công trình khác khá nổi tiếng về đồ thị ý niệm và các ứng dụng như [22], [23], [24] và [25] đã cung cấp một định hướng tiếp cận hợp lý và chặt chẽ, làm cơ sở cho các vấn đề nghiên cứu liên quan đến đồ thị ý niệm trình bày trong bài báo này

Từ các khảo sát phân tích trên, việc phát triển hệ thống hỏi đáp tiếng Việt trên cơ sở kế thừa có chọn lọc các ưu điểm của những nghiên cứu liên quan, đặc biệt là [1], [14], [21] và [27], để từ đó vận dụng hiệu quả cho ngôn ngữ tiếng Việt trong bài toán hỏi đáp và truy xuất thông tin nhằm hướng đến mục tiêu phát triển được một hệ thống truy vấn thông tin hỗ trợ ngữ nghĩa tốt hơn.

Bài báo này giới thiệu hướng xử lý suy diễn nội dung câu trả lời cho hệ thống hỏi đáp hướng đến ngữ nghĩa có hỗ trợ tiếng Việt dựa trên việc phân tích các phương án suy diễn xác định được từ cơ sở tri thức trên nội dung câu hỏi phân tích được dựa trên nền tảng Văn phạm Phụ thuộc ([16], [17] và [18]).

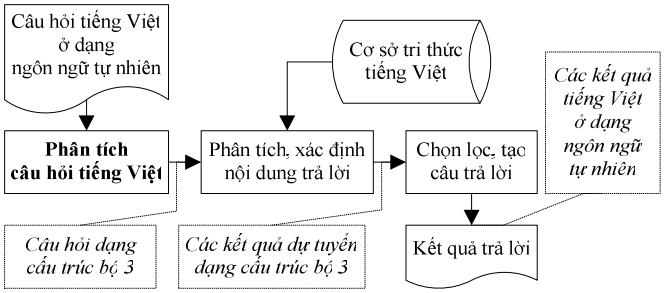
Nội dung bài báo này được trình bày như sau: Mục đầu tiên cung cấp một góc nhìn tổng quan về phương pháp đề xuất và các nghiên cứu liên quan ở trong và ngoài nước. Mô hình tổng quát về hệ thống QA được nêu trong mục II nhằm giới thiệu tiến trình xử lý câu hỏi và trả lời trong hệ thống đề xuất. Mục III trình bày nội dung chính của bài báo, một cách tiếp cận khác về đồ thị ý niệm và phương pháp suy diễn tự động dựa trên giải thuật CGBAR, và từ đó tóm lược kết quả thực nghiệm ở mục IV. Cuối cùng là phần kết luận và định hướng nghiên cứu kế tiếp của chúng tôi.

II. MÔ HÌNH HỆ THỐNG HỎI ĐÁP TIẾNG VIỆT

Về bản chất, hệ thống hỏi đáp tiếng Việt của chúng tôi hoạt động trong lĩnh vực thư viện điện tử nhằm hỗ trợ người dùng tra cứu các thông tin chuyên sâu liên quan đến các tài liệu khoa học kỹ thuật. Hệ thống này được mô tả trong Hình 1 ([17]).

Bảng 1. Tóm lược về các phần mềm hỗ trợ suy diễn công bố trên thế giới

	BaseVISor **	**Bossam ***	**FaCT ***	**FaCT++ ***	**HermiT ***	**Hoolet ***	**Jena ***	**KAON2 ***	**OntoBroker ****	**OWLIM ****	**Pellet ****	**RacerPro ***	**SweetRules ***
Giải thuật suy diễn	Rule-based, Rete	Rule-based	Tableau	Tableau	Hyper-tableau	First-order prover	Rule-based	Resolution & Datalog	OWL: Resolution & Datalog; F-logic: Rule-based (BottomUp, MagicSet, QSQ, DynamicFiltering)	Rule-based	Tableau	Tableau	Rule-based
Phiên bản	2.0	0.9b45	Không biết	1.1.8	1.2.4	Không biết	2.5.4	2008-06-29	6.0	2.x/3.x	2.0.2	2.0	2.1
OWL-DL Entailment	Không	Không biết	Có	Có	Có	Có	Không	Có	Có	Không	Có	Có	Không
Mức độ biểu hiện hỗ trợ cho suy diễn	R-entailment, OWL 2 RL	Không biết	SHIQ	SROIQ (D)	SROIQ (D)	Không biết	Các dạng khác nhau	SHIQ (D)	OWL: SHIQ (D) (for OntoBroker 6.1); F-logic	R-entailment, OWL 2 RL	SROIQ (D)	SHIQ (D-)	Không biết
Kiểm tra tính chắc chắn	Có	Không biết	Có	Có	Có	Có	Chưa xong cho OWL DL	Có	Có	Có	Có	Có	Không
Hỗ trợ DIG	Không	Không	Có	Có	Không	Không	Có	Có	Có	Không	Có	Có	Không
Hỗ trợ luật	Có,tự định dạng	Có, SWRL & tự định dạng	Không	Không	Có, SWRL - DL Safe Rules	Có, SWRL	Có, tự định dạng	Có, SWRL - DL Safe Rules	Có, SWRL, RIF, F-logic, ObjectLogic	Có, tự định dạng	Có, SWRL -DL Safe Rules	Có, SWRL& tự định dạng	Có, SWRL, RuleML, Jess



Hình 1. Mô hình hệ thống hỏi đáp tiếng Việt

Trong hệ thống này, ba khối chức năng chính là bộ phân tích câu hỏi tiếng Việt, bộ phân tích và xác định nội dung trả lời và bộ chọn lọc tạo câu trả lời.

Trong khối đầu tiên, câu hỏi tiếng Việt của người ở dùng dạng ngôn ngữ tự nhiên sẽ được phân tích theo trình tự sau đây:

- Phân đoạn từ và gán nhãn từ loại cho mỗi thành phần tương ứng trong câu hỏi. Bước xử lý này tạo tiền đề cho các tác vụ tiếp theo trong hệ thống.
- Phân tích dạng câu hỏi tương ứng dựa trên từ hỏi và các thông tin thu thập được ở bước xử lý trước đó. Từ đây, câu hỏi ban đầu được tổ chức lưu trữ dạng cấu trúc đồng thời được biểu diễn ở dạng cây phân tích.

Thông qua bước phân tích trong khối này, tùy theo câu hỏi nhập vào của người dùng có dạng đơn giản hay phức tạp mà kết quả xuất ra là một hay nhiều *bộ ngôn ngữ* (*linguistic tuple*) tương ứng. Ở đây khái niệm bộ ngôn ngữ biểu diễn cấu trúc lưu trữ của câu hỏi ban đầu, trong đó các thành phần gồm tác nhân, hành vi và đối tượng, như trong ví dụ <ai?, *viết*, sách AI>. Tùy thuộc vào từng dạng câu hỏi mà một hay các thành phần trong bộ ngôn ngữ này sẽ được biểu thị bằng từ hỏi (nghi vấn) hoặc dấu hỏi (ký hiệu " ? ").

Trong khối kế tiếp, *bộ ngôn ngữ* có thể đóng vai trò "tế bào" cơ bản trong cơ sở tri thức, nên còn được gọi là *bộ tri thức* (*knowledge tuple*) trong cơ sở tri thức, được hệ thống rút trích tương ứng phù hợp với các nội dung và từ hỏi trong *bộ ngôn ngữ*. Khái niệm bộ tri thức được dùng để diễn đạt nhóm thông tin liên quan đến tác nhân, hành vi và đối tượng, như mô tả trong ví dụ <Aho, *là tác giả*, sách Compiler>.

Bộ phận chọn lọc và tạo câu trả lời trong mô–đun cuối cùng giúp hệ thống cung cấp các câu trả lời tiếng Việt ở dạng ngôn ngữ tự nhiên cho người dùng một cách thân thiện hơn và dễ hiểu hơn. Đây cũng là một chức năng dự kiến trong hệ thống hỏi đáp tiếng Việt của nhóm tác giả (như [17] [18]).

III. SUY DIỄN NỘI DUNG TRẢ LỜI

Hướng tiếp cận cho việc suy diễn này được nhóm tác giả đề xuất dựa trên nền tảng lý thuyết đồ thị và bài toán tìm đường đi trên đồ thị (trong lĩnh vực trí tuệ nhân tạo và toán rời rạc). Bước đầu tiên trong hướng tiếp cận này là xây dựng và phát triển một đồ thị phân lớp các ý niệm. Đồ thị này đóng vai trò nền tảng cho phương thức suy diễn tự động ở bước tiếp theo, trong đó giải quyết vấn đề tìm nội dung trả lời cho một câu hỏi dựa trên các khả năng xảy ra tương ứng với các cung đường thích hợp trên đồ thị ban đầu.

1. Đồ thị các phân lớp ý niệm

Hướng nghiên cứu về đồ thị ý niệm (*Conceptual Graph*) đã được phát triển rất mạnh với nhiều nhóm nghiên cứu và công trình liên quan[2]. Trong bài báo này, nhóm tác giả trình bày khái niệm *Đồ thị các phân lớp ý niệm* để diễn đạt dạng đồ thị liên thông có hướng, với tập đỉnh E gồm các lớp ý niệm, tập cạnh V gồm các quan hệ ngữ nghĩa liên quan giữa các đỉnh trong E. Đồ thị này được ký hiệu là G^{cc} = <V, E>.

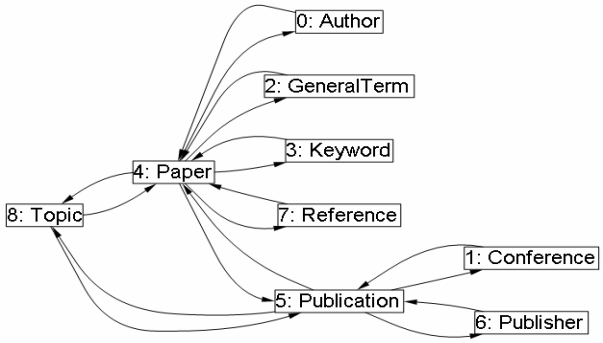
Đồ thị G^{cc} được minh họa trong Hình 2 gồm:

$E=\{e_{0:Author}, e_{1:Conference}, ..., e_{8:Topic}\}$ với tập các phân lớp ý niệm là {Author, Conference, ..., Topic}

$V=\{v_{8:Topic\text{-}4:Paper}, v_{0:Author\text{-}4:Paper}, ..., v_{5:Publication\text{-}6:Publisher}\}$

Đồ thị G^{cc}, trong đó các cạnh $v_{e1\text{-}e2}$ nối giữa đỉnh (đồng thời là lớp ý niệm) e_1 và e_2 được bổ sung thêm yếu tố trọng lượng, là giá trị xác suất điều kiện giữa một ý niệm thuộc lớp e_2 phụ thuộc vào một ý niệm thuộc lớp e_1, sẽ được gọi là đồ thị ý niệm có trọng lượng, ký hiệu là $G^{cc}_w = <V_w, E>$

[2] *Danh sách chi tiết trong en.wikipedia.org/wiki/Conceptual_graph*



Hình 2. Đồ thị các ý niệm phân lớp G^{cc}

Những dạng đồ thị trên có thể được ứng dụng trong các bài toán thuộc những lĩnh vực khác nhau, như xây dựng chỉ mục hướng ngữ nghĩa, hoặc thiết kế cấu trúc cơ sở tri thức, hay phục vụ bài toán tìm kiếm.

Trong đồ thị G^{cc}, với hai đỉnh e_1 và e_2 bất kỳ, luôn tồn tại một đường đi p từ e_1 đến e_2, đó là tập các cung liên kết (cạnh) nối giữa các đỉnh trung gian giữa chúng. Đường đi p này thể hiện một mối quan hệ ngữ nghĩa giữa một phần tử thuộc phân lớp e_1 với một phần tử thuộc phân lớp e_2 thông qua các quan hệ bắc cầu giữa các lớp trung gian.

Với k đỉnh e_1, e_2, …, e_k trong G^{cc}, tồn tại ít nhất một đường đi giữa các đỉnh này dựa trên các trường hợp:

- Ba đỉnh này cùng nằm "thẳng hàng" trên cùng một đường đi;
- Các đỉnh này thuộc những cung liên kết thuộc các đường đi khác nhau;

Ví dụ từ đồ thị trong Hình 3:

- Trường hợp thứ nhất xảy ra trên các đỉnh $e_{0:Author}$; $e_{5:Publication}$ và $e_{6:Publisher}$ vì các đỉnh này nằm trên cùng đường đi p: $e_{0:Author}$-$e_{4:Paper}$-$e_{5:Publication}$-$e_{6:Publisher}$
- Trường hợp thứ hai xảy ra với các đỉnh $e_{0:Author}$, $e_{3:Keyword}$, $e_{7:Reference}$ vì có hai đường đi liên quan p_1: $e_{0:Author}$-$e_{4:Paper}$-$e_{3:Keyword}$ và p_2: $e_{0:Author}$-$e_{4:Paper}$-$e_{7:Reference}$

Tương tự, đường đi p_w trong đồ thị $G^{cc}{}_w$ diễn đạt mối quan hệ ngữ nghĩa có trọng số dựa trên trị xác xuất tích hợp bởi các trị xác suất trung gian.

2. Dạng mẫu trong đồ thị G^{cc}

Từ đồ thị G^{cc} hay $G^{cc}{}_w$, giữa hai đỉnh bất kỳ e_1 và e_2 luôn tồn tại ít nhất một đường đi xuyên qua các đỉnh trung gian theo hướng từ e_1 đến e_2 hay ngược lại. Đường đi này dù theo hướng nào cũng tạo thành chuỗi các quan hệ ngữ nghĩa nối tiếp nhau. Chuỗi này biểu diễn quan hệ giữa các thành phần cốt lõi trong nội dung của một hay nhiều câu trong một văn bản. Dẫn đến, một câu hỏi hay một câu trả lời (liên quan đến nội dung văn bản) có thể gắn liền với một khung lõi các lớp thông tin liên quan bắc cầu nhau. Khung thông tin này được gọi là dạng mẫu. Ví dụ dạng mẫu *Author-Paper-Publication-Publisher* liên quan đến đường đi từ đỉnh *Author* đến *Publisher* (như Hình 2).

Thông qua đường đi trong một dạng mẫu xác định, ta có thể tìm được dạng mẫu ngược lại[3]. Ví dụ *Publisher-Publication-Paper-Author* là dạng mẫu ngược tìm được từ ví dụ trên.

Dựa trên đồ thị G^{cc} (hoặc $G^{cc}{}_w$), tập các dạng mẫu có thể được xác định thông qua phương án sau:

Phương án xây dựng mẫu từ đồ thị G^{cc}

Nhập: đồ thị G^{cc} đại diện cho một ontology O
Xuất: tập dạng mẫu biểu diễn các phân lớp thông tin trong O
Xử lý:
1. Với i=1..n (n là số đỉnh của G^{cc})
   1.1. Tìm tất cả các đường đi qua i đỉnh trong G^{cc}
   1.2. Xác định và lưu trữ dạng mẫu dựa trên các nút (phân lớp) có trong đường đi tìm được trên.
2. Trả về danh mục lưu trữ dạng mẫu xác định được.

Từ đồ thị ở Hình 2, tập mẫu được xác định như trong Bảng 2.

Những dạng mẫu này có thể được phát triển để xây dựng tập hợp các câu hỏi (kiểu Yes/No hay WH) hay câu trả lời trong hệ thống hỏi đáp.

Bảng 2. Danh sách mẫu cơ bản

STT	Các thành phần liên quan
1.	Author-Paper
2.	Author-Paper-Publication
3.	Author-Paper-Publication-Publisher
4.	Author-Paper-Reference
…	…
70.	Topic-Publication
71.	Topic-Publication-Publisher
72.	Topic-Paper-Reference

[3] *Để đơn giản, chúng ta chỉ cần nêu dạng mẫu và không trình bày dạng mẫu ngược.*

Ví dụ, một số câu hỏi liên quan dạng mẫu *Author-Paper-Reference* như:

- WH: Ai viết bài báo OPQ?
- Y/N: Tác giả ABC viết bài báo OPQ phải không?
- WH: Tác giả ABC viết bài báo OPQ có những tham khảo nào?
- Y/N: Tác giả ABC viết bài báo OPQ có tham khảo XYZ không?

Với một câu hỏi dạng chủ động hoặc bị động, bước phân tích (xem [18]) sẽ nhận dạng các thành phần tương ứng trong nội dung hỏi. Tùy thuộc vào thành phần nghi vấn trong câu hỏi mà dạng mẫu hay dạng mẫu ngược sẽ được chọn phù hợp.

3. Suy diễn nội dung câu trả lời

Sau quá trình phân tích những câu hỏi dựa trên dạng mẫu nêu trên, bước suy diễn tìm nội dung trả lời (trong Hình 1) được thực hiện dựa trên giải thuật suy diễn đường đi dựa trên đồ thị (*Conceptual Graph-based answering reasoning algorithm, CGBAR*). Giải thuật này được phát triển dựa trên giải thuật tìm đường đi trong lý thuyết đồ thị và trí tuệ nhân tạo[4].

Ở đây, q là câu hỏi của người dùng, t^i là một mệnh đề hỏi con trong q và đồng thời là cơ sở để xác định bộ ngôn ngữ, D_i là tập các bộ ngôn ngữ xác định được từ q; v^i_k là nội dung thành phần t_i trong q và tương ứng với phân lớp c^i_k và có thể có tính chất a^i_k; c^i_j là một đỉnh (điểm) trong đồ thị; n(i) là độ dài đường đi trong đồ thị tương ứng với t^i;

Giải thuật CGBAR

Nhập:

- câu hỏi *q* có các mệnh đề hỏi thành phần $Q=\{q_i\}_{i=1..n}$,
- ontology O chứa tập các phân lớp $C=\{c_j\}_{j=1..m}$

Xuất: nội dung trả lời

Xử lý:

1. Xây dựng đồ thị G^{cc} cho tập C của ontology O.
2. Với mỗi mệnh đề hỏi q_i (i=1, 2, …, n)
   2.1. $D_i \leftarrow \varnothing$
   2.2. Với mỗi thành phần t^i_k của mệnh đề hỏi q_i
      2.2.1. Xác định trị v^i_k, phân lớp c^i_k và tính chất a^i_k.
      2.2.2. $D_i \leftarrow D_i \cup \{<v^i_k, c^i_k, a^i_k>\}$
3. Với $D_i=\{<v^i_k, c^i_k, a^i_k> / k=1, 2, \ldots, n(i)\}$ (i=1..n)
   3.1. Sắp xếp thứ tự tên các lớp sao cho c^i_1 liên quan đến v^i_1 khác rỗng và khác từ nghi vấn.
   3.2. Thực hiện tìm kiếm *đường đi ngắn nhất p* giữa điểm đầu c^i_1 đến điểm cuối $c^i_{n(i)}$ trong đồ thị G^{cc}.
   3.3. Nếu n(i)=2, thực hiện suy diễn kết hợp giữa kết quả tìm được trong bộ t^i_1 liên quan v^i_1và t^i_2 liên quan v^i_2, từ đó trả về kết quả trung gian và chuyển sang bước 4.
   3.4. Nếu n(i)>2, thực hiện suy diễn kết hợp giữa kết quả tìm được trong bộ t^i_k liên quan v^i_k và t^i_{k+1} liên quan v^i_{k+1} (k<n(i)-1) theo lộ trình của đường đi.
      3.4.1. Nếu không tồn tại p, phân rã các đường đi con *ngắn nhất* thành phần $sp_1=c^1_1\ldots c^1_{l(1)}$, $sp_2=c^2_1\ldots c^2_{l(2)}$, …, $sp_x=c^x_1\ldots c^x_{l(x)}$ sao cho $c_1=c^1_1=c^2_1=\ldots=c^x_1$.
      3.4.2. Tạo các tập kết quả riêng phần s_1 từ sp_1, …, s_x từ sp_x.
      3.4.3. Liên kết các tập $s=s_1\otimes\ldots\otimes s_x$ trả về kết quả trung gian và chuyển sang bước 4.
4. Tạo và trả về nội dung câu trả lời dựa trên các tập kết quả trung gian của những bước trên.

Nếu dựa trên ontology O ta có thể xây dựng một đồ thị G^{cc}_w thì giải thuật CGBAR được cải tiến thành giải thuật *WCGBAR* bằng việc xử lý tìm kiếm trên đồ thị có trọng lượng G^{cc}_w ở các bước 1 và 3.2.

Hai giải thuật được đề xuất trên có sự khác biệt nhất định so với phương pháp của nhóm Salloum [14]. Nhóm này đã đề xuất một giải pháp đặc sắc thực hiện viện suy diễn dựa trên đồ thị ý niệm với kết quả khá thuyết phục. Tuy nhiên, nền tảng của phương pháp đó lại được xây dựng thủ công bởi chuyên gia. Điều này tạo sự giới hạn nhất định cho năng lực xử lý trong phương pháp của họ. Điểm khác biệt của hai giải thuật đề xuất nêu trên chính là cơ chế xử lý việc suy diễn tự động dựa theo bài toán tìm đường đi trên đồ thị. Điều này đã khắc phục các giới hạn trong phương pháp của [14] đồng thời cải tiến và nâng cao năng lực xử lý của hệ thống do nhóm tác giả đề xuất.

IV. THỰC NGHIỆM

1. Phát triển ontology phục vụ suy diễn hỏi đáp

Phân tích và khai thác dữ liệu các bài báo khoa học từ nguồn ACM (*www.acm.org*) gồm 31679 bài báo về công nghệ thông tin (dung lượng 149MB) như sau.

[4] *Một tham khảo trích từ en.wikipedia.org/wiki/A*_search_algorithm*

Bảng 3. Thống kê dữ liệu đã huấn luyện

Loại	Số lượng	Số lượng phân biệt chính xác	Tỉ lệ
Liên kết (*Links*)	144981	144978	100.00%
Tác giả (*Authors*)	111736	47458	42.47%
Từ khóa tổng quát (*GeneralTerms*)	222858	118995	53.39%
Từ khóa (*Keywords*)	559448	273360	48.86%
Nội dung bài báo (*Papers*)	27412	27412	100.00%
Tài liệu tham khảo (*References*)	309466	248540	80.31%
Chủ đề bài báo (*Topics*)	126997	7350	5.79+

2. *Thực nghiệm phương pháp suy diễn nội dung*

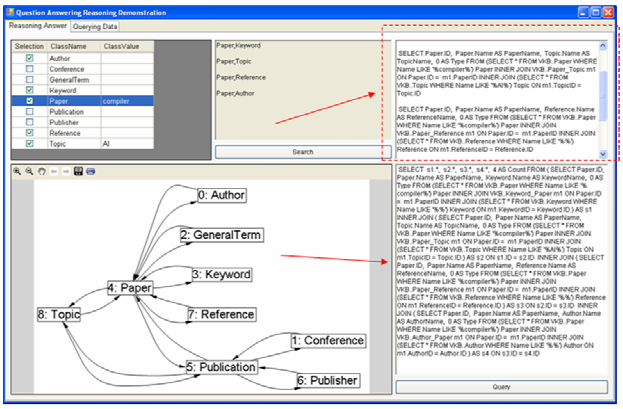
Trong giải thuật CGBAR, việc cài đặt các bước 3.3, 3.4.3 và 4 có thể được phát triển dựa trên những kỹ thuật khác nhau trong lĩnh vực công nghệ phần mềm. Hình 3 minh họa kết quả cài đặt và dạng sản phẩm trung gian của quá trình xử lý. Kết quả cài đặt hoàn chỉnh của động cơ suy diễn được thể hiện ở Hình 4.

Trong quá trình cài đặt của giải thuật CGBAR và động cơ suy diễn, một số vấn đề phát sinh như sau:

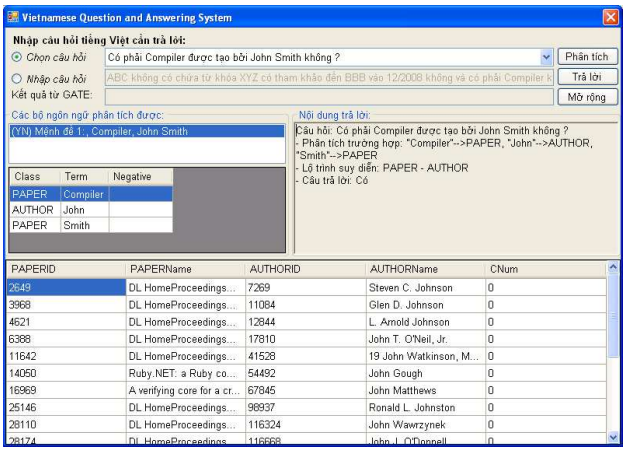
- (V_1) Tối ưu thời gian thực thi việc suy diễn trả lời vì các kết quả trung gian sinh ra quá phức tạp, khi dữ liệu nhiều thì truy vấn quá lâu.
- (V_2) Tối ưu nội dung bước suy diễn trả lời trong trường hợp có nhiều từ khóa xác định sẵn trong câu hỏi.
- (V_3) Một câu hỏi liên quan đến hành vi (động từ) bất kỳ, cần tìm giải pháp xử lý suy luận.
- (V_4) Mở rộng câu hỏi và tư vấn cho trường hợp câu hỏi mập mờ.

Phương án giải quyết cho V_1 là tối ưu kết quả trung gian thông qua việc tái tổ chức cấu trúc dữ liệu cho kết quả trung gian theo dạng bảng băm (*hash-table*) đã giúp tăng cường khả năng tìm kiếm và giảm thiểu xấp xỉ 60% thời gian xử lý, góp phần tăng hiệu suất hoạt động của động cơ suy diễn.

Một hướng xử lý cho vấn đề V_2 là giữ lại chỉ những từ khóa liên quan đến những thành phần nằm trên tập kết quả tổng hợp trong bước 4 hay bước 5.3 của giải thuật CGBAR.



Hình 3. Nội dung giải thuật CGBAR



Hình 4. Kết quả cài đặt hoàn chỉnh của động cơ suy diễn

Trong vấn đề V_3, việc xác định các mối quan hệ ngữ nghĩa (*is-a*, *part-of*, *similar*, *hypernymy* …) giữa động từ đó với một phân lớp của ontology O sẽ giúp xác định phân lớp chính mà hành vi đó có liên quan. Từ đó, những từ đồng nghĩa có trong phân lớp đó sẽ được sử dụng thay thế để phục vụ việc tìm kiếm nội dung trả lời.

Trường hợp xảy ra trong vấn đề V_4 khi có tối thiểu một thông tin trong câu hỏi không xác định được thuộc về phân lớp nào trong ontology O. Việc áp dụng phương pháp phản hồi liên quan (*relevant feedback*) trong mở rộng truy vấn là một chọn lựa hợp lý để giải quyết vấn đề này. Phương pháp này sẽ giúp động cơ suy diễn học được kiến thức từ các chuyên gia (người sử dụng), từ đó phân tích và thống kê được những chọn lựa có tần suất chọn lựa cao để tư vấn cho người dùng, góp phần giải quyết nhập nhằng của vấn đề này.

Trong thực nghiệm, bước phân tích câu hỏi để xác định ánh xạ giữa những nội dung trong câu hỏi với phân lớp trong ontology O được tiến hành trên 210 câu hỏi thuộc 5 nhóm thử nghiệm tương đồng nhau với 90.52% phân tích đúng. Tuy nhiên một số trường hợp một phần nội dung câu hỏi không nhận biết được do giới hạn số lượng đầu mục từ vựng trong ontology hay không nhận biết được chính xác từ tiếng Việt từ câu hỏi, dẫn đến số lượng kết quả đúng (theo chương trình và theo chuyên gia) vẫn còn ở mức giới hạn. Dựa trên kết quả xử lý trước này, việc việc suy diễn được tiến hành và thu được kết quả với số liệu tóm tắt như trong bảng sau.

Bảng 4. Kết quả suy diễn nội dung trả lời

Thực nghiệm	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Nhóm 1	31	31	29	30	93.548%	96.667%
Nhóm 2	38	38	36	37	94.737%	97.297%
Nhóm 3	54	54	52	54	96.296%	94.545%
Nhóm 4	40	40	37	39	92.500%	94.872%
Nhóm 5	47	47	44	46	93.617%	95.652%

Với:
(1) Số câu hỏi tiếng Việt phân tích đúng
(2) Số câu hỏi tiếng Việt suy diễn trả lời được
(3) Số câu hỏi tiếng Việt suy diễn trả lời được hợp lý
(4) Số câu hỏi tiếng Việt suy diễn trả lời được hợp lý theo chuyên gia
(5) Độ chính xác, (6) Độ bao phủ

Trong các nhóm thực nghiệm, một số trường hợp chưa đạt yêu cầu liên quan đến một trong những vấn đề $V_1 – V_4$ phân tích nêu trên và ngoài ra còn do vấn đề nhập nhằng chưa được giải quyết triệt để. Nhìn chung, việc suy diễn tạo kết quả trung gian cho thấy độ chính xác của bước xử lý tương đối tốt, dù các độ đo này có phụ thuộc vào kết quả phân tích trước đó.

V. KẾT LUẬN

Bài báo này trình bày hướng tiếp cận nhằm phát triển động cơ suy diễn cho hệ thống hỏi đáp có hỗ trợ tiếng Việt với trọng tâm là phương pháp suy diễn nội dung trả lời dựa trên giải thuật A* và lý thuyết đồ thị.

Các giải thuật CGBAR và WCGBAR được đề xuất trong nghiên cứu này với mong muốn định hướng xử lý tự động để tìm kiếm các phương án trả lời cho câu hỏi của người dùng. Việc tối ưu các giải thuật này là một trong các nghiên cứu kế tiếp của nhóm tác giả.

Kết quả thực nghiệm tuy bị hạn chế bởi dạng thức đơn giản của câu hỏi (dạng Y/N và WH một mệnh đề hỏi) và độ lớn của tập câu hỏi (chỉ 210 câu) cùng với cơ sở tri thức phục vụ thực nghiệm, nhưng cũng cho thấy tính khả thi của các phương pháp đề xuất thông qua các giá trị độ đo (độ chính xác, độ bao phủ) thu được. Từ đây, việc tối ưu phương pháp và mô hình đề xuất là sự nghiên cứu còn tiếp tục của nhóm nghiên cứu trong thời gian sắp tới với mục tiêu xây dựng một hệ thống hỏi đáp hỗ trợ tiếng Việt thực sự hiệu quả và hữu ích không chỉ trong lĩnh vực công nghệ thông tin mà còn trong các lĩnh vực khác.

TÀI LIỆU THAM KHẢO.

[1]. VANESSA LOPEZ, VICTORIA UREN, ENRICO MOTTA, MICHELE PASIN, "*AquaLog, An ontology – driven question answering system for organizational semantic intranets*", Journal of Web Semantics, 31 Mach 2007

[2]. START, start.csail.mit.edu

[3]. Lexxe, www.lexxe.com

[4]. Ask, www.ask.com

[5]. W5hanswers Q&A, www.w5hanswers.com

[6]. www.mshd.net

[7]. *Hesitation, www.hesitation.co.uk*

[8]. *Google Answers, knol.google.com/k/google-answers*

[9]. *Google Answers (Chinese), enda.tianya.cn/wenda*

[10]. *IBM Watson, ww.ibm.com/innovation/us/watson*

[11]. S. HANDSCHUH, S.STAAB, F.CIRAVEGNA, "*S-Cream, Sem–iautomatic CREAtion of metadata*", 13th Int. Conference on Knowledge Engineering and Management, 2002, Spain.

[12]. M. VARGAS–VERA, et. al., "*MnM: ontology driven semi–automatic supp for semantic markup*", 13th Int. Conference on Knowledge Engineering and Management, 2002, Spain.

[13]. HỒNG TRUNG DŨNG, CAO HOÀNG TRỤ, "*Dịch tự động truy vấn tiếng Việt sang đồ thị ý niệm*", Tạp chí Tin học và Điều khiển học, tập 23, số 3, 2007 (trang 272–283)

[14]. SALLOUM, WAEL, "*A Question Answering System based on Conceptual Graph Formalism*", 2nd Int. Symposium on Knowledge Acquisition & Modeling, IEEE CS Press, 2009.

[15]. JIRÍ MÍROVSKY, "*Netgraph Query Language for the Prague Dependentcy Treebank 2.0*", The Prague bulletin of Mathematical Linguistics, number 90, 12/2008, (pp.5–32)

[16]. TUOI PHAN, THANH NGUYEN, THUY HUYNH, "*Question Semantic Analysis in Vietnamese QA System*", ACIIDS 2010, Vietnam.

[17]. TUOI T.PHAN, THANH C.NGUYEN, "*Vietnamese knowledge base development and exploitation*", International Journal of Business Intelligence and Data Mining, 2010. ISSN: 1743-8195.

[18]. VÕ THANH HÙNG, "*Nghiên cứu và xây dựng tập các câu truy vấn phục vụ cho hệ thống hỏi đáp tiếng Việt*", Luận văn tốt nghiệp đại học, Khoa Khoa học và Kỹ thuật Máy tính, Đại học Bách khoa Tp.HCM, 2010.

[19]. ANH K. NGUYEN, HUONG T. LE, "*Natural Language Interface Construction using Semantic Grammars*", PRICAI08, Hanoi, 2008, Vietnam.

[20]. DAI Q. NGUYEN, DAT Q.NGUYEN, SON B. PHAM, "*A Vietnamese Question Answering System*", KSE09, IEEE CS, 2009.

[21]. CAO, T.H. & ANH, M.H, "*Ontology-Based Understanding of Natural Language Queries using Nested Conceptual Graphs*", 18th Int.Conference on Conceptual Structures, 2010, Malaysia, LNCS 6208.

[22]. PHILIP H. P. NGUYEN, DAN CORBETT, "*A basic mathematical framework for conceptual graphs*", Journal IEEE Transactions on Knowledge and Data Engineering Volume 18 Issue 2, February 2006.

[23]. CAO, T.H., "Conceptual Graphs and Fuzzy Logic: A Fusion for Representing and Reasoning with Linguistic Information", Studies in Computational Intelligence, Vol. 306, Springer-Verlag, 2010.

[24]. CAO, T.H., "*Fuzzy Conceptual Graph Programs for Knowledge Representation and Reasoning*", Tech. Report 400, University of Queensland, Australia, 1997.

[25]. CROITORU AND VAN DEEMTER, "*A Conceptual Graph Approach to the Generation of Referring Expressions*", IJCAI, 2007, Hyderabad, India.

[26]. DANG T. NGUYEN AND TRI PHI-M. NGUYEN, "*A Question Answering Model Based Evaluation for OVL (Ontology for Vietnamese Language)*", International Journal of Computer Theory and Engineering, Vol. 3, No. 3, June 2011.

[27]. THO THANH QUAN, SIU CHEUNG HUI, "*Ontology-based Natural Query Retrieval using Conceptual Graphs*", PRICAI08, Hanoi, 2008.

Nhận bài ngày: 18/05/2011

SƠ LƯỢC VỀ TÁC GIẢ

PHAN THỊ TƯƠI

Tốt nghiệp Đại học Kỹ thuật Tiệp khắc, ngành Máy tính, năm 1976. Tiến sĩ chuyên ngành Khoa học máy tính, năm 1985, Trường Đại học Charles, Cộng hòa Czech.

Hiện công tác tại Khoa KH và KT Máy tính, Trường Đại học Bách khoa Tp. HCM.

Hướng nghiên cứu: Xử lý ngôn ngữ tự nhiên- Xử lý văn bản; Truy xuất thông tin; Rút trích thông tin.

Email: tuoi@cse.hcmut.edu.vn

NGUYỄN CHÁNH THÀNH

Tốt nghiệp Đại học Sư phạm Tp. HCM, ngành Toán, năm 1994. Tốt nghiệp Kỹ sư, Thạc sĩ kỹ thuật và Tiến sĩ kỹ thuật ngành Máy tính năm 1998, 2003 và 2011, Trường Đại học Bách Khoa Tp. HCM.

Hướng nghiên cứu: Xử lý ngôn ngữ tự nhiên, Truy xuất thông tin, Rút trích thông tin, Web ngữ nghĩa.

Email: chanh.thanh@yahoo.com.vn